Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Immunology

Recent progress in the analysis of αβ T cell and B cell receptor repertoires

Mark M Davis[1,2,3] and Scott D Boyd[1,4,5]

T cell receptors (TCRs) and B cell receptors (BCRs) are vertebrate evolution's best answer to the threat of microbial pathogens that can evolve much faster than ourselves. These antigen receptors are generated during T cell or B cell development by combinatorial rearrangement of germline genome V, D and J gene segments, and with junctional residues capable of enormous diversity. For decades the complexity of these receptor repertoires has limited their analysis, but advances in DNA sequencing technology and an array of complementary tools have now made their study much more tractable, filling a major gap in our ability to understand immunology as a system. Here, we summarize the recent approaches and discoveries that are enabling these advances, with some suggestions as to what may lie ahead.

Addresses
[1] Institute for Immunity, Transplantation, and Infection, Stanford University School of Medicine, Stanford, CA, USA
[2] Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA
[3] The Howard Hughes Medical Institute, Chevy Chase, MD, USA
[4] The Sean N. Parker Center for Allergy and Asthma Research at Stanford University, Stanford, CA, USA
[5] Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA

Corresponding author: Davis, Mark M (mmdavis@stanford.edu)

Current Opinion in Immunology 2019, **59**:109–114

This review comes from a themed issue on **Special section on human immunology**

Edited by **Federica Sallusto**

For a complete overview see the Issue and the Editorial

Available online 18th July 2019

https://doi.org/10.1016/j.coi.2019.05.012

The αβ TCR repertoire

Ever since T cells were recognized as having diverse specificities in the 1970s, there have been efforts to understand their response repertoire, since that has much to tell us about their capabilities. But due to the limits of technology, these efforts could only estimate the overall diversity based on simple assumptions about the recombining units [1] or calculating the naïve frequency of a particular antigen specificity [2].

The modern era of repertoire studies began with the use of next generation DNA sequencing technologies by Robins and others to sequence hundreds of thousands of TCRs in a single experiment from populations of T cells in the blood [3,4]. These data confirmed the extreme complexity of TCRβ sequences, estimating approximately 5×10^{11} amino acid sequences but that only 3×10^{6} of these possibilities were present in the average individual [3]. Subsequent work has delineated pathogen-specific repertoires [5]. A number of interesting results have come from these early TCR sequencing studies, such as the common origins of memory resident versus peripheral $CD8^{+}$ T cells [6], the reduced TCR repertoire in older individuals [7] and the discovery that T cell clonal expansion and functional attributes are very broad at first in the context of a pathogen response, but then narrow to predominantly one cell type as the disease progresses [8].

But taking full advantage of the power of T cell repertoire analysis requires obtaining both chains of the TCR heterodimer from single cells, and Han *et al.* [9] developed the first relatively high throughput method, where hundreds of TCR pairs could be obtained in one set of experiments and this could be quickly built up to thousands. An alternative was also developed by Howie *et al.* [10] using a dilution strategy. These pairs of TCRs could then be transfected into reporter T cells and challenged with candidate antigens [11••] or expressed in a soluble form and used to screen yeast libraries that Garcia *et al.* have used to display up to one billion variant peptides bound to a given MHC molecule [12,13•]. This later approach is more arduous but is necessary when it is not clear what antigens to test. Other general screening methods have also been developed [14•], for more focused searches, such as surface molecules [14•] or cDNAs from particular cell types [15•].

It is important to recognize that there are two distinct types of T cell repertoire to consider. One is the TCR sequence repertoire, which is vast and consists of all possible sequences of TCR α and β, and their potential pairings. But the other, often more biologically relevant, repertoire is the ligand repertoire, which is ~1000× smaller. For the former repertoire, if every TCR α could pair with every TCR β, which is unlikely, one could imagine an upper limit for αβ TCR diversity of 10^{15}, far more than the number of T cells in a human being (2×10^{11}) or a mouse (3×10^{8}). This explains why even monozygotic twins have very few identical TCR

sequences, even though they have the same HLA alleles and are often recognizing the same antigens.

