Final Report: **Community Meeting:**
Analysis, Management and Sharing of Antigen Receptor Repertoire (ARR) Sequence Data
29 May – 1 June 2015
Morris J. Wosk Centre for Dialogue and Harbour Centre
Simon Fraser University
Vancouver, BC, Canada

This meeting brought together experts in the production and analysis of Antigen Receptor Repertoire (ARR) data with ethicists, data security personnel, IP and legal experts, and representatives from funding agencies to discuss the challenges involved with sharing and comparing this new and expanding type of data. About 70 people attended this Community Meeting, which was organized as a series of workshops addressing topics such as data production quality control, data analysis and integration, and the legal, ethical and IP considerations to sharing these data. This final report comprises the agenda for the meeting, reports from each workshop, and a report from the final day of the meeting which focused on setting up working groups and beginning a White Paper based on this initiative.

This Community Meeting was motivated by the fact that many academic labs, biomedical research institutions, and pharmaceutical companies are applying Next Generation Sequencing (NGS) technology to Antigen Receptor (Antibody/B-cell or T-cell) Repertoires. Given NGS technology, it is now possible to sequence millions of molecules from the ARR repertoire, describing this aspect of the immune response in great detail. However, storing and analyzing these data is a rapidly increasing challenge. These data are critical for studies of autoimmune diseases, development of vaccines, therapeutic antibodies and cancer immunotherapies, monitoring clinical trials, and other developments in immunological research and patient care. These data will be more valuable if researchers can share and compare data among studies and institutions, but at present there is no common data base format or common platforms for sharing these data. Beyond these bioinformatics challenges, IP, ethics and legal challenges thwart our ability to share and compare these data.

The main outcome of the meeting was a consensus that sharing these data is essential to the development of this field, and that the community needed to develop a common repository for these type of data. To that end we have formed working groups focused on: (1) determining a minimal set of metadata and associated data to include when publishing ARR data or submitting them to a common data base, (2) developing platforms for facilitating sharing of ARR data, (3) producing a White Paper explaining the goals of the ARR community, and (4) further development of this initiative, including organizing a meeting of the whole in about one year. The overall goal is to continue to self-organize the community involved in producing and analyzing Antigen Receptor Repertoires.

The Community Meeting was supported by CIHR, NIH, GenMab, The Antibody Society, CHAVI, the IRMACS Centre, and Simon Fraser University.

Community Meeting:
Analysis, Management and Sharing of Antigen Receptor Repertoire (ARR) Sequence Data
29 May – 1 June 2015
Morris J. Wosk Centre for Dialogue and Harbour Centre
Simon Fraser University
Vancouver, BC, Canada

Meeting Overview

Friday 29 May
Samuel & Frances Belzberg Atrium (6PM-10PM)
Asia Pacific Hall (6PM-10PM)

Saturday 30 May
Asia Pacific Hall (8AM-5PM)
Samuel & Frances Belzberg Atrium (8AM-5PM)
ICBC Concourse (10AM-2PM & 5PM-8PM)

Sunday 31 May
Asia Pacific Hall (8AM-5PM)
Samuel & Frances Belzberg Atrium (8AM-5PM)
ICBC Concourse (10AM-2PM)

Monday 1 June
Harbour Centre, Rm 1700 Labatt Hall (8AM-5PM)
Rm 2050 Alan & Margaret Eyre Boardroom (8AM-5PM)

Organizing Committee: Sohail Ahmed, Sanchita Bhattacharya, Felix Breden, Robert Cook-Deegan, Brian Corrie, Lindsay Cowell, Danny Douek, George Georgiou, Yvo Graus, Rob Holt, Yann Joly, Tom Kellam, Thomas Kepler, Marie-Paule Lefranc, Nishanth Marthandan, Tony Moody, Rik Rademaker, & Jamie Scott

Organizational Facilitators: Felix Breden, Thomas Kepler & Jamie Scott

Workshop Leaders:
Workshop 1: Sai Reddy & Danny Douek (Saturday AM)
Workshop 2: Lindsay Cowell & Steven Kleinstein (Saturday PM)
Rob Holt & Thomas Kepler (Sunday AM)

Workshop 3: Tania Bubela & Bob Cook-Deegan (Sunday PM)

Meeting Agenda

Friday 29 May
Samuel & Frances Belzberg Atrium
6:00 - 7:00 PM
Registration, Networking,
(Light snacks, non-EtOH drinks, 1 EtOH drink & cash bar 'til 8:00 PM)

Asia Pacific Hall
7:00 – 9:00 PM
Welcome and overview for the meeting, an action-item driven event
Introduce organizers, discussion leaders
Review the agenda: action items for discussion and consent
Keynote Address: Privacy, Governance & Innovation in the Era of Big Data
Paul Terry, CEO, PHEMI

Retire to local establishments for networking & fun (*e.g.,* Steamworks Brew Pub (375 Water Street), http://steamworks.com/brew-pub; Rogue Kitchen and Wetbar (601 West Cordova Street, http://www.roguewetbar.com/; or the Delta Hotel Lounge https://www.deltahotels.com/Hotels/Delta-Vancouver-Suites/Restaurants-Dining/Spencer-s-Resto-Lounge

Saturday 30 May
Samuel & Frances Belzberg Atrium
8:00 - 8:30 AM
Continental breakfast

Asia Pacific Hall
8:30 - 9:45 AM
Workshop 1: High-Throughput Sequencing **(**HTS) Technology and Antigen Receptor Repertoire (ARR) Data Generation

Samuel & Frances Belzberg Atrium
9:45 - 10:15 AM
Refreshment break

Asia Pacific Hall
10:15 AM – noon
Workshop 1: cont'd

ICBC Concourse
Noon – 1:30 PM
Networking Buffet Lunch and Posters

Asia Pacific Hall
1:30 – 3:00 PM
Workshop 2: Antigen Receptor Repertoire (ARR) Data Management and Analysis: Session 1, Management.

Samuel & Frances Belzberg Atrium

3:00 – 3:30 PM
Refreshment Break

Asia Pacific Hall
3:30 - 4:30 PM
Workshop 2, Session 1: cont'd

ICBC Concourse
5:00 – 7:00 PM
Demos of Tools and Platforms
(Light snacks, non-EtOH drinks, 1 EtOH drink & cash bar)

Sunday 31 May
Samuel & Frances Belzberg Atrium
8:00 - 8:30 AM
Continental breakfast

Asia Pacific Hall
8:30 - 10:00 AM
Workshop 2, Session 2: Antigen Receptor Repertoire (ARR) Data Management and Analysis: Analysis.

10:00 – 10:30 AM
Samuel & Frances Belzberg Atrium
Refreshment Break

Asia Pacific Hall
10:30 – noon
Workshop 2, Session 2: (cont'd)

ICBC Concourse
Noon – 1:30 PM
Networking Buffet Lunch and Posters

Asia Pacific Hall
1:30 – 3:00 PM
Workshop 3: Ethical and Legal Concerns in Sharing ARR Data

Samuel & Frances Belzberg Atrium
3:00 – 3:30 PM
Refreshment Break

Asia Pacific Hall
3:30 – 5:00 PM
Workshop 3: (cont'd)

Dinner at 7PM at Kamei Royale (Sushi restaurant). (For those who signed up at registration.)
211-1030 W. Georgia St., Vancouver, BC V6E 2Y3, Canada 604.687.8588

Monday 1 June
SFU Harbour Centre
Labatt Hall (Rm. 1700; across the street from the delta suites hotel)

8:00 - 8:30 AM
A light breakfast will be available in the meeting room. Coffee, *etc*. will be available throughout.

