

REVIEW

Sequence analysis of T-cell repertoires in health and disease

Daniel J Woodsworth^{1,2}, Mauro Castellarin^{1,3} and Robert A Holt^{1,2,3*}

Abstract

T-cell antigen receptor (TCR) variability enables the cellular immune system to discriminate between self and non-self. High-throughput TCR sequencing (TCR-seq) involves the use of next generation sequencing platforms to generate large numbers of short DNA sequences covering key regions of the TCR coding sequence, which enables quantification of T-cell diversity at unprecedented resolution. TCR-seq studies have provided new insights into the healthy human T-cell repertoire, such as revised estimates of repertoire size and the understanding that TCR specificities are shared among individuals more frequently than previously anticipated. In the context of disease, TCR-seq has been instrumental in characterizing the recovery of the immune repertoire after hematopoietic stem cell transplantation, and the method has been used to develop biomarkers and diagnostics for various infectious and neoplastic diseases. However, T-cell repertoire sequencing is still in its infancy. It is expected that maturation of the field will involve the introduction of improved, standardized tools for data handling, deposition and statistical analysis, as well as the emergence of new and equivalently large-scale technologies for T-cell functional analysis and antigen discovery. In this review, we introduce this nascent field and TCR-seq methodology, we discuss recent insights into healthy and diseased TCR repertoires, and we examine the applications and challenges for TCR-seq in the clinic.

The importance of T-cell repertoire analysis

The diversity and composition of the entire set of antigen receptors found within the T cells and B cells of any given individual has an extraordinary impact on health and disease. These repertoires are the product of a complex sequence of genomic events followed by cell- and organ-level selection (Box 1, Figure 1). In the case of T cells, the focus of this review, the T-cell receptor (TCR) repertoire has been found to affect a wide range of disease, including malignancy, autoimmune disorders and infectious diseases, and, given the broad involvement of the immune system in almost all of human health and disease, this reach should be expected to expand greatly. Characterizing TCR repertoires is a priority of great scientific interest and potential clinical utility, but this task is challenged by the enormous scope of TCR combinatorial diversity (Box 1).

As in many areas of investigation, advances in next generation sequencing (NGS) technologies, in which sequences

are decoded on arrays and many millions of sequences can be read simultaneously [1], have been transformative for immune repertoire analysis. Before the advent of NGS, immune repertoires were largely impenetrable, because it was not possible to enumerate a set of distinct T cells, or clonotypes, large enough to be a meaningful representation of the repertoire in its entirety. Nonetheless, foundational and hard-won insights into adaptive immunity, including the mechanisms of VDJ recombination and clonal selection (Box 1) were obtained without NGS. Recent advances and new insights into the properties and behaviors of immune repertoires enabled by high-throughput sequencing and single clonotype resolution have generally extended and clarified early observations and long-standing hypotheses, rather than introduce new paradigms.

A prodigious amount of immune repertoire data has now been obtained using NGS tools and these data have illustrated that our understanding of adaptive immunity is far from complete. However, given the extent to which cellular immunity affects human health, it is our view that T-cell repertoire sequencing will become a critical tool in both biomedical discovery (for example, profiling the TCRs of tumor-infiltrating lymphocytes to develop new diagnostics and prognostics) and clinical management of patients (for

* Correspondence: rholt@bcgsc.ca

¹BC Cancer Agency, Michael Smith Genome Sciences Centre, Vancouver, British Columbia V5Z 1L3, Canada

²Genome Sciences & Technology Program, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada

Full list of author information is available at the end of the article

Box 1 T-cell repertoire biology

In the genome, there are no loci that have greater complexity or extend a deeper and broader reach into human biology than those encoding the antigen receptors of T cells and B cells. The choreographed programs of stochastic recombination that unfold at these loci during T-cell and B-cell maturation provide each of us with the personalized armamentarium necessary for defining and defending our own cellular space. T cells, which mediate cellular immunity, express heterodimeric ($\alpha\beta$ or $\gamma\delta$) cell surface receptors (T-cell receptors, or TCRs). The vast majority of these are $\alpha\beta$ TCRs, which engage heterologous cells presenting peptide antigens bound to major histocompatibility complex (MHC) [22]. These peptide antigens are continuously produced by proteolytic turnover of the contents of a cell such that at any given time the population of MHC-presented peptides represents a diverse sampling of a cell's proteome. T cells develop in the thymus from progenitors originating from hematopoietic stem cells in the bone marrow. During this development new T cells are endowed with the ability, collectively, to recognize essentially any possible peptide, regardless of its origin. Because the diversity in protein sequence is nearly limitless, and our adaptive immune systems cannot know, *a priori*, what specific antigenic challenges lie ahead, so each of us must initiate and maintain a repertoire of T-cell clonotypes bearing receptor variants of enough diversity to recognize essentially any aberrant protein (derived from a pathogen or encoded by a mutated gene) that may be encountered during life. Fundamentally, this capacity is manifested as TCR structural diversity, which is in turn the result of VDJ recombination. This process, which occurs in T cells during their maturation in the thymus, is well characterized [22,58], and we review only a few key points relevant to the TCR β chain here.

Spanning 620 kb on chromosome 7, the TCR- β gene locus contains over 50 variable (V) gene segments, 2 diversity (D) segments, and 13 joining (J) gene segments. Single D and J segments are selected stochastically and recombined in a manner that introduces randomized, non-templated nucleotides at the recombination junction. This process is then repeated with a single randomly selected V segment joined to the DJ segment. The short (about 45 bp) region of the TCR subunit that spans the VD and DJ junctions is known as complementarity determining region 3 (CDR3). This is the region that has the most variability and that directly contacts peptide-MHC (pMHC). The CDR3 region is for most practical purposes unique to each TCR and therefore can be used as a TCR signature or barcode that can be decoded by sequencing.

The key to adaptive immunity is the capacity to discriminate, at the molecular level, self from non-self. T cells must be tolerant of self peptides but reactive toward non-self peptides introduced by infection, or altered self peptides originating from the protein products of mutated genes. This is achieved in three steps, the first of which is the generation of TCR diversity that maximizes the probability that a TCR will exist that matches any foreign or mutated antigen. This is followed by positive selection, in which T cells that do not display sufficient affinity for MHC are eliminated, followed by negative selection, which depletes the population of self-reactive T cells, such that T cells that exit the thymus are an educated, self-tolerant subset of the original population.

