Questions posed during Prof. Victor Greiff's webcast:

Steps in data processing and analysis of adaptive immune receptor repertoires: best practices, pitfalls, and future directions.

Broadcast date: April 6, 2021.

Disclaimer: Some answers given here may differ from those given during the webcast. If so, please regard the answers given here as more complete and accurate. Questions may have been altered so that acronyms are spelled out on first use in this document.

Q1: Sometimes multiple V genes are identified in B cells upon sequencing a cultured B-cell line. Is that a technical mistake or is it normal?
A1: It depends on how the cell line was produced. It seems as if in most of the cases, multiple different B cells are used to produce a given cell line. Therefore, finding different V genes may be normal.

Q2: Is the clonal overlap small also after the immune system has been challenged? (e.g., vaccination or in chronic inflammation)
A2: The overlap depends on the antigen given. The overlap may even be higher than in the naive repertoire (see for example, Greiff et al., Cell Reports 2017).

Q3: What is the best method to completely avoid next-generation sequencing mistakes in the sequencing of massive repertoires? Strict forward/reverse merging of completely overlapping reads? Massive oversampling and throwing everything with low n away?
A3: I think it cannot be avoided completely. But yes, oversampling, Unique Molecular Identifier (UMI) correction and computational filtering, and error correction (e.g., via mixcr) should go a long way. Unfortunately, a standardized approach to error correction does not exist yet, as far as I know.

Q4: What seq-depth is recommended if using UMI?
A4: My general rule is to plan for 5-10 times more reads than the number of cells you input. And for at least 10 times more than that if you are planning to use UMIs. This means 100 times more reads than you input cells. I cannot cite a paper for this, but I think it is a good rule to get you started. Prior to any big experiment, this should be tested with representative samples to make sure one is operating with the right sampling depth.

Q5: What can be the reason for undersampling of UMIs?
A5: UMIs tag each RNA/cDNA molecule. Therefore, a high diversity of UMIs is needed – requiring, in turn, a high sequencing depth to reach the criterion of at least three reads with the same UMI for consensus read building.

Q6: Knowing that only a few software packages are Adaptive Immune Receptor Repertoire (AIRR) Community-compliant, a lot of the work in the field is done with custom programs. In this case, what do you recommend for code sharing and reproducibility? How much should we openly share?
A6: It is true that only a few software packages so far are *certified* AIRR-compliant (https://docs.airr-community.org/en/stable/swtools/airr_swtools_compliant.html). However, many more tools than those that are certified readily allow for AIRR-compliant data input, etc. (https://github.com/airr-community/airr-standards/blob/73e9f4a1829596980ae4ac412cda9e0213784c80/docs/resources/tables/rearrangement_support.tsv). Furthermore, independent of AIRR-compliance, I think most software in the AIRR domain is at least to some degree accessible via various repositories. More generally, I think sharing enables faster progress.

Q7: What is the importance of the sequencing length except from the sequencing depth in B-cell receptor (BCR) analysis? Also, which tool would you recommend for the VDJ alignment (e.g., VDJpuzzle, Basic,BraCeR). Thank you!
A7: Sequencing length may enable the coverage of some portion of the constant region enabling demultiplexing of different isotypes/subclasses.

Q8: If the absence of clonal overlap is a sequencing-based issue, is there a reliable machine-learning model to solve this issue?
A8: No, I don't think there is. The lack of overlap across individuals is also biological – due to the high diversity of immune repertoires. To address this issue and to be able to compare samples, several approaches such as k-mer encoding and diversity profiles (see for example Greiff Trends in Immunology 2015 and Miho, Frontiers in Immunology 2018) can be used.

Q9: Also, what overlap percentage should we expect between mice (in the case of antigen specific B cells)?
A9: See Greiff et al. Cell Reports 2017.

Q10: Can you comment on V gene polymorphisms in mice?
A10: Within one mouse strain, by definition, ideally there should not exist any polymorphisms. Across strains, germline gene repertoires may vary (https://royalsocietypublishing.org/doi/full/10.1098/rstb.2014.0236).

Q11: Is there are special distance matrix for Hamming distance calculation for BCRs, because mutations that appear due to hypermutation seems to be biased?
A11: Have a look at these papers: https://www.frontiersin.org/articles/10.3389/fimmu.2018.02149/full and https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008391 (e.g., scaled BLOSUM).

Q12: Do you consider clonal diversity based on vh-vl pairing or only focusing on vh?

A12: Depends on the data. In case VH-VL data is available, paired data should be used for clonal definition. That said, it remains a challenge to define clones based on paired data.

Q13: UMI correction + UMI decontamination + MIXCR correction is a standard pipeline for BCR/T-cell receptor (TCR) data in the Chudakov lab for the last 5 years, FYI. There is a number of papers and publicly available datasets, both human and murine data.

A13: Yes. I think what I meant during my webinar is that a more general analysis of the pros and cons of UMI vs computational error correction is missing. As the Barennes et al. Nat Biotechnol 2021 study (doi: 10.1038/s41587-020-0656-3) had clearly shown, UMI may lead to drastic diversity reduction, which may impact the biological conclusiveness of a study.

Q14: For single-cell TCR seq for instance, some just run packages such as scRepertoire with the standard cellranger outputs without corrections you mentioned. Is it a bad practice? Thanks for very insightful talk!

A14: I think in general it is advisable to be critical vis-à-vis the data obtained and to work with strict quality cutoffs.

Q15: How do you take into account naive B cells when you filter out singletons from data? Probably, a lot of naive B cells might be singletons.