8:30 – 9:30 AM
Restating and Reaching Consensus on Action Items: Workshop 1 - Danny Douek & Sai Reddy

9:15 – 10:30 AM
Restating and Reaching Consensus on Action Items: Workshop 3 - Tania Bubela & Bob Cook-Degan

10:30 - 10:45 AM
Refreshment Break

10:45 AM – 12:15 PM
Restating and Reaching Consensus on Action Items: Workshop 2 (2 sessions) - Lindsay Cowell, Steven Kleinstein, Tom Kepler & Rob Holt

12:15 PM
Conclusion of Meeting

Outline for Workshop 1:
High-Throughput Sequencing (HTS) Technology
and Antigen Receptor Repertoire (ARR) Data Generation
(Saturday morning)

Workshop Leaders: Sai Reddy and Danny Douek

The generation of ARR data involves the following steps: (A) experimental design including acquisition of samples from a donor; (B) isolation of one or more lymphocyte subset(s) to analyze; (C) library generation; (D) HTS sequencing; (E) data filtering and error correction; (F) data analysis and sharing. The entire process is technically demanding and is constantly evolving as technology progresses. For ARR data to be meaningful, reproducible and amenable to meta-analysis, each and every step in the list above needs to be performed by following detailed, well-documented SOPs and informatics tools. Built-in quality assurance metrics should be considered. Frequent re-evaluation and versioning of SOPs will likely be needed to keep up with technology advances.

Problems:

A. Experimental design: No standards exist for annotation of experiments and samples (what tissue is being analyzed, how many cells, what SOPs are used, etc.)
B. Isolation of lymphocyte subsets: Cell isolation is critically dependent on reagents (cell surface marker antibodies used), instrumentation and resolution metrics, none of which is standardized.
C. Library generation: There are numerous variations in methods for library generation, e.g. starting with RNA or gDNA, primers used, barcoding strategy (if employed); libraries of single chain or paired immune receptor amplicons etc.
D. Sequencing platform: Sequencing technologies are evolving very rapidly and each has distinct advantages and disadvantages. Sequencing depth also needs to be considered.
E. Data filtering and error correction: The pertinent issues here depend on steps (C.) and (D.) above.
F. Lack of infrastructure for data sharing. Data are not deposited. Data analysis packages are either not publically available or if they are, the code is not available and the programs have to be taken at face value. Lack of documentation. Obsolete data analysis packages are not archived and hence older data and metadata cannot be re-evaluated.

Examples:

A. Even if data are publically available they are meaningless if the experiment and the procedures used to generate the data have not been properly annotated.
B. FACS techniques vary from lab to lab. Reagents used for cell separations have a huge impact on the quality of the sort. New lymphocyte subsets are constantly proposed or previously established ones re-classified.
C. Nearly every lab uses a different procedure for library generation. Same goes for primer sets. The quantitation of clonal expansions is problematic. Library contamination is an issue unless each library is separately barcoded.
D. Different HTS platforms vary with respect to read lengths, errors and cost (which affects sequencing depth).
E. There is no consensus on error correction or on sequence clustering to account for sequencing errors.
F. It is impossible to analyze many sets of published data either because the data are not available, or the tools for analysis are not available and/or transparent.

Solutions:

A. Agree on standard annotation fields to be used for every experiment. Burden to researcher needs to be considered though, because if the annotation becomes too detailed compliance will end up being lax.
B. No easy solution is available. We believe that establishing SOPs for FACS analysis is beyond the scope of this group.
C. Develop validated primer sets for B and T cell repertoires for human, mouse. Develop SOPs for library construction from mRNA, gDNA. Establish recommendations with respect to barcodes.
D. Document in (A.) above the platform used. Establish recommendations for sequence depth required based on cell number.
E. Establish error correction standards.
F. Create a mechanism for depositing data. Analysis tools (and codes) should become publically available.

Short-term Actions:

A. Write a position paper to major journals to highlight the problem and elicit interest by funding agencies and publishers on finding solutions.
B. Working group to develop interim guidelines for publication before final recommendations are established as below.

Specific short-term Actions:

1. Establish list of mandatory annotation fields and provide options.
2. Do not address for the time being.
3. Working group needs to be set up to propose library construction SOPs.
4. Provide recommendations for sequencing depth desired relative to number of cells. Develop and use internal standards (pre-defined cell line mixtures) to allow error assessment.
5. Working group needs to be set up to agree on best practices for HTS sequencing error accounting.
6. Establish mechanisms for deposition and access of data. Set up working group to evaluate how to make analysis tools transparent and accessible.

Long-term Actions:

A. Coordinate with funding agencies to support efforts for ARR data standardization.
B. Make final proposals to publishers regarding standards for publication of ARR data

Notes: Workshop 1 (Notetaker: Ramy Arnaout)

Theme: be as precise as possible.

The vision for our repository is a curated home for searchable metadata, and a UI for searching it. It can have pointers to where raw (and other) data actually lives, and a sandbox for code to find the data. But the search, and having the metadata to search, is key.

Key action item: checklist/metadata: minimal standards for data and publication (recommendations for publication, requirements for repository).

Cell subsets. Cell-surface markers vs. names. At minimum, store markers. Have facility to search by markers or names. Google/Wikipedia/parsing will give the correspondence. My sketch of a search box and results page.

Record kinetics. Don't feel compelled to hew to 3, 5, 7, 30 days, 1 year ("round" numbers).

Many sources of bias: type of starting material (e.g., plasma-cell RNA), amount of starting material (incl. number of cells added vs. that got amplified), method (multiplex vs. 5'RACE), platform, depth of coverage, filtering, and (workshop 2 stuff) annotation and analysis. Biases include nucleotide errors, chimeras, contamination, unknown reference genomes (consider sequencing germline with sample? by Sanger?).

We need standards for all these things.

Eventually we can put forth best practices that may well include:

--Ask questions first, sequence later
--Sort subsets
--For every figure, it should be clear where the data is that led to it.
--Best-practices for quantitation e.g. spiking with known sequences
--Think about controls and replicates for various stages of process.
--Deposit raw reads (but how raw?)/make it available
--Make code available (not for reviewers to run ahead of publication, but as a standard for future comparison if necessary) and use version control (GitHub)
--Demand (disclosure of method of) error correction
--List uncertainties in everything you submit
--Define terms (e.g., "clones")

Outline for Workshop 2:
ARR Data Management & Analysis,
Session 1, Management (Saturday afternoon)

Workshop Leaders: Lindsay Cowell and Steven Kleinstein

The management of ARR data presents several important issues around data exchange standards and nomenclature; all are exacerbated by the volume of data to be managed.