But what about the ligand repertoire? Here the best numbers come from studies using peptide-MHC tetramers [16] or higher order multimers. Using the advanced enrichment techniques used by Jenkins [17], where frequencies of approximately one in one hundred thousand and one in one million are seen in humans [18,19], and are only modestly less rare in mice [17], indicating an intrinsic limitation, possibly positive selection in the thymus. Multiple TCR motifs can be seen binding to the same peptide-MHC tetramer (an average of five noted by Glanville *et al.* [11••]). It seems that the ligand repertoire for a given MHC allele might easily be in the low millions. This is a much more manageable figure for most purposes, although given the thousands of MHC alleles in human beings, it still covers a lot of ground! But how to convert the relatively easy to obtain sequence data on TCRs-thousands to millions-into the much more biologically relevant ligand information typically desired? Here two bioinformatic approaches have been developed that can accomplish this task. One is the program GLIPH (Grouping of Lymphocyte Interactions by Paratope Hotspots, [11••]) which used tetramer gathered T cell sequence data to identify short 3–4 aa motifs and other parameters as key determinants of specificity, presaged in part by the kmer work of Chain *et al.* [20] and the extensive Tdist analysis of Dash *et al.* [21••]. These programs take large raw TCR sequence datasets, alpha or beta, or both, and use them to produce clusters of TCRs which share the same peptide-MHC specificity, and also to predict the relevant MHC allele, where the input population has sufficient diversity.

The validation of these results indicates that one can use these programs to interrogate any collection of TCR sequences and reduce at least a fraction of the sequences into meaningful specificity groups. These groups in turn can be queried for whether they are important in any given biological or clinical situation. For example, one can ask whether a cohort of individuals who respond successfully to a particular vaccine differs significantly in their T cell repertoire response from those who do not. Given the same distribution of MHC alleles in the cohort, one might find that the diversity of the responders TCR repertoire is significantly greater, or that specific TCRs are expressed in one group or the other. In either case, one does not have to know what the antigen or MHC is to see what a critical trend might be. This is critical, since as the data sets get larger and larger, you can be dealing with thousands of specificity groups, with no hope of finding ligands for more than a fraction.

Progress should also be noted in the use of peptide MHC multimers to analyze T cell responses to many different ligands at once. Here the development of combinatorial labeling strategies, first using fluorescent probes [22,23] and more recently with metal labels and mass cytometry [24,25] or with oligonucleotide barcodes [26] enables hundreds of different T cell specificities to be analyzed simultaneously, together with detailed phenotyping.

In TCR analyses, there has been frequent speculation about 'public' versus 'private' TCRs, where public refers to TCR identities between individuals, either at the DNA or protein sequence levels. A recent paper by DeWitt *et al.* [27], analyzes 80 million TCR sequences from over 660 people, and finds that ~14% are 'public'. While some of these might represent TCRs that have been repurposed for important stereotypical interactions, as the γδ TCRs in the dendritic epithelial cells in the skin of mice [28], most are probably just the result of chance and less complex rearrangements, as shown by Ethanati *et al.* [29••], although they do contain many pathogen related sequences [21••].

For a more extensive review of TCR repertoire work than space allows here, please see Bradley and Thomas [30•].

BCR and antibody analysis

High-throughput DNA sequencing (HTS) methods now enable the analysis of millions of Ig gene rearrangements in a single experiment. Early efforts were applied to the analysis of malignant B cell populations, zebrafish antibodies, healthy donor Ig repertoires and phage display libraries [31–34] and then a wide range of autoimmune, infectious disease, immunodeficiency, and vaccination conditions. Scaled-up sequencing of physically joined Ig heavy and light chain cDNA from single cells followed, enabling synthesis and characterization of native monoclonal antibodies [35,36]. More recently, BCRs have been analyzed from single cell RNASeq data [37], or enriched Ig-containing cDNA libraries, analogous to published TCR approaches [38]. A new addition to this toolbox is DNA barcoding of reagent monoclonal antibodies for protein phenotyping on single cells via DNA sequencing of the antibody tags [39,40]. Similar methods could label antigens and enable highly multiplexed analysis of B cell antigen specificity and BCR sequences. New analysis algorithms have been developed to address the wealth of new data, with tradeoffs between their ability to handle very large datasets, their sophistication of modeling the structure of Ig germline genes and rearrangements, and their applicability to single-cell data [37,41–45]. Equally important are new initiatives formulating data standards, data structures and mechanisms to improve quality assurance and the ability to share and reanalyze Ig sequence data [46–48]. In parallel with the genetic tools, improved protein fragment analysis using mass spectrometry has enabled comparison of serum antibody proteins and BCR gene sequences in an individual, to characterize dominant clones in influenza vaccine responses [49,50].