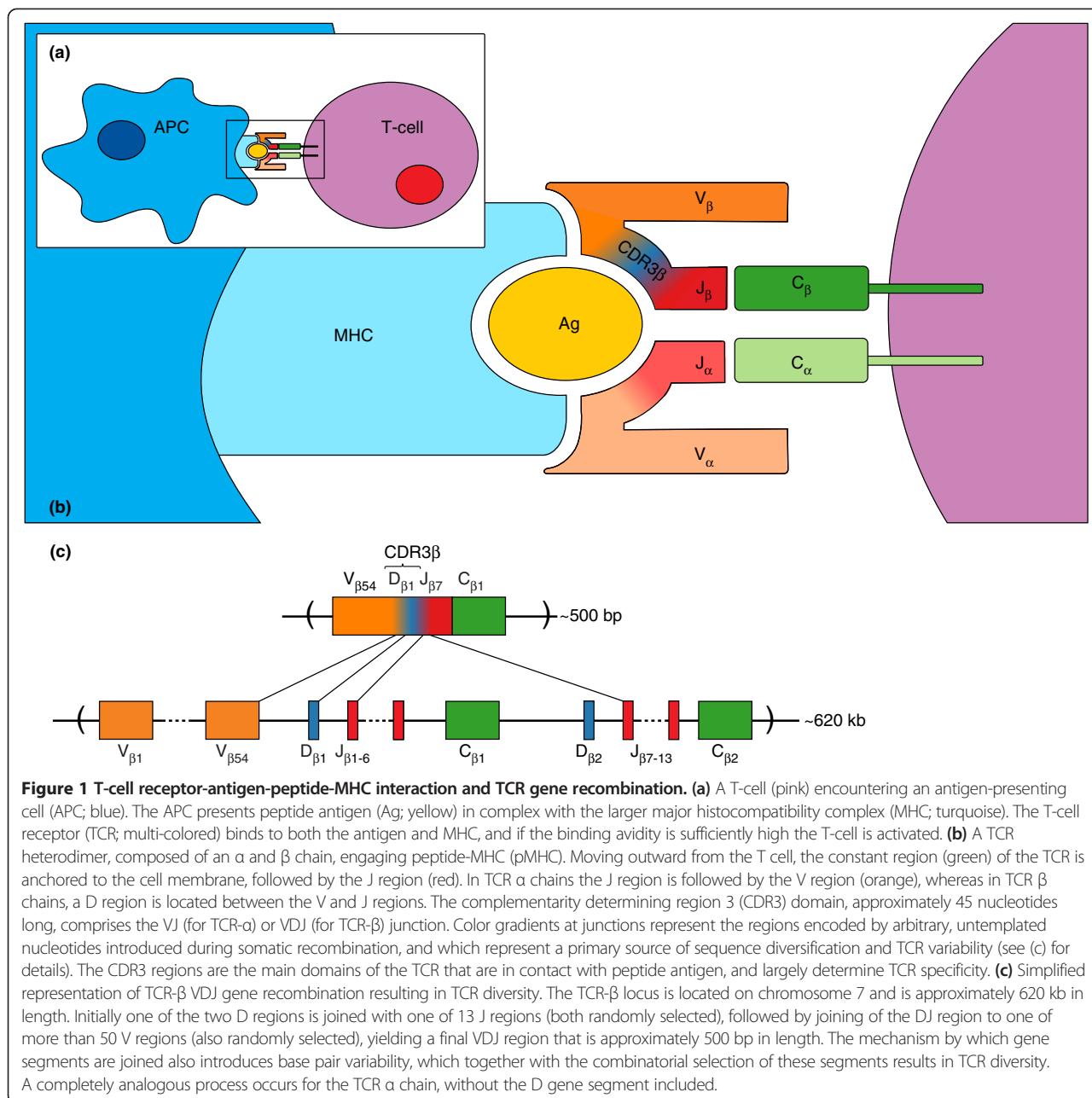
The T-cell repertoire is not static. Binding of a naïve T cell's TCR to a structurally compatible pMHC on an antigen-presenting cell will, with the appropriate interaction of co-stimulatory molecules, initiate rapid clonal expansion to generate a population of effector cells carrying identical TCRs. 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 [54].

example, using the TCR repertoire diversity to follow and manage patients after hematopoietic stem cell transplantation). Here, we first examine the various methodologies that have been used to profile TCR repertoires using NGS technologies. We then discuss some of the insights that have been gained in the context of both healthy and clinical populations, and finally we discuss several challenges and opportunities for the field.

TCR-seq methodology

Reduced to its current, simplest form, TCR-seq involves PCR amplification and sequencing of the CDR3 region from one of the TCR subunits. So far, for several reasons,

most TCR-seq has focused on the TCR β chain. First, the TCR β chain locus contains a D gene component that is missing from the α chain locus, as the α chain locus comprises V and J gene segments only. Thus, the TCR β chain has potential for greater combinatorial and junctional diversity than the α chain. Second, as a result of allelic exclusion mechanisms [2], each $\alpha\beta$ T cell expresses only a single β chain variant, such that the number of distinct β chain sequences observed is an indication of the number clonotypes present in a sample. It is well recognized, however, that using TCR content to specify T-cell clonotypes is imperfect, given that some $\alpha\beta$ T cells will express both of the recombined α chains, each of which can pair



with the expressed β chain, such that the number of distinct TCRs in a repertoire can exceed the number of T cells. Finally, the TCR β chain tends to be the most heavily interrogated because in peripheral blood, which is the most accessible source of T cells, more than 90% of T cells are $\alpha\beta$ T cells rather than $\gamma\delta$ T cells. Importantly, however, $\gamma\delta$ T cells comprise a large fraction of T cells in other tissue compartments, such as skin and the gastrointestinal tract. The functions of $\gamma\delta$ T cells, including their antigen recognition properties, remain poorly understood, and we expect TCR-seq will be instrumental in their future characterization.

The forerunner of the T-cell repertoire deep sequencing methodology is the spectratyping assay [3,4]. Spectratyping involves the use of one or more V and J gene segment-specific primer pairs for PCR amplification of CDR3 from some source of T-cell DNA or RNA, usually peripheral blood. CDR3 amplicons are separated according to size by polyacrylamide gel electrophoresis, which typically yields six or so distinct amplicons per primer pair, spaced at three nucleotide intervals in accordance with reading frame (only sequences resulting from productive rearrangements that encode functional receptors will survive thymic selection; Box 1). The relative intensities of the bands reveal the

proportion of CDR3 amplicons of each length, but the method is blind to the underlying nucleotide variation within each size class. TCR-seq, the deep sequencing of a T-cell repertoire, is not conceptually different from spectratyping. The distinguishing feature is one of scale, whereby all CDR3 amplification products move directly from PCR to library preparation and bulk sequencing. Alignment of the resulting CDR3 sequence reads reveals the abundance of each distinct CDR3 that was present after PCR amplification, which is, in turn, an indication of the number of CDR3 sequences and hence the number distinct T-cell clonotypes present in the original sample. This is, however, an idealized scenario and in reality there are imperfections at each methodological step that must be considered when undertaking immune repertoire analysis.

The source of starting material, T-cell DNA or RNA, is the first major consideration. DNA is often preferable starting material for sequencing applications because of its abundance, ease of isolation and long-term stability. In addition, because for each TCR subunit there are two chromosomal loci per cell, the number of DNA template molecules in the PCR reaction indicates the number of T cells. The main drawbacks of DNA-based analyses are twofold. First, because the TCR loci are single copy the vast majority of DNA in the PCR reaction will be irrelevant. At least for very deep sequencing, the reaction size and number would need to be scaled up in order to obtain sufficient TCR template to capture the diversity of the sample. Second, for DNA analysis, V and J gene-specific primers are combined in a highly multiplexed PCR reaction in order to capture the entire repertoire. Because annealing and amplification efficiency cannot be perfectly matched among primer pairs, read counts can reflect PCR bias in addition to *bona fide* differences in the abundance of TCR templates. This can be a significant concern, particularly when striving for absolute quantification of TCRs. However, it is has been recognized that this mode of PCR bias is highly reproducible and, therefore, at least for comparative studies, its impact can be mitigated by maintaining consistent amplification methods and conditions.

For RNA-based analyses, standard 5'-RACE (rapid amplification of cDNA ends) techniques support comprehensive coverage of TCR templates using a single primer pair. A common 5'-RACE approach is the use of template switching cDNA synthesis methodology [5,6] to incorporate a priming site at the 5' end (preceding the V gene) of a TCR template, and pair this with a C-gene-specific 3' primer for PCR amplification. Incidentally, a common artifact in this system is spurious template switching within the typically GC-rich CDR3 region, giving a truncated PCR product that must be removed before library construction and sequencing. Technically, RNA analysis is more amenable than DNA analysis to broader repertoire coverage from a given amount of starting template, given the larger proportion of

actual input TCR template molecules. Further, the use of a single primer pair for amplification avoids the bias that can be incurred when using multiple sets of primers concurrently, as is done for most DNA-based TCR analysis, described above. The obvious drawback of RNA analysis is that variation in TCR expression levels among T cells means that the copy number of TCR template molecules is not strictly proportional to cell count.