Problems:
A. Lack of metadata and data standard for sharing ARR data associated with publications.
B. Lack of common file format for sharing ARR data between analysis tools.
C. Lack of standard for describing analysis methods applied to generate results.
D. Lack of API standard for querying ARR sequence databases.
E. Lack of standard IRB protocol language to permit deposition and sharing of ARR data.

Solutions:
A. Draft recommendations on what level(s) of data should be made available with publications or upon submission to a common data repository.
B. Develop standards for data deposition, along with data templates that can be filled in by researchers to support submission of ARR data to repositories. Disseminate the standards to promote adoption by researchers, publishers, *etc*.
C. Recommend ARR data file format, and make available open-source APIs to read/write this format.
D. Rally around a common ARR repository that will implement the proposed standards.

Actions:
A. Establish a list of needed standards, the scope of each and how they relate to existing standards.
B. Draft an outline of proposals for each of the standards: brainstorm list of metadata and data fields to include.
C. Convene working groups to flush out these proposals, and establish procedures for reaching agreement on standards.
D. Define list of use-cases and requirements for a common data repository.
E. Identify online forum to host documents, and collaborate on these efforts.
F. Work with existing and emerging ARR repositories to conform to proposed standards and implement use-cases.

Notes Workshop 2, Session 1 (Notetaker: Adrian Thorogood)

SRA - Problems:
- Both filtering and creation of consensus reads are described
- In manuscripts, often mismatch between samples in SRA and in manuscript
- Unknown amount of aggregration (1 file to 1 sample? More?)
- Insufficient sample descriptions, spread over many different pages
- SRA – describe experiment – platform, layout, (are choices appropriate for repertoire studies?) – drop down v.s. free text.
- Mostly people only provide what is required.

Sharing ARR Data

- Lack of common file format – input output between analysis pools
- 4 chunks: sequence preprocessing, rearrangement inference, repertoire characterization, repertoire comparisons
- VQuest, IgBlast - all these packages read in and read out different file formats.
- **File formats** at each junction – a challenge for the community.
- Lack of standards for describing analysis methods:
    - Raw reads
    - Processing pipelines
- Lack of a standard for querying ARR sequence databases (metadata): study, experiment, sample, read, repertoire.
- Lack of standard IRB protocol language to permit deposition and sharing of ARR data.
- *Terminological debate – what is the difference between 'data', 'metadata', data about (e.g. experiment), data derived)*.

Kleinstein

- Reused/reanalysis existing data v.s reproducing experiments?
- Related projects:
    - Human Immunology Project Consortium (HIPC) – developing data standards
    - Immport data repository:
    - NIH Big Data 2 Knowledge: metadata for IMM studies
    - *Its really hard to get the metadata down – how can we automate the process?*
    - bioCaddie
    - *How can we work with these existing initiatives?*
- We could propose a data template for people to fill out when submitting ARR data (say to an existing database)
- Leveraging existing ontologies (e.g. cell types)
    - Makes the data MACHINE READABLE
- *But are human cells easily amenable to cell ontologies – can't have a pure population of more than 1 cell.*
- *Could we have a category and a justification?*
- *SRA – in terms of metadata – asks for cell isolation protocol and targeted cell type.*
- Gating definition – list of markers to define cell type – and then choose the name as best they can?
- Why have a simple name? On one side, it may be confusing. But from a discovery point of view, maybe a general definition helpful?
- Even +- might not be enough – need complete flow data?
- Gold standard is what it is, not what it means.

Formal mechanism for settling on recommendations? Come up with a use case document?

Minimum Information Standards (e.g. Microarray)

- Critical elements of MIAME: raw data, processed, essential sample design, annotation of the array, laboratory and data processing protocols.
- Determining this: what is our goal. Literal reproduction of experiment? Or just reanalysis/reuse? (Different consumers – meta-analysis v carrying out similar study)
- What about for believing the results? Having confidence?

ImmPort:

- Has a number of categories and tables, then provides a link to the raw data.
- *Does anything need to be added to this template?*
- Currently taking a lot of time to ensure submissions are high quality
- Can't do a meta-analysis without a software infrastructure.
- This will allow us to build on these tools and focus on establishing an external repository.

Part 2:

List out some of the use cases community cares about. Then use these to see what kind of metadata we need.

- MS studies – collate all the individuals b/c they aren't that big.
- (USE cases getting put into document)

Next step: what data do we need to do these use cases? Think of 2 formats: pre and post VDJ assignment.

Proposal: Define a set of entity tags to specify annotations. V,D,J Segment calls?
Proposal: Define standard xml scheme (or column headers).

(Brainstorm on meta-data)

- Next step widdle list down and then go to SRA and tell them what we want people to submit, or see if we have to build our own system.
- Justification – journals / databases could provide the use cases as justification when requiring metadata
- This is a lot of work for researchers! What's in it for me.
- Need for champions, iterations, evolution.
- Discussion of flybase – don't necessarily need a central database –
- Evidence review in clinical research – requires reading the literature related to the use cases.
- Another approach – take all other existing standards – review of existing requirements.
- Lot of debate about whether to pare down this list to minimal requirements.
- MIAME – too onerous, not adhered to?
- This full list is about best practices.
- Carrots: interoperability of tools is a carrot. (example of bibliographies for tool interoperability – need to support 4 or 5)
- Attribution is another important carrot – a creditable academic output – contribution of data or tools.

Outline for Workshop 2:
ARR Data Management & Analysis,
Session 2, Analysis

Workshop Leaders: Rob Holt and Tom Kepler

The analysis of ARR data is typically broken down into at least two stages: inference of the rearrangement parameters ("VDJ annotation") and downstream analysis according to the relevant design. The latter step often involves the inference of clonal kinship.

Problems:
A. Lack of agreement on the appropriate quantification of uncertainty in all aspects of data analysis: base calls, read annotations, clonal assignments, subset abundances.
B. Lack of "gold-standard" datasets for testing data-analytic methods.
C. Lack of comprehensive databases of germline genes.

Examples:
A. Substantial variation in results from applying existing tools.
B. Substantial variation revealed when directly estimated.
C. Discovery of novel alleles directly from ARR HTS data.

Solutions:
A. Draft recommendations for the use of statistical tools in ARR.
B. Support the development of gold-standard datasets or surrogates for them and facilitate their use for testing.
C. Offer standardized tests for the evaluation of new tools.
D. Draft recommendations for the approval of germline gene segments inferred using statistical evidence.

Actions:
A. Establish a set of desired data for the proffered solutions.
B. Organize a working group or groups to draft the recommendations.
C. Circulate and gain approval for the recommendations.
D. Meet with publishers and funding agencies to offer recommendations.

Notes on Data analysis for ARR Workshop 2 Session 2 (Notetaker: Nishanth Marthandan)

<u>Tom Kepler</u>

- MIAME was discussed for determining what to do and what not to do
  - Not prescribe or endorse a set of tools
  - Use and include statistics in analysis approaches
- Do statements about ARR have unambiguous interpretations?
  - Uncertainties should be explicitly mentioned
  - Erik brought up that dissimilarity of models' assumptions will affect comparisons of probabilities of results from different analysis tools and should be stated along with the probability values, especially during comparisons. Tom agrees.

