Genes and Immunity (2016), 1–12

www.nature.com/gene

REVIEW

Current status and recent advances of next generation sequencing techniques in immunological repertoire

X-L Hou, L Wang, Y-L Ding, Q Xie and H-Y Diao

To ward off a wide variety of pathogens, the human adaptive immune system harbors a vast array of T-cell receptors (TCRs) and B-cell receptors (BCRs), collectively referred to as the immune repertoire. High-throughput sequencing (HTS) of TCR/BCR genes allows in-depth molecular analysis of T/B-cell clones, providing an unprecedented level of detail when examining the T/B-cell repertoire of individuals. It can evaluate TCR/BCR complementarity-determining region 3 (CDR3) diversity and assess the clonal composition, including the size of the repertoire; similarities between repertoires; V(D)J segment use; nucleotide insertions and deletions; CDR3 lengths; and amino acid distributions along the CDR3s at sequence-level resolution. Deep sequencing of B-cell and T-cell repertoires offers the potential for a quantitative understanding of the adaptive immune system in healthy and disease states. Recently, paired sequencing strategies have also been developed, which can provide information about the identity of immune receptor pairs encoded by individual T or B lymphocytes. HTS technology provides a previously unimaginable amount of sequence data, accompanied, however, by numerous challenges associated with error correction and interpretation that remain to be solved. The review details some of the technologies and some of the recent achievements in this field.

Genes and Immunity advance online publication, 10 March 2016; doi:10.1038/gene.2016.9

INTRODUCTION

The adaptive immune system drives the immune response via specific hypervariable molecules: B-cell-generated immunoglobulins (Igs) and Ig-like T-cell receptors (TCRs) on T lymphocytes. These molecules are formed by genomic recombination enabling them to recognize a multitude of potential pathogens. The TCR is critical for peptide/major histocompatibility (pMHC) recognition, and B-cell receptors (BCRs) are necessary to bind diverse antigens and produce an effective humoral immune response. According to the type of TCR, T cells are classified into αβ T cells and γδ T cells. The most common form of TCR comprises a α and β chain and is present on over 90% of T cells in humans. BCRs comprise two identical heavy-chains (IgHs) and two identical light-chain proteins. TCR/BCR diversity is generated in a number of ways, and TCR genes are organized similar to Ig genes. The Ig and TCR gene loci contain many different variable (V), diversity (D) and joining (J) gene segments, which are subject to rearrangement processes during early-lymphoid differentiation.[1,2] In addition, trimming and addition of non-template nucleotides at the V(D)J junction sites (N-diversity mechanisms) further increases the diversity.[3,4] Take the αβ-TCR, for example, in humans, the TRA locus (position 14q11.2) comprises 47 TRA (T-cell receptor alpha) V genes, 57 TRAJ genes and a single TRAC gene; VJ recombination can rearrange these 105 genes into 2679 unique α-chain VJC gene combinations. The T-cell receptor beta (TRB) locus (position 7q35) contains 54 TRBV genes, 2 TRBD genes, 13 TRBJ genes and 2 TRBC genes. VDJ recombination can rearrange these 71 genes into 2808 unique β-chain VDJC gene combinations. Merging all α- and β-chain gene combinations, an impressive 7 522 632 gene combinations are possible. Subsequent deletions and insertions of nucleotides at the junctions result in a theoretical repertoire of 10^{15}–10^{20} different TCRs that could be generated in humans.[5] Unlike TCRs, rearranged BCR genes are further diversified by helper T-cell-mediated somatic hypermutation—a process of stepwise incorporation of single nucleotide substitutions into the V gene.[6] Through clonal affinity selection for enhanced antigen binding, non-germ-line somatic hypermutation-mediated variation contributes significantly to the diversification of the mature B-cell repertoire.[7,8] Theoretically, the potential for TCR/BCR diversity is nearly infinite, but actual diversity in a biological repertoire is restricted by deletion of over- and under-reactive cells during thymic maturation and is molded continuously by the clonal expansion of antigen responsive cells in the periphery.[9,10] A normal adult polyclonal T-cell compartment comprises an estimated 2.5×10^7 different αβ T-cell clones each expressing a unique TCR.[11] However, studies by other investigators revealed that this figure is considered a conservative estimate, with the upper bounds potentially comprising 10^8–10^{11} unique αβ TCR structures per individual,[11,12] and a much more diverse B-cell repertoire. For both TCRs and BCRs, much of the diversity is focused in the third complementarity-determining region (CDR3), which interacts most closely with the antigenic peptide (Figure 1).[13] The diversity of CDR3 amino acid sequences provides a measure of T/B-cell diversity in an antigen-selected T/B-cell repertoire.

Over the past two decades, several strategies have been developed to probe human TCR diversity. Fluorescence activated cell sorting is a powerful tool for analysis of T lymphocytes, including TCR expression. However, it is strongly restricted by the availability and the specificity of anti-TRB antibodies and no TCR sequence information can be gained.[14] To overcome the

State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, China. Correspondence: Dr H-Y Diao, State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, Zhejiang 310003, China.
E-mail: diao.hy@163.com
Received 7 October 2015; revised 20 January 2016; accepted 20 January 2016

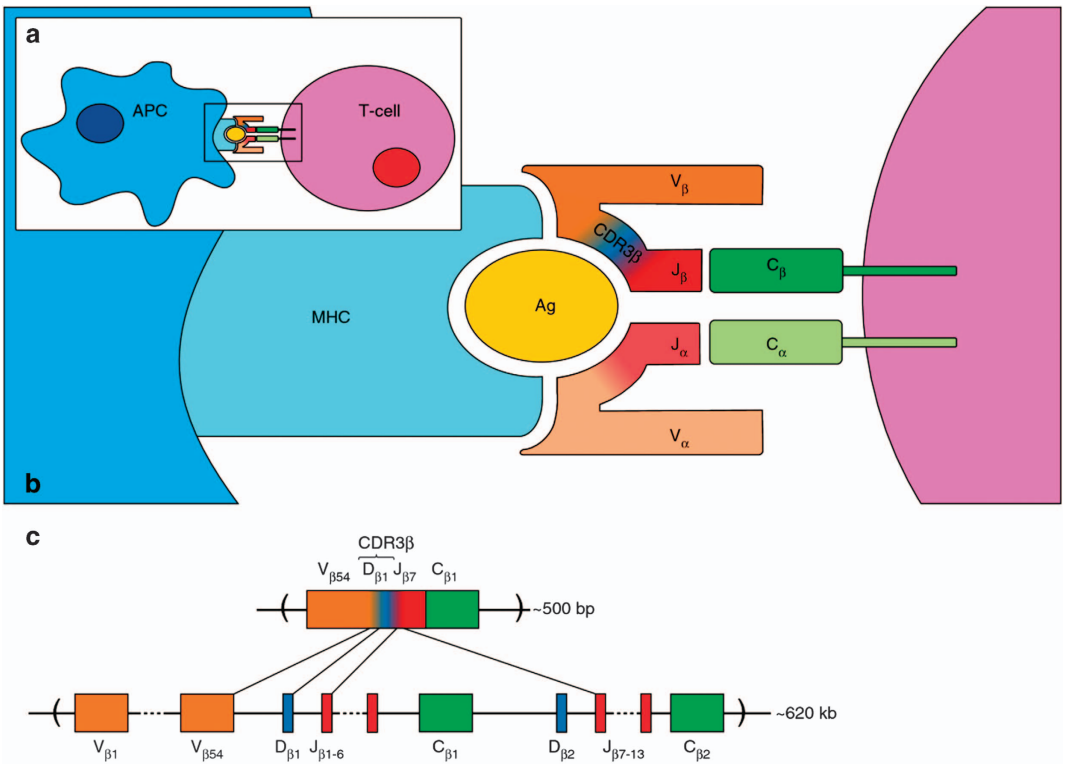

Figure 1. T-cell receptor–antigen–peptide–MHC interaction and TCR gene recombination. (**a**) A T-cell (pink) encountering an antigen-presenting cell (APC; blue). The APC presents peptide antigen (Ag; yellow) in complex with the larger major histocompatibility complex (MHC; turquoise). The T-cell receptor (TCR; multi-colored) binds to both the antigen and MHC, and if the binding avidity is sufficiently high the T-cell is activated. (**b**) A TCR heterodimer, composed of an α and β chain, engaging peptide–MHC (pMHC). Moving outward from the T cell, the constant region (green) of the TCR is anchored to the cell membrane, followed by the J region (red). In TCR α chains, the J region is followed by the V region (orange), whereas in TCR β chains a D region is located between the V and J regions. The complementarity-determining region 3 (CDR3) domain, ~45 nucleotides long, comprises the VJ (for TCR-α) or VDJ (for TCR-β) junction. Color gradients at junctions represent the regions encoded by arbitrary, untemplated nucleotides introduced during somatic recombination, and which represent a primary source of sequence diversification and TCR variability (see **c** for details). The CDR3 regions are the main domains of the TCR that are in contact with peptide antigen and largely determine TCR specificity. (**c**) Simplified representation of TCR-β VDJ gene recombination resulting in TCR diversity. The TCR-β locus is located on chromosome 7 and is ~620 kb in length. Initially one of the two D regions is joined with one of 13 J regions (both randomly selected), followed by joining of the DJ region to one of more than 50 V regions (also randomly selected), yielding a final VDJ region that is ~500 bp in length. The mechanism by which gene segments are joined also introduces base pair variability, which together with the combinatorial selection of these segments results in TCR diversity. A completely analogous process occurs for the TCR α chain, without the D gene segment included.[13]