Currently, most large-scale TCR-seq uses Illumina sequencers, which can generate extensive sequence data at low cost, although other platforms, in particular 454, have also been used successfully for smaller-scale analyses. In the earliest days of TCR-seq the key limitation of the Illumina platform was the short length of sequences (36 to 50 bp) obtainable and the precipitous drop-off of sequence quality toward the ends of reads. Because the rearranged TCR β chain is approximately 500 bases in length, and the informative CDR3 region is positioned nearest the 3' end (Figure 1), early studies that used very short reads relied on either the limited number of adequate quality reverse sequence reads that primed in the J gene region [7], or they used more elaborate and inconvenient amplicon fragmentation and shotgun assembly methods [8,9]. Currently, the Illumina HiSeq platform yields a vast quantity (billions) of 150 bp reads per run, which is adequate for deep coverage of CDR3 by reverse reads primed in the J or C gene region. The Illumina MiSeq, which currently generates tens of millions of paired 250 base reads, is also becoming a useful tool for TCR-seq.

A particular vulnerability of TCR-seq that became readily apparent in early studies was the unique sensitivity of the method to sequencing errors inherent in NGS. Strictly speaking, due to the stochasticity of VDJ recombination, a TCR sequence that differs from others by even a single nucleotide could represent a legitimate, low-frequency clonotype. While the per base error rate of Illumina sequencing is very low (substantially less than 1%), the TCR amplicon interrogated by TCR-seq is short, such that high-copy clonotypes may show extreme coverage. The chance occurrence of errors at a single nucleotide position becomes non-negligible and gives the appearance of a population of novel but ultimately erroneous low-frequency clonotypes. This issue is now well recognized and error mitigation approaches have been described. Stringent removal of low quality sequence is the first step, but for deep TCR sequencing this is inadequate. Because artifactual TCR sequences appear as low-abundance clonotypes, it is possible to simply remove from analysis all TCR sequences that fall below an abundance threshold, typically up to a few percent [9,10]. Alternatively, each low-frequency clonotype can be clustered together with whatever more highly abundant clonotype it closely resembles, under the assumption that the more highly abundant version is the correct sequence. Optimal error correction can be achieved with

algorithms that integrate these types of error recognition and correction modalities [11], and in all cases it is highly beneficial to define error accumulation and error filtering efficacy in each experiment by assessing the error profile of either relatively invariant J or C gene sequences flanking CDR3, or of a known TCR sequence spiked into the original sample [9,12]. Interestingly, current rates of sequencing error accumulation fundamentally preclude the complete sequencing of an immune repertoire, because upon very deep sequencing it becomes impossible to distinguish with certainty the rare *bona fide* clonotypes from sequence errors.

Exhaustive sequencing is rarely a goal for T-cell repertoire analyses, and sequencing of extreme depth is unnecessary for comparative studies aiming to elucidate differences among samples. For these studies, although appropriate sequence error handling must be implemented, 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. The problem is the same regardless of whether samples are from different individuals or obtained from a single individual as either a time-course study or before and after a particular intervention or from different tissue compartments. This issue is probably best addressed by establishing the confidence with which one can assert that the representation of clonotypes in one sample versus another is different from what is expected by chance. Fundamentally, if a given TCR is not observed in a dataset, that may be because it was not present in the sample or because sequencing depth was inadequate to reveal it. Abundant clonotypes are easily revealed at low sequence depth, and the probability of observing rarer clonotypes increases with the number of reads gathered. Thus, to compare two or more samples, it is essential to normalize the input data, for example, by unbiased removal of reads from larger datasets until they are of the size of the smallest comparator. After appropriate normalization, it has been demonstrated that methods widely used in ecology have potential utility in comparing immune repertoires. Examples include the Simpson diversity index for comparing diversity between samples [13], or the Morisita-Horn similarity index for determining the similarity, or overlap, between samples [14]. The Gini coefficient, which in economics is used to describe the distribution of a commodity among individuals, has been shown to be another useful means by which to compare the content of T-cell repertoires [15]. However, the comparison of large TCR-seq datasets is computationally intensive, and scaling up presents a challenge for conventional statistical tools. The field awaits a robust, standardized and easily accessible set of computational tools for TCR-seq data processing and analysis along the lines of the common whole genome and transcriptome alignment programs that have proliferated

since the advent of NGS. The V-quest package [16,17] created and hosted by ImMunoGeneTics is a useful tool for medium-scale TCR-seq data handling and annotation. The new tool for TCR-seq data processing called MiTCR [18] is a recent and welcome addition, as is CIG-DB [19], a new repository for TCR and immunoglobulin sequences observed in cancer studies. The field will benefit from further development of these types of community resource. In particular, a centralized and comprehensive repository of TCR sequences compiled from public domain TCR-seq studies would be a valuable community resource that would facilitate interpretation of new data (for example, has this interesting clonotype been seen before?) and support early exploration of the species-wide TCR meta-repertoire.

So far, large-scale T-cell repertoire analysis has been limited to interrogation of a single TCR subunit per sequencing run. Most $\alpha\beta$ TCR profiling studies have targeted the TCR β chain, for reasons described above. Functional antigen-engaging TCRs are, however, undeniably heterodimeric proteins comprising both an α and a β chain and, therefore, for any meaningful functional analysis of TCRs, both subunits must be defined. TCR-seq studies begin with cell lysis, which instantly eradicates α and β chain pairing specificity. Recently, however, pairwise Illumina sequencing of $\alpha\beta$ TCRs has been demonstrated. This method relies on overlap-extension RT-PCR-mediated fusion of α and β chain mRNA, executed in a parallelized manner within droplets of water in oil emulsion each containing a single cell [20]. Using this approach, hundreds of $\alpha\beta$ TCR sequences could be identified from starting populations of greater than 1 million peripheral blood mononuclear cells (PBMCs). Likewise, it was recently demonstrated that thousands of immunoglobulin heavy and light chain pairs can be obtained by bead capture of single B-cell mRNA followed by linkage PCR in single-bead-containing emulsion droplets [21]. Although the yield of each of these approaches is modest, and the techniques themselves are technically challenging, these studies represent very important advances towards the goal of deep, cheap and fast profiling of dimeric antigen receptors.

Still missing from most TCR repertoire analyses are demonstrations, by independent methods, of the validity of specific TCR clonotypes of interest. Verification of the identity and abundance of notable clonotypes, using methods such as targeted V-J subset resequencing or clonotype-specific quantitative PCR, is important for guarding the integrity of this rapidly developing field.

New insights from T-cell repertoire analysis by TCR-seq

Although much of the work in this nascent field has focused on methodological development as described above, these techniques have already been applied to yield

significant insights into the TCR repertoire in both health and disease.

Properties of a healthy repertoire

Within an individual repertoire, there is a profound disparity in the frequencies of distinct clonotypes, which can vary in abundance by many orders of magnitude [9]. In terms of distinct TCR sequences, the number of theoretically possible TCRs is almost incomprehensibly large. For example, the theoretical maximum number of unique, approximately 45 bp TCR β CDR3 nucleotide sequences is 4^{45} . Models that place informed constraints on theoretical diversity still project more than 10^{11} potential β chains [22,23]. The extent of the diversity, or size, of actual biological repertoires falls dramatically short of these theoretical estimates because of bias in generation frequency and the depletion by thymic selection of T cells bearing non-productively rearranged TCRs (Box 1). In a seminal pre-NGS study, Arstila and co-workers [24] estimated actual β chain diversity in an individual to be approximately 10^6 by extrapolating from the total number of sequences found within the scarce V β 18 and J β 1.4 subset. They further predicted 2.5×10^7 total $\alpha\beta$ TCRs by determining the number of V β sequences in V α 12 $^+$ sorted T cells. A revised estimate of β chain repertoire size derived from deep TCR-seq data using the unseen species model [25] was 3 to 4 million [7], and the deepest TCR-seq experiment so far recovered 1.3 million [9] distinct TCR β chain sequences from a single individual, which places a directly measured lower limit on diversity. Unfortunately, as previously discussed, a direct measure of total diversity remains out of reach owing to the practical and ethical limitations on obtaining very large numbers of T cells from research subjects and the difficulty in distinguishing rare clonotypes from sequencing errors.

