High-throughput sequencing of B- and T-lymphocyte antigen receptors in hematology

Edus H. Warren,1,3 Frederick A. Matsen IV,2 and Jeffrey Chou1,3

1Program in Immunology, Clinical Research Division, and 2Program in Computational Biology, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, U.S.A.; 3Department of Medicine, University of Washington, Seattle, WA, U.S.A.

Running head: Deep sequencing of lymphocyte antigen receptors

Corresponding author and address for correspondence:

Edus H. Warren, MD, PhD
Program in Immunology
Fred Hutchinson Cancer Research Center
1100 Fairview Ave N, D3-100
P.O. Box 19024
Seattle WA 98109-1024
Telephone: 206.667.6441
E-mail: ehwarren@u.washington.edu

Keywords: Immunobiology
Abstract

Application of high-throughput DNA sequencing to the analysis of B- and T-lymphocyte antigen receptors has great potential for improving the monitoring of lymphoid malignancies, assessing immune reconstitution after hematopoietic cell transplantation, and characterizing the composition of lymphocyte repertoires. Current technology can define the number and frequency of IGH, TCRα, TCRβ, or TCRγ chains expressed in a population of lymphocytes; techniques for determining the number of antigen receptor heterodimers, such as TCRαβ pairs, expressed in the population are under development.
B and T lymphocytes are unique amongst somatic cells because much of their fundamental biology is determined by DNA sequences that are not encoded in the germline. The genes that encode B- and T-cell antigen receptors undergo rearrangement during lymphocyte development in a process that concatenates noncontiguous variable (V), diversity (D), and joining (J) gene segments to assemble sequences encoding functional receptors whose antigenic specificity is determined by 3 complementarity-determining regions (CDRs) (Figure 1; Supplementary File 1). The V gene segments encode the first and second CDRs of both B- and T-cell receptors, while the V, D, and J segments collectively encode the third CDR (CDR3), which is the most critical determinant of antigenic specificity. Extremely diverse repertoires of B- and T-cell antigen receptors are created by exploiting both combinatorial diversity – made possible by the existence of many distinct V, D, and J gene segments in the 4 T-cell and 3 B-cell antigen receptor loci – as well as junctional diversity. The latter is created by the unique molecular mechanism that mediates V-D-J rearrangement, and involves template-independent insertion and deletion of nucleotides at the V-D, D-J, and V-J splice junctions. The diversity of the B-cell receptor (BCR) repertoire is further increased by somatic hypermutation (SHM) of previously rearranged BCR genes during affinity maturation after initial antigen encounter. SHM is not limited to the V-D-J region that encodes the CDR3, and occurs throughout the V gene segment, potentially affecting the sequence of CDR1 and CDR2 as well as the intervening framework regions.

The antigen receptor repertoires created by the processes of V-D-J rearrangement and SHM are far too diverse to be amenable to comprehensive exploration and definition by conventional capillary-based DNA sequencing. Over the past 5 years, high-throughput
DNA sequencing has been applied to the analysis of BCR and T-cell receptor (TCR) CDR3 region sequence repertoires, enabling analysis of the repertoires realized in any given individual with unprecedented depth, resolution, and accuracy. The CDR3 region in the vast majority of successfully rearranged TCR β chains and immunoglobulin heavy (IGH) chains comprises no more than 66 or 90 base pairs (bp), encoding no more than 22 or 30 amino acids, respectively. This region is therefore ideally suited to comprehensive definition by current sequencing platforms that can generate ~1 x 10^9 reads of the requisite length in a single run. Although these platforms cannot yet capture the entire V gene segment in addition to the V-D-J region, which limits their utility for analysis of SHM in BCR genes, SHM can be studied using other platforms that provide longer read lengths, albeit with less depth per sequencing run (Figure 2). Both genomic DNA and cDNA from lymphocytes have been used as the template for sequencing, but genomic DNA is preferred for studies in which the relative frequency of specific BCR or TCR sequences is to be assessed. Molecular and computational strategies based on high-throughput DNA sequencing have made it possible to explore BCR and TCR sequence repertoires of any degree of complexity, and thereby to address biological questions that were never before amenable to direct experimental analysis.

