Current Opinion in Biotechnology 2020

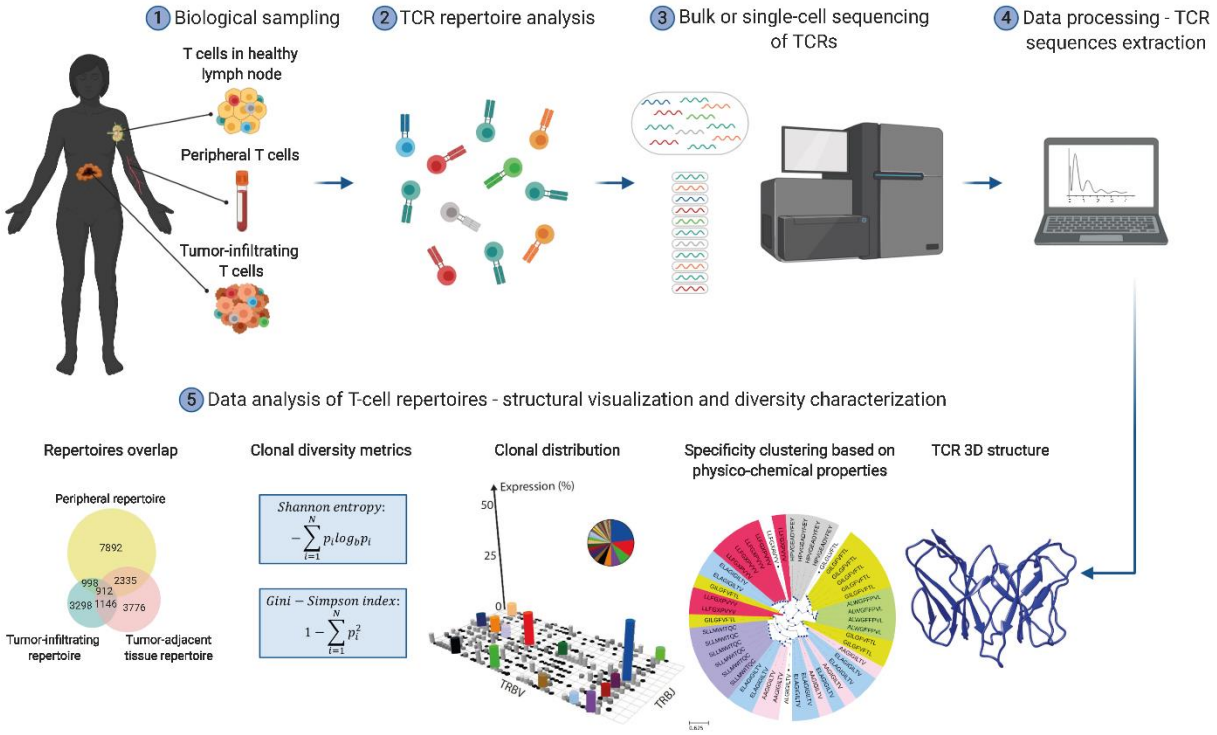
T-cell repertoire analysis and metrics of diversity and clonality

Johanna Chiffelle[1], Raphael Genolet[1], Marta A. S. Perez[2,3], George Coukos[1], Vincent Zoete[2,3], Alexandre Harari[1,*]

[1]Department of Oncology, Ludwig Institute for Cancer Research Lausanne, Lausanne University Hospital (CHUV) and University of Lausanne (UNIL), Lausanne, Switzerland
[2]Department of Oncology, Ludwig Institute for Cancer Research Lausanne, University of Lausanne (UNIL), Lausanne, Switzerland
[3]Molecular Modelling Group, Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

*Corresponding author: Alexandre Harari; alexandre.harari@chuv.ch

Graphical abstract



Abstract

The recent developments of high-throughput bulk and single-cell sequencing technologies accelerated the understanding of the complexity of immune repertoire dynamics combined to transcriptomics. Also, profiling of cellular repertoires in health or disease requires statistical metrics to capture clonal diversity characterized by clones frequency, repertoire richness and convergence. Here we present the common technologies of bulk and single-cell sequencing of T-cell receptors (TCRs), discuss current knowledge regarding computational tools clustering and predicting specificity of TCR repertoires based on shared structural motifs and review main indices for repertoire diversity and convergence analyses. These tools represent potential biomarkers to decipher the fitness of immune repertoires in diseased or treated patients but also the presages and promises of computational approaches to revolutionize personalized immunotherapy.

