Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Repertoire sequencing and the statistical ensemble approach to adaptive immunity

Curtis G. Callan Jr. [1], Thierry Mora[2] and Aleksandra M. Walczak[3]

Abstract

The recent advent of high-throughput sequencing of immune receptors allows for the study of immune repertoires in unprecedented depth. This should eventually lead to a better understanding of basic immune function and the development of valuable new diagnostic tools. However, the interpretation of these new sequence data can be difficult because the relationship between receptor sequence and immune specificity is generally unknown. In particular, phenotypically similar repertoires will in general be completely different at the sequence level because of cross-reactivity. Here we argue that sequence repertoires need to be considered statistically to overcome this functional degeneracy. New tools are needed to extract the functionally relevant statistical features from sequence data, separating them from individual-specific, stochastic, and other non-reproducible effects.

Addresses

[1] Joseph Henry Laboratories, Princeton University, Princeton, NJ 08544 USA

[2] Laboratoire de physique statistique, UMR8550, CNRS and École normale supérieure, 24, rue Lhomond, 75005 Paris, France

[3] Laboratoire de physique théorique, UMR8549, CNRS and École normale supérieure, 24, rue Lhomond, 75005 Paris, France

Corresponding author: Callan Jr, Curtis G (ccallan@princeton.edu)

Current Opinion in Systems Biology 2017, **1**:44–47

This review comes from a themed issue on **Future of systems biology**

Edited by **Arnold Levine**

For a complete overview see the Issue and the Editorial

Available online 19 December 2016

http://dx.doi.org/10.1016/j.coisb.2016.12.014

Keywords

Immunology, Statistical physics, Next generation sequencing.

The adaptive immune system exploits the diversity of receptors encoded on the surface of B- and T-cells to recognize unknown pathogens and protect the host organism from infections. The set of all these receptors in an individual, termed the immune repertoire, is a unique example of a biological system with a very high level of somatic genetic diversity. Each time a new immune cell is created, the gene for its surface receptor is quasi-randomly generated from combinatorially chosen genomic templates [1] and diversity is further increased by random nucleotide deletions from, and insertions between, the templates. From the generation process onward, each lymphocyte cell and its clonal descendants will carry a unique surface receptor gene and have a unique pathogen recognition specificity. The diversity of this generation process, called VDJ recombination, is so great that, while any one individual has many millions of B- or T cell clones [2,3], two different individuals have almost completely non-overlapping immune repertoires, while being protected against the same pathogens. The great heterogeneity of individual repertoires makes classical association studies unfit to make predictions about the functioning of immune repertories. Instead, we must look for key statistical properties of the repertoire, rather than its detailed sequence content, to identify what is functionally relevant and common to different individuals behind the apparent diversity of their repertoires.

Newly produced lymphocytes undergo an initial selection step, where they are screened against proteins that are naturally produced by the host organism. Receptors that bind too strongly to these self-proteins could trigger auto-immune disease and are eliminated, whereas those that do not bind to any self-protein are suspected of being poor receptors and are also eliminated. Cells that pass these tests are released into the periphery, where they form the naive repertoire. Those that recognize pathogens proliferate and a fraction of their offspring is kept as memory cells that are governed by their own homeostasis and used for a faster response to recurring antigens. B-cells also acquire hypermutations while they proliferate, which are further selected upon, in a process called affinity maturation. In summary, the immune repertoire is made of naive and memory subsets, with clones of different sizes that share the same receptor protein in each subset.

Modern high-throughput sequencing has vastly expanded the scope of the information that we can obtain about immune repertoires. It is now possible to obtain a list of the genomic DNA or expressed mRNA of all the T cell or B cell receptors in a given biological sample (blood, thymus, spleen, tumor) [4–6] (reviewed in [7–11]). In addition to sequence, the data can include the number of times the sequenced molecule occurred in the original sample, thus giving access to the abundance of each clonotype [12]. In this article, we will

discuss the challenges inherent in using these rich data sets to uncover the underlying statistical features that are responsible for how the immune system works, and might eventually be useful for medical diagnostics and therapeutics [13—15].