disadvantage of fluorescence activated cell sorting analysis, PCR-based methods (including multiplex PCR and RACE-based PCR methods) have been developed in many laboratories. However, specific primers were not available for all the V segments defined previously. In addition, different amplification efficiencies among individual primers and cross-reactivity between subfamilies have hampered the estimation of precise frequency of individual TCRB (T-cell receptor beta-chain) V families.[15] An immunoscope technique, CDR3 spectratyping that analyzes CDR3 length polymorphisms, is more widely used. Unfortunately, this approach has a keyhole perspective that has important drawbacks.[16] First, it does not give a full-repertoire perspective. Second, it cannot quantitatively compare the clonal aberrations found across the repertoire. Third, it is prone either to miss clonal aberrations in the first screening step or experience false positive aberrations due to low input. Although Sanger sequencing technique can deliver single clone resolution, it is laborious and generally limits data to a few hundred, or in rare cases a few thousand, TR sequences per investigation.[17] Nowadays, with the advent of NGS technologies, it is now possible to sequence millions of receptor clones representing the entire immune repertoire in a routine experiment. Immune Repertoire Sequencing (IR-SEQ) refers to a method to evaluate the diversity of immune system by amplifying the CDR of BCR or TCR using multiple-PCR or 5′-rapid amplification of complementary DNA (cDNA) ends (RACE) methods, followed by HTS, which can be used to investigate the association between immune repertoire and diseases.[18] However, most of existing IR-SEQ technologies yield data on only one of the two chains of immune receptors and thus cannot provide information about the identity of immune receptor pairs encoded by individual B or T lymphocytes.[19] Fortunately, recently, paired sequencing strategies have also been developed, which includes high-throughput VH:VL pairing technique,[20] single-cell paired TCR sequencing method.[21] In these methods, some need isolated single cell, some do not need. In summary, advanced IR-SEQ technologies are powerful tools to dissect the BCR and TCR populations at high resolution; however, the enormous quantity of reads generated by NGS technologies necessitates cautious interpretation. Potential errors during the sequencing process may skew the interpretation. In the present article, we aimed to present recent progress in understanding the immune repertoire and to discuss the methods used to recognize and eliminate PCR and sequencing errors.

IR-SEQ IN FEATURE MINING OF IG AND TCR PROFILING

HTS of lymphoid receptor genes is an emerging technology that can comprehensively assess the diversity of the immune system. The kinds of analysis enabled by high-throughput DNA

sequencing of TCR or Ig rearrangements can be classified into three main categories.[22] First, this method can be used to measure overall repertoire features, including: V, D and J segment usage frequencies; CDR3 length, VD indel length and DJ indel length; the pattern of amino acid usage in the CDR3 region; and the number of distinct sequences present, which can be used to estimate repertoire diversity. Second, the receptors expressed by clonally expanded B cells or T cells can be detected and characterized, whether or not one knows the antigen specificity or other functional features of the expanded clones. Third, B-cell or T-cell clones of interest that have previously been identified and correlated with known function can be tracked. Each of these types of analysis can yield insights into lymphocyte populations and the features of Ig and TCR repertoires.

Estimation of immune repertoire diversity

The diversity of immune repertoire is vitally important for health. The more subtypes of immune proteins one has, the more powerful the immune system is, and *vice versa*. In addition, many other factors can also affect the diversity of the immune repertoire, including age, environment, diseases and medicine and so on. Relevant to this investigation, Robins *et al.*[23] demonstrated that the number of unique TCRβ CDR3 sequences in the adult repertoire significantly exceeded previous estimates based on first generation sequencing technology.[11] In their study, the sum of the calculated number of unique TCR CDR3 sequences from the four flow cytometrically defined T-cell compartments (CD4+CD45RO+, CD4+CD45RO-, CD8+CD45RO+, CD8+CD45RO-) in the peripheral blood of the two healthy male donors is three to four million, which was at least fourfold higher than previous estimates. In a recent study, Warren *et al.*[24] sequenced the TCR repertoire, and successfully obtained >1 billion raw reads from a single blood sample, with a yield of about 200 million TCR-β nucleotide sequences. In addition, they confirmed that a 20 ml blood sample captured only a portion of the diversity present within an individual's peripheral blood repertoire, which indicated that TCRB diversity was greater than that captured by a single library or a single blood sample. Sequencing is a numbers game. To try and reach the best coverage of the immunological repertoire, we aim to reach an SI (sequenced Igs): TI (total number of Igs) ratio of 1. When this SI: TI ratio has been reached, an account for the entire repertoire can be obtained. Smaller model organisms, therefore, provide a better starting point from which to reach this ratio. In zebrafish, the equivalent would be a 1:1 sampling; that is, current machinery enables the sampling of all the T or B cells in the zebrafish. However, sampling of the repertoire in humans could be thought of as sampling a swimming pool using a single tablespoon. Cell count intervals that are calculated according to the Poisson model suggest that a comprehensive analysis of a population containing a diversity of 10^7 (estimated individual TCR beta diversity) requires analysis of a sample of at least 10^8 cells.[23] The corresponding number of T cells is contained in ~50–70 ml of blood from a healthy individual. It is also necessary to mention that the T-cell repertoire is not static. As we know, immune system is a dynamic system. Binding of a naive T cell's TCR to a structurally compatible pMHC on an antigen-presenting cell will initiate rapid clonal expansion to generate a population of effector cells carrying identical TCRs. Ordinarily, once the antigen that initiated the immune response has been cleared, the expanded pool gradually contracts and persists as a smaller number of memory cells, which are poised for another potential encounter with the antigen. Thus, the T-cell repertoire is continuously molded by the input of new T cells and response to immune challenge.[9,13] Therefore, a direct measure of total diversity remains out of reach owing to following reasons: first, the practical and ethical limitations on obtaining very large numbers of T cells from research subjects; second, the sampling is just a snapshot. No matter how well and exhaustive we sample we probably will never comprehensively profile the repertoire; third, it is difficult to distinguish rare clonotypes from sequencing errors. In basic and clinical research, the principal concern is how to determine whether a given sample has different TCR content from another sample, which is difficult when sampling is incomplete and sampling depth is often variable. In addition, it is important to recognize that the required depth of TCR repertoire analysis could vary across a wide range according to the questions posed. Much smaller bottleneck limits can be estimated for the analysis of a narrower T-cell population, such as a fraction of effector or memory T cells or a pre-sorted subpopulation of specific TCR V beta family T cells, Treg cells or MHC-tetramer sorted cells and so on.[25,26] Overall, a clear understanding of the parameters that determine the sample preparation bottlenecks and their minimal required size is important for an adequate experimental design.

The diversity of the TCR/BCR repertoire can be calculated based on the Simpson index of diversity (*Ds*),[27] Shannon–Wiener index (H')[28,29] and the diversity 50 (D50) value. *Ds* and H' are function of both the relative number of clonotypes present (richness) and the relative abundance or distribution of each clonotype (evenness),[27,30] which are calculated as follows.

$$Ds = 1 - \sum_{i=1}^{C} \frac{ni(ni-1)}{n(n-1)}$$

$$H' = -\sum_{i=1}^{S} \frac{ni}{N} \ln \frac{ni}{N}$$

In both *Ds* and H', *ni* is the clonal size of the ith clonotype (that is, the number of copies of a specific clonotype). C (in *Ds*) and S (in H') are the number of different clonotypes, and *n* (in *Ds*) and *N* (in H') are the total number of TCR/BCR sequences analyzed. *Ds* uses the relative frequency of each clonotype to calculate a diversity index ranging from 0 to 1, where 0 represents the minimum and 1 represents the maximum diversity. A sample of TCRs/BCRs containing more than one clonotype will have maximum diversity when all clonotypes have equal clone size.[30] TCR/BCR repertoire diversity can be calculated at different resolutions of distinct DNA sequences, amino acid sequences and V–J combinations. In addition, D50 is defined as the minimum percentage of distinct clonotypes or CDR3 peptides accounting for at least half of the total clonotypes or CDR3 peptides in a population or subpopulation of immune system cells; the higher the number, the greater the level of diversity.[30] Moreover, many other methods widely used in ecology have also potential utility in comparing immune repertoires. Examples include the Morisita–Horn similarity index for determining the similarity, or overlap, between samples.[31] Earlier studies also explored various bioinformatic tools for TCR-seq data processing and analysis. ImMunoGeneTics (IMGT)/HighV-QUEST[32] (http://www.imgt.org) is a useful tool for medium-scale TCR-seq data handling and annotation. Another new tool for TCR-seq data processing called MiTCR (developed by MiLaboratory; http://mitcr.milaboratory.com/downloads/)[33] is a recent and welcome addition, as is the Cancer-related Immunological Gene Database (CIG-DB),[34] a new repository for TCR and Ig sequences. This program has an automate adjust mechanism for errors introduced by sequencing, PCR and so on. It will feedback alignment statistic information like CDR3 expression and INDEL.

In addition to TCR structural diversity as mentioned above, the diversity of immune repertoire includes the T-cell functional diversity.[9] Such diversity was observed when populations of antigen-specific T cells (as judged by the binding of specific pMHC tetramers) were found to have heterogeneous proliferation; cytokine and chemokine secretion; cytotoxic T lymphocyte activity; and expression of natural killer-cell receptors; chemokine receptors and integrins; as well as migratory patterns and other

features.[35–38] Similar functional heterogeneity was also observed in T cells expressing the same TCR, indicating that at least part of the functional heterogeneity could be independent of TCR structural features.[37–40] However, it remains unclear whether and how TCR structural diversity influences functional diversity, and what is the impact of functional diversity on pathogen resistance. In conclusion, many important initial steps have now been made in quantifying TCR diversity in different lymphoid compartments and across various biological processes, providing an excellent platform from which to continue studies of role of TCR diversity in immune defense. Despite such encouraging advances, the diversity of the immune repertoire requires further investigation.