Beyond repertoire size considerations, numerous recent TCR-seq studies have provided intriguing and valuable insights into the characteristics of T-cell repertoires. The notion that the diversity of the naïve 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 [12]. Further, detection of identical TCRs within numerous carefully sorted T-cell subsets, including naïve, memory, cytotoxic, T-helper and T-regulatory subsets, suggests that T-cell specificity determination precedes the differentiation of nascent T cells into distinct phenotypic subsets, helping to resolve this long-standing question of chronology in T-cell maturation [26].

Public T cells, which are identical T-cell clonotypes shared among individuals, have been a curiosity for some time given the incredibly low likelihood of identical TCRs being generated in separate individuals by chance. TCR-seq studies have revealed that public T cells are actually commonplace [9,23,27,28] and result from the increased

generation probability of these shared TCR specificities across individuals [29], as well as the fact that different TCR nucleotide sequences can code for the same TCR amino acid sequence, because of the degeneracy of the genetic code. The proportion of an individual's TCR repertoire that is public has been shown to be as high as 14%, and the true extent of the public repertoires could be much higher still [9]. Interestingly, based on TCR-seq data, sharing of major histocompatibility complex (MHC) class I alleles does not seem to strongly influence TCR sharing. The evaluation of antigen-specific anti-viral CD8 $^+$ repertoires in healthy populations by conventional sequencing has confirmed that MHC class I is not a driver of TCR diversity; rather, antigen-specific repertoires are shaped in a peptide-dependent manner [30].

Thus, in summary, a picture of a typical immune repertoire is emerging whereby a healthy individual may be expected to harbor several million readily measurable TCR β chain clonotypes that vary widely in abundance, some subset of which originate as recombinants with high generation probability and which can be shared across different T-cell compartments, and indeed among individuals.

T-cell repertoires affected by disease

Following on from studies in healthy individuals, the TCR profile of individuals affected by disease has been investigated using various TCR-seq methodologies (summarized in Table 1). These studies have yielded novel insight into the disease biology, as well as demonstrated the potential for TCR-seq applications in the clinic (both of which are summarized in Table 2). So far, much of the work done to profile T-cell repertoires in disease populations has focused on post-hematopoietic stem cell transplant (HSCT) patients. This therapy is used in conditions in which the patient's immune system has become dysfunctional, and involves pre-transplant ablation of the patient's own lymphocytes and bone marrow, followed by transplantation of HSCs, which regenerate the patient's immune system [31,32]. The therapeutic benefit derives from the elimination of the dysfunctional immune cells from the patient, as well as, in the setting of allogeneic transplants in malignancy, a 'graft versus tumor effect' in which the incoming donor cells are reactive against the patient's tumor, resulting in a therapeutically beneficial regression of the tumor [31,32]. The process of immune reconstitution takes months, and during this time the low abundance and diversity of lymphocytes place patients at serious risk of developing infection and cancer [31,32].

Numerous T-cell repertoire considerations are relevant to this procedure, including: (i) the source of the HSCs - bone marrow, peripheral blood or umbilical cord, autologous or allogeneic; (ii) the risk for, incidence of, and potential methods for mitigation and treatment of 'graft versus host disease' (GVHD) post-transplant (wherein donor T cells are

Table 1 Summary of methodology used by clinical TCR profiling studies

Disease	Patients	Controls/comparison	Sample	Platform	Sequencing depth	Diversity metric	References
Hematological malignancy	Post-HSCT (n = 28)	Healthy donors	RNA from PBMCs	454 (Roche)	Variable	Inverse Simpson's diversity index (1/D _g)	[34]
Post-HSCT GI GVHD	Steroid refractory (n = 7)	Steroid responsive (n = 8)	DNA from PBMCs and GI biopsy	GA IIx or MiSeq (Illumina)	10 ³ -10 ⁶	Bhattacharyya coefficient and Gini coefficient	[36]
Post-HSCT pediatric neuroblastoma	Early T-cell infusion (n = 32)	Late T-cell infusion (n = 12)	DNA from PBMCs	ImmunoSeq (Illumina Platform)	Not reported	Number of clones and V _B gene usage	[75]
Ankylosing spondylitis	Post-HSCT (n = 1)	Pre-HSCT (n = 1; same patient)	RNA from PBMCs	454 GS FLX (Roche)	Not reported	Number of clones and V _B gene usage	[38]
Rheumatoid arthritis	Recent onset (n = 6)	Established (n = 6)	RNA from synovium and PBMCs	454 GS FLX Titanium (Roche)	>6 × 10 ⁴ reads per sample	Frequency distribution	[40]
CMV and EBV infection	Kidney transplant (n = 2)	EBV ⁺ /CMV ⁺ donors (n = 5)	RNA from PBMCs	454 GS FLX Titanium (Roche)	>10 ⁴ reads per sample	Frequency distribution	[76]
Recurrent HSV-2 Infection	Comparison of samples from various infection sites and times, and PBMCs in n = 1 individual		DNA from skin biopsy and PBMCs	HiSeq (Illumina)	Not reported	Frequency distribution	[77]
TALL	Post-treatment (n = 43)	Pre-treatment (n = 43) and healthy donors (n = 6)	DNA from bone marrow and PBMCs	HiSeq (Illumina)	Not reported	Frequency distribution	[41]

Abbreviations: CMV, cytomegalovirus; EBV, Epstein-Barr virus; GI, gastrointestinal; GVHD, graft versus host disease; HSCT, hematopoietic stem cell transplant; HSV, Herpes simplex virus; PBMC, peripheral blood mononuclear cells; TALL, T-cell acute lymphoblastic leukemia.

Table 2 Examples of insights derived from TCR-seq in clinical populations

Disease	Insights	References
Hematological malignancy	TCR repertoire diversity significantly higher in patients who received HSCT using DUCB as stem cell sources, as compared with conventional and TCD-derived stem cells	[34]
Post-HSCT GI GVHD	Highly expanded indicator clones identified in GI biopsy at time of diagnosis significantly expand over time as measured in PBMC samples. Furthermore, the degree of expansion is much greater in steroid refractory patients, raising the potential that this could be used to stratify patients for treatment protocols	[36]
Post-HSCT pediatric neuroblastoma	Early infusion of expanded T-cell product in post-HSCT pediatric neuroblastoma patients results in significantly improved TCR repertoire diversity recovery as opposed to late administration	[75]
Ankylosing spondylitis	Many highly expanded autologous clones survive HSCT pre-conditioning regimen and HSCT therapy itself in a patient with ankylosing spondylitis. This suggests that the therapeutic effects of HSCT in this disease are due to an immune system 'reset' resulting from attrition of the T-cell compartment, rather than complete ablation	[38]
Rheumatoid arthritis	1. TCR repertoire in synovium of patients with newly diagnosed RA is dominated by small number of highly expanded T-cell clones, much more so than in patients with established RA 2. Significant overlap in TCR profile within affected joints in the same patient, with most expanded clone common to most joints 3. No overlap in TCR profile between synovium and PBMC within patient or between patients	[40]
TALL	1. TCR-seq can identify minimal residual disease in TALL at higher sensitivities than flow cytometry, the current gold standard 2. It is possible that the CDR3 sequence may be adapted as a biomarker for risk stratification of minimal residual disease.	[41]

Abbreviations: CDR3 complementarity determining region 3, DUCB dual cord umbilical blood, GI gastrointestinal, GVHD graft versus host disease, HSCT hematopoietic stem cell transplant, PBMC peripheral blood mononuclear cell, RA rheumatoid arthritis, TALL T-cell acute lymphoblastic leukemia, TCD T-cell depleted, TCR T-cell receptor.

reactive against host antigens); (iii) the prevention and treatment of post-transplant infection and malignancy resulting from lymphopenia and immunosuppression; and (iv) the timescale and nature of the post-transplant reconstitution of the TCR repertoire. There has been much historical interest in these factors, and although efforts have been made to profile patient T-cell repertoires after HSCT using spectratyping, they have always yielded an incomplete view [33]. Now, using TCR-seq, investigators can examine more comprehensively the repertoires of HSCT patients, yielding much more incisive and reliable insights into post-HSCT repertoire reconstitution.