The published literature using these new methods includes many proof-of-concept papers featuring relatively few subjects. Looking ahead, studies of larger cohorts will be needed to characterize the variation associated with population groups, age, sex, and health status. Analysis of tissues other than blood will also expand understanding of human immune responses, following the lead of rhesus macaque studies using serial fine-needle aspiration of lymph nodes to study germinal center B cell reactions after vaccination [51].

BCR repertoire formation

Human Ig germline loci (heavy, kappa, and lambda on chromosomes 14, 2, and 22, respectively) are poorly characterized in most populations. These repetitive loci defy conventional mapping or assembly, and conceal large-scale insertions, deletions, and complex rearrangements. Recent sequencing of functionally haploid hydatidiform mole cell lines has begun to address these deficiencies [52,53]. Unsurprisingly, few genome-wide association studies (GWAS) have examined linkage between Ig locus variation and immunological diseases or phenotypes. A recent study of rheumatic heart disease due to *Streptococcus pyogenes* infection in Oceanic populations showed the potential for such associations, finding disease-susceptibility in subjects with the *02 allele of heavy chain gene segment IGHV4-61 [54•]. Prior work showed that IGHV1-69 alleles with phenylalanine at position 54 preceded by a hydrophobic residue at position 53 in CDR-H2, are preferentially used in broadly neutralizing antibodies that bind the influenza hemagglutinin stalk [53]. Thorough characterization of Ig loci in diverse populations will enable the search for other disease or therapeutic response associations.

It is currently unclear how diverse the Ig repertoire needs to be to provide competent humoral immunity over a human lifespan. An estimate of 10^7–10^8 IgH clonotypes and 10^{16}–10^{18} potential heavy-light chain pairs in healthy donor blood B cells was recently published, based on HTS from multiple biological replicates per individual, as previously described for TCRβ measurements [55,56]. Non-circulating B cells in secondary lymphoid organs, the gastrointestinal tract and other sites could further increase these estimates. Key questions include whether the Ig repertoire diversity decreases in old age, as reported for the TCRβ repertoire, and whether decreased Ig diversity is associated with poorer vaccine responses [56].

Selection

Naïve B cell repertoires in human infants encounter an onslaught of antigens from microbiota, foods, vaccinations, and pathogens, triggering clonal expansion, isotype switching, somatic mutation, and differentiation to memory, plasma cell and other fates. We recently used IgH HTS to analyze B cell maturation in the Stanford STORK birth cohort of infants and young children with small-volume blood samples from years 1, 2, and 3 of life [57•]. Correlations with environmental or pathogen exposures, and clinical symptoms showed that infants with disrupted skin barriers due to eczema had increased IgE somatic mutation (SHM), while upper respiratory infection was associated with increased IgM and IgD SHM. Antigen exposure contributes to the development, or prevention, of peanut allergy [58]; increased SHM in IgE-expressing B cells could reflect sensitization occurring via the disrupted skin barrier. Further analysis of antigen-specific B cells will be needed to clarify other childhood immunological events, such as the effects of the types of initial influenza virus infections on responses in adulthood [59].

Despite the diversity of the Ig repertoire, HTS has identified highly similar sequences in people exposed to the same antigens, such as Dengue virus [60], influenza [61,62], *Haemophilus influenzae* type b, tetanus [63], and HIV [64]. Strikingly, patients infected with *Plasmodium falciparum* exhibit a novel kind of convergent antibody, with portions of the LAIR1 gene (from chromosome 19) inserted between IGHV and IGHD-J gene segments, or into the switch region downstream from IGHJ gene segments [65••]. LAIR1 insertions confer binding to the RIFIN surface antigens of the malaria parasite. Although an individual's humoral responses consist mostly of private clones, convergent clonotypes could be diagnostically or prognostically useful by revealing prior antigen exposures. Ongoing work will add to reference databases of convergent Ig sequences, and test any clinical associations.

Conclusion

These new tools for characterizing TCR and BCR repertoires and functions are already populating the literature and databases such as VDJB at an exponential rate, increasing our systems-level understanding of immunological diseases, vaccinations and other interventions. These data will continue to expand our ability to discern broad trends and common patterns in these types of immune responses, as well as potential interactions, such as pathogen-specific TCR motifs in autoimmune repertoires, as already noted by Bradley *et al.* [26].

