Review

CellPress

Studying the antibody repertoire after vaccination: practical applications

Jacob D. Galson, Andrew J. Pollard, Johannes Trück, and Dominic F. Kelly

Oxford Vaccine Group, Department of Paediatrics, University of Oxford and the NIHR Oxford Biomedical Research Centre, Oxford, UK

Nearly all licensed vaccines have been developed to confer protection against infectious diseases by stimulating the production of antibodies by B cells, but the nature of a successful antibody response has been difficult to capture. Recent advances in next-generation sequencing (NGS) technology have allowed high-resolution characterization of the antibody repertoire, and of the changes that occur following vaccination. These approaches have yielded important insights into the B cell response, and have raised the possibility of using specific antibody sequences as measures of vaccine immunogenicity. Here, we review recent findings based on antibody repertoire sequencing, and discuss potential applications of these new technologies and of the analyses of the increasing volume of antibody sequence data in the context of vaccine development.

Vaccination and the antibody repertoire

Since the development of the first vaccine by Edward Jenner in 1796, vaccination has become one of the most important public health interventions worldwide. Where effective vaccines are used, they have greatly reduced the burden of infectious disease, preventing millions of deaths each year [1]. Most licensed vaccines confer protection against disease by stimulating B cells to produce pathogen-specific antibodies. The number of circulating B cell clones with distinct antibody heavy chain rearrangements is estimated to be at least 10^6 [2–4], although there is the theoretical potential for every one of a human adult's 10^{11} B cells to produce a unique antibody variant [5]; such diversity is a key element in the recognition of any potential pathogen. The sum of all circulating antibodies is known as the antibody repertoire.

Antibodies are large molecules, consisting of paired heavy and light polypeptide chains. These form a variable (V) antigen-binding region (known as V_H and V_L for the heavy and light chains respectively), as well as a constant (C) region [6]. Variability is concentrated in the three complementarity-determining regions (CDR1–3), which form a variety of tertiary structures to bind different antigens [7]. Initial $V_{H/L}$ diversity is generated during B cell development by somatic recombination of multiple variable (V), diversity (D; V_H only) and joining (J) gene segments, to form functional V_H and V_L regions [8]. Diversity is increased at the junctions between segments by the addition of palindromic (P) and nontemplated (N) nucleotides, and exonuclease activity leading to potential nucleotide deletion. During response to an antigen, further $V_{H/L}$ diversification occurs through rounds of somatic hypermutation, followed by selection of B cells for improved antigen binding in the germinal center (GC) [9].

An individual's antibody repertoire may be determined by a variety of factors, including their genotype and chromatin structure [10,11], antigen exposure history [12] and age [12–17]. The diversity of the antibody repertoire makes it difficult to study, but improvements in NGS over the past decade now allow parallel sequencing of millions of antibody sequences (Box 1), making in-depth studies of the repertoire possible. There is an increasing body of data characterizing changes in the antibody repertoire following vaccination {influenza [4,13,16,18–21], tetanus [18,22,23], *Haemophilus influenzae* type b (Hib) [24,25], and *Streptococcus pneumoniae* [13,26]}, and natural infection {including influenza virus [27–32], rotavirus [33–36], HIV [37–44], hepatitis C virus [45], cytomegalovirus (CMV) [12], Epstein–Barr virus (EBV) [12], *Staphylococcus aureus* [46], and dengue virus [47–51] among others}. The majority of the vaccine studies are based on low-resolution methods for characterizing the repertoire, but there are six recent studies, that use NGS for in-depth characterization (Table 1).

Despite the increasing number of studies of the antibody repertoire, there is little mention of the application of these data, with most publications instead focusing on methods development [18,52]. Here, we summarize recent advances in repertoire sequencing, and the insight this has given into the antibody repertoire. We discuss both the use of vaccines as a tool with which to investigate the antibody repertoire, and also the use of antibody repertoire sequencing as a tool for vaccinology that can potentially be applied to aid the development and assessment of new vaccines and vaccine schedules.

Methods for studying the antibody repertoire

Low-resolution methods

The earliest studies of the antibody repertoire used isoelectric focusing of antibodies on polyacrylamide gels [53], resolving the antibodies into patterns of discrete bands based on their isoelectric pH values. CDR3 size spectratyping, a

Corresponding authors: Galson, J.D. (jacob.galson@paediatrics.ox.ac.uk); Pollard, A.J. (andrew.pollard@paediatrics.ox.ac.uk).

Keywords: antibody repertoire; B cell receptor repertoire; immunoglobulin repertoire; VDJ; next generation sequencing; vaccination.

1471-4906/

 http://dx.doi.org/10.1016/j.it.2014.04.005

Box 1. Sequencing platforms used to study the antibody repertoire

The use of Sanger sequencing for characterization of single antibody genes from small numbers of cells is highly robust, as there is a high signal-to-noise ratio, low error rates, and a long read length. For characterizing whole repertoires of antibody genes from large numbers of cells, Sanger sequencing is too labor intensive, so NGS is used instead. NGS platforms allow multiple different DNA fragments to be sequenced simultaneously but generally have higher error rates, and shorter read lengths compared to Sanger sequencing. There are a number of different NGS platforms currently on the market, each utilizing different template amplification strategies, sequencing chemistries, and detection methods [97]. This leads to different types and degrees of sequence error (indels and substitutions) in the output data from each platform as well as different limitations such as read length, depth of sequencing (number of sequences that can be sequenced simultaneously), and cost.