Recently, Pamer and colleagues [34] investigated how both the source of the HSCs as well as the transplant protocol affect the post-transplant TCR repertoire and subsequent clinical outcomes. Historically, bone marrow harvested from MHC-matched donors was the main source for HSCs, but this has been supplanted by HSCs derived from peripheral blood or umbilical cord blood. In the case of unmatched allogeneic transplants, the risk of GVHD can be mitigated by depleting the donor HSCs of T cells before transplant [34,35]. Pamer and colleagues [34] investigated the TCR profile in allogeneic HSCT patients who received either conventional peripheral blood-derived HSCs, T-cell-depleted HSCs or double unit umbilical cord HSCs. Their findings included confirmation of previous reports that the TCR diversity is markedly restricted after HSCT, as well as measuring a 50-fold greater diversity in CD4⁺ over CD8⁺ T cells, across all HSC sources [34]. Most notably, this study clearly demonstrated significant deficiencies in post-

HSCT TCR diversity in T-cell-depleted patients as compared with patients receiving HSCs derived from other sources. Moreover, GVHD, steroid treatment and viral infection all had a negative impact on TCR diversity. Finally, the TCR profiles of several patients who had especially low diversity were reassessed 2 years after transplant. Significant reconstitution was found in one patient, while no improvement was observed in the others [34]. Together, these results represent the first steps towards using TCR profiling to select HSCT protocols, and to stratify risk and guide treatment in post-allogeneic-HSCT patients.

GVHD is a frequent significant complication in HSCT patients, and gastrointestinal GVHD accounts for most mortality [31]. Recent work by Negrin and colleagues [36] focused on comparing the TCR profiles of post-HSCT GVHD patients who either responded to first-line steroid therapy for their gastrointestinal GVHD or did not (steroid refractory patients). Importantly, no TCR sequences conserved across patients were found in notable abundance. However, when TCR sequences of T cells isolated from biopsies of various gastrointestinal sites were compared, there was measurable repertoire similarity across biopsy sites within patients with steroid refractory disease, much more than in patients who responded to steroid treatment. Moreover, when patient-specific high-abundance 'indicator clones' that were initially identified in gastrointestinal biopsies were tracked over time in peripheral blood samples using their TCR sequences, the frequency of these indicator clones was observed to expand in the steroid refractory patients and contract in the responsive patients. It is

possible that these results could be refined to develop a biomarker for identifying patients at risk for steroid refractory GVHD, as well as for guiding and following treatment.

Although hematological malignancy is the main indication for HSCT, the procedure is used in an ever-expanding array of diseases, including other cancers, autoimmune disease and immunodeficiency disorders [31,32]. In one of the more radical departures from conventional paradigms, HSCT is being used as treatment for advanced autoimmune disorders such as ankylosing spondylitis, an inflammatory arthritis that can result in fusion of the spine [37]. The exact mechanism of HSCT efficacy in this disorder is unknown, and it was investigated using TCR profiling [38]. This approach yielded the important observation that a significant fraction of host T-cell clones survived pre-HSCT ablation, suggesting that the efficacy of HSCT in the context of this disease is due to a 'reset' of the immune system, rather than complete replacement. Moreover, the pre-HSCT conditioning seems to act through attrition as higher-frequency clones were more likely to survive conditioning and expand further thereafter. Together, these results demonstrate that, in the context of autoimmunity, a reduced dose conditioning regimen might still achieve sufficient attenuation of lymphocytes to allow an HSCT-mediated 'reset', while also lowering conditioning-associated toxicity.

In the field of autoimmune diseases in general, there is an ongoing intensive search for antigens that initiate T-cell cross-reactivity [39]. This effort is confounded by the fact that such an antigen might be present on a transient foreign pathogen that is cleared, or in a tissue compartment separate from the one that is diseased. Identifying autoreactive T cells by TCR-seq can provide an alternative, albeit indirect, path to identifying auto-antigens, although one that is technically challenging (Box 2). In the case of rheumatoid arthritis (RA), de Vries and colleagues [40] examined the TCR profile of the synovium from affected joints of patients with either recent onset or established RA, and found that the repertoire in the synovium of recent onset patients was dominated by a small number of highly expanded clones, much more so than in established RA. Moreover, there was a large overlap in the TCR profiles of affected joints in the same patient, but minimal overlap between affected joints and peripheral blood, or between patients. Further investigation of the overlap between affected joints showed that the most expanded clone in each joint was in fact the same. These results indicate that although a patient-specific approach may be required, in RA it is indeed reasonable to focus the search for auto-antigens on the synovium.

Finally, a recent paper by Robins and colleagues [41] suggests that TCR-seq may become another basic tool, analogous to conventional flow cytometry, both for the study of disease and for clinical diagnostics. In this study, the authors [41] used TCR sequences as markers to track minimal residual disease in T-cell acute

lymphocytic leukemia (TALL). Before treatment, malignant T cells were first identified by the overabundance of their respective TCRs. Following treatment, patient TCR repertoires were probed periodically for evidence of any malignant clones. The authors found that in several cases, malignant T cells that had survived chemotherapy were detected at lower levels than when using conventional flow cytometry, which might enable earlier intervention and perhaps better relapse management. Moreover, the authors [41] found patterns in the TCR sequence of certain transformed T cells that correlated with the more malignant early T-cell precursor TALL subtype. Importantly, this subtype necessitates a more aggressive treatment. This work highlights the potential of using TCR profiling as both a diagnostic tool and for tracking disease and guiding treatment.

In our view, two limitations were found across several of the clinically oriented TCR-seq studies that have been reviewed here. First, the frequency distribution of TCR clones is often used to quantify repertoire diversity. Although informative, this requires investigators to make interpretations and decisions regarding where to set thresholds during data analysis, and given that there are no standardized thresholds, comparison among studies is difficult. Second, many of these studies show disparities in TCR repertoire diversity among different groups of patients. However, so far, there has been minimal correlation between these molecular findings and clinical variables of interest, such as survivorship, incidence of infection or cancer relapse, or disease progression. We encourage investigators to consider these aspects in the design of future studies, as this could be expected to dramatically increase the clinical utility of TCR repertoire profiling. Nevertheless, we feel that the impact of TCR-seq on medicine will be significant, and eventually broad. If paired with a high-throughput method for identifying TCR cognate antigen, TCR-seq could potentially revolutionize antigen discovery (Box 2). Once combined with longer-term and more clinically relevant data, we feel that TCR-seq will become a useful biomarker for diagnostic, prognostic and treatment stratification applications in a wide range of diseases. As the technique matures and achieves widespread recognition, validation and accessibility, we believe that antigen receptor profiling will have an impact on many areas of medicine.

Conclusions and future directions

High-throughput sequencing applied to profiling TCR repertoires provides a new high-resolution view of cellular immunology, and has yielded new insights into the properties of normal T-cell repertoires in ordinary, healthy individuals, plus a view of T-cell repertoires affected by disease and/or modified by transplantation.

Box 2 What do all these T cells recognize?