    - Felix raised the need for precise definitions and characterizations of diversity metrics.
    - Jessica from Vanderbuilt: There is also need for specific definitions/characterizations of other aspects such as binding specificities.
- Outline
    - Problems
        - How to handle the dissemination of putative new genes/alleles?
            - MP: Based on the history of IMGT and its beginnings the process of adding new alleles was contentious and evolved into a deliberately curated process and after agreement during once a year workshops/meetings. Felix summarized: IMGT is deliberate and careful. Community needs to decide how to deal with inferred alleles?
            - Jamie brought up the issue of what would be the best practices for reporting those putative alleles in journals
            - There was general agreement that the rich amount of information from NGS data of repertoire needs to be leveraged for putative alleles
            - Corey brought up sanger sequencing for validation and Amgen participant seconded the suggestion
            - Jamie was enquiring if there is a computational solution to partly eliminate some of the putative alleles from candidates
            - Steve Kleinstein: keep assessing and comparing the inferences from different analysis tools/groups. Also validate (computationally and via molecular biology techniques) for new alleles to move forward.
            - MP: IMGT needs ATG to VRS sequence for the full allele and its functionality. **No separate dB in IMGT for potential alleles.** Community can come up with a location/dB to keep track of those inferred alleles and the labs reported, number of times the putative alleles were observed.
            - MP: define primer for the new allele at genomic level and validate
            - Stockholm commenter Q: How to deal with duplication that is not on the same locus/position of the genome? The point was duly noted.
            - **FB summarizing: seems like consensus on supplementary dB and gold standard IMGT dB**

Uncertainity in germline gene composition

- Tom mentioned that one issue to be aware of is that from adding more alleles there will be more false assignments
- **SK: Was checking if IMGT will support the putative dB list? Like the UNSWIg database**
- **MP: Would be ok with links to the putative genes/alleles dB like the UNSWIg resource**

<u>Rob Holt</u>

- TCR diversity relevant to Cancer Biology

    - Patients with T-cell infiltrates in tumors have better outcomes
- Need to specify antigen specificities and characterize the cell populations in specific details especially while comparing groups/individuals
- Uncertainty due to limited sampling
    - As you sequence deeply one sees more gaps in the shared clones getting filled (concordance reached near 1.7 million reads for some samples)
    - How does one know one has exhaustively sequenced?
        - Try to estimate using rarefaction analysis (accumulation curves)
            - Should strive to achieve flat accumulation curves before comparing the 2 samples
            - Use tools from ecology
                - RH: If there are other methods/tools to address the problem?
                - Andrew Bradbury questioned if the difference is just quantitative when leveraging the tools from Ecology (i.e. 1000s of species vs millions in repertoire)?
                    - RH: Microbial would be good example for comparable levels of species diversity
                - RH: Statisticians are skeptical of tackling these problems due to lack of good models and to deal with errors in PCR, etc
                - RH: suggested use of asymptotic distribution
                - AR: The problem is the uncertainties in the expected underlying distribution. Chou has done work on good estimators if you know the underlying distribution. Need to iteratively observe distributions and refine the tools

Part deux

Rob Holt

- TCGA resource had data sharing and analysis co-ordination. Good resource for data source. Would like to have something similar for immune repertoire community
- Impact of optimizing parameters
    - Problems:
        - For example, a tool's approach to annotate sequence as CDR3 based on the anchored/conserved sites such as the Cys (C) and Phe(F) would result in bogus results from a genome sample. These problems are magnified, especially, when tools are usually run with default parameters.
    - Needs positive and negative controls in experiment design (with the analysis tools in mind)
- Finding relevant T cell specificities is difficult against a background of bystanders
    - In previous instances, without proper positive controls and deep sequencing large number of spurious shared clones were reported
    - To find TCR specificities the appropriate reference dB needs to be used
- Key analysis considerations
    - Evaluating control data sets
    - Version control

Action items

Tom Kepler

- Solutions:
  - Gold standard dataset
    - Danny Douek: History/precedence of work in flow cytometry should be leveraged. Gold standard set was sent to 10 labs and the results were compared.
    - Jamie: Different gold datasets may be needed for different questions. Use paired VH-VL data for testing results of clonal lineage tools
    - DD and TK: seconds the idea of different standard sets for different questions
    - Rob enquired Cindy (of Adaptive) about the use of synthetic templates as controls and if it was spiked as internal controls. Cindy concurred that synthetic templates were spiked in as internal controls
    - TK: how many synthetic antibody genes can be produced for testing? Implications on Cost etc. So a solution should take those into considerations
    - Andrew Bradbury: it could be done and other groups are doing. Simulate somatic mutations in synthetic templates and use for testing
    - Ramy Arnout: Cannot be deterministically determined for the induced variations.
    - Christine: How much diversity needs to be induced in the synthetic set
    - Erik: what about substitutions or insertions/deletions between V and D
    - Sai: consider dsDNA or RNA for those synthetic molecules. Scale of price is important consideration.
    - DD: need real molecules for sequencing testing, for analysis an in silico would be needed
    - TK: data insilico are generated under some assumptions and the analysis tool would be biased towards that.
      - May be the goal would be to delineate different types of datasets.
      - How to go about soliciting funds (RO1, letters to NIH, etc)?
    - AR: the need depends on different questions.
      - One subset could be vial of biological sample and roughly the same in the actual sample and reduce annotation variability
      - For analysis tool there is the need for obtaining same results from different tools
      - RH: From history of microbiome project needs both
        - Several bias was discovered when for example only 20 species was included etc
        - Similar could be done for repertoire community
    - TK: summarizing: needs both: real and synthetic. Now the question is what questions to address, what datasets to have, etc
      - FB: have a resource with an insilico dataset and gets iteratively via crowd sourcing/community it gets evaluated
      - SK: difficult to chose which in silico dataset to use?
      - Erik: build single simulation engine for insilico dataset?
        - SK: seconds but shouldn't be only solution
      - Adam: Seconds the both options and diversity of insilico datasets

- RM: echoing TK, if one tool has enough parameters to test all the unknown. Seconds a multiparameter model via a mechanism from community inputs
- JS: engineered cell line, not hybridoma, single cell line without a gene for example would be an example
- Christopher from Amgen: delineate finite use cases/conditions and then

- Recommendations
  - What would be the Process for recommendations?
  - Erik Q: beyond providing datasets, what about evaluation process
  - TK: seconds. Do we have to have an event
    - Erik: history of such events: Assembler community (a nucleotide site, Docker?), John Burton would be reference for that tool
    - RA: seconded Docker
    - Jessica: Rosetta community, have had competitions, doesn't need winners and losers, just identify the tools to be specific for certain conditions, etc
    - JS: if another event in a year where the competitions could be held, progress made, compare analysis tools etc
  - TK: probably have in silico datasets right now that could be used to the existing tools
    - SK: seconds TK suggestion, Docker, also brought up the file format for output
    - TK: was there agreements on the input files?
      - There was discussion of fastq or VDJ annotation results (output that serves as input in another)
      - VDJ results should probably have probabilistic values in the output format
      - Jessica seconded VDJML, also seconds work with what we have
      - Lindsay: requests comments on the VDJML that was online
      - Marie Paul: individual files with lots of details are usually not downloaded by the user. Only the synthesized results.
      - FB: needs further work with working groups. Needs directions for those. Statistical methods weren't discussed. What does one mean by statistical method?
        - RH: currently mostly observing repertoire data and lacking statistical methods, p-values for inference, etc
      - TK: work from this group in determining standard datasets, etc are already helpful towards the goals of having statistical methods in the analysis processes
      - JS: What needs to be included in a paper for facilitating meta-analyses
      - FB: Microbiome, has a 100 microbiome etc. Is this needed for the repertoire?
        - RH: tools present in Microbiome to compare population 1 to population 2
        - Erik: Microbiome and TCR are similar but not to BCR
        - FB, Erik, Christopher from Amgen: Alabama group doing the repertoire of many samples but are currently funded by industry stakeholders. The data from that effort is not

currently shared and looks like to be the case for a long time to come.