The probability distribution from which immune cell receptor sequences are drawn is one of these key statistical properties. This distribution is impossible to estimate directly because it is only sparsely sampled by even the largest datasets. We can solve this problem by recognizing that each immune receptor sequence reflects a series of hidden events—which genomic templates were joined, what nucleotide deletions and insertions were made to generate the observed sequence. These events are stochastic and their distributions, which reflect the biochemistry of VDJ recombination, are *a priori* unknown. The genesis of a given receptor sequence can only be inferred probabilistically. Given a large enough set of sequences, statistical inference methods allow us to infer the underlying distributions of the hidden variables, to quantitatively describe the generation process of new receptors, and to quantitatively characterize repertoire diversity and individual-to-individual variability [16]. This probabilistic estimate can be further refined by inferring the broad features of receptor sequences that make them functional, by quantitatively comparing productive receptors to failed recombinations [17].

Once the hidden variable distributions are known, we can assign to any individual sequence its probability of generation in a single recombination event. Remarkably, the generation probability of specific human T-cells varies over 20 orders of magnitude, and even more for B-cells [18]. This analysis allows us to quantify our level of surprise at observing any specific receptor sequence in a given repertoire, knowledge that is essential when describing selection or evolution of repertoires in response to infections and vaccines.

Gauging our surprise also proves very important when discussing so called public repertoires—sets of receptors that are shared by many individuals. It had been suggested that public receptors might occur simply as a result of convergent recombination [19,20]. Using hidden scenarios to quantify the probability of generating a given receptor, we predicted the probability that the same receptor be generated in two unrelated individuals purely by chance, in excellent numerical agreement with data [16,17]. Such shared receptors are found to have a much higher-than-average probability of generation, as predicted quantitatively by the theory. Although the phenomenon of convergent recombination has been identified as a major source of public sequences for some time, we would argue that being able to *calculate* its exact extent, as afforded by a high-throughput sequence data combined with statistical modeling, constitutes an important advance.

In this connection, it has been pointed out [21,22] that sequences with small numbers of insertions are more likely to be generated independently multiple times because of their lower diversity and higher-than-average generation probability. The enzyme TdT that is responsible for insertions in VDJ recombination is known to be down-regulated in prenatal life, both in humans and mice [23]. This fact has some interesting consequences. First, since identical twins share a circulatory system *in utero*, one would expect them to have, due to long-lived shared clones created before birth, more shared T cell sequences than unrelated individuals. Since the shared clones were created when the insertion enzyme was inactive, one would expect the shared sequences to have very few insertions, and this is what is observed [24]. Hidden variable models make it possible to directly trace the change in the insertion distribution in developing mice and quantify the upregulation of the insertion enzyme.

The general lesson is that we need a full probabilistic description of the repertoire in order to discriminate between antigen driven responses and chance events. While it is tempting, and often useful, to look for a restricted set of key sequences that can be linked to certain conditions and disorders and are widely shared for functional reasons [25—27], sharing is meaningful only if it is statistically unlikely. The picture we are proposing is that the immune repertoire functions as a statistical *ensemble*: everyone's immune system is different at the level of actual sequences, but different repertoires are functionally equivalent in a statistical sense [28]. The search for a core public repertoire, defined as a minimal list of sequences that must be present in each individual for protection against common pathogens, may miss important aspects of how each individual responds to infections in a personalized manner.

This diversity of responses is directly related to the phenomenon of cross-reactivity—the fact that one receptor recognizes more than one antigen [29]—and its counterpart—that one antigen is recognized by more than one receptor [30]. Cross-reactivity is almost certainly an important, perhaps even central, aspect of the immune repertoire. In particular, it has been pointed out that a non-cross-reactive mouse immune repertoire would require a number of cells that would take up more volume than the whole mouse [31]. Cross-reactivity also explains how individuals with apparently different repertoires may be protected against the same infections.