Distribution characteristics of CDR3 length

IgH CDR3 loops can vary in both length and sequence, allowing for the ability to recognize diverse antigens;[41] however, such variation must also be constrained to prevent the accumulation of poorly functional or autoreactive Igs. Experiments in mice showed that the average length of the IgH CDR3 loop increases during murine B-cell development.[42] However, long HCDR3 loops have been associated previously with antibody auto-reactivity and polyreactivity that are removed from the human repertoire during B-cell development.[43–45] To understand better how selection processes balance the benefits of Ig repertoire diversity with the risks of non-functionality and auto-reactivity of highly variable IgH CDR3s, Larimore *et al*.[46] collected millions of rearranged germline IgH CDR3 sequences by deep sequencing of DNA from mature human naive B cells purified from four individuals and analyzed the data computationally. They found that human IgH rearrangement had evolved several mechanisms to generate longer CDR3 loops for the repertoire, including the use of long D gene segments, insertion of large N regions and usage of multiple D gene segments in tandem. However, via a comparison of productive and out-of-frame IgH rearrangements, they observed a selection bias against long HCDR3 loops, which agreed with previous findings. In addition, they identified that at least 69% of initial productive IgH rearrangements were removed from the repertoire during B-cell development. A previous study demonstrated that 55–75% of antibodies cloned from human early immature B cells were self-reactive based on *in vitro*-binding assays, with the majority of this self-reactivity removed during development,[43] which supported their calculation of the preselection repertoire size based on the extensive data set from peripheral B cells. Taken together, these results suggest that self-reactivity, rather than non-functionality, might be the major reason for loss of productive IgH rearrangements during development. In addition, some studies reported shorter CDR3 lengths in relation to CD4 single-positive cells in the thymus.[47] A previous study showed that splenic BCR CDR3 length distributions are characterized by low standard deviations and few local maxima, compared with peripheral blood distributions, and established a supervised machine learning model, based on CDR3 length distribution features, can detect myelodysplastic syndromes with $\sim$93% accuracy.[48] Taken together, the CDR3 length distribution contains important information regarding the immune system's condition, the details regarding differences and functions of the longer and shorter CDR3 groups require further research.

Nucleotide insertions and deletions bias

Much of the CDR3 diversity in the TCR-β chains is created by the template-independent insertion of nucleotides at the Vβ–Dβ and Dβ–Jβ junctions by terminal deoxynucleotidyl transferase (Tdt). The frequency at which Tdt inserts each of the four nucleotides has been estimated. Robins *et al*.[23] found there were nucleotide insertion bias at the Vβ–Dβ and Dβ–Jβ junctions, such that Tdt is biased toward insertion of C and G over A and T. This confirmed the results from the study by Freeman *et al*.[49] who found that non-templated bases at the V–D junction are 62.9% GC and the non-templated bases at the J–D junction are 54.3% GC. Interestingly, they also observed an interesting phenomenon that the number of insertions and deletions at the Vβ–Dβ and Dβ–Jβ junctions were the features that were most closely correlated with frequency, and demonstrating an inverse correlation. Indeed, recent studies have suggested that higher frequency clonotypes were more commonly shared between compartments (that is, naive and memory compartments) and individuals, and high-frequency TCRβ CDR3 sequences with fewer insertions and deletions have receptor sequences that are closer to the germline sequence.[25,50] It was also interesting to note that extensive N-nucleotide addition is a mechanism that contributes to IgH CDR3 length.[46] In summary, nucleotide insertion/deletion on CDR3 creates extreme diversity in the antigen recognition regions of TCRs or BCRs. Previously, we thought that this process is random, however, based on the findings detailed above, we appreciate that the process is not completely random, and the mechanism needs to be clarified further.

The distribution of amino acids along the CDR3 region

Amino acid residues within CDRs can contribute to antigen binding directly, via contribution of a side group that makes contact(s) with the antigen. In addition, the amino acids can have an indirect effect on the conformation of the peptide backbone in a manner that facilitates direct interaction of neighboring amino acid side groups.[51] Therefore, it is important to analyze the amino acid composition along the CDR3 region. Wu *et al*.[52] used deep sequencing technologies to study human B-cell Ig heavy chain repertoires, and compared the characteristics of amino acid composition of transitional, naive, IgM memory and switched memory B cells. They found that average CDR3 sizes were comparable between transitional and naive cells, but there was a decreased proportion of positively charged amino acids in naive B cells, resulting from a difference in arginine composition. In addition, there was also a decrease in the aliphatic index in naive B cells compared with transitional cells, with an accompanying downward trend in hydrophobicity. Moreover, they also observed that IgM memory cells had fewer negatively charged amino acids than switched memory cells, while the levels of positively charged amino acids did not appear to vary. Thus, the amino acid composition varies among these different B-cell populations. However, it is difficult to predict what kinds of antigen might select for this characteristic. Liaskou *et al*.[53] used HTS to determine if disease-associated TCRs could be identified in the non-viral chronic liver diseases primary biliary cirrhosis (PBC), primary sclerosing cholangitis and alcoholic liver disease. They showed that 8–42 clonotypes were detected uniquely in each of the 3 disease groups ($\geqslant$30% of the respective patient samples). Notably, they identified that disease-associated clonotypes shared common amino acid characteristics. The presence of one hydrophobic residue glycine (G) among two polar residues threonine (T) and/or serine (S) was present in the CDR3 region of 62.5% of the primary sclerosing cholangitis-associated clonotypes (T-G-T, TS-GG-T, GG-T, TS-GG-T, S-G-T). An aspartic acid (D) or glutamic acid (E) was present in 87.5% of primary sclerosing cholangitis disease-associated clonotypes. The presence of glycine–asparagine (G–N) was evident in 62.5% of the alcoholic liver disease-associated clonotypes. In summary, it may be that different cell populations and disease-associated clonotypes share similar protein characteristics. To study the protein (or amino acid) characteristics will help us to clarify the interaction between the antigen and the T/B-cell receptor more deeply, and provide the basis for the development of antibodies and vaccines.

V(D)J segment use and combination

Recently, a study by Freeman *et al*.[49] used a 5′-RACE and Illumina sequencing strategy to sample CDR3β diversity in normal human peripheral blood leukocytes (PBL) pooled from 550 individuals. Their laboratory data revealed that TRBV gene usage ranged from 0.01% for TRBV17 to 24.6% for TRBV20-1. TRBJ gene usage ranged from 1.6% for TRBJ2-6 to 17.2% for TRBJ2-1. In addition, they identified 562 TRBV–TRBJ combinations, among which TRBV20-1 to TRBJ2-1 was the most frequent pairing, accounting for 4.1% of all pairings. These data reflected the TRB (V/D/J) family usage pattern in a population, but it could not provide an insight into individual repertoire features because their samples were derived from blood pooled from multiple individuals. However, for individual repertoires, a large body of work has also demonstrated that certain TRBV and TRBJ genes are utilized commonly while others are relatively rare, and the pairing of TRBV and TRBJ is not random.[54,55] Relevant to this investigation, Robins *et al*.[50] assessed the realized CD8+TCRβ CDR3 sequence repertoire in the blood of seven healthy adults. They found that the frequency with which specific Vβ–Jβ combinations were utilized was highly variable in each of the seven individuals; the frequency with which specific combinations were observed varied by $>10\,000$-fold. In addition, they provided evidence that rearrangement between Vβ and Dβ gene segments was random, while that between Dβ and Jβ gene segments was not. They hypothesized that the apparent non-random association between specific Dβ and Jβ gene segments was likely attributable to the organization of the TCRβ locus, in which Dβ1 lies 5′ of all 13 Jβ segments, while Dβ2 lies 3′ of the 6 members of the Jβ1 cluster but 5′ of the 7 members of the Jβ2 cluster. The Dβ1 segment was observed at roughly equal frequency with all 13 Jβs, while Dβ2 was much more frequently paired with members of the Jβ2 compared with the Jβ1 family. In this regard, Boyd *et al*.[56] observed preferential pairwise segment associations for at least three combinations (D2-2with J6, D3-3with J6 and D3-22 with J3) across the group of individuals. Over-representation of these D/J combinations was observed in 122/138, 119/138 and 113/138 sequenced aliquots, respectively. Preferential TRBV/TRBD/TRBJ recombination is a prerequisite for the emergence of a public response. The reasons for the bias are not clearly understood but are probably caused by a combination of proximity effects and recombination signal sequence compatibilities that influence initial TCR development, plus thymic selection and immune challenge that modify the representation of selected clones in the extant repertoire.[57]

Public T/B-cell repertoires

Public T-cell responses, in which T cells bearing identical TCRs are observed to dominate the response to the same antigenic epitope in multiple individuals, have long been a focus of immune T-cell repertoire studies. Nowadays, public TCRs have been observed for a variety of T-cell responses in many different species.[58–60] Previous experiments from our laboratory showed that any two individuals in the healthy control group share $4.85 \pm 2.50\%$ of their DNA sequences and $12.17 \pm 0.81\%$ of their expressed CDR3 amino acid sequences (non-redundant sequences for each subject).[18] In addition, studies in syngeneic mice showed that up to 27% of the peripheral repertoire of one naive mouse overlapped with that of another.[61] What is the molecular basis for public T-cell responses? Several lines of evidence support the hypothesis that recombination biases, convergent recombination and T-cell selection play important roles in shaping the observed pattern of CDR3 sharing.[62] First, convergent recombination results in a range of production frequencies, with some TCRs being produced rarely, some TCRs being produced at an intermediate frequency and other TCRs being produced frequently. Public repertoires are produced more efficiently than private ones by the recombination machinery. Subsequently, other factors, such as TCR affinity for the pMHCI and stochastic events, further influence clonal dominance. In addition, Warren *et al*.[24] recently reported a strong association between the sharing of HLA class 1 alleles and the proportion of shared TCRB sequences ($P < 1 \times 10^6$). This aspect of the public repertoires may have serious implications for our understanding of the initial ability of an individual to fight incoming threats. Thus, understanding the basis of public T-cell responses is not only important for our understanding of immune repertoire and diversity and hierarchy, but also has implications for immune control of pathogens and vaccine design.