Platform	Mechanism	Error rate/type	Cost/Mb	Read length	Depth
454	Pyrosequencing	1.4% of reads homopolymer associated indels	~$31	500 bp	10^5–10^6 reads
Illumina	Dye terminator sequencing	3.2% of reads random substitutions	~$0.5	2 × 300 bp	>10^7 reads
Ion Torrent (314 chip)	Semiconductor sequencing	1.2% of reads homopolymer associated indels	~$22.5	400 bp	10^5–10^6 reads

High-throughput antibody repertoire sequencing was initially conducted using Roche 454 sequencing technology, because this was the only platform with a sufficient read length (~500 bp) to cover the entire V_H region [3,64]. Recent advances in the Illumina chemistry now permit 300 bp paired-end sequencing (http://www.illumina.com/systems/sequencing.ilmn, accessed April 2014) on their MiSeq platform, so more laboratories are now moving towards this technology instead [4,18]. The large output from Illumina sequencing means that 96 samples can be multiplexed in a single run, while still giving >100 000 reads for each sample, thus offering significant cost benefits over 454 sequencing. Recently, the Ion Torrent has also been used to sequence the antibody repertoire, where the emphasis was on speed of data acquisition rather than depth of sequencing [98]. By using whole blood rather than sorted B cells to characterize the repertoire, the authors were able to generate sequence data from a sample in a single day using the Ion Torrent. Illumina, 454 and Ion Torrent, were recently compared by Bolotin *et al.* for their application in profiling the T cell receptor repertoire [99]. Data from this study was used to estimate the error rate of each of these platforms, as shown below. Costs are derived from [100].

NGS technology is rapidly advancing, with the major vendors frequently releasing updates to their systems, which reduce error and increase read length and depth. Costs are also constantly decreasing, making these technologies more accessible for use by laboratories with different specializations. As the technology develops, allowing deeper sequencing of the antibody repertoire, it is likely that more of the potential applications of this technology will be realized.

PCR, and electrophoresis-based method for determining CDR3 diversity based on nucleotide length distribution, can also give some insight into the antibody repertoire after vaccination [14]. The advent of Sanger sequencing allowed the exact nucleotide sequences coding for specific antibodies to be determined, albeit in small numbers [54]. Lymphocytes can be isolated after vaccination, followed by production of immortalized cell lines. Rearranged immunoglobulin DNA can then be amplified from these cell lines, and sequenced [24,55]. Additionally, cloning amplified DNA into expression vectors allows functional characterization of sequences for antigen-specificity [56]. Fluorescence-activated single cell sorting (FACS) has allowed more precise definition of the B cell subsets being studied. Sufficient DNA for sequencing or cloning can be generated from single cells by clonal expansion using culture [33], or single cell reverse transcription PCR [20,57–59].

Such methods have been used to investigate small numbers of monoclonal antibodies (mAbs) generated against a variety of antigens, and although limited by cell numbers, we now know some antigen-specific sequences used in response to vaccines against influenza, tetanus, Hib, and some serotypes of *S. pneumoniae* (Table 1), as well as natural infection (reviewed in [56]). These studies have shown that the B cell response to an antigen with a simple biochemical structure, such as Hib polysaccharide, appears markedly oligoclonal, with similar clones (predominantly V_H3–23 [24,55,60,61], with a conserved 'GYGMD' CDR3 amino acid motif) dominating the repertoire in different individuals (although it should be noted that these studies were limited to a total of 19 individuals, and 25 sequences between them) [62]. Repertoire diversity is also restricted for protein antigens [e.g., tetanus toxoid (TT), and influenza haemagglutinin (HA)] [20,22,23], but the expanded clones differ more between individuals (again limited to at most 14 individuals, and <500 sequences per individual in any study) [20,22]. Even in a single individual, there appears to be little similarity in the response after repeated TT vaccination, with only one-third of the clones sequenced being shared between vaccination events [22].

NGS methods

The advent of NGS (Box 1) enabled the sequencing of antibody genes from millions of cells simultaneously, giving greater insight into the entire antibody repertoire. An in-depth antibody repertoire study was initially conducted in zebrafish; this is a useful model system due to its small number of antibody-producing B cells (~300 000), which is five orders of magnitude lower than in humans [63]. This study gave the first insights into the entire antibody repertoire of an organism, showing that zebrafish use 50–86% of all possible VDJ combinations, and the VDJ frequency distribution is similar between individual zebrafish. It is not possible to sequence every B cell in humans, so representative samples (generally derived from peripheral blood) are taken. The first human studies sought to discover frequencies of VDJ segment usage, the similarity in VDJ segment usage, and the CDR3 sequence between individuals, and to estimate the size of the antibody repertoire [2,3,5,11,64–67]. Size estimates are difficult because