Highlights

- Immune fitness in health and disease depends notably on the diversity of cellular repertoires
- Single-cell sequencing is fostering the understanding of clonal diversity of T-cell repertoires combined to transcriptomic profiling
- Structural TCR clustering is a promising tool for epitome mapping of immune repertoires
- Plethora of diversity metrics are used as markers of TCR repertoire fitness, yet with no consensus regarding their relevance and overlap
- Undersampling is a caveat in the attempt to capture TCR repertoires diversity with sensitivity

Estimating T-cell repertoire diversity by computational and mathematical modeling

Unlike the innate immune system, which is mobilized by general threats, adaptive immunity is highly specific to antigens and plays a central role in the fight against pathogens and cancer as well as in autoimmune or inflammatory diseases. Recognition of nonself- or self-antigens is mainly driven by T and B cells. The efficacy of T-cell immunity in identifying peptide fragments of antigens bound to the major histocompatibility complex (MHC) molecule depends on the diversity of its repertoire. The development of next-generation sequencing (NGS) and single-cell approaches brought a revolution in the characterization of immune repertoires allowing massive parallel TCR sequencing [1,2]. This led to the development of a wide range of computational and mathematical tools to model interactions between TCR and peptide-MHC (pMHC) and describe repertoire diversity. In the present review, we describe NGS approaches allowing structural characterization of TCRs, which is the basis of clustering models inferring shared antigen specificity of immune repertoires [3]. Aside from these specificity-based clustering models, we also present the different mathematical indexes currently used to interpret TCR diversity and convergence of immune repertoires [4•]. However, diversity measures comprehending the number of distinct clones and their frequencies in a repertoire is not trivial. Thus, different diversity measures are available, each capturing slight differences, giving distinct weights to the relative clonotypes frequency. Moreover, experimental sampling only partially estimates the diversity of repertoires [5•]. Therefore, caution must be taken when interpreting and comparing immune repertoire diversity within and across studies.

TCR structural diversity driving antigen specificity

The structure of the majority of human T-cell receptor is a disulfide-linked α/β heterodimer, each chain composed of a constant and a variable domain [6]. These chains are formed by somatic rearrangements of the variable (V), diversity (D), and joining (J) gene segments together with random addition or deletion of nucleotides [7]. These diversification mechanisms yield a huge variety of TCRs [3]. TCR diversity is confined to six variable hairpin loops located in the α/β variable domains, named complementarity-determining regions (CDRs), with three CDRs per chain (CDR1α, CDR2α and CDR3α and CDR1β, CDR2β and CDR3β, respectively). The process of V(D)J recombination leads to CDRs 1 and 2 entirely encoded in germline DNA segment, whereas the CDR3α and CDR3β loops are products of junctional diversity, consequently being the most variable [8,9]. The binding between TCRs and peptide antigens displayed by MHC is of relatively low-affinity [10,11] and is degenerate, meaning that many TCRs recognize the same peptide antigen and many peptide antigens are recognized by the same TCR [12,13].

During recognition events, CDR1α, CDR1β, CDR2α and CDR2β contact the MHC [14,15], while CDR3α and CDR3β directly communicate with the peptide antigen [16,17••] (**Figure 1**). However, all six CDRs might be involved in antigen recognition [18,19]. As shown on **Figure 1C**, the direct contact between the peptide and CDR1α provides an exception to the rule claiming that CDR3s are responsible for peptide specificity, while the limited contacts exchanged between Trp5 and Met4 and CDR3β opposes the idea that CDR3β, above all, is driving the peptide recognition. This shows that taking all CDRs, as well as detailed structural aspects [20], into account in TCR clustering approaches might be necessary to achieve the highest efficacy [21,22].