THE APPLICATION OF IR-SEQ IN BASIC RESEARCH

Immune repertoire features of lymphocyte subpopulations

Deep sequencing enables detailed repertoire analysis for different lymphocyte subsets. Wang *et al*.[26] used HTS to evaluate the TCR distribution among key T-cell developmental subsets (naive and transitional T cell (Tn+t), activated T cell (Ta) and memory T cell (Tm)) and effector subsets (T helper cell 1 (Th1), T helper cell 2 (Th2), T regulatory cell (Tr) and T cytotoxic cell (Tc)) from a single donor. They found that the various T-cell subsets examined exhibited many common features in terms of Vα, Jα, Vβ, Dβ or Jβ utilization, CDR3 length, number of N-nucleotide additions, nibbling at ends of germline gene segments and amino acid usage at the CDR3 intervals, which suggested that T-cell specificity determination precedes the differentiation of nascent T cells into distinct phenotypic subsets, helping to resolve this longstanding question of chronology in T-cell maturation. By estimating the number of unique CDR3 intervals for both the TCRα and TCRβ repertoires as a measure of the diversity of the CDR3 repertoire of different T-cell subsets, they found the Ta subset exhibited the least diversity among the developmental subsets. Among the various effector populations, the Th2 subset appeared least diverse and the Tr population exhibited the greatest diversity. In addition, there was a significant amount of sharing of TCR sequences among the Th1, Th2 and Tr populations, which suggested that the choices of Th1, Th2 and Tr outcomes were stochastic and could be driven by the same or highly similar antigenic stimuli.

In addition, NGS technologies provide an opportunity to understand the dynamic relationship between the naive and memory T-cell repertoires. Venturi *et al*.[25] reported that the CDR3 length distributions and clonotype size distributions differed between the memory and naive TCRβ repertoires, and a subset of TCRβ amino acid clonotypes was common to the memory and naive pools. The notion that the diversity of the naive repertoire greatly exceeds that of the memory repertoire has been challenged by the observation that the memory subset, particularly the CD4+ memory subset, mainly comprises a broad diversity of low-frequency clonotypes.[24,63] The human naive T-cell repertoire is the repository of a vast array of TCRs. On episodic or persistent antigenic stimulation, individual naive T cells have multiple fates and can differentiate into effector and memory T-cell subsets. These findings have important implications for our understanding of the dynamic relationship between the naive and memory T-cell repertoires. Further studies are needed to investigate the factors that steer maturation of the memory phenotype and determine the size of memory clones.

Similarly, a recent study used deep sequencing technologies to analyze human B-cell Ig heavy chain repertoires, and compared the characteristics of transitional, naive, IgM memory and switched memory B cells.[52] They found that the memory B-cell repertoires differed from the transitional and naive repertoires, and that the IgM memory repertoire was distinct from that of class-switched memory. Based on these findings, they concluded that a large proportion of IgM memory B cells were not derived from the same developmental pathway as switched memory cells.

In addition, to identify the origin and antigen specificity of intestinal Treg cells, Cebula *et al.*[64] performed single-cell and HTS of the TCR repertoires of CD4+Foxp3+ and CD4+Foxp3 − T cells, and analyzed their reactivity against specific commensal species. They showed that thymus-derived Treg cells comprised the highest proportion Treg cells in all lymphoid and intestinal organs, including the colon, where their repertoire was heavily influenced by the composition of the microbiota. Their results suggested that thymic Treg cells, and not induced Treg cells, mediated the tolerance to antigens produced by intestinal commensals dominantly. Further investigation of these topics will be greatly enhanced by the use of HTS, and by a more comprehensive measurement of TCR or Ig rearrangements present in different lymphocyte subpopulations.

Immune repertoire features in healthy subjects of various ages

HTS techniques also greatly enhance the analysis of how lymphocyte diversity declines with age. Immunity declines with age,[65,66] and declining T-cell repertoire diversity associated with aging is a contributing factor to the impaired ability of aged individuals to mount effective immune responses to infections and vaccines.[67,68] Boyd *et al.*[56] detected the diversity of clonal B-cell expansions in healthy subjects of various ages (ranging from 19 to 79 years) by sequencing six-fold replicate samples of peripheral blood Ig heavy-chains (IgHs) from each individual. The sizes of those larger clones were estimated by the expanded clonal sequence's proportion of total sequences obtained from these samples. They found that for the 54-year-old patient this value was 0.15%, while for the 68-year-old patient the value was 1.5% of the total sequences. These results revealed that the degree of clonal expansion in the elderly group was significantly greater, and the T-cell repertoire of the elderly group was relatively restricted. It is likely that accumulated infections throughout the lifespan have caused the reduction in TCR diversity. In turn, restrictions in the peripheral TCR repertoire can lead to impaired immune responses. Yager *et al.*[69] took advantage of the well-characterized influenza virus model to address the relationship between the naturally occurring age-associated decline in repertoire diversity and the response of aged animals to a newly encountered pathogen. Their data demonstrated directly the impact of an age-associated decline in T-cell repertoire diversity on the capacity to respond to newly encountered antigens. Importantly, the data showed that the age-associated decline in CD8 T-cell repertoire diversity could be so profound for responses of those with low naive precursor frequencies as to result in the development of 'holes' in the repertoire for normally immunodominant epitopes, possibly to compromised protective immunity. Herein, it was also interesting to note that BCR CDR3 length distributions of elderly people were found to be positively skewed toward short CDR3s.[48] This may be explained by the finding that memory cells have shorter CDR3,[52] combined with the reports in the literature that older people have more B-cell memory cells.[70] Thus, therapeutic approaches for improving survival and maintenance of naive T cells, prolonging thymic output and reconstituting the repertoire of the elderly through hematopoietic stem cell reconstitution should also be considered.[69,71,72] It may be also desirable to prime cellular immunity before severe loss of thymic output, suggesting that more vaccinations during middle age may be indicated. Furthermore, newer strategies need to focus on boosting preexisting memory T-cell responses present within aged individuals.

NOTES ON IR-SEQ

As illustrated above, NGS has established itself as a highly useful platform in the study of the immune repertoire. However, performing deep, unbiased and quantitative analysis of millions of CDR3 sequences that include highly homologous variants is quite challenging, because of primer dimerization, accumulation of PCR and sequencing errors, and ratio bias. Altogether, these technical challenges lead to the loss of the original TCR/BCR repertoire of an analyzed T/B-cell sample, generation of huge artificial TCR/BCR diversity and the inability to interpret sequence information in a quantitative way, thus challenging intelligent analysis and cross-comparison of acquired datasets and generally complicating adaptive immunity studies. Here, we compare the three leading platforms (Illumina, 454 and Ion Torrent) for individual TCR profiling, and present platform-specific approaches to error correction.

Different NGS platforms for IR-SEQ

Although this technology is rapidly developing, most published work on HTS of Igs and TCRs to date has used the leading platforms: 454, Illumina and Ion Torrent. Different platforms provide different advantages and disadvantages, which are summarized in Table 1.[73] The 454 GS Junior generates the longest reads (up to 600 bases) and most contiguous assemblies but has the lowest throughput (70 Mb per run, 9 Mb h^{-1}). Run in 100-bp mode, the Ion Torrent PGM, has the highest throughput (80–100 Mb h^{-1}).[74] The Illumina platform comprises the HiSeq and MiSeq sequencing systems. MiSeq is based on the existing Solexa sequencing-by-synthesis chemistry but has dramatically reduced run times compared with Illumina HiSeq (fastest run 4 h versus 1.5 days for 36-cycle sequencing or 16 h versus 8.5 days for 200-cycle sequencing), made possible by a smaller flow cell, reduced imaging time and faster microfluidics.[74] The key variables in high-throughput DNA sequencing are read length, throughput, accuracy and cost. Most published work on HTS of Igs and TCRs to date has used either the 454 platform or the Illumina platform. The 454 instrument can capture a full Ig heavy chain V(D)J sequence in a single read, which is very helpful when studying patterns of hypermutation in clonally related IgHs.[7,52,56,75,76] TCR sequences can be captured by shorter reads covering the V(D)J junction, and can take advantage of the Illumina platform's throughput, which is higher for comparable cost.[49,56]

Different starting materials and PCR amplification methods

Genomic DNA or mRNA can be used as the starting material for TCR/BCR profiling. mRNA is more commonly used for four reasons: First, splicing of the TCR constant region at the mRNA level simplifies amplification strategies because all rearranged receptor genes can be captured with a single primer,[77] which avoids the use of the complex multiplex primer sets and thus, decreases PCR bias. Second, mRNA is less complex and multiple copies are

Table 1. Advantages and disadvantages of sequencing methods noted by Bolotin *et al.*[73]