- DD: Questions NIAID program officer for funding such event
- NIAID program officer: RO1 doesn't seem to be appropriate for this. Need to scope the request out and it would depend on it
- TK: seconds that and use working groups to delineate it out
- TK: funding is secondary for now
- RH: is there an appetite for 1000 genomes like project. There are many sub fields, sub sets etc
    - JS: different sites looking into flu vaccine responses (before and after)
    - Jessica from Vanderbuilt: There was a proposal for immunome like project. Looks like its going to happen. Details are being sketched out. Not sure who is spearheading (RA thinks Wayne was working on that)
    - RH: any enthusiasm for such an effort?
    - DD and Tony: The design in those efforts may not be clearly designed and specific to the needs of the ARR community. Thus need to clearly delineate and define the design for an Immunome like project for ARR.
    - Lindsay: Requests information on Repertoire 10K project
    - Adam: currently involved in the 10K project. Industry sponsors. Protected data. Post 6 months of the completion of the project. There is data for 100 individuals (TCR beta).

Outline for Workshop 3:
Ethical and legal concerns in sharing antigen-receptor repertoire (ARR) data
(Sunday afternoon)

Workshop Leaders: Tania Bubela and Bob Cook-Deegan

The sharing of ARR data presents several important issues beyond data production, analysis, and management. This Workshop will discuss and pose solutions to legal, cultural, and ethical challenges in sharing data and research materials. Challenges include variable institutional policies and practices, inadequate incentives to share, and variability in national and international laws, policies and practices for participant consent, data security and privacy, data sharing requirements, sustainability of sharing infrastructure, and commercialization.

The session will consider issues in the context of international policies and best practices in the sharing of 'omics data. It will address challenges in the context of both legacy data and data collection, and storage and access to ARR data moving forward. The latter may benefit from harmonization of policies and practices. The session will raise questions about stakeholder interests, including the research community; funders; infrastructure providers; research institutions; publishers (e.g., journals); pharma, device, diagnostic and biotech firms; and most importantly the interests of patients and publics. Considerations of the latter must include the special interests of vulnerable communities, such as the rare diseases communities and indigenous communities.

The workshop will build on existing policies and best practices for 'omics data. For each of the following topics, the discussion will:

1. Briefly outline the challenges and seek concrete examples from the ARR community;
2. Prioritize the challenges;
3. Discuss potential solutions for each problem based on national and international policies and best practices;
4. Identify key stakeholders for the development and implementation of each solution;
5. Prioritize each potential solution based on feasibility;
6. Discuss next steps in the near and long-term.

Discussion Topics

A. Creating a broad-based community culture of sharing and reporting of ARR data.
This requires:

1. Incentives for participation in community infrastructure projects, especially at the institutional level;
2. Addressing concerns over publication/research priority;
3. A realistic assessment of commercialization potential, including formal intellectual property rights.

B. Developing ethical and legal standards for collection, storage, accessibility, and use of ARR data.
Requires the development of policies and practices that are:

1. Compliant with laws relevant to the protection of 'omics research participants;
2. Appropriately balance participant's privacy interests with research interests. Special attention arises from the fact that participants may be identified from ARR data alone. Vulnerable populations have specific concerns;
3. Account for participant consent over legacy data;

4. Develop harmonized consent policies and practices for the collection, sharing and use of ARR data.

C. Developing appropriate governance structures to ensure compliance with legal and ethical policies and best practices.
Requires the development of legal structures that can:

1. Contribute to funding and long-term sustainability of data sharing infrastructure;
2. Ensure compliance with collection policies and best practices, especially for:
    a. Informed consent
    b. Protection of privacy;
3. Ensure data security for sharing infrastructure.
4. Control access to data in compliance with participant consent and privacy concerns (gatekeeper for access to data and other resources).

The workshop will close with a discussion of next steps, both short and long-term, including plans for further and expanded stakeholder engagement.

Background Resources

Deborah Mascalzoni, Edward S Dove, Yaffa Rubinstein, Hugh J S Dawkins, Anna Kole, Pauline McCormack, Simon Woods, Olaf Riess, Franz Schaefer, Hanns Lochmüller, Bartha M Knoppers and Mats Hansson, **International Charter of principles for sharing bio-specimens and data Eur J Hum Genet 23: 721-728; advance online publication, September 24, 2014; doi:10.1038/ejhg.2014.19; http://www.nature.com/ejhg/journal/v23/n6/pdf/ejhg2014197a.pdf**

Framework for Resp onsible Sharing of Genomic and Health-Related Data, Global Alliance for Genomics and Health (Sept 2014): http://genomicsandhealth.org/about-the-global-alliance/key-documents/framework-responsible-sharing-genomic-and-health-related-data

Jalayne J. Arias, Genevieve Pham-Kanter, and Eric G. Campbell, **The growth and gaps of genetic data sharing policies in the United States,** J Law Biosci (February 2015) 2 (1): 56-68 doi:10.1093/jlb/lsu032, see esp. Table 1 on data-sharing agreements

Regulatory and Ethics Working Group, Global Alliance for Genomics and Health, policy documents: http://genomicsandhealth.org/our-work/working-groups/regulatory-and-ethics-working-group/work-products . (Consent Policy, Consent Tools, Privacy and Security Policy)

For general background on Elinor Ostrom's work on the construction and maintenance of common resources:
Governing the Commons (1990)
Understanding Institutional Diversity (2005)
Understanding Knowledge as a Commons (2011)

Notes Workshop 3: Ethical and legal concerns in sharing antigen-receptor repertoire (ARR) data [Tania Bubela and Bob Cook-Deegan] (Notetaker: Chaim Schramm)

Initial presentation:

- Reasons for not sharing data: time/effort, liability, priority, etc
- Main issue for research commons is underuse – network effects enhance value (knowledge is "non-rivalrous") but "commons" is global so governance is hard
- Why a "tragedy"? Hardin only imagined two possible solutions: government control or property rights/markets. Ostrom realized shared information can lead to cooperative solutions in the real world.
- Who owns data and materials? Researcher, institutions, public, patient groups?
- Data Access-Transparent Analysis (DA-TA):
    - Acess to:
        - Personal right to access data about ourselves (interoperable)
            - Not just genomics, any lab report/value
            - As of 10/06/2014
        - Scientific replication/verification
        - Clinical interpretation
        - Algorithms as well as data
    - Analysis for:
        - Independent verification
        - Evidence-based medicine
        - Disease models/interpretive frameworks
- Bermuda principles vs. Solera (Cech report)
    - 30% increase in citations
    - 30% increase in products
    - Solera eventually deposited data publically after all
- Many possible models for agreeing to a framework and set of rules (Ft Lauderdale, OneMind, etc)
    - Movement to stop promising anonymity and to retain linked info with requisite data security
    - Possible export constraints
- Distinctive features of ARRs
    - Designed to be variable/unstable (different meaning of "reference genome")
    - Uniquely identifiable in a very different way from other parts of the genome
    - Subject to strong selective pressure over short times
    - Infrastructure developed for more stable genomic regions
- Implications of distinctive features:
    - Relevance to disease
    - String commercial potential
        - Need sophisticated IP/licensing strategies
    - Need to retain links to individuals
        - Need privacy protections

    - Forensic uses
    - Information about exposures
- Policy/practice response based on risk-benefit analysis
    - Legal Backdrop: consent, privacy and IP laws
    - Differences between data and materials need to be considered
    - May need memoranda of understanding or contractual arrangements for federated databases
    - Researchers don't have legal authority –need involvement of institutions
    - Simple agreements are probably best
- Creating a culture of sharing
    - Need incentives (attribution/citations) for participation in infrastructure projects
    - Address concerns of publication/research priority (embargoes?, data-sharing plans with funding agencies)
    - Realistic assessment of commercialization potential/protection of IP
        - Often written in as potential exemption to data sharing plan (but needs to be tightly worded)
        - IP-free zones have worked in some cases (structural genomics) to reduce friction
- Develop ethical and legal standards
    - Compliant with local laws
    - Balance privacy with research
    - Account for restrictions on legacy data
    - Develop harmonized consent policies
- Developing appropriate governance structures
    - Must contribute to funding and long-term sustainability
        - international issues (local cost/global benefit)
        - alternatively a federation of dispersed national (?) databases
            - data management concerns
    - Ensure compliance with collection policies and best practices
        - Especially informed consent and privacy protection
    - Protect data security within sharing infrastructure
    - Control access in compliance with consent and privacy

Discussion:

- How do we rewrite consent forms to allow deposit in open databases?
    - If we are starting our own repository (as a community), can write whatever consent form within legal limits
    - With existing repository, might be constrained by their requirements
    - Currently shift in practice to more broad consent for reuse
        - How do we square this with IRB requirements to specify research purpose
        - Possible example with UK BioResource
        - Need to make community aware and possibly provide template language to encourage getting broad consent ahead of time

        - GlobalAlliance is working on a lot of these issues and can possibly be a template
            - Privacy and security policy with appendix defining terms used in various countries for different levels of anonimization
            - Web resources with generic consent clauses for data sharing (?)
    - Need governing structures to foster trust, can make it easier to get broad consent (not necessarily IRB)
    - May be inevitable for “interesting” samples –better to come up with an infrastructure that can work with it
        - Nagoya Protocol and the Convention on Biological Diversity
        - dbGap: two layers of gate-keeping (IRB and NIH)
- How does a 6 month embargo differ from a 6 month delay for deposit?
    - When does clock start (generation, publication)
    - Can others use it in the meantime?
    - Currently some pushback against pure Bermuda Principles because of free-rider problems
    - Vagueness in Bermuda Principles (6 months from when exactly?) is probably inevitable in “soft law”
- What are the sanctions?
    - Not sharing in Human Genome Project data was enough of a stick for Bermuda Principles
    - Reputation within community likely a string incentive
- On over-patenting and protection of IP:
    - MTAs by default as a CYA mechanism
    - They are necessary sometimes, though, to enforce restrictions on access
- Can an IP-free zone work?
    - Sometimes we’re looking for different things (eg mutational patterns for engineering, not specific Abs)
    - Can there be protected and not-protected sequences within same database?
        - Signing up to share data doesn’t mean sharing all data or with everyone
    - Patentability (eg for bNabs) may become more difficult due to other legal requirements (non-obvious, not found in nature/adding function). Already sequence alone is probably not enough for a patent, now need to include paratope data
    - Patents don’t necessarily block creation of commons
        - Can be defensive
        - Management tool that can be handled further through MTA/contractual agreement
- What would be the goal of building our own infrastructure?
    - Can we set up a database that answers question without exposing underlying data?
    - Broker model? (If db can’t answer question, can give contact for negotiating MTA directly)
    - Also possible to “use” data on site, without necessarily allowing download

    - Can we negotiate a customized permission set with (eg) dbGap?
        - Multiple possible levels of restrictions (“zones of control”)
        - Want to minimize friction
    - Very difficult to get new infrastructure funding – need business case
        - Agency support (very difficult)
        - Fee/payment system (need industry support/partnership)
- Thinking about different zones of control:
    - Can an individual be identified from ARR? (some controversy)
        - If not, a pledge not to re-identify may be sufficient
        - Genomics now operates on the principle that anonymity of data cannot be guaranteed
    - Dealing with vulnerable populations?
        - Maybe by disease status, not just ethnicity
    - Useful to maintain identifying information/metadata for future studies and possible recontact
        - Legal regime will depend if link is to info collected at time of sampling under IRB or to other/external HER etc
- Are there privacy/security issues with submitting data to IMGT?
    - Need to think about precisely these issues when writing consent
    - Also need to consider implications of “shipping” data internationally
- What type of legal entity are we imaging for governance of a repository?
    - Non-profit
    - Absorbed within institution (dbGap is in NIH)
- Heuristics and rules
    - Heuristics refers to rules of thumb/best practices/norms
    - Rules imply that there is a consequence to not obeying
    - “Constitution” – who gets to make rules and about what

ARR Conference Wrap-up and Action Items Monday June 1
Tom Kepler facilitating (Notetaker: Felix Breden)

Tom and Jamie and Felix met together Sunday at noon, and then met with Workshop Leaders at end of Sunday's session. We all agreed to change from the published agenda (which emphasized wrapping up each of the Workshops) and instead to concentrate on several Action Items.

Thus the agenda for Monday was changed to the following agenda that had been circulated to the whole (to be be followed by writing workshop for White Paper):

New Agenda for Monday a.m.: Outcomes & Action Plans for Community Meeting

I. Minimal standards of ARR data: http://b-t.cr
- (a) recomendations for submission for journals and
- (b) requirements for deposition to database.
  - Start the list for 10 min (e.g., Interoperability for file formats (VDJML))
  - Formulate a working group
  - Group populates the list for 1 month
  - Working group tunes up the list and submits to group for comments
  - Get approval of final edited list from the group

II. Ask for use cases to be submitted as motivation for the white paper
- Start with Tom's list from Workshop 2
- Submission by the group to google doc over ?? months
- White paper group will curate the list

III. Practices for generating ARR data for commons/sharing
- Common virtual site where protocols and tools are located.
- Resources: RNA, plasmids, cell lines
- Analysis (evaluating tools)
  1. *in silico* libraries/benchmarks
  2. "real" data sets, paired read (VH & VL, TcR🚲 & TcR✔, TcR■ & TcR❤) libraries
  3. Place and where inferred GLGs are located and standards for calling them.