The use of new sequencing technologies has enabled TCR repertoire analysis at unprecedented resolution. T-cell antigen discovery methodologies have not kept pace, and there is now an extreme mismatch between our ability to recognize unique antigen-experienced T cells by TCR-seq and our ability to identify their cognate antigens. There are numerous biological and technical issues that hamper T-cell antigen discovery. First, the immunological synapse, where pMHC and TCR interact, is immensely complex. It is the site of convergence of all TCR repertoire diversity, MHC allelic heterogeneity and peptide variation encoded in the human and human microbiome proteomes. Further, TCR-pMHC engagement tends to be low affinity [59] compared with other biomolecular interactions such as antibody binding, and there is extensive cross-reactivity, whereby a given TCR can recognize many pMHC targets, and a given pMHC can be recognized by numerous TCRs [54,60]. The net result is that a molecular space that is almost unfathomably large must be screened for low-affinity hits that, if observed, may or may not represent natural epitopes.

Various strategies for T-cell epitope discovery have been developed. Naturally presented peptides can be acid-eluted from MHC and identified by mass spectrometry [61-63], or MHC binding peptides can be predicted computationally based on the presence, within a peptide, of favorable MHC-interacting residues [64-66]. Unfortunately, neither of these approaches are immediately informative regarding the identity of the interacting TCRs. In situations in which a T-cell clone has been identified that bears a TCR of interest, conventional T-cell antigen discovery efforts have tended to rely on evaluating candidate antigens by, for example, ELISPOT screening of antigen-presenting cells loaded with custom-synthesized minimal peptides or transiently transfected with RNA or cDNA encoding these peptides, or by tetramer sorting [67]. Tetramers are multivalent pMHC molecules assembled *in vitro* [67]. The scale of these experiments is, however, clearly inadequate for unbiased, *de novo* antigen discovery. Here, systematic screening of combinatorial encoded tetramers [60] and multiplexed peptide pools [61,62] has shown promise, as has the screening of random peptide encoding DNA libraries with reporter T cells expressing recombinant TCRs [68,69]. TCR tetramers [70] and pMHC display methodologies [71-74] have not gained widespread use for antigen discovery, and they tend to be hampered by a combination of low solubility and binding affinity issues, as well as the difficult logistics of large-scale screens. New, innovative methods for rapid and comprehensive T-cell antigen screening are urgently needed, and this need will only escalate as the TCR repertoires continue to be unveiled.

From a scientific standpoint, a critical avenue for future research is to address the pressing need for methodological advances that will enable routine profiling of dimeric TCRs and T-cell antigen discovery (Box 2).

Clinically, the main developments that have arisen from TCR-seq applied to medical problems have been blood-based biomarkers for diagnosis, prognosis, treatment and risk stratification in various autoimmune disorders, post-HSCT transplant patients and those with hematological malignancy. Importantly, given the personalized nature of immune repertoires, clinical applications of TCR-seq will require patient-specific approaches, and therefore cost reduction will increasingly become a driving factor. Another challenge is building data acquisition and analysis pipelines that are fast enough to provide actionable information in clinically relevant timeframes (often on the order of days), and sufficiently accurate for regulatory approval and for acceptance by clinicians and patients. These issues are already being encountered and addressed in the nascent field of personalized onco-genomics [42,43]. Increasingly, high-quality clinical and research-oriented TCR-seq services are being offered by biotechnology companies [44-47], and this sector is poised for growth.

There remain numerous as yet untapped applications for TCR-seq and follow-on technologies. In oncology, the presence of tumor-infiltrating lymphocytes in solid tumors is a

well recognized correlate of favorable outcome [48], and yet we and others are just beginning to explore the clonal diversity and antigen specificities of these T cells [49,50]. Likewise, using TCR-seq to characterize host-pathogen interactions and immunodeficiency disorders has great potential. Given that CD8⁺ T cells are crucial in suppressing human immunodeficiency virus (HIV)-infected cells (predominantly CD4⁺ T cells), it seems that defining T-cell dynamics at the level of individual clonotypes will have utility in studies of HIV control. This is especially the case given that this field is in considerable flux [51]. Early studies indicated that TCR diversity correlated with viral control, but more recently these findings have been countered with reports that public TCR sequences are more important and are a superior predictor of patient response to HIV infection. So far, these studies have been hampered by low sample sizes, heterogeneous patient populations and disease stages, as well as incomplete snapshots of the TCR repertoire due to the relatively low throughput methods employed - issues that may be ameliorated with TCR-seq.

The use of TCR-seq to study vaccine response, in HIV or for that matter any other vaccine intended to elicit a cellular (as opposed to humoral) immune response, remains an alluring but as yet unfulfilled prospect that is complicated by the polyspecific nature of T-cell activation. Beyond pathogen and vaccine responses, TCR-seq will also be instrumental in

unraveling the complex interplay between cellular immunity and the broader human microbiome, as exemplified by a recent TCR-seq study showing that central regulatory T cells (rather than peripherally induced regulatory T cells, as previously thought) constitute most of the regulatory T cells in the gut and mediate tolerance to antigens produced by the commensal gut microbiota [52]. T-cell repertoire analysis is also well positioned to help refine our understanding of the dramatic changes in cellular immunity that transpire through the course of normal aging [53], which result from inescapable age-dependent involution of the thymus, as well as memory pool expansion driven by reactivation by chronic persistent viral infections [54].

Regarding the future of TCR-seq as applied to problems in the clinic, we feel that the main limitation so far has been that most studies have used TCR sequences as markers, with little or no emphasis on functional data or clinical correlates and outcomes. Our view is that there is a need to push past the current focus on providing clinically useful information and towards new treatments. Although a vast amount of discovery oriented work remains, it is clear that TCR repertoires are negatively affected by disease and that low TCR diversity is often associated with poor clinical outcomes, and it is time to start addressing these issues at the point of care. For example, initial efforts in this regard could focus on improving TCR diversity in post-HSCT patients, perhaps by developing more sophisticated pre-conditioning regimens or stimulating T-cell regeneration after HSCT [55]. Further into the future, given recent advances in TCR and chimeric antigen receptor engineering [56,57], it is possible that suites of TCRs will be available, such that a patient-specific suite may be selected off the shelf and administered to patients via transduced autologous cells to fill holes in their TCR repertoires. Finally, in the case of autoimmunity, the identification of expanded autoreactive TCRs may allow administration of TCR-specific inhibitors in both acute and chronic autoimmune disease, or eventually a new paradigm of targeted T-cell depletion for these disorders.

Abbreviations

CAR: Chimeric antigen receptor; CDR3: Complementarity determining region 3; GVHD: Graft versus host disease; HIV: Human immunodeficiency virus; HSC: Hematopoietic stem cell; HSCT: Hematopoietic stem cell transplant; MHC: Major histocompatibility complex; NGS: Next generation sequencing; pMHC: Peptide major histocompatibility complex; RA: Rheumatoid arthritis; RACE: Rapid amplification of cDNA ends; RT-PCR: Reverse transcription polymerase chain reaction; TALL: T-cell acute lymphocytic leukemia; TCR: T-cell receptor.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

DJW is supported by a Vancouver Coastal Health-CIHR-UBC MD/PhD Studentship Award and a Canadian Institutes of Health Research (CIHR) Frederick Banting and Charles Best Canada Graduate Scholarship Doctoral Research Award. Research by the Holt laboratory is supported by the Canadian Institutes of Health Research, Genome British Columbia & Genome Canada, the Canadian Breast Cancer foundation, the British Columbia Cancer Agency and the Department of Defense, USA.

Author details

¹BC Cancer Agency, Michael Smith Genome Sciences Centre, Vancouver, British Columbia V5Z 1L3, Canada. ²Genome Sciences & Technology Program, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada. ³Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada.