**Defining the characteristics of B- and T-cell repertoires**

High-throughput sequencing of rearranged BCR and TCR genes in lymphocytes from peripheral blood has begun to define the essential characteristics of the B- and T-cell repertoires in healthy adults. Sequencing of the IGH locus in naïve B cells from 4 individuals revealed an average of ~86,000 unique CDR3 sequences per person. Up to
650,000 unique TCRβ CDR3 sequences were observed in naïve CD8+ T-cells from single individuals,7 and exhaustive TCRβ sequencing of pooled naïve and memory T cells from a single healthy adult revealed > 1 x 10^6 unique sequences.10 Statistical methods developed for estimating total species diversity in ecological and microbial communities11,12 have been used to estimate total CDR3 sequence diversity, including the sequences that are directly observed as well as those that remain “unseen” due to the fact that such diverse repertoires cannot be sampled completely. Such efforts have suggested that the naïve and memory CD8+ and CD4+ TCR β chain CDR3 repertoires in healthy adults collectively contain at least 3-4 x 10^6 unique sequences,6 with at least as many unique TCR α chain CDR3 sequences.8 The IGH CDR3 repertoire is estimated to contain at least 10^6 distinct sequences.2 The IGH and TCRβ repertoires are both characterized by an extremely large number of CDR3 sequences that are present at low frequencies.2,5-7 It has therefore proven difficult, despite the application of methods for estimating unseen species diversity, to determine the true diversity in individual BCR and TCR repertoires with great accuracy, and the published estimates are best viewed as lower bounds for repertoire diversity.

Although most of the deep sequencing studies published to date have focused on the TCRβ and IGH loci, the loci that encode the α chains of αβ T cells and the γ chains of γδ T cells have also been explored. Analysis of TCRα CDR3 sequences in several subsets of αβ T cells from a single individual revealed 1.2 to 2.4-fold more unique TCRα than TCRβ CDR3 sequences in each the subsets studied.8 Sequencing of the TCRγ locus has provided valuable insight into the structure of the γδ T-cell repertoire. Although the theoretical diversity of γδ TCRs rivals that of αβ TCRs, the antigen receptors expressed by peripheral blood γδ T cells demonstrate very limited TCR CDR3 sequence diversity.
Indeed, deep sequencing of the rearranged TCRγ genes expressed in peripheral blood γδ T cells from 3 healthy adults revealed that 45% of the TCRγ CDR3 sequences from the 3 individuals were identical to a previously described sequence found in a “public” γδ TCR, found in most individuals, that is specifically reactive with non-peptide prenyl pyrophosphate antigens.13

Deep sequencing of the TCRγ locus in αβ and γδ T cells from peripheral blood of healthy adults has also provided valuable insight into the process by which T cells are committed to the αβ or γδ lineages during thymic development. The vast majority of both αβ and γδ T cells in peripheral blood contain rearranged TCRγ genes, suggesting that rearrangement of the TCRγ locus occurs before αβ/γδ lineage commitment.14-16 In contrast, only a very small fraction (<4%) of γδ T cells appear to have rearranged TCRβ loci, suggesting that rearrangement of the TCRβ locus only occurs in T cells that have committed to the αβ lineage.16

Comparison of the TCR repertoires carried in specific T cell subsets within single individuals is the focus of several current studies, and has great potential for defining the relationships between subsets as well as illuminating the developmental pathways that lead to each subset. Analysis of TCRα and TCRβ diversity in the CD4+ regulatory T cell (Treg) compartment in a single individual, for example, demonstrated higher TCR diversity in this subset than in any other major subset.8 This observation is consistent with the results of previous studies of both murine17 and human18 Tregs, performed before the advent of high-throughput sequencing, which demonstrated that CD4+ Treg TCR diversity is at least comparable to, if not in fact greater than, that observed in the conventional CD4+ compartment.
Reconstitution of B- and T-cell repertoires after hematopoietic cell transplantation