Sequencing approaches to capture TCR diversity

If the diversity of immune repertoires was difficult to appreciate in the past, the arrival of NGS created a revolution in the field of TCR analysis and promoted the emergence of several high-throughput TCR sequencing (TCR-Seq) assays to characterize T-cell repertoires. The first factor to consider for TCR sequencing is the source of material, *i.e.* DNA or RNA. DNA was largely used owing to its stable number of copies per cell, thus allowing straightforward quantification of clonotypes frequency. However, DNA-based methods are less sensitive and do not consider allelic exclusion, therefore overestimating diversity. Conversely, RNA is less stable and expression level may vary from cell to cell therefore impacting TCR quantification [1]. However, RNA-based methods are more sensitive, circumvent the allelic exclusion issue and allow implementation of unique molecular identifiers (UMI) that correct for amplification and sequencing errors [23].

Bulk TCR sequencing

Among the latest high-throughput sequencing methods for the analysis of bulk immune populations, three main technical concepts have emerged: 1) TCR amplification by multiplex PCR [24], 2) addition of common adapters prior to PCR amplification [25–27] and 3) TCR amplification following gene capture [28]. Multiplex PCR is the most commonly used but heterogeneity in primers efficiency introduces bias during amplification, leading to misrepresentations in the relative proportion of clones [29]. Next to multiplex amplifications, strategies adding a common adapter to the 5' end for the amplification were developed, such as the 5'RACE PCR [25]. As other ligation-based methods, 5'RACE is limited by a suboptimal ligation efficiency of the adapter [30]. This impacts quantification accuracy and low frequency TCRs detection and could explain why 5'RACE was shown to be less reproducible than multiplex PCR [1,31]. Altogether, biases introduced by current bulk methods affect repertoire analyses and weaken the pairing of α/β chains required for functional analyses and therapeutic applications. To this end, Howie and colleagues introduced a new concept to pair TCR chains based on multiple sequencings of the same sample and combinatorial analyses [32]. This high-throughput method, called pairSEQ, requires a large number of cells from a given clone to allow chains pairing, thus limiting its application to large samples and highly represented clones.

Single-cell TCR sequencing

In the last years, several single-cell based approaches emerged allowing α and β chains pairing, also potentially associated with transcriptional profiling [2,33•]. Originally, physical single-cell isolation conjugated to multiplex PCR and Sanger sequencing [34] or high-throughput sequencing [35] was developed to obtain paired TCRαβ sequences. Han and colleagues, by using a single-cell barcoding strategy could increase the scalability of sequenced cells and, in addition to the combinative determination of both TCR chains, could sequence specific genes linked to T-cell functionality [35]. However, these methods only allowed hundreds to few thousands of cells to be sequenced.

A major improvement in the throughput of single-cell TCR sequencing (scTCR-Seq) came with emulsion-based approaches. Using microfluidics, water-in-oil emulsion droplets containing a single cell trapped with small volumes of reagents are created, multiplexing the number of cells analyzed. A method using emulsion-trapped cells paired the α and β transcripts by overlap extension reverse-transcription PCR directly within the droplet [36]. Despite using high number

of cells as starting material, the yield was low, therefore affecting the detection of rare clones. A few years later, an updated version came out as a new platform for T-cell repertoire analysis [37]. This low-cost technology allowed, for the first time, a full high precision profiling of TCR sequences from millions of cells. Recently, Spindler and colleagues presented a high-throughput method linking TCR identification with direct functional testing to determine TCR reactivity and avidity using a microfluidics-based system [38]. Currently, a commercially available and easy-to-use system is widely used for single-cell profiling of immune cells, for instance for intratumoral immune populations characterization or clonal changes upon anti-PD-1 therapy [39••,40]. This microfluidics technology, developed by 10x Genomics, generates so-called Gel Beads-in-emulsion containing bead-attached primers with DNA barcodes capturing polyadenylated mRNA and resulting in barcoded cDNA. Although 10x Genomics approach is detecting fewer genes than other single-cell RNA sequencing (scRNA-Seq) methods, it can cover up to 15'000-20'000 cells and can combine scTCR-Seq with transcriptional profiling of T-cell subsets.

Other commercially available single-cell encapsulation methods are being developed but a major drawback of these technologies is the need for microfluidics devices that are not always accessible by research laboratories as well as the high cost of ready-to-use assays, such as 10x Genomics. Moreover, the scalability is often limited as compared to bulk sequencing, due to the microfluidics technology itself and the yield can be low: 10x Genomics reaches 50-60% of successful cell encapsulation. An overview of the applications and limitations of the aforementioned bulk and single-cell TCR-sequencing methods is presented in **Figure 2**. Despite being attractive for multiplexed data, single-cell transcriptomic profile analyses require high cellular viability material and significant computational analyses need to be handled afterwards. However, the major developments in single-cell immune repertoires sequencing coupled to transcriptomic signature are shedding a new light on the description of T-cells' clonality and dynamics within a wide range of applications such as the development and improvement of immunotherapeutic treatments for cancer research.