Conflict of interest statement

Nothing declared.

Acknowledgement

MMD is grateful for support from the Howard Hughes Medical Institute, USA.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Davis MM, Bjorkman PJ: **T-cell antigen receptor genes and T-cell recognition**. *Nature* 1988, **334**:395-402.

2. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P: **A direct estimate of the human alphabeta T cell receptor diversity**. *Science* 1999, **286**:958-961.

3. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS: **Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells**. *Blood* 2009, **114**:4099-4107.

4. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, Carlson CS, Warren EH: **Overlap and effective size of the human CD8+ T cell receptor repertoire**. *Sci Transl Med* 2010, **2**:47ra.

5. Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, Desmarais C, Klinger M, Carlson CS, Hansen JA *et al.*: **Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire**. *Nat Genet* 2017, **49**:659-665.

6. Gaide O, Emerson RO, Jiang X, Gulati N, Nizza S, Desmarais C, Robins H, Krueger JG, Clark RA, Kupper TS: **Common clonal origin of central and resident memory T cells following skin immunization**. *Nat Med* 2015, **21**:647-653.

7. Goronzy JJ, Qi Q, Olshen RA, Weyand CM: **High-throughput sequencing insights into T-cell receptor repertoire diversity in aging**. *Genome Med* 2015, **7**:117.

8. Becattini S, Latorre D, Mele F, Foglierini M, De Gregorio C, Cassotta A, Fernandez B, Kelderman S, Schumacher TN, Corti D *et al.*: **T cell immunity. Functional heterogeneity of human memory CD4(+) T cell clones primed by pathogens or vaccines**. *Science* 2015, **347**:400-406.

9. Han A, Glanville J, Hansmann L, Davis MM: **Linking T-cell receptor sequence to functional phenotype at the single-cell level**. *Nat Biotechnol* 2014, **32**:684-692.

10. Howie B, Sherwood AM, Berkebile AD, Berka J, Emerson RO, Williamson DW, Kirsch I, Vignali M, Rieder MJ, Carlson CS *et al.*: **High-throughput pairing of T cell receptor alpha and beta sequences**. *Sci Transl Med* 2015, **7**:301ra131.

11. •• Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, Ji X, Han A, Krams SM, Pettus C *et al.*: **Identifying specificity groups in the T cell receptor repertoire**. *Nature* 2017, **547**:94-98.
GLIPH, a computational method to cluster TCR sequences according to their likely peptide-MHC specificity. See also Dash [20]. GLIPH, a CDR3 motif driven algorithm to group raw TCR sequences into likely specificity peptide-MHC specificity groups.

12. Birnbaum ME, Mendoza JL, Sethi DK, Dong S, Glanville J, Dobbins J, Ozkan E, Davis MM, Wucherpfennig KW, Garcia KC: **Deconstructing the peptide-MHC specificity of T cell recognition**. *Cell* 2014, **157**:1073-1087.

13. • Gee MH, Han A, Lofgren SM, Beausang JF, Mendoza JL, Birnbaum ME, Bethune MT, Fischer S, Yang X, Gomez-Eerland R *et al.*: **Antigen identification for orphan T cell receptors expressed on tumor-infiltrating lymphocytes**. *Cell* 2018, **172**:549-563 e516.
Refs. [12] and [13•] describe a very powerful yeast display library technology to find peptide-MHC ligands for TCRs of interest.

14. • Joglekar AV, Leonard MT, Jeppson JD, Swift M, Li G, Wong S, Peng S, Zaretsky JM, Heath JR, Ribas A *et al.*: **T cell antigen discovery via signaling and antigen-presenting bifunctional receptors**. *Nat Methods* 2019, **16**:191-198.
An alternative approach to ligand discovery using oligonucleotide tagged tetramers.

15. • Kisielow J, Oberman F-J, Kopf M: **Decifering CD4$^+$ T cell specificity using novel MHC_TCR chimeric receptors**. *Nat Immunol* 2019, **20**:652-662.
A method to focus TCR ligand discovery efforts on the genes expressed by a particular cell type.

16. Altman JD, Moss PA, Goulder PJ, Barouch DH, McHeyzer-Williams MG, Bell JI, McMichael AJ, Davis MM: **Phenotypic analysis of antigen-specific T lymphocytes**. *Science* 1996, **274**:94-96.