While sequencing is a powerful tool, it tells us little about the *functional* diversity indicative of the recognition phenotype, i.e. the set of antigenic targets the repertoire is specific to. The space of these functional responses is too high-dimensional and difficult to access experimentally to be explored exhaustively. Experimental techniques based on directed evolution can find the best antibody binders to a specific antigenic target [32,33]. However, repertoire-level studies of responses to specific antigens show that many clones change their frequency, and also show a large variability of responding clones between even genetically identical individuals. As a result of these studies, we have learned that, at the receptor sequence level, the response to the same antigen is sparse and non-overlapping between different individuals [34–37]. Linking these sequencing-based assays to phenotypic-response, or *in vitro* deep-mutational scanning assays [38,39], and to models of self-tolerance [40], is an extremely important and unsolved problem. We note that the ability to assign to sequences their generation probability could give an instructive window into this problem: if the clone responding to an antibody in one individual has a very low generation probability, the same clone is unlikely to occur in another individual and the antibody response, if any, should be via a different sequence clone.

As we observe different responses to vaccines in different individuals, can we nevertheless try to identify common features of the responding receptors? In a recent study where mice were immunized with a killed *Mycobacterium tuberculosis* antigen, it was shown that although immunization altered the T cell repertoire, it did not lead to repertoire convergence [37]: a certain number of clones could be identified in the different immunized mice, but their frequencies differed tremendously and the researchers could not identify a set of relevant responding clones. However, they were able to identify amino acid patterns that were overrepresented in the immunized mice compared to random expectations, and machine learning approaches were proposed to identify such patterns. Albeit preliminary, these results suggest that responding sequences in different individuals may share learnable features despite being distinct, giving hope for the characterization of an effective phenotypic public repertoire.

The functioning of the immune repertoire currently remains a mystery. While high-throughput sequencing technology provides us with a wealth of data on receptor sequence repertoires, we argue that a probabilistic or ensemble view of the repertoire that accounts for the cross-reactivity of its receptors and the functional diversity that it entails, is necessary for understanding the diversity of immune responses to common challenges, and will prove instrumental in making useful prediction and diagnostic tools.

Acknowledgements

The work of TM and AW was supported in part by grant ERCStG n. 306312. The work of CC was supported in part by NSF grant PHY-1305525.

References

1. Hozumi N, Tonegawa S: **Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions**. *Proc Natl Acad Sci* 1976, **73**:3628–3632.
2. Qi Q, *et al.*: **Diversity and clonal selection in the human T-cell repertoire**. *Proc Natl Acad Sci* 2014.
3. Arstila TP, *et al.*: **A direct estimate of the human alphabeta T cell receptor diversity**. *Science* 1999, **286**:958–961.
4. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR: **High-throughput sequencing of the zebrafish antibody repertoire**. *Science* 2009, **324**:807–810.
5. Boyd SD, *et al.*: **Measurement and clinical monitoring of human lymphocyte clonality by massively parallel {VDJ} pyrosequencing**. *Sci Transl Med* 2009, **1**:12ra23.
6. Robins HS, *et al.*: **Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells**. *Blood* 2009, **114**:4099–4107.
7. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S: **Rep-Seq: uncovering the immunological repertoire through next-generation sequencing**. *Immunology* 2012, **135**:183–191.
8. Baum PD, Venturi V, Price DA: **Wrestling with the repertoire: the promise and perils of next generation sequencing for antigen receptors**. *Eur J Immunol* 2012, **42**:2834–2839.
9. Six A, *et al.*: **The past, present and future of immune repertoire biology – the rise of next-generation repertoire analysis**. *Front Immunol* 2013, **4**:413.
10. Georgiou G, *et al.*: **The promise and challenge of high-throughput sequencing of the antibody repertoire**. *Nat Biotechnol* 2014, **32**:158–168.
11. Calis JJ, Rosenberg BR: **Characterizing immune repertoires by high throughput sequencing: strategies and applications**. *Trends Immunol* 2014:1–10.
12. Shugay M, *et al.*: **Towards error-free profiling of immune repertoires**. *Nat Methods* 2014, **11**:653–655.
13. Warren EH, Matsen Fa, Chou J: **High-throughput sequencing of B- and T-lymphocyte antigen receptors in hematology**. *Blood* 2013, **122**:19–22.
14. Robins H: **Immunosequencing: applications of immune repertoire deep sequencing**. *Curr Opin Immunol* 2013, **25**:646–652.
15. Galson JD, Pollard AJ, Trück J, Kelly DF: **Studying the antibody repertoire after vaccination: practical applications**. *Trends Immunol* 2014, **35**:319–331.
16. Murugan A, Mora T, Walczak AM, Callan CG: **Statistical inference of the generation probability of T-cell receptors from sequence repertoires**. *Proc Natl Acad Sci* 2012, **109**:16161–16166.
17. Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM: **Quantifying selection in immune receptor repertoires**. *Proc Natl Acad Sci* 2014, **111**:9875–9880.
18. Elhanati Y, *et al.*: **Inferring processes underlying B-cell repertoire diversity**. *Philos Trans R Soc Lond B Biol Sci* 2015, **370**. 20140243.
19. Venturi V, *et al.*: **Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination**. *Proc Natl Acad Sci* 2006, **103**:18691–18696.
20. Venturi V, Price DA, Douek DC, Davenport MP: **The molecular basis for public T-cell responses?** *Nat Rev Immunol* 2008, **8**:231–238.
21. Fazilleau N, *et al.*: **V and V public repertoires are highly conserved in terminal deoxynucleotidyl transferase- deficient mice**. *J Immunol* 2004, **174**:345–355.