Recent TCRβ and IGH sequencing studies have characterized reconstitution of the αβ T-cell and B-cell compartments in recipients of allogeneic hematopoietic cell transplantation (HCT). One study of αβ T cell reconstitution in 10 allogeneic HCT recipients, in which a uniform sequencing depth of > 1 x 10^6 reads per sample was achieved, demonstrated reduced TCRβ complexity at 100 days posttransplant, with modest improvement noted by 1 year posttransplant. A different study of 28 allogeneic HCT recipients, in which the sequencing depth was non-uniform and 50-200-fold lower (< 1.5 x 10^4 reads per sample), evaluated reconstitution of TCRβ diversity in the CD8+ and CD4+ compartments of cord blood and T cell-depleted (TCD) HCT recipients. At 6 months posttransplant the cord blood recipients demonstrated CD8+ and CD4+ TCRβ diversity comparable to healthy adults, but the TCD recipients had > 10-fold lower diversity; by one year posttransplant the CD4+ but not the CD8+ TCRβ diversity in the TCD recipients had improved. Serial sequencing of the IGH locus with a uniform number of 15,000 reads/sample in 6 CLL patients after allogeneic HCT revealed that reconstitution of IGH diversity to levels observed in the HCT donor typically required more than 1 year. Although repertoire diversity was a primary end point in all 3 of these studies, they did not address the important issues of sampling depth, the existence of “unseen” CDR3 sequence diversity, or possible confounding by inter-sample variation in absolute T or B lymphocyte count. Nonetheless, these 3 pilot studies have provided valuable initial data on the reconstitution of lymphocyte repertoire diversity after allogeneic HCT, and have laid the foundation for more comprehensive future studies.
Tracking malignant or therapeutically administered lymphocytes

High-throughput sequencing can be exploited to detect the rearranged CDR3 sequences carried in malignant B and T cells with unprecedented sensitivity and specificity. *IGH*\(^{1,2,9,21-24}\) and *TCRβ/γ*\(^{25}\) sequencing is already being utilized to identify the malignant clone(s) in patients with B- and T-lymphoid malignancies, respectively, at the time of diagnosis, and to track them during and after therapy. Deep sequencing of antigen-receptor loci has multiple advantages when compared with multi-parameter flow cytometry or custom-designed patient- and clone-specific PCR assays for the monitoring of lymphoid malignancies. Deep sequencing demands less time and labor, has superior sensitivity\(^{22,23,25}\), and can simultaneously track all of the clones that comprise the malignant population. Its utility for monitoring disease burden in patients with CLL\(^1,2,9,21\), pediatric B-lineage ALL\(^{22,23}\), and T-lineage ALL\(^{25}\) has been compellingly demonstrated, and it will undoubtedly have utility for monitoring other lymphoid malignancies. Serial sequencing of the *IGH* locus in pediatric B-ALL has revealed surprisingly dynamic evolution of the locus in some patients,\(^{22}\) demonstrating that this technology may also generate valuable insights into the biology of B- and T-cell cancers.

Deep sequencing of antigen receptor genes has numerous other potential applications. It will likely become the technique of choice for tracking the fate of therapeutically administered lymphocytes, as its sensitivity, precision, and capacity for monitoring multiple clones simultaneously are far superior to that which can be achieved with clone-specific PCR or tetramer-based assays. Comprehensive profiling of specific lymphocyte populations, such as tumor infiltrating lymphocytes or lymphocytic infiltrates
associated with infections, graft rejection, or autoimmune disease, is also readily achieved with deep sequencing.

**Limitations of antigen receptor sequencing**

Current strategies for analysis of BCR and TCR repertoires using high-throughput sequencing have notable limitations. All protocols use multiplex PCR to amplify the rearranged variable regions of antigen receptor loci to provide sufficient template DNA for sequencing, which introduces bias due to more efficient amplification of some templates compared with others. Such bias can be addressed molecularly and computationally if its magnitude can be accurately determined. Analysis and interpretation of antigen receptor sequence data also poses a number of lingering computational challenges, one of the most difficult of which is inferring the most likely sequence of molecular events that produced each antigen receptor sequence. Addressing these challenges has provided the rationale for the nascent field of computational immunology.

**Conclusions and future research directions**

Application of high-throughput sequencing to the analysis of B- and T-cell antigen receptor repertoires is elucidating the fundamental structure of lymphocyte repertoires and seems poised to transform the diagnosis and monitoring of B- and T-lymphoid malignancies. The scientific and clinical utility of this innovative technology will strongly depend on the accuracy, precision, and reproducibility with which it can define both the number of unique antigen receptor sequences in a given clinical specimen and their relative frequency. Future studies that aim to compare the relative frequency of specific BCR or
TCR sequences in two or more samples must take into account the many potentially confounding factors that could influence such comparisons, prominent among which are the proportion of lymphocyte DNA in the samples, the sequencing depth, and the underlying clonal structure of the samples. Studies that aim to determine the diversity of antigen receptors expressed in lymphocyte populations should explicitly take into account the existence of “unseen” diversity and the possibility that its magnitude may differ significantly from sample to sample.