17. Moon JJ, Chu HH, Pepper M, McSorley SJ, Jameson SC, Kedl RM, Jenkins MK: **Naive CD4(+) T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude**. *Immunity* 2007, **27**:203-213.

18. Yu W, Jiang N, Ebert PJ, Kidd BA, Muller S, Lund PJ, Juang J, Adachi K, Tse T, Birnbaum ME *et al.*: **Clonal deletion prunes but does not eliminate self-specific alphabeta CD8(+) T lymphocytes**. *Immunity* 2015, **42**:929-941.

19. Su LF, Kidd BA, Han A, Kotzin JJ, Davis MM: **Virus-specific CD4 (+) memory-phenotype T cells are abundant in unexposed adults**. *Immunity* 2013, **38**:373-383.

20. Thomas N, Best K, Cinelli M, Reich-Zeliger S, Gal H, Shifrut E, Madi A, Friedman N, Shawe-Taylor J, Chain B: **Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence**. *Bioinformatics* 2014, **30**:3181-3188.

21. •• Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB, Nguyen THO, Kedzierska K *et al.*: **Quantifiable predictive features define epitope-specific T cell receptor repertoires**. *Nature* 2017, **547**:89-93.
Tdist, a statistical method to cluster TCRs into specificity groups from raw sequence data. See also Ref. [11].

22. Newell EW, Klein LO, Yu W, Davis MM: **Simultaneous detection of many T-cell specificities using combinatorial tetramer staining**. *Nat Methods* 2009, **6**:497-499.

23. de Visser KE, Cordaro TA, Kioussis D, Haanen JB, Schumacher TN, Kruisbeek AM: **Tracing and characterization of the low-avidity self-specific T cell repertoire**. *Eur J Immunol* 2000, **30**:1458-1468.

24. Newell EW, Sigal N, Nair N, Kidd BA, Greenberg HB, Davis MM: **Combinatorial tetramer staining and mass cytometry analysis facilitate T-cell epitope mapping and characterization**. *Nat Biotechnol* 2013, **31**:623-629.

25. Cheng Y, Zhu YO, Becht E, Aw P, Chen J, Poidinger M, de Sessions PF, Hibberd ML, Bertoletti A, Lim SG *et al.*: **Multifactorial heterogeneity of virus-specific T cells and association with the progression of human chronic hepatitis B infection**. *Sci Immunol* 2019, **4**.

26. Kwong GA, Radu CG, Hwang K, Shu CJ, Ma C, Koya RC, Comin-Anduix B, Hadrup SR, Bailey RC, Witte ON *et al.*: **Modular nucleic acid assembled p/MHC microarrays for multiplexed sorting of antigen-specific T cells**. *J Am Chem Soc* 2009, **131**:9695-9703 Oligonucleotide tagged tetramers.

27. DeWitt WS 3rd, Smith A, Schoch G, Hansen JA, Matsen FAT, Bradley P: **Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity**. *eLife* 2018, **7**.

28. Allison JP, Havran WL: **The immunobiology of T cells with invariant gamma delta antigen receptors**. *Annu Rev Immunol* 1991, **9**:679-705.

29. •• Elhanati Y, Sethna Z, Callan CG Jr, Mora T, Walczak AM: **Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination**. *Immunol Rev* 2018, **284**:167-179.
Statistical analysis of TCR recombination showing that many 'public' specificities are the result of coincidental joining events.

30. • Bradley P, Thomas PG: **Using T cell receptor repertoires to understand the principles of adaptive immune recognition**. *Annu Rev Immunol* 2019, **37**:547-570.
A much more extensive review of TCR repertoire analysis.

31. Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, Follows GA, Green AR, Futreal PA, Stratton MR: **Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing**. *Proc Natl Acad Sci U S A* 2008, **105**:13081-13086.

32. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD *et al.*: **Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing**. *Sci Transl Med* 2009, **1**:12ra.

33. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, Ni I, Mei L, Sundar PD, Day GM *et al.*: **Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire**. *Proc Natl Acad Sci U S A* 2009, **106**:20216-20221.

34. Weinstein JA, Jiang N, White RA 3rd, Fisher DS, Quake SR: **High-throughput sequencing of the zebrafish antibody repertoire**. *Science* 2009, **324**:807-810.

35. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, Varadarajan N, Giesecke C, Dorner T, Andrews SF *et al.*: **High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire**. *Nat Biotechnol* 2013, **31**:166-169.

36. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, Kuroda D, Ellington AD, Ippolito GC, Gray JJ *et al.*: **Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires**. *Proc Natl Acad Sci U S A* 2016, **113**:E2636-2645.

37. Upadhyay AA, Kauffman RC, Wolabaugh AN, Cho A, Patel NB, Reiss SM, Havenar-Daughton C, Dawoud RA, Tharp GK, Sanz I *et al.*: **BALDR: a computational pipeline for paired heavy and light chain immunoglobulin reconstruction in single-cell RNA-seq data**. *Genome Med* 2018, **10**:20.

38. Khosravi-Maharlooei M, Obradovic A, Misra A, Motwani K, Holzl M, Seay HR, DeWolf S, Nauman G, Danzl N, Li H *et al.*: **Crossreactive public TCR sequences undergo positive selection in the human thymic repertoire**. *J Clin Invest* 2019, **130**.

39. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P: **Simultaneous epitope and transcriptome measurement in single cells**. *Nat Methods* 2017, **14**:865-868.

40. Shahi P, Kim SC, Haliburton JR, Gartner ZJ, Abate AR: **Abseq: ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding**. *Sci Rep* 2017, **7**:44447.

41. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH: **Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles**. *Proc Natl Acad Sci U S A* 2015, **112**:E862-E870.

42. Lefranc MP: **Immunoglobulin and T cell receptor genes: IMGT ((R)) and the birth and rise of immunoinformatics**. *Front Immunol* 2014, **5**:22.

43. Ralph DK, Matsen FAT: **Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation**. *PLoS Comput Biol* 2016, **12**:e1004409.

44. Ralph DK, Matsen FAT: **Likelihood-based inference of B cell clonal families**. *PLoS Comput Biol* 2016, **12**:e1005086.

45. Ye J, Ma N, Madden TL, Ostell JM: **IgBLAST: an immunoglobulin variable domain sequence analysis tool**. *Nucleic Acids Res* 2013, **41**:W34-W40.

46. Bukhari SAC, O'Connor MJ, Martinez-Romero M, Egyedi AL, Willrett D, Graybeal J, Musen MA, Rubelt F, Cheung KH, Kleinstein SH: **The CAIRR pipeline for submitting standards-compliant B and T cell receptor repertoire sequencing studies to the national center for biotechnology information repositories**. *Front Immunol* 2018, **9**:1877.

47. Breden F, Luning Prak ET, Peters B, Rubelt F, Schramm CA, Busse CE, Vander Heiden JA, Christley S, Bukhari SAC, Thorogood A *et al.*: **Reproducibility and reuse of adaptive immune receptor repertoire data**. *Front Immunol* 2017, **8**:1418.

48. Ohlin M, Scheepers C, Corcoran M, Lees WD, Busse CE, Bagnara D, Thornqvist L, Burckert JP, Jackson KJL, Ralph D *et al.*: **Inferred allelic variants of immunoglobulin receptor genes: a system for their evaluation, documentation, and naming**. *Front Immunol* 2019, **10**:435.

49. Lee J, Boutz DR, Chromikova V, Joyce MG, Vollmers C, Leung K, Horton AP, DeKosky BJ, Lee CH, Lavinder JJ *et al.*: **Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination**. *Nat Med* 2016, **22**:1456-1464.

50. Lee J, Paparoditis P, Horton AP, Fruhwirth A, McDaniel JR, Jung J, Boutz DR, Hussein DA, Tanno Y, Pappas L *et al.*: **Persistent antibody clonotypes dominate the serum response to influenza over multiple years and repeated vaccinations**. *Cell Host Microbe* 2019, **25**:367-376 e365.

51. Havenar-Daughton C, Carnathan DG, Torrents de la Pena A, Pauthner M, Briney B, Reiss SM, Wood JS, Kaushik K, van Gils MJ, Rosales SL *et al.*: **Direct probing of germinal center responses reveals immunological features and bottlenecks for neutralizing antibody responses to HIV Env trimer**. *Cell Rep* 2016, **17**:2195-2209.