Specificity clustering of TCR based on sequence similarity architecture

The prediction of epitopes recognized by a repertoire of T-cells (*i.e.* the epitome) from TCR sequences remains one of the biggest challenge of cellular and computational immunologists. Identifying TCR by deep sequencing of immune repertoires allows discovery of receptor patterns that might be linked to antigen specificity or to clinical outcomes. Recent computational studies demonstrated that common patterns can be inferred among TCR sequences interacting with the same epitope [17••,41••,42••], opening the perspective of *in silico* prediction of targets, diversity and complexity of TCR repertoires obtained experimentally [3,21,43••,44].

Global and local motifs similarity: the GLIPH algorithm

Analysis of 52 TCR-pMHC structures highlighted the possible determination of pMHC contact sites in CDR3s, notably in CDR3β, as an opportunity to cluster with a high probability TCRs on the basis of the prediction of shared specificity [17••]. Based on this assumption, the authors developed a clustering algorithm, called GLIPH (grouping of lymphocyte interactions by paratope hotspots), built on global and local TCR sequences similarity. GLIPH specificity groups, likely to recognize the same or very similar MHC ligands, are scored based on the enrichment of common V-genes, the CDR3 lengths, clonal expansions, shared HLA alleles among contributors, motif significance and cluster size. When benchmarking GLIPH on a

training set of 2,068 unique sequences spanning eight pMHC specificities, 94% of TCRs were correctly grouped in clusters of TCRs with common specificity, even when originating from different donors. Such an approach could be used to predict the specificity of a new TCR, by verifying its affiliation to a specificity group determined by GLIPH. Essentially, it also provides information regarding a given immune response and its complexity through the analysis of the number and size of the clusters determined by GLIPH.

Distance measure: the TCRdist algorithm

Also based on sequence similarity, Dash and colleagues defined a novel distance measure on the space of TCRs, TCRdist, allowing for clustering and visualization of repertoire diversity [41••]. This quantitative measure of similarity is obtained by listing the residues belonging to the CDR1, 2 and 3 loops, all known to possibly contact the pMHC, and by computing a similarity-weighted mismatch distance defined based on the BLOSUM62 substitution matrix, with a gap penalty to capture variations in the length of CDRs. Of note, a higher weight was given to the CDR3 sequence in view of its prominent role in epitope binding. This distance can then be calculated for each possible pair of TCRs belonging to a given repertoire, generating a so-called distance matrix. It can be used for TCRs clustering or the construction of hierarchical distance trees to analyze the diversity and complexity of TCR repertoires. The high-dimensional TCR landscape can also be projected into two dimensions plots, with each dot representing a TCR, through the dimensionality reduction of this distance matrix. Thanks to these analytical tools based on their definition of the distance between two TCRs, the authors found that TCR repertoires often contain dominant clusters of TCRs whose sequence similarity is generated partially from the use of common V- and J-regions and from the similarity of CDR3 motifs. Moreover, each epitope-specific repertoire enclosed a clustered group of receptors with strong sequence similarities, together with divergent non-clustered receptors, both providing different solutions to the pMHC binding challenge. Finally, they highlighted key conserved residues driving TCR binding to pMHC.

Clustering based on TCRs biophysicochemical properties

Recently, Ostmeyer and colleagues introduced a novel class of methods for analyzing immune repertoires of patients in order to cluster and identify disease-associated TCRs [42••]. Their approach consists in feeding machine-learning techniques, based on logistic regressions, with biophysicochemical descriptors of the TCR interface, rather than with TCR sequences. The biophysicochemical characteristics of sliding windows of four consecutive residues of CDR3β (*i.e.* so-called 4-mers), excluding the first four and last three residues, are described using five Atchley factors encoding for codon diversity, secondary structure, molecular size, polarity and electrostatic charge of the residues. The method identified a short list of preferred values for these descriptors at key positions in TCRs present in tumors, which permitted the identification of disease-associated TCRs. Although this approach leads to the hypothesis that these TCRs share the same specificity, this was however not validated. In addition, restricting the analysis of 4-mers of CDR3β, a choice resulting from the analysis of a small number of TCR-pMHC structures, constitutes a limitation of the method. Nevertheless, it represents a first step in the direction of physics-based predictors that can potentially fit the extremely large sequence diversity of immune receptors into a limited number of quantitative characteristics at key positions. Although the method needs to be retrained for each set of TCRs and remains restricted to CDR3β only limiting its predictive ability, this type of sequence-based 'property'-based approach could circumvent some of the drawbacks of purely sequenced-based