Method	*Advantages*	*Disadvantages*
454	Longest read lengths	Lowest read number, resulting in bottlenecks frame-shift errors
Illumina	Greatest read number	Highest error rate; shortest read length
Ion Torrent		Frameshift errors; short fragment length requires highly multiplexed PCR, resulting in amplification bias

present in each cell, making it easier to amplify all rearranged receptor genes from any given sample. Third, when starting from genomic DNA, the entire sample isolated from certain PBMC aliquots must be amplified to gain a comprehensive representation of the TCR repertoire. This inclusion may be technically challenging when large T-cell populations are studied. For example, a starting sample of 10 million T cells requires amplification of ~100 µg of PBMC-derived genomic DNA. In contrast, a reasonable aliquot (5–10 µg) of an RNA sample obtained from the same cell population is sufficient to sample the diversity. In addition, at the DNA level, each T cell carries two rearranged TCR beta genes, and one of them is non-functional. In contrast, out-of-frame mRNA molecules are efficiently degraded by the nonsense-mediated decay mechanism;[78,79] thus, these non-functional molecules are not sampled. However, the drawback of the use of mRNA is that the ratio of rearranged receptor genes can be skewed if different cells harbor different numbers of mRNA copies, due either to different levels of transcription or decay.[80] For example, active B cells and plasma cells produce vastly increased amounts of mRNA compared with resting B cells. Given that our aim is to derive the structure of the repertoire as it is defined per cell in the immune system, these different quantities of RNA may introduce a major bias toward sequences expressed by cells that are more actively producing RNA. Also, if mRNA is used, information about the number of cells represented in the amplification template is lost. A given amount of mRNA may represent a small amount of mRNA from many cells or a large amount of mRNA from a few cells, potentially leading to distortions in the estimate of repertoire abundance. Fortunately, the problem of absolute copy counting can be overcome using a barcoded template-switch primer.[81] In this technique, the template-switching effect[82,83] is used to introduce a 5′-adaptor (5′-AAGCAGUGGTAUCAACGCAGAGUNN NNUNNNNUNNNNUCTTrGrGrGrG-3′), which carries a molecular identifier (12 random 'N' nucleotides) and dU nucleotides (U). As a result, each synthesized cDNA molecule is specifically labeled with one of 4^{12} (>16.7 million) unique identifier variants. Such molecular identifiers allow robust estimation of the number of cDNA templates in a deeply sequenced library.[81] This approach minimizes the impact of individual donor blood characteristics or unavoidable bottlenecks and biases during blood sampling, library preparation and sequencing. DNA templates have the advantage of not requiring a reverse transcription step, which can affect yield and introduce sequence errors. In addition, one can infer the number of cells represented in the assay because there is only one DNA copy of a rearranged receptor gene per cell. The downside is the lower abundance of DNA templates. Moreover, the lack of a uniform constant region sequence means that highly multiplexed PCR strategies are required (Figure 2).[84] Multiplexed PCR reactions using large numbers of primers specific to the V and J segments gene families have the advantage of relatively efficiently capturing sequences for amplification; however, a strong bias is expected toward specific V and J segments; thus, the observed relative sequence abundances may not accurately reflect the real amounts. A major finding of Bolotin *et al*.[73] was that such highly multiplexed PCR strategies are associated with distortions in the relative abundance of TCR Vβ families as compared with less multiplexed amplification strategies and antibody staining. While the 5′ RACE PCR introduced few errors, probably because of the use of only one primer, high-fidelity polymerases and low cycle numbers, recent studies established that the majority of errors in TR deep sequencing occur during the solid-phase steps.[85] In summary, using mRNA as a starting material and applying a 5′ RACE PCR approach is more reliable for TCR/BCR repertoire analysis.

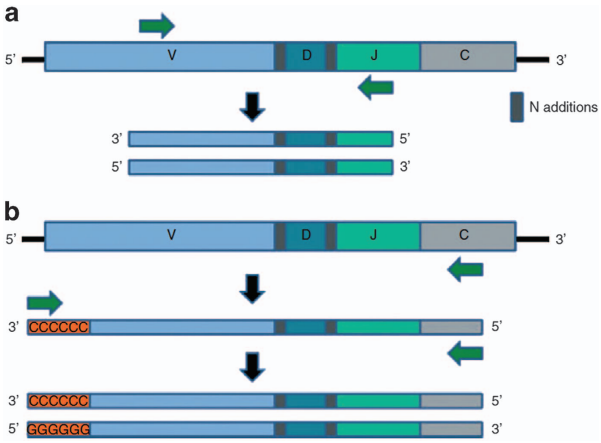

Figure 2. Different PCR amplification techniques. (**a**) Multiple primers—two primers are designed to complement regions within the V and J segments. (**b**) 5′ RACE—only one primer is designed to complement the constant region of the cDNA. After the first amplification round, a homopolymer is synthetically added to the 3′. The cDNA is again amplified with the first specific primer and another primer targeting the homopolymer.[84]

Methods to recognize and eliminate PCR and sequencing errors

The enormous quantity of reads generated by NGS technologies necessitates cautious interpretation. Potential errors during the sequencing process may skew interpretation. Therefore, the reliability of repertoire analysis depends on sequencing depth and coverage, but also on sequencing accuracy. Errors in the primary sequencing data are derived primarily from two sources: (1) nucleotide misincorporation that occurs during the PCR amplification of TCR/BCR CDR3 template sequences, and (2) errors in base calls introduced by the genome analyzer during sequencing of the PCR-amplified library of CDR3 sequences.[23] These errors lead to the generation of artifactual sequence variants that could complicate the estimate of the true diversity of an Ig or TCR library. Therefore, recognition and mitigation of sequencing errors is essential for accurate repertoire enumeration. How can we improve our ability to differentiate true variants from sequencing artifacts? As suggested by Bolotin *et al*.,[73] the more we know about the rates and types of sequence errors produced by a particular method, the better we can tailor our analysis. There are several approaches for obtaining this information. One approach is to analyze a 'gold standard' of known, fixed sequence, whereby variants can be attributed unequivocally to sequencing errors. Another approach is to quantify changes in the germline-encoded regions of the somatically variable template, such as the outer parts of the V and J segments distinct from the CDR3 core of the TCR, which should remain invariant after VDJ recombination. In addition, the use of known standards and repeated measurements are critical to understand the magnitude, distribution and variability of sequencing errors obtained with different methods, which in turn are critical for modeling appropriate error correction.[86] In the sequencing process, we should use a well-defined DNA fragment or synthetic TCR/BCR library (internal control) in the control lane to monitor the sequencing quality and estimation of error rates. In this respect, Control libraries generated from the PhiX virus can serve as an effective control in sequencing runs. Characteristics of the PhiX genome provide several benefits:[87] First, PhiX is a small genome, which enables quick alignment and estimation of error rates. Second, the PhiX genome contains ~45% GC and 55% AT. PhiX has a well-defined genome sequence. Third, Illumina cluster generation algorithms are optimized around a balanced representation of A, T, G and C nucleotides.

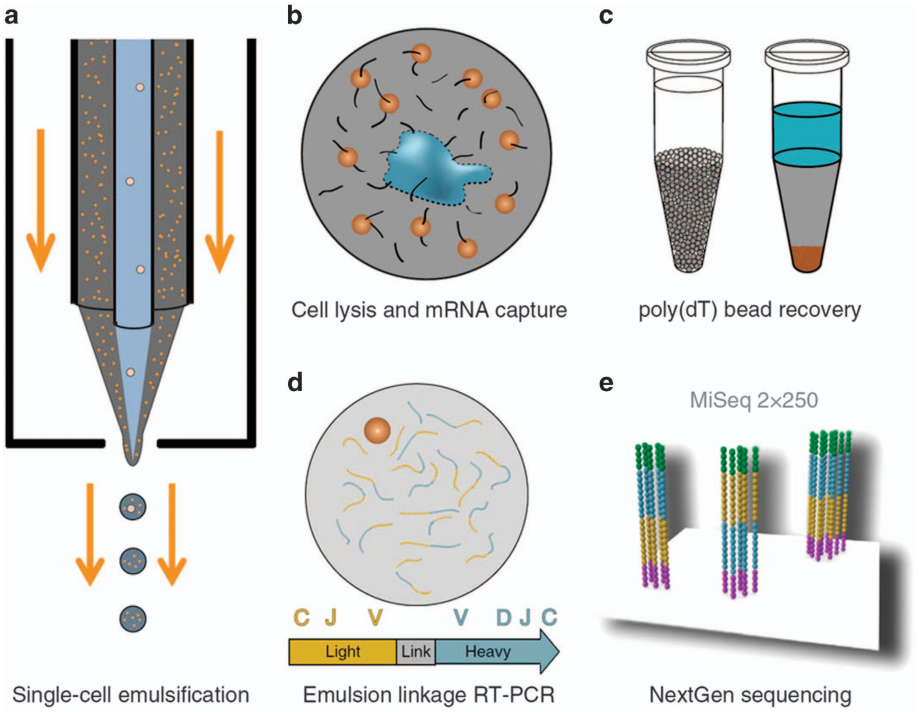

Figure 3. Technical workflow for ultra-high-throughput VH–VL sequencing from single B cells. (**a**) An axisymmetric flow-focusing nozzle isolated single cells and poly(dT) magnetic beads into emulsions of predictable size distributions. An aqueous solution of cells in PBS (center, blue with pink circles) and cell lysis buffer with poly(dT) beads (gray with orange circles) exited an inner and outer needle and were surrounded by a rapidly moving annular oil phase (orange arrows). Aqueous streams focused into a thin jet, which coalesced into emulsion droplets of predictable sizes, and cells mixed with lysis buffer only at the point of droplet formation (Supplementary Figure 1). (**b**) Single-cell VH and VL mRNAs annealed to poly(dT) beads within emulsion droplets (blue figure represents a lysed cell, orange circles depict magnetic beads and black lines depict mRNA strands). (**c**) poly(dT) beads with annealed mRNA were recovered by emulsion centrifugation to concentrate aqueous phase (left) followed by diethyl ether destabilization (right). (**d**) Recovered beads were emulsified for cDNA synthesis and linkage PCR to generate an ~850-bp VH–VL cDNA product. (**e**) Next generation sequencing of VH–VL amplicons was used to analyze the native heavy and light-chain repertoire of input B cells.[90]