52. Watson CT, Steinberg KM, Graves TA, Warren RL, Malig M, Schein J, Wilson RK, Holt RA, Eichler EE, Breden F: **Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity**. *Genes Immun* 2015, **16**:24-34.

53. Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS, Qin K, Fu Y, Huang CY, Beigel JH *et al.*: **IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity**. *Sci Rep* 2016, **6**:20842.

54. • Parks T, Mirabel MM, Kado J, Auckland K, Nowak J, Rautanen A, Mentzer AJ, Marijon E, Jouven X, Perman ML *et al.*: **Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania**. *Nat Commun* 2017, **8**:14946.

Associations between germline immunoglobulin gene sequences and disease risks, as demonstrated in this study, remain an understudied area that should benefit from better characterization of these loci.

55. Briney B, Inderbitzin A, Joyce C, Burton DR: **Commonality despite exceptional diversity in the baseline human antibody repertoire**. *Nature* 2019, **566**:393-397.

56. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee JY, Olshen RA, Weyand CM, Boyd SD, Goronzy JJ: **Diversity and clonal selection in the human T-cell repertoire**. *Proc Natl Acad Sci U S A* 2014, **111**:13139-13144.

57. • Nielsen SCA, Roskin KM, Jackson KJL, Joshi SA, Nejad P, Lee JY, Wagar LE, Pham TD, Hoh RA, Nguyen KD *et al.*: **Shaping of infant B cell receptor repertoires by environmental factors and infectious disease**. *Sci Transl Med* 2019, **11**.

Infant and childhood immunity is understudied, but likely affects responses throughout adult life; applying modern phenotyping tools such as the Ig repertoire analysis used in this study can reveal the associations between antigen exposures and immune system shaping in early life.

58. Du Toit G, Roberts G, Sayre PH, Bahnson HT, Radulovic S, Santos AF, Brough HA, Phippard D, Basting M, Feeney M *et al.*: **Randomized trial of peanut consumption in infants at risk for peanut allergy**. *N Engl J Med* 2015, **372**:803-813.

59. Fonville JM, Wilks SH, James SL, Fox A, Ventresca M, Aban M, Xue L, Jones TC, Le NMH, Pham QT *et al.*: **Antibody landscapes after influenza virus infection or vaccination**. *Science* 2014, **346**:996-1000.

60. Parameswaran P, Liu Y, Roskin KM, Jackson KK, Dixit VP, Lee JY, Artiles KL, Zompi S, Vargas MJ, Simen BB *et al.*: **Convergent antibody signatures in human dengue**. *Cell Host Microbe* 2013, **13**:691-700.

61. Jackson KJ, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, Marshall EL, Gurley TC, Moody MA, Haynes BF *et al.*: **Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements**. *Cell Host Microbe* 2014, **16**:105-114.

62. Joyce MG, Wheatley AK, Thomas PV, Chuang GY, Soto C, Bailer RT, Druz A, Georgiev IS, Gillespie RA, Kanekiyo M *et al.*: **Vaccine-induced antibodies that neutralize group 1 and group 2 influenza A viruses**. *Cell* 2016, **166**:609-623.

63. Truck J, Ramasamy MN, Galson JD, Rance R, Parkhill J, Lunter G, Pollard AJ, Kelly DF: **Identification of antigen-specific B cell receptor sequences using public repertoire analysis**. *J Immunol* 2015, **194**:252-261.

64. Setliff I, McDonnell WJ, Raju N, Bombardi RG, Murji AA, Scheepers C, Ziki R, Mynhardt C, Shepherd BE, Mamchak AA *et al.*: **Multi-donor longitudinal antibody repertoire sequencing**

reveals the existence of public antibody clonotypes in HIV-1 infection. *Cell Host Microbe* 2018, **23**:845-854 e846.

65. •• Pieper K, Tan J, Piccoli L, Foglierini M, Barbieri S, Chen Y, Silacci-Fregni C, Wolf T, Jarrossay D, Anderle M *et al.*: **Public antibodies to malaria antigens generated by two LAIR1 insertion modalities**. *Nature* 2017, **548**:597-601.

It has long been known that antibody gene rearrangements can contain small indels; this study revealed that much larger insertions derived from other chromosomal loci can be added in the VDJ rearrangement, or in the switch region downstream of the IGHJ gene, and give rise to strong selection of such B cell clones due to binding of pathogen-derived antigens.