Published: 30 October 2013

References

- Metzker ML: Sequencing technologies - the next generation. *Nat Rev Genet* 2010, **11**:31–46.
- Khor B, Sleckman BP: Allelic exclusion at the TCR β locus. *Curr Opin Immunol* 2002, **14**:230–234.
- Pannetier C, Cochet M, Darche S, Casrouge A, Zöller 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 U S A* 1993, **90**:4319–4323.
- Gorski J, Yassai M, Zhu X, Kissela B, Kissella B, [corrected to Kissela B], Keever C, Flomenberg N: Circulating T cell repertoire complexity in normal individuals and bone marrow recipients analyzed by CDR3 size spectratyping. Correlation with immune status. *J Immunol* 1994, **152**:5109–5119.
- Peters DG, Kassam AB, Yonas H, O'Hare EH, Ferrell RE, Brufsky AM: Comprehensive transcript analysis in small quantities of mRNA by SAGE-lite. *Nucleic Acids Res* 1999, **27**:e39.
- 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.
- Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS: Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 2009, **114**:4099–4107.
- Warren RL, Nelson BH, Holt RA: Profiling model T-cell metagenomes with short reads. *Bioinformatics* 2009, **25**:458–464.
- Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, Webb JR, Holt RA: 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.
- 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**:1–13.
- Bolotin DA, Mamedov IZ, Britanova OV, Zvyagin IV, Shagin D, Ustyugova SV, Turchaninova MA, Lukyanov S, Lebedev YB, Chudakov DM: Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur J Immunol* 2012, **42**:3073–3083.
- Klarenbeek PL, Tak PP, van Schaik BDC, Zwintzman AH, Jakobs ME, Zhang Z, van Kampen AHC, van Lier RAW, Baas F, de Vries N: Human T-cell memory consists mainly of unexpanded clones. *Immunol Lett* 2010, **133**:42–48.
- 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.
- 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.
- Thomas PG, Handel A, Doherty PC, La Gruta NL: Ecological analysis of antigen-specific CTL repertoires defines the relationship between naïve and immune T-cell populations. *Proc Natl Acad Sci U S A* 2013, **110**:1839–1844.
- Brochet X, Lefranc M-P, Giudicelli V: IMGT/V-QUEST: the highly customized and integrated system for Ig and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* 2008, **36**:W503–W508.
- Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc M-P: IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Res* 2012, **8**:1–15.
- Bolotin DA, Shugay M, Mamedov IZ, Putintseva EV, Turchaninova MA, Zvyagin IV, Britanova OV, Chudakov DM: MiTCR: software for T-cell receptor sequencing data analysis. *Nat Methods* 2013, **10**:813–814.
- 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.

20. Turchaninova MA, Britanova OV, Bolotin DA, Shugay M, Putintseva EV, Staroverov DB, Sharonov G, Scherbo D, Zvyagin IV, Mamedov IZ, Linnemann C, Schumacher TN, Chudakov DM: **Pairing of T-cell receptor chains via emulsion PCR.** *Eur J Immunol* 2013, **43**:1–9.
21. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, Varadarajan N, Giesecke C, Dörner T, Andrews SF, Wilson PC, Hunnicke-Smith SP, Wilson CG, Ellington AD, Georgiou G: **High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire.** *Nat Biotechnol* 2013, **31**:166–169.
22. Davis MM, Bjorkman PJ: **T-cell antigen receptor genes and T-cell recognition.** *Nature* 1988, **334**:395–402.
23. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, Carlson CS, Warren EH: **Overlap and effective size of the human CD8+ T cell receptor repertoire.** *Sci Transl Med* 2010, **2**:47ra64.
24. 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.
25. Efron B, Thisted R: **Estimating the number of unseen species: How many words did Shakespeare know?** *Biometrika* 1976, **63**:435–447.
26. Wang C, Sanders CM, Yang Q, Schroeder HW, Wang E, Babrzadeh F, Gharizadeh B, Myers RM, Hudson JR, Davis RW, Han J: **High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets.** *Proc Natl Acad Sci U S A* 2010, **107**:1518–1523.
27. Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T, Asher TE, Almeida JR, Levy S, Price DA, Davenport MP, Douek DC: **A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing.** *J Immunol* 2011, **186**:4285–4294.
28. Venturi V, Kedzierska K, Price DA, Doherty PC, Douek DC, Turner SJ, Davenport MP: **Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination.** *Proc Natl Acad Sci U S A* 2006, **103**:18691–18696.
29. Murugan A, Mora T, Walczak AM, Callan CG: **Statistical inference of the generation probability of T-cell receptors from sequence repertoires.** *Proc Natl Acad Sci U S A* 2012, **109**:16161–16166.
30. Koning D, Costa AI, Hoof I, Miles JJ, Nalnlohy NM, Ladell K, Matthews KK, Venturi V, Schellens IMM, Borghans JAM, Kesmir C, Price DA, van Baarle D: **CD8+ TCR repertoire formation is guided primarily by the peptide component of the antigenic complex.** *J Immunol* 2013, **190**:931–939.
31. Copelan EA: **Hematopoietic stem-cell transplantation.** *N Engl J Med* 2006, **354**:1813–1826.
32. Passweg JR, Halter J, Bucher C, Gerull S, Heim D, Rovó A, Buser A, Stern M, Tichelli A: **Hematopoietic stem cell transplantation: a review and recommendations for follow-up care for the general practitioner.** *Swiss Med Weekly* 2012, **142**:w13696.
33. Murphy WJ: **Raising the spectra of T-cell profiling.** *Blood* 2008, **112**:3008–3009.
34. Van Heijst JWJ, Ceberio J, Lipuma LB, Samilo DW, Wasilewski GD, Gonzales AMR, Nieves JL, van den Brink MRM, Perales MA, Pamer EG: **Quantitative assessment of T cell repertoire recovery after hematopoietic stem cell transplantation.** *Nat Med* 2013, **19**:372–377.
35. Antin JH: **T-cell depletion in GVHD: less is more?** *Blood* 2011, **117**:6061–6062.
36. Meyer EH, Hsu AR, Lilienthal J, Löhr A, Florek M, Zehnder JL, Strober S, Lavori P, Miklos DB, Johnson DS, Negrin RS: **A distinct evolution of the T cell repertoire categorizes treatment refractory gastrointestinal acute graft-versus-host disease.** *Blood* 2013. doi:10.1182/blood-2013-03-489757.
37. Sieper J, Braun J, Rudwaleit M, Boonen A, Zink A: **Ankylosing spondylitis: an overview.** *Ann Rheum Dis* 2002, **61**:iii8–iii18.
38. Mamedov IZ, Britanova OV, Bolotin DA, Chkalina AV, Staroverov DB, Zvyagin IV, Kotlobay AA, Turchaninova MA, Fedorenko DA, Novik AA, Sharonov GV, Lukyanov S, Chudakov DM, Lebedev YB: **Quantitative tracking of T cell clones after haematopoietic stem cell transplantation.** *EMBO Mol Med* 2011, **3**:201–207.
39. Backes C, Ludwig N, Leidinger P, Harz C, Hoffmann J, Keller A, Meesse E, Lenhof HP: **Immunogenicity of autoantigens.** *BMC Genomics* 2011, **12**:1–13.
40. Klarenbeek PL, De Hair MJH, Doorenspleet ME, van Schaik BDC, Esveldt REE, van de Sande MGH, Cantaert T, Gerlag DM, Baeten D, van Kampen AHC, Baas F, Tak PP, de Vries N: **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.
41. Wu D, Sherwood A, Fromm JR, Winter SS, Dunsmore KP, Loh ML, Greisman HA, Sabath DE, Wood BL, Robins H: **High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia.** *Sci Transl Med* 2012, **4**:134ra63.
42. Jones SJ, Laskin J, Li YY, Griffith OL, An J, Bilenky M, Butterfield YS, Cezard T, Chuah E, Corbett R, Fejes AP, Griffith M, Yee J, Martin M, Mayo M, Melnyk N, Morin RD, Pugh TJ, Severson T, Shah SP, Sutcliffe M, Tam A, Terry J, Thiessen N, Thomson T, Varhol R, Zeng T, Zhao Y, Moore RA, Huntsman DG, et al: **Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors.** *Genome Biol* 2010, **11**:R82.
43. De Palma M, Hanahan D: **The biology of personalized cancer medicine: facing individual complexities underlying hallmark capabilities.** *Mol Oncol* 2012, **6**:111–127.
44. Adaptive Biotechnologies Corporation. <http://www.adaptivebiotech.com/>.
45. iRepertoire. <http://www irepertoire com/>.
46. Sequentia. <http://sequentiainc.com/>.
47. Gigagen. <http://gigagen.com/>.
48. Gooden MJM, de Bock GH, Leffers N, Daemen T, Nijman HW: **The prognostic influence of tumour-infiltrating lymphocytes in cancer: a systematic review with meta-analysis.** *Br J Cancer* 2011, **105**:93–103.
49. Andersen RS, Thrué CA, Junker N, Lyngaa R, Donia M, Ellebæk E, Svane IM, Schumacher TN, Thor Straten P, Hadrup SR: **Dissection of T-cell antigen specificity in human melanoma.** *Cancer Res* 2012, **72**:1642–1650.
50. Emerson RO, Sherwood AM, Rieder MJ, Guenthoer J, Williamson DW, Carlson CS, Drescher CW, Tewari M, Bielas JH, Robins HS: **High-throughput sequencing of T cell receptors reveals a homogeneous repertoire of tumor-infiltrating lymphocytes in ovarian cancer.** *J Pathol* 2013, **10.1002/path.4260**.
51. Miconnet I: **Probing the T-cell receptor repertoire with deep sequencing.** *Curr Opin HIV AIDS* 2012, **7**:64–70.
52. Cebula A, Seweryn M, Rempala GA, Pabla SS, McIndoe RA, Denning TL, Bry L, Kraj P, Kisielow P, Ignatowicz L: **Thymus-derived regulatory T cells contribute to tolerance to commensal microbiota.** *Nature* 2013, **497**:258–262.
53. Boyd SD, Liu Y, Wang C, Martin V, Dunn-Walters DK: **Human lymphocyte repertoires in ageing.** *Curr Opin Immunol* 2013, **25**:511–515.
54. Nikolich-Zugich J, Slifka MK, Messaoudi I: **The many important facets of T-cell repertoire diversity.** *Nat Rev Immunol* 2004, **4**:123–132.
55. Tagoh H, Kishi H, Okumura A, Kitagawa T, Nagata T, Mori K, Muraguchi A: **Induction of recombination activating gene expression in a human lymphoid progenitor cell line: requirement of two separate signals from stromal cells and cytokines.** *Blood* 1996, **88**:4463–4473.
56. Curran KJ, Pegram HJ, Brentjens RJ: **Chimeric antigen receptors for T cell immunotherapy: current understanding and future directions.** *J Gene Med* 2012, **14**:405–415.
57. Molloy PE, Sewell AK, Jakobsen BK: **Soluble T cell receptors: novel immunotherapies.** *Curr Opin Pharmacol* 2005, **5**:438–443.
58. Bassing CH, Swat W, Alt FW: **The mechanism and regulation of chromosomal V(D)J recombination.** *Cell* 2002, **109**(Suppl):S45–S55.
59. Matsui K, Boniface JJ, Reay PA, Schild H, Fazekas de St Groth B, Davis MM: **Low affinity interaction of peptide-MHC complexes with T cell receptors.** *Science* 1991, **254**:1788–1791.
60. Mason D: **A very high level of crossreactivity is an essential feature of the T-cell receptor.** *Immunol Today* 1998, **19**:395–404.
61. Cox AL, Skipper J, Chen Y, Henderson RA, Darrow TL, Shabanowitz J, Engelhard VH, Hunt DF, Slingluff CL: **Identification of a peptide recognized by five melanoma-specific human cytotoxic T cell lines.** *Science* 1994, **264**:716–719.
62. Singh-Jasuja H, Emmerich NPN, Rammensee HG: **The Tübingen approach: identification, selection, and validation of tumor-associated HLA peptides for cancer therapy.** *Cancer Immunol Immunother* 2004, **53**:187–195.
63. Kasuga K: **Comprehensive analysis of MHC ligands in clinical material by immunoaffinity-mass spectrometry.** *Methods Mol Biol* 2013, **1023**:203–218.
64. Parker KC, Bednarek MA, Coligan JE: **Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains.** *J Immunol* 1994, **152**:163–175.
65. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanović S: **SYFPEITHI: database for MHC ligands and peptide motifs.** *Immunogenetics* 1999, **50**:213–219.
66. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B: **The immune epitope database 2.0.** *Nucleic Acids Res* 2010, **38**:D854–D862.
67. Altman JD, Moss PA, Goulder PJ, Barouch DH, McHeyzer-Williams MG, Bell JI, McMichael AJ, Davis MM: **Phenotypic analysis of antigen-specific T lymphocytes.** *Science* 1996, **274**:94–96.
68. Siewert K, Malotka J, Kawakami N, Wekerle H, Hohlfeld R, Dornmair K: **Unbiased identification of target antigens of CD8+ T cells with combinatorial libraries coding for short peptides.** *Nat Med* 2012, **18**:824–828.