Notably, error rate assessment can also be performed by aligning the obtained sequences to TCR J segment germline sequences and by filtering the erroneous reads based on sequence redundancy. For example, Warren *et al.*[24] aligned raw reads to the known TCR J gene segments to assess sequence accuracy. From the aligned raw single pass reads, they observed 9.4 errors per kb, but when they added the requirements of (1) double-strand coverage, (2) minimum quality score of Q30 and (3) no high-quality discrepancy between strands at any position, the error rate fell to 2.2 errors per kb. Nguyen *et al.*[85] analyzed specific transgenic TCRs obtained from RAG-deficient mice, allowing them to express a single germline-rearranged TCR and therefore to compare the sequenced receptor with the original DNA. Their findings showed that the overall frequencies of erroneous CDR3β sequences for the three TCRs were similar at 5.23 ± 0.21%, 5.24 ± 0.12% and 6.00 ± 0.34% for the 5C.C7, OT-1 and DO11.10 TCRs respectively, which were greatly reduced after the filtering process. At a phred (q) values = 30 and cutoff of 1% of single nt mismatch sequences, only 0.0086 ± 0.002%, 0.030 ± 0.006% and 0.057 ± 0.008% of the total sequences were erroneous for the three CDR3s respectively. Therefore, additional filtering of sequence based on quality scores and filtering single nt mismatch sequences have the potential to dramatically decrease overall error rates in CDR3s acquired by NGS. Moreover, it is also important to recognize that PCR amplification per se is non-linear, and there are differences in the efficiency of PCR amplification of CDR3 regions using different Vβ and Jβ gene segments. To estimate the magnitude of any such bias, Robins *et al.*[23] compared the number of observations of 30 000 sequences in the 25-cycle lane with the number of observations in the 40-cycle lane. They found that of the TCRβ CDR3 sequences observed in the 25-cycle PCR lane, 97% were also observed in the 40-cycle PCR lane, and identified that each cycle of PCR amplification potentially introduces a bias of average magnitude $1.5^{1/15} = 1.027$. Accordingly, efforts to limit the number of amplification cycles will reduce quantitative distortions, as well as error rates. Finally, filtering and error rate assessment should be performed with extreme caution because the rare receptor sequences that are presented at very low levels in an individual might be mistaken for error-containing sequences and ignored, leading to underestimation of certain TCR/BCR clonotypes. Based on these considerations, Bolotin *et al.*[73] proposed an advanced error correction algorithm for processing Illumina TCR profiling data. To some extent, this algorithm provides efficient and safe elimination of PCR and sequencing errors with minimal information loss.

Overall, rational analysis of the TCR repertoire by NGS to make firm and clear statements based on the data obtained demands intelligent design of the whole experimental pipeline, including blood sampling, DNA/RNA purification, cDNA synthesis, PCR amplification, pre-sequencing preparation, sequencing depth, platform choice and intelligent interpretation of the NGS output with regard to all potential errors accumulated above. We believe that as the technology develops, improvements in the accuracy of the sequencing technologies themselves and the reliability of the base-calling quality information provided will make the task of analysis easier.

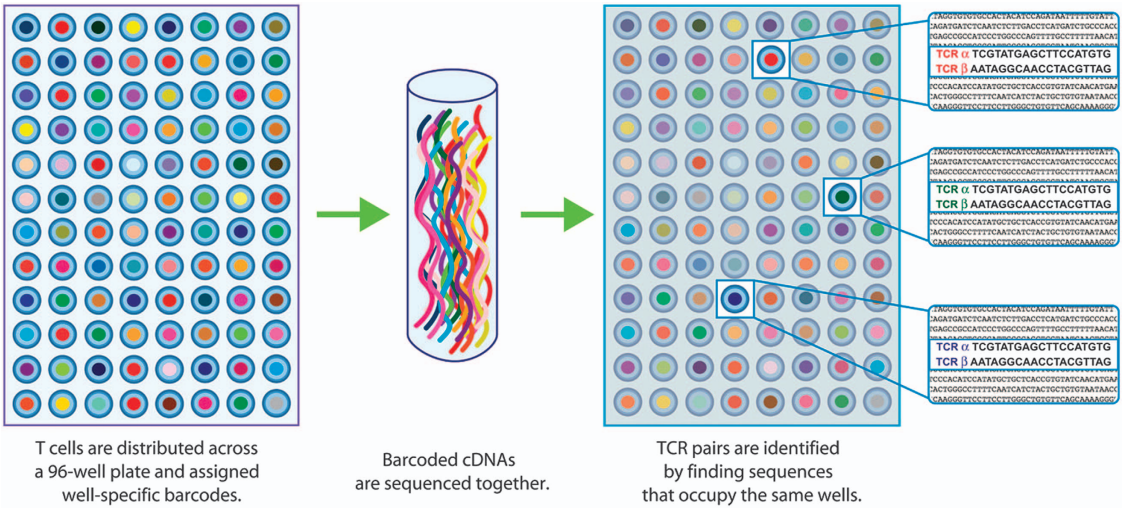

Figure 4. Schematic of the pairSEQ approach. A fixed number of T cells are randomly allocated to each well on a 96-well plate, and their mRNA is extracted, converted to cDNA, and amplified by TCR-specific primers. Well-specific barcodes are attached, and the TCR molecules are pooled for sequencing, followed by computational demultiplexing to map each TCR sequence back to the wells in which it originated. The immune repertoire is highly diverse, and the probability that two clones will occupy exactly the same sets of wells is miniscule, so any pair of TCRA and TCRB sequences that uniquely shares a set of wells can be inferred to have come from the same clone.[91]

ADVANCES IN PAIRED TCR/BCR SEQUENCING

It is worth noticing that most published TCR/BCR repertoire studies have been limited to a single chain (for example, TCR β-chains). This approach disrupts cognate pairing of TCR alpha and beta chains (or heavy and light chains) and cannot provide information about the identity of immune receptor pairs encoded by individual T or B lymphocytes, so that they revealed only 'half the truth'. To reconstitute TCRs/BCRs for functional analysis, therapeutic use or modeling of receptor–antigen binding, the TCRA (T-cell receptor alpha chain) and TCRB chains from a complete TCR (or the heavy and light chains from a complete BCR) must be identified as a pair. There have been multiple attempts to pair heavy and light chains in B, and α and β chains in T cells using single-cell technology. One approach is to isolate individual lymphocytes and physically link the chains by bridge PCR before sequencing; alternatively, the heavy and light chains can be barcoded at the single-cell level. However, these approach are limited by low efficiency or low cell throughput and require fabrication and operation of complicated microfluidic devices.[21,88] Chudakov and co-workers[88] recently reported the use of one-pot cell encapsulation within water-in-oil emulsions, achieving cell lysis by heating at 65 °C concomitant with reverse transcription of the genes encoding TCRα and TCRβ and linking by overlap extension PCR to determine TCRα–TCRβ pairings, albeit only for TCRβV7 and with a very low efficiency (~700 TCRα–TCRβ pairs recovered from 8×10^6 PBMCs). Likewise, it was recently demonstrated that thousands of Ig heavy- and light-chain pairs could be obtained by bead capture of single B-cell mRNA followed by linkage PCR in single-bead-containing emulsion droplets.[20] Although the yield of each of these approaches is modest, and the techniques themselves are technically challenging, these studies represent very important advances toward the goal of deep, cheap and fast profiling of dimeric antigen receptors. With the in-depth study and advancement of experiment technology, paired sequencing strategy has also been improved. Inspired by methods for the production of highly monodisperse polymeric microspheres for drug delivery purposes,[89] DeKosky *et al.*[90] developed a low-cost, single-cell, emulsion-based technology for sequencing antibody VH–VL repertoires from $>2\times10^6$ B cells per experiment with demonstrated pairing precision >97%. The experimental procedure is shown in Figure 3a—simple flow-focusing apparatus is used to sequester single B cells into emulsion droplets containing lysis buffer and magnetic beads for mRNA capture; subsequent emulsion reverse transcription–PCR generate VH–VL amplicons for next-generation sequencing. The workflow presented here is high throughput, which permits sequence analysis of the entire population of human B cells contained in a 10-ml blood draw, or, if needed, even in a unit of blood (450 ml) in a single-day experiment. In addition, Howie *et al.*[91] presented a technology to pair lymphocyte receptor sequences at high throughput without the need for isolated single-cell methods. Their strategy uses combinatorics, rather than physical isolation, to match the pairs. Figure 4 shows how these ideas are implemented in a pairSEQ experiment. Moreover, they demonstrated the high accuracy and throughput of pairSEQ by identifying TCR pairs from a wide range of clonal frequencies in multiple sample types, including more than 2 00 000 pairs from a single 96-well plate.

Taken together, as DNA-sequencing technologies continue to progress, low-cost high-throughput single-cell antibody sequencing can enable paired antibody repertoire analysis at great depth in large study cohorts and clinical patients, which enables rapid interrogation of the immune response and can be applied to investigate B-cell maturation, vaccine efficacy, immune system health and autoimmunity in clinical and research settings. As indicated earlier, pairSEQ can be used to identify the TCRs of tumor-infiltrating lymphocytes, and the resulting information can be used to engineer T cells to express tumor antigen-targeting receptors.