analyses. Indeed, very large numbers of disease-associated TCR sequences for training are not necessary anymore and the possibility to detect potential antigen-binding TCRs with divergent sequences from those previously encountered exists. This approach can also be used to cluster and analyze TCRs repertoires, by defining a possible distance between two receptors as the difference between the five Atchley factors of the most similar pair of 4-mers taken from their respective CDR3β (clustering tree example in the abstract figure).

Quantifying clonality, diversity and convergence of TCR repertoires

Aside from the diversity in antigenic specificity of T-cell repertoires (*i.e.* the epitome), clonotype diversity can capture immune fitness during disease development or in response to treatment. Numerous computational algorithms analyzing sequence reads of TCRs and characterizing repertoire clonality were established [45]. The broad structural diversity characterizing TCRs renders the analysis of immune repertoires challenging but allows fingerprinting of T-cell clones that can be tracked within different tissues (peripheral blood, tumor tissue, adjacent normal tissue, etc.) at different time-points in immune profiling studies. In the past years, several studies centered their analyses on TCR repertoire dynamics as indicators of immune monitoring in inflammatory diseases such as multiple sclerosis [46], autoimmune diseases [47], viral infection [48,49] or cancer [43••,50–52] as well as as biomarkers of response to immunotherapy [40,53–55••]. Therefore, models for immune repertoires visualization and statistically-derived descriptive indices to estimate repertoire diversity and homology with no described consensus analytical method have emerged [4•]. In the following section, we recapitulate the main indices characterizing diversity and similarity of T-cell repertoires and discuss their limitations.

Diversity measures: Hill numbers and Rényi entropy

Most of diversity indices are mathematically derived from the information theory widely used in ecology to quantify ecosystems biodiversity [5•,56]. In T-cell repertoires, diversity takes into account the clonal composition, equivalent to the number of unique TCR sequences referred from now on as richness and the distribution spectrum of these sequences (*i.e.* their relative abundance) hereafter referred to as evenness. Diversity relates to the level of uncertainty that a TCR sequence would be sorted from a repertoire and would belong to a certain T-cell clone (*i.e.* unique TCR sequence). Commonly used measures of diversity are related to the Hill numbers also referred to as effective numbers of species, from which one can retrieve the effective number of distinct clonotypes (*i.e.* number of equally abundant sequences producing the given value of diversity) in the dataset [57,58]:

$$Hill\ numbers = D_\alpha = \left(\sum_{i=1}^{N} p_i^\alpha\right)^{1/(1-\alpha)} \quad (1)$$

where p_i is the frequency of sequence i in the repertoire and N is the total number of unique sequences. The order α parametrizes the diversity index and allows to calculate different features of immune repertoire diversity.

The Hill diversity numbers are based on the generalized measure of entropy, the Rényi entropy, quantifying the diversity or randomness of a system [4•,58,59]:

$$Rényi\ entropy = H_\alpha = \frac{1}{1-\alpha}\log_b\left(\sum_{i=1}^{N} p_i^\alpha\right) \quad (2)$$

where *b*, the base of the logarithm, determines the choice of units of the entropy measure.