69. Seitz S, Schneider CK, Malotka J, Nong X, Engel AG, Wekerle H, Hohlfeld R, Dommar K: Reconstitution of paired T cell receptor alpha- and beta-chains from microdissected single cells of human inflammatory tissues. *Proc Natl Acad Sci U S A* 2006, **103**:12057–12062.
70. Subramanian RA, Moriya C, Martin KL, Peyerl FW, Hasegawa A, Naoi A, Chhay H, Autissier P, Gorgone DA, Lifton MA, Kuus-Reichel K, Schmitz JE, Letvin NL, Kuroda MJ: Engineered T-cell receptor tetramers bind MHC-peptide complexes with high affinity. *Nat Biotechnol* 2004, **22**:1429–1434.
71. Brophy SE, Holler PD, Kranz DM: A yeast display system for engineering functional peptide-MHC complexes. *J Immunol Methods* 2003, **272**:235–246.
72. Vest Hansen N, Ostergaard Pedersen L, Stryhn A, Buus S: Phage display of peptide / major histocompatibility class I complexes. *Eur J Immunol* 2001, **31**:32–38.
73. Le Doussal J, Piqueras B, Dogan I, Debré P, Gorochov G: Phage display of peptide/major histocompatibility complex. *J Immunol Methods* 2000, **241**:147–158.
74. Crawford F, Jordan KR, Stadinski B, Wang Y, Huseby E, Marrack P, Slansky JE, Kappler JW: Use of baculovirus MHC/peptide display libraries to characterize T-cell receptor ligands. *Immunol Rev* 2006, **210**:156–170.
75. Grupp SA, Prak EL, Boyer J, McDonald KR, Shusterman S, Thompson E, Callahan C, Jawad AF, Levine BL, June CH, Sullivan KE: Adoptive transfer of autologous T cells improves T-cell repertoire diversity and long-term B-cell function in pediatric patients with neuroblastoma. *Clin Cancer Res* 2012, **18**:6732–6741.
76. Klarenbeek PL, Remmerswaal EBM, ten Berge IJM, Doorenspleet ME, van Schaik BDC, Esveldt REE, Koch SD, ten Brinke A, van Kampen AHC, Bemelman FJ, Tak PP, Baas F, de Vries N, van Lier RAW: Deep sequencing of antiviral T-cell responses to HCMV and EBV in humans reveals a stable repertoire that is maintained for many years. *PLoS Pathog* 2012, **8**:e1002889.
77. Zhu J, Peng T, Johnston C, Phasouk K, Kask AS, Klock A, Jin L, Diem K, Koelle DM, Wald A, Robins H, Corey L: Immune surveillance by CD8aa(+) skin-resident T cells in human herpes virus infection. *Nature* 2013, **497**:494–497.

doi:10.1186/gm502

Cite this article as: Woodsworth et al.: Sequence analysis of T-cell repertoires in health and disease. *Genome Medicine* 5:98.