CONCLUSIONS

NGS technologies have revolutionized the study of immune repertoires. These methods provide a previously unimaginable amount of sequence data. As in many areas of biology, the rate limiting steps now seem to be data management and analysis rather than acquisition. Therefore, the rapid changes and development in the field of repertoire sequencing call for new databases, computational algorithms and software to analyze whole repertoires, and for the comparisons between species, to produce meaningful data. Bashford-Rogers *et al.*[92] developed a novel computational approach for BCR repertoire analysis using established NGS methods coupled with network construction and

population analysis. Based on this approach, they showed the short-term effect of therapy on the B-cell repertoire in chronic lymphocytic leukemia. Moreover, enormous quantity of reads generated by NGS technologies necessitates cautious interpretation. Potential errors during the sequencing process may skew interpretation. For antigen receptors, the challenges posed by the vast quantities of data generated by NGS technologies will require dedicated solutions beyond those developed for genome sequencing, which may differ depending on the sequencing technology used and the purpose of the experiment. Many investigators are developing such methods, based on different sequencing platforms, but critical details of protocol and performance are proprietary. The field will move forward when these methods are shared and standardized, and when the accuracy, sensitivity and reproducibility of various sequencing and analytic methods are evaluated using standardized samples in comparative experiments. Moreover, it is worth noticing that bulk deep sequencing while offer advantages of generating massive amount of data they do not offer paired structure, while recent Paired-Seq offers to resolve complete repertoire, which should have been widely applied to promote research into the nature and antigen specificity of many medically important protective or pathological T/B-cell responses, allowing for the development of novel diagnostic, therapeutic or preventive strategies. For example, recently published results have shown that the pairSEQ technology has the potential to rapidly identify sequence pairs of tumor-infiltrating lymphocytes, which can then be used to reconstruct TCR receptors within T cells engineered for cancer immunotherapy.[91]

In conclusion, currently, the major issues are: how best to prepare immune-receptor-sequence libraries, which sequencing technologies to use, how to analyze the data, and how to relate sequence data to the functional activities of the Ig or TCR complexes. We believe that as the technology develops rapidly, scientific discoveries in TCR/BCR repertoire studies will form the basis for new clinical applications in personalized medicine and will provide a deeper understanding of immune behavior and immune response.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by funds received from the National Natural Science Foundation of China (No. 81271810, 81571953), 12-5 state S&T Projects for infectious diseases (2012ZX10002-007), Doctoral Fund of Ministry of Education of China (20120101110009), and Zhejiang medical science and technology project (2015118507).

REFERENCES

1 Tonegawa S. Somatic generation of antibody diversity. *Nature* 1983; **302**: 575–581.
2 Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. *Nature* 1988; **334**: 395–402.
3 Pannetier C, Cochet M, Darche S, Casrouge A, Zoller M, Kourilsky P. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor beta chains vary as a function of the recombined germ-line segments. *Proc Natl Acad Sci USA* 1993; **90**: 4319–4323.
4 Cabaniols JP, Fazilleau N, Casrouge A, Kourilsky P, Kanellopoulos JM. Most alpha/beta T cell receptor diversity is due to terminal deoxynucleotidyl transferase. *J Exp Med* 2001; **194**: 1385–1390.
5 Miles JJ, Douek DC, Price DA. Bias in the αβ T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunol Cell Biol* 2011; **89**: 375–387.
6 Li Z, Woo CJ, Iglesias-Ussel MD, Ronai D, Scharff MD. The generation of antibody diversity through somatic hypermutation and class switch recombination. *Genes Dev* 2004; **18**: 1–11.
7 Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. *Science* 2009; **324**: 807–810.
8 Batrak V, Blagodatski A, Buerstedde JM. Understanding the immunoglobulin locus specificity of hypermutation. *Methods Mol Biol* 2011; **745**: 311–326.
9 Nikolich-Zugich J, Slifka MK, Messaoudi I. The many important facets of T-cell repertoire diversity. *Nat Rev Immunol* 2004; **4**: 123–132.
10 Harty JT, Badovinac VP. Shaping and reshaping CD8 T cell memory. *Nat Rev Immunol* 2008; **8**: 107–119.
11 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.
12 Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. Diversity of human alpha beta T cell receptors. *Science* 2000; **288**: 1135.
13 Woodsworth DJ, Castellarin M, Holt RA. Sequence analysis of T-cell repertoires in health and disease. *Genome Med* 2013; **5**: 1–13.
14 Maciejewski JP, O'Keefe C, Gondek L, Tiu R. Immune-mediated bone marrow failure syndromes of progenitor and stem cells: molecular analysis of cytotoxic T cell clones. *Folia Histochem Cytobiol* 2007; **45**: 5–14.
15 Diu A, Romagne F, Genevee C, Rocher C, Bruneau JM, David A *et al.* Fine specificity of monoclonal antibodies directed at human T cell receptor variable regions: comparison with oligonucleotide-deriven amplification for evaluation of Vβ expression. *Eur J Immunol* 1993; **23**: 1422–1429.
16 Klarenbeek PL, de Hair MJ, Doorenspleet ME, van Schaik BD, Esveldt RE, van de Sande MG *et al.* Inflamed target tissue provides a specific niche for highly expanded T-cell clones in early human autoimmune disease. *Ann Rheum Dis* 2012; **71**: 1088–1093.
17 Li S, Lefranc MP, Miles JJ, Alamyar E, Giudicelli V, Duroux P *et al.* IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat Commun* 2013; **4**: 2333.
18 Sui W, Hou X, Zou G, Che W, Yang M, Zheng C *et al.* Composition and variation analysis of the TCR-chain CDR3 repertoire in systemic lupus erythematosus using high-throughput sequencing. *Mol Immunol* 2015; **67**: 455–464.
19 Wilson PC, Andrews SF. Tools to therapeutically harness the human antibody response. *Nat Rev Immunol* 2012; **12**: 709–719.
20 DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM *et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* 2013; **31**: 166–169.
21 Furutani S, Nagai H, Takamura Y, Aoyama Y, Kubo I. Detection of expressed gene in isolated single cells in microchambers by a novel hot cell-direct RT-PCR method. *Analyst* 2012; **137**: 2951–2957.
22 Maecker HT, Lindstrom TM, Robinson WH, Utz PJ, Hale M, Boyd SD *et al.* New tools for classification and monitoring of autoimmune diseases. *Nat Rev Rheumatol* 2012; **8**: 317–328.
23 Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O *et al.* Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood* 2009; **114**: 4099–4107.
24 Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R *et al.* Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res* 2011; **21**: 790–797.
25 Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T *et al.* A mechanism for TCR sharing between T-cell subsets and individuals revealed by pyrosequencing. *J Immunol* 2011; **186**: 4285–4294.
26 Wang C, Sanders CM, Yang Q, Schroeder Jr HW, Wang E, Babrzadeh F *et al.* High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T-cell subsets. *Proc Natl Acad Sci USA* 2010; **107**: 1518–1523.
27 Venturi V, Kedzierska K, Turner SJ, Doherty PC, Davenport MP. Methods for comparing the diversity of samples of the T cell receptor repertoire. *J Immunol Methods* 2007; **321**: 182–195.
28 Keylock C. Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy. *Oikos* 2005; **109**: 203–207.
29 Stewart JJ, Lee CY, Ibrahim S, Watts P, Shlomchik M, Weigert M *et al.* A Shannon entropy analysis of immunoglobulin and T cell receptor. *Mol Immunol* 1997; **34**: 1067–1082.
30 Wu J, Liu D, Tu W, Song W, Zhao X. T-cell receptor diversity is selectively skewed in T-cell populations of patients with Wiskott-Aldrich syndrome. *J Allergy Clin Immunol* 2015; **135**: 209–216.
31 Venturi V, Kedzierska K, Tanaka MM, Turner SJ, Doherty PC, Davenport MP. Method for assessing the similarity between subsets of the T cell receptor repertoire. *J Immunol Methods* 2008; **329**: 67–80.
32 Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGTs tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/ HighV-QUEST for NGS. *Methods Mol Biol* 2012; **882**: 569–604.