Diversity of order 1: Shannon entropy

The order *α* sets the degree of sensitivity of the diversity index to species abundance in the system. When *α*→0, all species are weighted equally and (1) is equivalent to species richness meaning the number of unique sequences in a repertoire, independently of their abundance. When *α*→1, the generalized form of the entropy (2) is equivalent to the Shannon entropy or Shannon diversity index [60]:

$$Shannon\ entropy = H_1 = -\sum_{i=1}^{N} p_i log_b p_i \quad (3)$$

Figure 3A shows that monoclonal (*i.e.* 1 TCR) and oligoclonal repertoires (*i.e.* emergence of a few dominant clones) have a Shannon's index closer to 0. Moreover, when there is a unique dominant clone and the other clones are evenly represented, the Shannon index is higher than in case of oligoclonality due to a higher uncertainty of the possible outcome of picking one sequence in the repertoire in the first case. Thus, when a repertoire is composed of sequences evenly distributed, the Shannon entropy reaches his maximum (*i.e.* maximal diversity), which is the logarithm of the number of unique sequences. This index being widely described, it is often used in immune studies. For example, when profiling dynamic changes in peripheral T-cell repertoire upon cervical carcinogenesis, the use of Shannon entropy index revealed a drop in diversity in patients with advanced cancer, thus potentially reflecting the emergence of expanded clones [50]. Shannon entropy was also used to discriminate diversity changes in melanoma-bearing mice receiving different combinations of immunotherapy [61] and was linked to clinical prognosis in patients with advanced lung cancer [62].

Diversity of order 2: Gini-Simpson index

Finally, when *α*→2:, the generalized entropy formula (2) becomes:

$$H_2 = -log_b \sum_{i=1}^{N} p_i^2 = -log_b(\lambda) \quad (4)$$

where *λ* represents the Simpson's index [63], the probability of two entities being chosen randomly in a system (sampling with replacement) to belong to the same species. To follow the intuitive principle that a high index expresses high diversity, people commonly use the unity minus the Simpson's index, referred to as Simpson diversity index or Gini-Simpson index:

$$Gini - Simpson\ index = 1 - \lambda = 1 - \sum_{i=1}^{N} p_i^2 \quad (5)$$

with value close to 0 characterizing a repertoire with no diversity (*i.e.* highly oligoclonal) and 1 representing infinite diversity (*i.e.* polyclonal repertoire with equivalent representation of each clone). In **Figure 3A**, the highly diverse scenarios (#4, #7, #10 and #13) have a Gini-Simpson index that increases with higher richness to get closer to 1. Along with Shannon entropy, the Gini-Simpson index decreases with appearance of dominant clones since the probability of two selected sequences to be different drops. Rather than the Shannon entropy, several studies of repertoire diversity use the Gini-Simpson index. Lately, it was applied to describe the clonal architecture of patients with adult T-cell leukemia/lymphoma [64] or to assess the

clinical prognostic value of T-cell repertoires from peripheral blood or metastases in patients with primary melanoma [51•].

Entropy-based diversity indices limitations

As mentioned, the *α* order determines the indices' sensitivity to rare or common species. Orders lower than 1 reflect a diversity measure highly affected by the number of rare species whereas increasing *α* orders tend to be more sensitive to abundant species and when *α*=1, each species is weighted by its proportional abundance [65]. Therefore, the Shannon diversity index encounters higher variation upon addition of low frequency clones than the Gini-Simpson index. In **Figure 3A**, the Gini-Simpson index, in contrast to the Shannon entropy, is barely affected by the increasing number of unique TCRs in the repertoire. Moreover, within a repertoire composed of equal numbers of unique TCRs, the Shannon entropy is more impacted by the presence of low frequency clones than the Gini-Simpson index. Most of the studies do not mention the rationale behind the choice of the diversity indices. Moreover, all these diversity indices behaving non-linearly, caution should be taken when correlating them to biological interpretation and statistical tests should be adapted. The best way to correctly interpret these entropy-derived measures would be to analyze them simultaneously (*i.e*. "diversity profiles") to be able to derive any biological meaning from the observed differences [66•].

Evenness measure: Pielou's index

Aside from the degree of uncertainty and heterogeneity of a system, description of the equivalency in species abundance can also be used. This measures the dominance of clones in a repertoire thus referred to as clonal diversity or clonal evenness. In a study describing changes in peripheral blood TCR diversity upon ipilimumab treatment in metastatic melanoma, the authors characterized clonal diversity defined as the ratio between the number of sequences accounting for 50% of the total repertoire abundance (*i.e.* cumulative frequency of each of these sequences) and the repertoire richness [67]. This measurement, referred to as diversity evenness 50 (DE_{50}), was used to describe increasing oligoclonal responses in TILs from melanoma-bearing mice treated with optimal combinative immunotherapy [61]. In parallel to DE_{50}, clonal evenness of a repertoire can be calculated using Pielou's index, which is itself derived from the ratio between the Shannon entropy and the maximization of the diversity distribution of species within a sample [68]:

$$Pielou's\ index = -\frac{\sum_{i=1}^{N} p_i log_b p_i}{log_b(\mathrm{N})} = \frac{H_1}{H_{1max}} \quad (6)$$