A15: Not if you have sufficient sequencing depth. Independently of the number of cells per clone that are in a given sample, each of these clones, with sufficient oversampling, would be attributed multiple reads – thus, it would not be a singleton. So, the ideal dataset is the one that does not have singletons. [Here, I define by singleton as complementarity-determining region 3s (CDR3s) or V-CDR3J that have only one sequencing read].

Q16: Do you have any recommendations for acquiring peripheral blood mononuclear cell (PBMC) samples for deep repertoire sequencing? Are there any pitfalls to be aware of?

A16: PBMC samples are generally tricky since one doesn't know the exact number of T cells/B cells one is working with (it depends on the research question though whether this is a drawback or not). I am not sure if I fully understand the question. Please feel free to follow up via email.

Q17: Wonderful talk! How do you expect the problem of lack of clonal overlap to eventually be solved in the field? Better seq tech, limit sampling to diseased tissue over PBMCs, other options?

A17: We expect there to be only small clonal overlap across individuals due to the sequence diversity. That said, in order to experimentally determine the true extent of clonal overlap, oversampling and cell and tissue-based sampling will go a long way.

Q18: Could you comment on long read sequencing technologies for BCR? Are they advanced enough to cover the repertoire accurately?
A18: Long-read sequencing is generally not widely used due to the required throughput. However, it is increasingly being used for specialized applications such as this one here: https://www.frontiersin.org/articles/10.3389/fimmu.2020.02136/full#h3

Q19:Awesome talk! In Libra-Seq, does a high Libra seq score mean higher affinity to that antibody?
A19: Based on the Libra-seq paper (Setliff et al., 2019, Cell): "For each cell, the LIBRA-seq scores for each antigen in the screening library were computed as a function of the number of unique molecular identifiers (UMIs) for the respective antigen barcode (STAR Methods)."

Q20: What is the antibody frequency and how it is computed in a repertoire?
A20: Sorry for not being clear. In my webinar, I defined frequency as the relative number of reads that map to a certain clone (e.g., defined by CDR3).

Q21: What are some good ways to visualize repertoires?
A21: It depends on the research question: e.g., for diversity (diversity profiles), for similarity (network), for phylogenetics (trees). See the respective sections of my webinar for references for inspiration. The On Demand version is found here: https://webinars-antibodysociety.org/store/seminar/seminar.php?seminar=167291

Q22: Is there a software pipeline for the processing of amplicon sequencing data from raw data to table of counts?
A22: Yes for example, mixcr or the immcantation suite.

Q23: What is your favourite way to visualize TCR repertoire networks?
A23: Igraph/networkx for building graphs and their analysis, but for pure visualization cytoscape is clearly superior to igraph/networkx.

Q24: Thanks for a great overview! Would you recommend the best go-to resource for single-cell TCR analysis (esp. in combination with transcriptome analysis) for beginners? Thank you!
A24: Have a look at this tutorial: https://icbi-lab.github.io/scirpy/tutorials/tutorial_3k_tcr.html. Overall, not many tutorials are out there yet for this specific application.

Q25: Awesome talk! Is there a known mechanism explaining why some AIRR degree distributions generate power laws? Maybe somatic hypermutation (plasma-cells) or perhaps even sequencing errors?
A25: Yes, in B cells certainly somatic hypermutation contributes. Overall, I think this is linked to generation probabilities and VDJ recombination statistics. More details on this can be found here: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000314

Q26: In human studies, like you mentioned, we don't usually have high n to really use sex as a variable, would you comment on how to adjust for sex differences? Is there a set of sex-related genes used in TCR analysis? Thank you!
A26: Very tough – it's best to adjust for differences in the study design.

Q27: Is there scope for moving forward with this kind of computational learning to the structural 3D level?
A27: Yes, we briefly touched on this here: https://www.sciencedirect.com/science/article/pii/S2452310020300524. But, please also look at these excellent review articles: https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbz095/5581643, https://www.jbc.org/article/S0021-9258(17)48926-5/abstract and https://www.mdpi.com/2073-4468/9/2/12.

Q28: Regarding the antibody-antigen prediction field, can some kind of metadata be used with a motif vocabulary to improve antibody-antigen prediction ? Thank you.
A28: If we define metadata as physico-chemical properties (PCP), yes, PCP may be incorporated into the prediction process but I think it remains to be conclusively shown that they really help with prediction accuracy. This deserves further attention.

Q29: Is the almost exclusive focus on CDR3s due to short read lengths problematic for network analysis and machine learning approaches?
A29: The focus on CDR3 for network analysis is because most of the variability is located there. Any two sequences are most likely to differ in the CDR3. For BCRs, the entire sequence may be taken into account – to account for somatic hypermutation.

Q30: Just to confirm your point, Professor Greiff: can we take as a fact that both "PCR error correction mechanisms" and "higher sequencing depths" are inevitably necessary criteria for minimally representative repertoire data? And if it's true, do you see any remaining utility for non-corrected data, or corrected-but-with-less-depth data?
A30: Yes, I would agree with your statement. I think one can still try to work with non-optimal data (no data is perfect) but one should be very careful with drawing big conclusions. For any repertoire study, the study design is crucial (as for any study). It is nearly impossible to correct bad data after the fact.

Q31: What about using Deep Learning techniques?
A31: Please have a look at these recent reviews: https://www.sciencedirect.com/science/article/pii/S2452310020300524 and https://www.mdpi.com/2073-4468/9/2/12.

Q32: Can we use immuneML platform to implement some immune-based solution (e.g., Artificial immune systems) to a non-immune dataset?
A32: I would tend to say no. immuneML has been built with immune receptor biology and related research questions in mind.