33 Bolotin DA, Shugay M, Mamedov IZ, Putintseva EV, Turchaninova MA, Zvyagin IV *et al.* software for T-cell receptor sequencing data analysis. *Nat Methods* 2013; **10**: 813–814.
34 Nakamura Y, Komiyama T, Furue M, Gojobori T, Akiyama Y. CIG-DB: the database for human or mouse immunoglobulin and T cell receptor genes available for cancer studies. *BMC Bioinformatics* 2010; **11**: 398.
35 Belz GT, Xie W, Doherty PC. Diversity of epitope and cytokine profiles for primary and secondary influenza a virus-specific CD8+ T cell responses. *J Immunol* 2001; **166**: 4627–4633.
36 Echchakir H, Dorothée G, Vergnon I, Menez J, Chouaib S, Mami-Chouaib F. Cytotoxic T lymphocytes directed against a tumor-specific mutated antigen display similar HLA tetramer binding but distinct functional avidity and tissue distribution. *Proc Natl Acad Sci USA* 2002; **99**: 9358–9363.
37 Manjunath N, Shankar P, Stockton B, Dubey PD, Lieberman J, von Andrian UH. A transgenic mouse model to analyze CD8+ effector cell differentiation in vivo. *Proc Natl Acad Sci USA* 1999; **96**: 13932–13937.
38 Gudmundsdottir H, Wells AD, Turka LA. Dynamics and requirements of T cell clonal expansion in vivo at the single-cell level: effector function is linked to proliferative capacity. *J Immunol* 1999; **162**: 5212–5223.
39 Slifka MK, Whitton JL. Functional avidity maturation of CD8+ T cells without selection of higher affinity TCR. *Nat Immunol* 2001; **2**: 711–717.
40 Manjunath N, Shankar P, Wan J, Weninger W, Crowley MA, Hieshima K *et al.* Effector differentiation is not prerequisite for generation of memory cytotoxic T lymphocytes. *J Clin Invest* 2001; **108**: 871–878.
41 Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* 2000; **13**: 37–45.
42 Ivanov II, Schelonka RL, Zhuang Y, Gartland GL, Zemlin M, Schroeder HW Jr. Development of the expressed Ig CDR-H3 repertoire is marked by focusing of constraints in length, amino acid use, and charge that are first established in early B cell progenitors. *J Immunol* 2005; **174**: 7773–7780.
43 Wardemann H, Yurasov S, Schaefer A, Young JW, Meffre E, Nussenzweig MC. Predominant autoantibody production by early human B cell precursors. *Science* 2003; **301**: 1374–1377.
44 Aguilera I, Melero J, Nuñez-Roldan A, Sanchez B. Molecular structure of eight human autoreactive monoclonal antibodies. *Immunology* 2001; **102**: 273–280.
45 Ichiyoshi Y, Casali P. Analysis of the structural correlates for antibody polyreactivity by multiple reassortments of chimeric human immuno-globulin heavy and light chain V segments. *J Exp Med* 1994; **180**: 885–895.
46 Larimore K, McCormick MW, Robins HS, Greenberg PD. Shaping of human germline IgH repertoires revealed by deep sequencing. *J Immunol* 2012; **189**: 3221–3230.
47 Yassai M, Ammon K, GovermanJ, Marrack P, Naumov Y, Gorski J. A molecular marker for thymocyte-positive selection: selection of CD4 single-positive thymocytes with shorter TCRB CDR3 during T cell development. *J Immunol* 2002; **168**: 3801–3807.
48 Pickman Y, Dunn-Walters D, Mehr R. BCR CDR3 length distributions differ between blood and spleen and between old and young patients, and TCR distributions can be used to detect myelodysplastic syndrome. *Phys Biol* 2013; **10**: 056001.
49 Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res* 2009; **19**: 1817–1824.
50 Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR *et al.* Overlap and effective size of the human CD8+ T-cell receptor repertoire. *Sci Transl Med* 2010; **2**: 47ra64.
51 Hardiman G. Next-generation antibody discovery platforms. *Proc Natl Acad Sci USA* 2012; **109**: 18245–18246.
52 Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* 2010; **116**: 1070–1078.
53 Liaskou E, Henriksen EK, Holm K, Kaveh F, Hamm D, Fear J *et al.* High-throughput T-cell receptor sequencing across chronic liver diseases reveals distinct disease-associated repertoires. *Hepatology* 2015; e-pub ahead of print 7 August 2015; doi:10.1002/hep.28116.
54 Even J, Lim A, Puisieux I, Ferradini L, Dietrich PY, Toubert A *et al.* T-cell repertoires in healthy and diseased human tissues analysed by T-cell receptor beta-chain CDR3 size determination: Evidence for oligoclonal expansions in tumours and inflammatory diseases. *Res Immunol* 1995; **146**: 65–80.
55 Manfras BJ, Terjung D, Boehm BO. Non-productive human TCRβchain genes represent V-D-J diversity before selection upon function: Insight into biased usage of TCRBD and TCRBJ genes and diversity of CDR3 region length. *Hum Immunol* 1999; **60**: 1090–1100.
56 Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B *et al.* Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci Transl Med* 2009; **1**: 12ra23.
57 Krangel MS. Gene segment selection in V(D)J recombination: Accessibility and beyond. *Nat Immunol* 2003; **4**: 624–630.
58 Cibotti R, Cabaniols JP, Pannetier C, Delarbre C, Vergnon I, Kanellopoulos JM *et al.* Public and private Vβ T cell receptor repertoires against hen egg white lysozyme (HEL) in nontransgenic versus HEL transgenic mice. *J Exp Med* 1994; **180**: 861–872.
59 Trautmann L, Rimbert M, Echasserieau K, Saulquin X, Neveu B, Dechanet J *et al.* Selection of T cell clones expressing high-affinity public TCRs within human cytomegalovirus-specific CD8 T cell responses. *J Immunol* 2005; **175**: 6123–6132.
60 Boudinot P, Boubekeur S, Benmansour A. Rhabdovirus infection induces public and private T cell responses in teleost fish. *J Immunol* 2001; **167**: 6202–6209.
61 Bousso P, Casrouge A, Altman JD, Haury M, Kanellopoulos J, Abastado JP *et al.* Individual variations in the murine T cell response to a specific peptide reflect variability in naive repertoires. *Immunity* 1998; **9**: 169–178.
62 Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W *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.
63 Klarenbeek PL, Tak PP, van Schaik BD, Zwinderman AH, Jakobs ME, Zhang Z *et al.* Human T-cell memory consists mainly of unexpanded clones. *Immunol Lett* 2010; **133**: 42–48.
64 Cebula A, Seweryn M, Rempala GA, Pabla SS, McIndoe RA, Denning TL *et al.* Thymus-derived regulatory T cells contribute to tolerance to commensal microbiota. *Nature* 2013; **497**: 258–262.
65 Linton PJ, Dorshkind K. Age-related changes in lymphocyte development and function. *Nat Immunol* 2004; **5**: 133–139.
66 Murasko DM, Jiang J. Response of aged mice to primary virus infections. *Immunol Rev* 2005; **205**: 285–296.
67 Naylor K, Li G, Vallejo AN, Lee WW, Koetz K, Bryl E *et al.* The influence of age on T cell generation and TCR diversity. *J Immunol* 2005; **174**: 7446–7452.
68 Messaoudi I, Lemaoult J, Guevara-Patino JA, Metzner BM, Nikolich-Zugich J. Age-related CD8 T cell clonal expansions constrict CD8 T cell repertoire and have the potential to impair immune defense. *J Exp Med* 2004; **200**: 1347–1358.
69 Yager EJ, Ahmed M, Lanzer K, Randall TD, Woodland DL, Blackman MA. Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus. *J Exp Med* 2008; **205**: 711–723.
70 Siegrist CA, Aspinall R. B-Cell responses to vaccination at the extremes of age. *Nat Rev Immunol* 2009; **9**: 185–194.
71 Nikolich-Zugich J. T cell aging: naive but not young. *J Exp Med* 2005; **201**: 837–840.
72 van den Brink MR, Alpdogan O, Boyd RL. Strategies to enhance T-cell reconstitution in immunocompromised patients. *Nat Rev Immunol* 2004; **4**: 856–867.
73 Bolotin DA, Mamedov IZ, Britanova OV, Zvyagin IV, Shagin D, Ustyugova SV *et al.* Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur J Immunol* 2012; **42**: 3073–3083.
74 Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 2012; **30**: 434–439.
75 Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci USA* 2008; **105**: 13081–13086.
76 Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C *et al.* Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* 2011; **333**: 1593–1602.
77 Douek DC, Betts MR, Brenchley JM, Hill BJ, Ambrozak DR, Ngai KL *et al.* A novel approach to the analysis of specificity, clonality, and frequency of HIV-specific T cell responses reveals a potential mechanism for control of viral escape. *J Immunol* 2002; **168**: 3099–3104.
78 Bhalla AD, Gudikote JP, Wang J, Chan WK, Chang YF, Olivas OR *et al.* Nonsense codons trigger an RNA partitioning shift. *J Biol Chem* 2009; **284**: 4062–4072.
79 Wang J, Vock VM, Li S, Olivas OR, Wilkinson MF. A quality control pathway that down-regulates aberrant T-cell receptor (TCR) transcripts by a mechanism requiring UPF2 and translation. *J Biol Chem* 2002; **277**: 18489–18493.
80 Li H, Ye C, Ji G, Wu X, Xiang Z, Li Y *et al.* Recombinatorial biases and convergent recombination determine interindividual TCRβ sharing in murine thymocytes. *J Immunol* 2012; **189**: 2404–2413.
81 Britanova OV, Putintseva EV, Shugay M, Merzlyak EM, Turchaninova MA, Staroverov DB *et al.* Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J Immunol* 2014; **192**: 2689–2698.
82 Matz M, Shagin D, Bogdanova E, Britanova O, Lukyanov S, Diatchenko L *et al.* Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Res* 1999; **27**: 1558–1560.

83 Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* 2001; **30**: 892–897.
84 Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* 2012; **135**: 183–191.
85 Nguyen P, Ma J, Pei D, Obert C, Cheng C, Geiger TL. Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics* 2011; **12**: 106.
86 Warren RL, Nelson BH, Holt RA. Profiling model T-cell metagenomes with short reads. *Bioinformatics* 2009; **25**: 458–464.
87 Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand Genomic Sci* 2015; **10**: 18.
88 Turchaninova MA, Britanova OV, Bolotin DA, Shugay M, Putintseva EV, Staroverov DB *et al.* Pairing of T-cell receptor chains via emulsion PCR. *Eur J Immunol* 2013; **43**: 2507–2515.
89 Berkland C, Pollauf E, Pack DW, Kim K. Uniform double-walled polymer microspheres of controllable shell thickness. *J Control Release* 2004; **96**: 101–111.
90 DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD *et al.* In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat Med* 2015; **21**: 86–91.
91 Howie B, Sherwood AM, Berkebile AD, Berka J, Emerson RO, Williamson DW *et al.* High-throughput pairing of T cell receptor α and β sequences. *Sci Transl Med* 2015; **7**: 1–11.
92 Bashford-Rogers RJ, Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA *et al.* Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res* 2013; **23**: 1874–1884.

Supplementary Information accompanies this paper on Genes and Immunity website (http://www.nature.com/gene)