The antigen receptors in B cells, αβ T cells, and γδ T cells comprise heterodimers whose 2 constituent chains are generated by independent rearrangement events at distinct chromosomal loci. The true diversity of BCRs or TCRs expressed in an individual is thus defined by the number of distinct pairs of chains that occur, which is likely to be higher than the simple number of distinct IGH or TCRβ monomers that are generated. The development of molecular and computational strategies for estimating the number and frequency of distinct heavy/light, αβ, and γδ pairs in the B- and T-cell repertoires is therefore an important research priority. Techniques for simultaneous sequencing of the variable regions of the heavy/light or αβ gene pairs expressed in single lymphocytes have been developed, and extension of these techniques to the analysis of millions of lymphocytes is likely achievable.

Future applications of deep sequencing to antigen-receptor repertoires will hopefully address a broad range of critical questions in hematology and immunology. Perhaps most importantly: are specific characteristics or dimensions of the B- or T-cell repertoires significantly correlated with clinical immunity and pathogen resistance? Can one or more of these characteristics serve as a reliable surrogate for immune reconstitution?
after HCT? Can serial analysis of the B- or T-cell repertoire in an individual reliably identify changes that, in turn, are associated with specific clinical events or interventions, such as infections or immunizations? The current pace of research in this field suggests that answers to these questions will soon be at hand.
Acknowledgments

This work was supported by grants from the National Institutes of Health CA015704-37, AI033484, and HL110907.

Authorship

EHW wrote the manuscript. All authors read and revised the manuscript.

Conflicts of Interest

The authors report no conflicts of interest.
References


30. The PyMOL Molecular Graphics System, Version 1.5.0.5 Schrödinger, LLC.
Figure Legends

**Figure 1:** Crystal structures of T- and B-cell receptors in complex with their cognate antigens illustrate the critical role of complementarity-determining regions (CDRs) in antigen recognition. (A) Structure of the 1G4 TCR in complex with the NY-ESO-1157-165 peptide and HLA-A*02:01.28 The CDR1, CDR2, and CDR3 regions of the 1G4 α chain are colored magenta, orange, and red, respectively, to enable easier visualization, and the corresponding regions of the 1G4 β chain are colored blue, cyan, and green. (b) Structure of the broadly neutralizing 10E8 anti-HIV gp41 antibody Fab fragment in complex with a gp41-derived peptide.29 The CDR1, CDR2, and CDR3 regions of the 10E8 heavy chain and the corresponding regions of the 10E8 light chain are colored magenta, orange, red, blue, cyan, and green, respectively. The structures in (A) and (B) were rendered with PyMol.30

**Figure 2:** Generic structure of a locus encoding the variable region of a TCRβ or IGH chain after VDJ rearrangement, illustrating the intervals that encode the 4 framework (FR1-FR4) and 3 complementarity-determining (CDR1-CDR3) regions, the portions contributed by the V, D, and J gene segments, and the “N” nucleotides inserted at the V-D and D-J junctions. The maximum length of FR1-FR3, CDR1, and CDR2, and the modal length of CDR3, in base pairs (bp) and amino acids (aa), are indicated by the numbers at the top of the figure. The left-pointing arrows below the block structure indicate the approximate fraction of the locus that would be sequenced with reads of the indicated length (65, 130, or ≥400 bp) initiated by sequencing primers annealing to the 5’ region of
the J gene segment, typical of current deep sequencing protocols. The scale bar at the bottom of the figure indicates the length of the generic locus in bp.
Figure 1
Figure 2

bp    78   36   51   30   117   45
aa    26   12   17   10   39    15

FR1   FR2   FR3   FR4
CDR1  CDR2  CDR3  

V      NDN    J

0      100    200    300    400 bp

65     130    400
High-throughput sequencing of B- and T-lymphocyte antigen receptors in hematology

Edus H. Warren, Frederick A. Matsen IV and Jeffrey Chou