22. Venturi V, *et al.*: **A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing**. *J Immunol* 2011, **186**:4285–4294.

23. Bogue M, Gilfillan S, Benoist C, Mathis D: **Regulation of N-region diversity in antigen receptors through thymocyte differentiation and thymus ontogeny**. *Proc Natl Acad Sci U S A* 1992, **89**:11011–11015.

24. Pogorelyy MV, *et al.*: **Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires**. *arXiv qbio* 2016:1–21. www.arxiv.org:1602.03063.

25. Madi A, *et al.*: **T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity**. *Genome Res* 2014, **24**:1603–1612.

26. Truck J, *et al.*: **Identification of antigen-specific B cell receptor sequences using public repertoire analysis**. *J Immunol* 2015, **194**:252–261.

27. Galson J, Kelly D, Trück J: **Identification of antigen-specific B cell receptor sequences from the total B cell repertoire**. *Crit Rev Immunol* 2016, **35**:463–478.

28. Mayer A, Balasubramanian V, Mora T, Walczak AM: **How a well-adapted immune system is organized**. *Proc Natl Acad Sci* 2015, **112**:5950–5955.

29. Wooldridge L, *et al.*: **A single autoimmune T cell receptor recognizes more than a million different peptides**. *J Biol Chem* 2012, **287**:1168–1177.

30. Birnbaum ME, *et al.*: **Deconstructing the peptide-MHC specificity of t cell recognition**. *Cell* 2014, **157**:1073–1087.

31. Mason D: **A very high level of crossreactivity is an essential feature of the T- cell receptor**. *Immunology today* 1998, **19**:395–404.

32. Hoogenboom HR: **Selecting and screening recombinant antibody libraries**. *Nat Biotechnol* 2005, **23**:1105–1116.

33. Boyer S, *et al.*: **Hierarchy and extremes in selections from pools of randomized proteins**. *Proc Natl Acad Sci* 2016, **113**:3482–3487.

34. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR: **Genetic measurement of memory B-cell recall using antibody repertoire sequencing**. *Proc Natl Acad Sci* 2013, **110**:13463–13468.

35. Jiang N, *et al.*: **Lineage structure of the human antibody repertoire in response to influenza vaccination**. *Sci Transl Med* 2013, **5**:171ra19.

36. Laserson U, *et al.*: **High-resolution antibody dynamics of vaccine-induced immune responses**. *Proc Natl Acad Sci* 2014, **111**:4928–4933.

37. Thomas N, *et al.*: **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.

38. Fowler DM, Fields S: **Deep mutational scanning: a new style of protein science**. *Nat Methods* 2014, **11**:801–807.

39. Adams RM, Kinney JB, Mora T, Walczak AM: **Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves**. *arXiv* 2016. qbio:1601.02160, www.arxiv.org:1601.02160.

40. Kosmrlj A, Jha AK, Huseby ES, Kardar M, Chakraborty AK: **How the thymus designs antigen-specific and self- tolerant T cell receptor sequences**. *Proc Natl Acad Sci U S A* 2008, **105**:16671–16676.