Sharing Data & Platforms

A. Common repository
1. Least restrictions that apply to the data type based on uses
   a. Enable industry access? And under what terms (if we set our own DB)
2. Explore existing repositories (look into dbGaP? Open and restricted access)
3. Central repository vs. federated/distributed repositories
4. Legal agreements (MTAs, DTAs, AAs, etc. among institutions)
5. Shared/harmonized consent forms (including data reuse, recontact)
   Biobanking example; partner with Global Alliance
6. Identify and include all stakeholders

A. Future meetings

B. Virtual meetings

1. Working groups
2. Seminar series

C. Shared projects (Use list)

End of new agenda for Monday am

The following are notes from Monday's discussion (note taker: Felix Breden):

I. Minimal standards for ARR data – We will not be able to hammer out all the minimum standards exactly this morning but we could get close.
Goal is to come up with a list of minimal requirements for metadata. I.e., common set of metadata that should be described in any experiment. We want to focus on metadata specific to ARR data.
The purpose of this morning is to go over the google doc. Everyone will have a chance to comment and to add/subtract
Danny: reminded us that these are both the requirements for the data commons, and recommendations for publication in journal.

Andrew: has to take into account different types of data

Lindsay: levels of annotation, that determine then the next levels of metadata

Chaim: absolute requirements have to change with types of data

Tom: goes through the various larger sections of the list – people make corrections/suggestions – E.g., Sequencing metadata = annotation metadata

Chaim: have to have a tree for samples, e.g., often have more than one sequence per sample

Rik: - Number of clonotypes can't be part of the minimum data set

Several people agreed we are looking for the minimum set of data for submission to journal

Tom: the plan is for the group to go away and make additional comments, then hand this over to a working group.

Note added July 06, 2015: go to **http://tinyurl.com/q8ou2jh** to see working list of proposed metadata.

Discussion of Working Groups in general: How do we form a working group? People could be asked now. Plus people can contact Tom, Jamie or Felix as organizers, or contact the working groups once they are started.
Possible working groups?
Metadata list Working Group: Danny Douek volunteers
Tool evaluation and resources comparison working group
Repository working group
Whitepaper working group
Process for Working groups - Choose chair, choose treasurer, secretary – do work period over a few months, come up with drafts, these get circulated, entire body can make comments.

The Minimal Standards working group (initial composition) – Uri Hershberg, Danny Douek, Marie-Paule Lefranc, Nina, Brian Corrie, Christian Chaim, Steve Kleinstein, Chris Murawsky, Florian Rubelt

Note added July 6, 2015: from google doc at **http://tinyurl.com/q8ou2jh**,
minimum standards working group consists of:

Minimal Standards Working Group: Draft minimal standards of ARR data: (a) recommendations for submission to journals, and (b) requirements for deposition to database
Members: Steve Kleinstein, Uri Hershberg, Danny Douek, Brian Corrie, Nina Luning Prak, Christian Busse, Chris Murawsky, Florian Rubelt

II. Use Cases – we worked on a list of Use Cases during the regular sessions, but we wanted to get a broader swath of use cases. These can be used to motivate the white paper, to help further the organization, (e.g., drafting of grant proposals for asking for resources), etc.
The White Paper working group will be curating the Use Cases document.
In one month please have the Use Cases submitted
There is a use cases list started as part of the Google Doc

Group addresses Sharing Data and Platforms Working Group

Felix: questions whether this working group can really tackle the job "identify and include stakeholders"
Jamie explains that this involves identifying the vulnerable populations that will have to be involved in consent –
Tom: we will be communicating with the group, so if people need to add and subtract issues they can

Working group – Holly Longstaff, Nina, Felix Breden, Adrian Thorogood, Corey Watson, Tony Moody

Note added July 6, 2015: from google doc at **http://tinyurl.com/q8ou2jh**,
Repository working group consists of:

Repository Working Group: Problems to be addressed: explore existing repositories such as dbGaP; shared/harmonized consent forms (get input from biobanks); inter-entity agreements; identify stakeholders and their roles and needs in the process

Members: Holly Longstaff, Nina Preto, Felix Breden, Brian Corrie, Adrian Thorogood, Tony Moody, Corey Watson

III. (back to following order of agenda) Practices for generating data and sharing – becomes "Tools and resources" Working group

Tom goes over the tasks listed
Tom: this group needs to look at how to determine how to infer genes from NGS
Unidentified : this group needs to decide on inter-operability
Steve K: should we split this up to biological resources versus in silico resources?
Many people in room say no, experimentalists need to talk to analysis people.

If tasks need to be split up, they can be done within the group

At Meeting, Working group: Erick Matsen, Jessica Finn, Chaim Schramm, Chris Tipton, Martin Corcoran, Steve Kleinstein, Danny Douek, Sai Reddy, Tony Moody, Davide Bagnara, David Kipling, Christian Busse
George Georgiou?

Joe Breen – would like to be a resource for this Working Group – can't be a formal member of the group – could be called ex officio – will be available as a resource

Note added July 6, 2015: from google doc at **http://tinyurl.com/q8ou2jh**,
Tools and Resources Working Group: Tools and resources for generating ARR data for commons/sharing; common virtual site where protocols and tools are located; resources (such as RNA, plasmids, cell lines etc.); Analysis (evaluating tools) such as in silico libraries/benchmarks, "real" data sets, place and where inferred GLGs are located and standards for calling them; determining how inferred germline genes are associated with analysis; common inputs/outputs and data file formats; version control!
Working document at: http://j.mp/arr-tools-resources

Members: Erick Matsen, Jessica Finn, Chaim Schramm, Chris Tipton, Martin Corcoran, Steve Kleinstein, Danny Douek, Sai Reddy, Tony Moody, Davide Bagnara, David Kipling, Christian Busse, Anna Fowler, Rion Dooley, Ramy Arnaout, Jake Galson, Johannes Trück, Martin Corcoran, George Georgiou

Final Group to form is White Paper Working Group – task is to get a first draft out – 3 facilitators, Tom Jamie and Felix, will get together this afternoon and start – anyone is invited especially other Workshop Leaders.

Future directions

Do we want to be a group/consortium/society?
Do we want to have more meetings?
Do we want to form alliances with other groups?

Danny Douek – Can we be a religious sect? Who is the messiah?

Andrew: can we work with The Antibody Society?
That would hurt the effort to connect to T-cell people

Tom: should we form a society ourselves, or should we be a type of group within a group like the TABS

Jamie: should we have another meeting in a year? Then we can evaluate whether the working groups are doing their job? What is the level of commitment? But we could flip the meeting compared to this May 2015 Community meeting – i.e., have working groups and workshops in the background, but emphasize sessions where people are talking about the tools, where the talks are more data and science focused, show how the science is progressing.

Tom: one possibility, have virtual seminars from working groups, that would morph into a new meeting committee.

Danny; in a year, have updates of where we are – e.g., we agreed to have a bunch of t-cell libraries for standards, did we do this?

Jamie: the working groups could help organize the meeting, form a Coordination Committee –

Have to approach NIH soon

Organize virtual seminar group – Chaim volunteers to help organize virtual seminar group

Should we have a consortium? Andrew: we should affiliate ourselves with another group

Jamie: we have a list of about 220 contacts, but I am sure we missed a lot of people - So we could put up the contact list on the web, and we could then ask people to add to the contact list

Jamie: should we ask for addition to the working groups, to the white paper list, etc?

Tom: need to look for signatories for the White Paper (open it up beyond the May 2015 meeting)

Jamie: meet next year, looks like it will be organized according to the working groups

Chaim: the meeting should include a half day of review of where we are

Whitepaper session – starts 10:42 in Labatt Hall in SFU Harbour Centre

Should we go for a longer paper or a short commentary?

We could go for a whitepaper with recommendations/guidelines – Genome Medicine is interested in a whitepaper

Sai: we need to be prepared to have a link to the information; if we publish something we need to have this information worked out

Tom: when the time is right we approach Nature or Nature Medicine; we tell them, we are launching the site, here is the link to the recommendations, etc.

Jamie: it would be good to get something out soonish, such as something around minimum publishing standards. We could introduce ourselves and the initiative, as well as get the minimum publishing standards out.

Tom: we need to figure out what we want to get out; we largely have the minimums worked out now, we can draft the paper about the technology and what the purpose is; we can write 75% of the paper now. We got a big group, let get something out now, we already said 1 month deadline on minimum information.

Minimum standards working group is essential for this WP (White paper).

Tom asks Lindsay: how do these set of recommendations get done? can you publish before the final set of recommemndations are agreed upon?

Lindsay: before the publication you need a period of public comment, need to have the community have an opportunity to comment.

Steve: shows link to the HIPC document; a WP for HIPC in Nature Medicine was a letter to the editor, stating “we are going to establish standards for the data; we are going to establish common resources” etc.

Danny: talking to Nature Medicine editor, he will talk to the editor about this initiative

Nina; have to have a website, announce group, announce who we are, then we need
A – letter to editor initial announcement
B – a more substantial document

Lindsay: first move is to announce the consortia; we had this meeting, we exist, we are developing standards

Rob: if we become a society, this has a legal implications

Jamie: definition of a consortium is a group of people working together.

Tony: we do need to decide on a name and get a presence on the web.

Rob: do we intend to incorporate?

Tom: let us do this informally to start with, then discuss incorporation;B ob talked about this, the need for a legal entity to enter into agreements, but not worry about this now?

Lindsay: not sure if we should be an organization of individuals or an organization of institutions; for example, only people who have an institution log on id could be part of this if it is an organization of institutions.

Tom: is our purpose to enter into agreements? then we need to be a certain type of organization.

Bob: isnt your first step to set up your minimum standards? Don’t need to be a legal organization to recommend a set of minimal standards.

Tony: why do we need a society/consortium? Why isn’t this just genomics, why don’t we fit in to other established groups? Why are we different?

Steve: one possible society is the functional genomics data society; are there other groups, whose toes we could step upon?

Tom: I don’t think it is a difficult task to show that we are unique and we present unique new problems

Tony: We should draft something, get people to look at it, and if they don’t get it, we need to refine that.

Tom: we need to get writing to define why we are different

Steve: let’s put Tom’s white paper start in the google doc

There was lots of agreement that Tom’s start is good

Steve: need to be more explicit about problems

Jamie: are we Immune Repertoire or Antigen Repertoire? ARR (antigen receptor repertoire more exact, but Immune receptor repertoire more catchy, and we might want to include KIR and toll-like receptors in the future.

Jamie: the uniqueness is the somatic recombination

Sai: Adaptive immune Repertoire – that is unique

Continue on Working on details on Letter to Editor – break for lunch at noon

White Paper Session starts up again at 1:30 pm in Board room in SFU Harbour Centre

Jamie: we need outreach for next meeting, funding for next meeting, where is next meeting?

Danny – Office of AIDS Research – good contacts for OAR
Danny will talk to OAR
Danny talked to Joe Breen – meeting space at NIH -
Summarize meeting and getting Joseph looking into meeting
Rob and Genome Canada meeting – Tania Bubela – IP access, open sharing, co-funding – contacts for letters for her Network –
Adaptive – Felix will help get letters for Tania's grant proposal

For Paper – Intro, Community building

Discuss long time: what are the incentives for sharing…valication
Keeping up with the field
Learning from others
Tania – make the compliance for publication easier – stick easier
Jamie – pride in their data, responsibility, community virtues
Tom – responsibility to human subjects, global health, vulnerable communites

Discuss for a long time – Use Cases

Tom mentions – a physician wants to use deep sequencing in clinical comparison, needs to have validated tools and sequencing protocols to get there

Tania – use cases, what is the value of sharing the data, go back to medical reasons

Tony – by increase number of controls, increase power of test

Steve – another microarray consortium – make quality control tools available, Maxi consortium – Steve will send this out –

Tom – have to do due diligence, to see other initiatives that are doing similar initiatives

Paragraphs for Commentary:

Front end – Tania and Tom and Bob

Short history - Felix and Jamie

Shared repository - Corey and Felix and Jamie

Minimal standards – steve and brian Corrie and Marie-Pule

use case HIV – Tony and Tom

use case TIL – Danny

use case Humanization – Sai

Resources – Sai and George

Ethics etc. Tania and Bob produced 2 paragraphs already

Summary of Working Groups (notes added July 9 2015 from http://tinyurl.com/q8ou2jh)

Minimal Standards Working Group: Draft minimal standards of ARR data: (a) recommendations for submission to journals, and (b) requirements for deposition to database
Members: Steve Kleinstein, Uri Hershberg, Danny Douek, Brian Corrie, Nina Luning Prak, Christian Busse, Chris Murawsky, Florian Rubelt

Tools and Resources Working Group: Tools and resources for generating ARR data for commons/sharing; common virtual site where protocols and tools are located; resources (such as RNA, plasmids, cell lines etc.); Analysis (evaluating tools) such as in silico libraries/benchmarks, "real" data sets, place and where inferred GLGs are located and standards for calling them; determining how inferred germline genes are associated with analysis; common inputs/outputs and data file formats; version control!
Working document at: http://j.mp/arr-tools-resources

Members: Erick Matsen, Jessica Finn, Chaim Schramm, Chris Tipton, Martin Corcoran, Steve Kleinstein, Danny Douek, Sai Reddy, Tony Moody, Davide Bagnara, David Kipling, Christian Busse, Anna Fowler, Rion Dooley, Ramy Arnaout, Jake Galson, Johannes Trück

Repository Working Group: Problems to be addressed: explore existing repositories such as dbGaP; shared/harmonized consent forms (get input from biobanks); inter-entity agreements; identify stakeholders and their roles and needs in the process

Members: Holly Longstaff, Nina Preto, Felix Breden, Brian Corrie, Adrian Thorogood, Tony Moody, Corey Watson

White Paper Working Group: Handles use cases document once list is closed; task of drafting white paper will be circulated to the rest of the group;

Members: Felix Breden, Thomas Kepler, Jamie Scott

Organizing Committee: have a representative from each working group who serves as a point of contact; organizing virtual and face-to-face meetings; coordinate the dissemination of information from the working groups; founding documents of our group; outreach to current and potential members?

Members: Chaim Schramm, Felix Breden, Jamie Scott, Thomas Kepler, Danny Douek