As shown in **Figure 3A**, the complement of clonal evenness (1-Pielou's index) is often used to get a clonality score of 0 representing a maximally diverse population with even frequencies and values close to 1, a repertoire driven by clonal dominance. As shown in **Figure 3A**, even though the abundance of dominant clones in repertoires #3, #6, #9 and #12 is identical, clonal evenness increases since the dominance of these oligoclonal sequences is more important in the case of a repertoire with high richness. In examining peripheral and tumoral T-cell clonality in patients with metastatic melanoma treated with immunotherapy drugs, an association between clonal expansion represented by 1-Pielou's index and clinical response was highlighted [53]. Recently, T-cell repertoires obtained from 236 NSCLC patients showed higher TCR clonality measured by 1-Pielou's evenness in healthy tumor-adjacent tissue compared to tumor tissue suggesting an impaired antigenic response [43••].

Inequality measure: Gini coefficient

Another index, the Gini coefficient (not to be mistaken with the Gini-Simpson index) is sometimes used to represent clonal distribution of a repertoire. It is a measure of inequality that is widely used in economics to study wealth distribution [69]. It quantifies the balance of a system (*i.e.* evenness of distribution) rather than its variety (*i.e.* species richness) [70]:

$$Gini_c = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|p_i - p_j|}{2N^2\bar{p}} \quad (7)$$

with p_i, p_j the frequency of the respective i^{th} and j^{th} sequences in the repertoire and $\bar{p}$ the average of clone frequencies. Gini coefficient ranges from 0, maximal diversity of the repertoire (*i.e.* equal abundance of each sequence) to 1, with high value representing extreme inequality (*i.e.* high clonality towards one sequence). Thus, in **Figure 3A**, the Gini coefficient increases as the number of abundant clones rises, thus further reducing the frequency of less represented clones (*i.e.* higher inequality). Moreover, with increasing richness of repertoires, the inequality between dominant and sub-dominant (low frequency) clones gets wider, leading to a small rise in the Gini coefficient. In a recent study interpreting T-cell evolution upon checkpoint inhibitors treated melanoma patients, repertoire clonality was assessed using the Gini coefficient [55••]. Moreover, a linear discriminant analysis was built to distinguish patients based on their clinical response using clonal dominance (*i.e.* Gini coefficient) and diversity (*i.e.* Rényi entropy with α=1) as repertoire features.

Repertoires overlap measures

Aside from measures of diversity and clonality that are applied on a unique repertoire, TCR sequencing data also call for similarity analyses allowing comparison of overlap between T-cell repertoires. A first similarity indices, the Jaccard index, is defined as the size of overlapping species divided by the size of the union of both compared samples [71]:

$$Jaccard\ index = J(i,j) = \frac{c_{ij}}{N_i + N_j - c_{ij}} \quad (8)$$

with c_{ij} being the number of overlapping sequences and N_i and N_j the total number of sequences in repertoire i and j respectively. Its related indices, the Sorensen index or Sorensen-Dice coefficient differs by counting twice the shared sequences (once in both the numerator and the denominator) [72,73]:

$$Sorensen\ index = S(i,j) = \frac{2c_{ij}}{N_i + N_j} \quad (9)$$

Both indexes vary from 0 (no similarity) to 1 (total similarity between repertoires). From the Sorensen index, the Bray-Curtis index of dissimilarity can be deduced as the complement of the Sorensen index (*i.e.* Bray-Curtis index = 1-Sorensen index) [74]. All these similarity indices are based on the presence or absence of specific sequences therefore retaining sensitivity in more heterogeneous repertoires but not taking into consideration the relative abundance of the overlapping sequences. Thus, repertoire homology between healthy tumor-adjacent tissue and tumor tissue only based on Jaccard index is not robust enough to drive any conclusion and other metrics should be used in parallel [43••]. To this extend, the Morisita-Horn overlap index considering the relative frequency of species in compared samples is a widely used measure of dispersion [75,76]:

$$Morisita - Horn\ index = MH(i,j) = \frac{2\sum_{i=1}^{S} n_{1i}n_{2i}}{(\sum_{i=1}^{S} f_i^2 + \sum_{i=1}^{S} g_i^2)n_1 n_2} \quad (10)$$

with $f_i=n_{1i}/n_1$ and $g_i=n_{2i}/n_2$, n_{1i} and n_{2i} being the clone sizes of the i^{th} sequence (*i.e.* entities representing a sequence) and n_1 and n_2, the total number of entities in sample 1 and 2 respectively. *S* is the total number of unique sequences found in both samples. The indices goes from 0 (*i.e*. no overlap between repertoires) to 1, repertoires identical in terms of richness and evenness. The Morisita-Horn index can be used to compare immune repertoires during viral infection [49], among different T-cell compartments in cancer patients [43••], to observe T-cell repertoire turnover upon treatment [53] or to track persistence of clones from an immune therapeutic product in peripheral blood after adoptive cell transfer [54].

Undersampling – "unseen species" problem

All aforementioned metrics are widely used to profile T-cell repertoires. However, due to the high diversity of TCR sequences and limitations in sequencing methods, the frequency distribution of clones in a repertoire and its richness is largely biased by the fact that only a fraction of repertoires is analyzed, leading to undersampling (*i.e.* "unseen species" problem) [5•]. This translates into biases in diversity measures, as shown in **Figure 3B**, where 18 cells were sampled out of a repertoire composed of 180 cells with 10 unique clones. Undersampling was repeated ten million times to get the frequency of occurrence of the most probable scenarios. The top five and five additional randomly selected ones based on the Monte Carlo approach are shown. Strikingly, we observe that the probability of each subsampled scenarios is low, even for the #1 scenario, recapitulating the richness and evenness of the total repertoire, showing the heterogeneity in clones distribution obtain by sampling a large TCR repertoire. The fold changes between each undersampled scenario and the total repertoire for four diversity metrics are highlighted. In the didactic example shown, clonal evenness (*i.e.* 1-Pielou's evenness) represents the index that is the most affected by undersampling, as clonal distribution is biased relative to the total repertoire. The Gini coefficient, also relying on clone distribution, can be less sensible to undersampling since unique TCR sequences present in the total repertoire disappear, balancing the inequality brought by changes in frequency distribution such as in scenarios #7 and #8. Between the two diversity indices derived from Rényi entropy, Shannon entropy is more sensitive to undersampling than the Gini-Simpson index, mostly due to changes in low frequency clone numbers (*i.e.* repertoire richness). In addition, scenarios #4 and #5 present a case of homogeneity between the indices because the number and frequency of each TCR sequence is stable. However, we miss the information of different sequences sampled from the original repertoire, each scenario capturing another structural diversity. To address the issue of underestimating TCR repertoire diversity by sampling only few cells, several estimators of species richness were developed and can be applied for immune repertoire analyses: Rempala *et al.*, 2011 [77], Chao and Jost, 2012 [78], Chao and Jost, 2015 [58], Greiff *et al.*, 2015 [45], Laydon *et al.*, 2015 [5•], Koch *et al.*, 2018 [79]..

Conclusions

The diversity of clonotypes composing a repertoire is a major feature of the immune system and reflects the epitome of naïve as well as antigen-experienced T-cells. Even though scRNA-Seq methods now allow coupling TCR sequences with transcriptomics, predictions of antigen specificity of a given repertoire remains challenging. In theory, computational approaches

based on structural modeling rise opportunities for epitome mapping and prediction of TCR cross-reactivity of completely different sequences. Moreover, even though deep repertoire profiling magnifies the capacity to capture TCR diversity, sampling of repertoires commonly leads to an inaccurate estimation of diversity. This limits the interpretation of dynamic clonality changes of immune repertoires captured with diversity metrics. Various methods are now being developed to accurately estimate true diversity of cellular repertoires. Moreover, a gold standard method for immune repertoire analysis has not yet been described, revealing the caution that need to be taken when comparing studies using various measurement methods. However, deep profiling of T-cell repertoires represents a potential biomarker to characterize immune fitness in diseased or treated patients and development of computational tools to measure diversity changes could foster immunology research such as cancer immunotherapy.

Figure 1

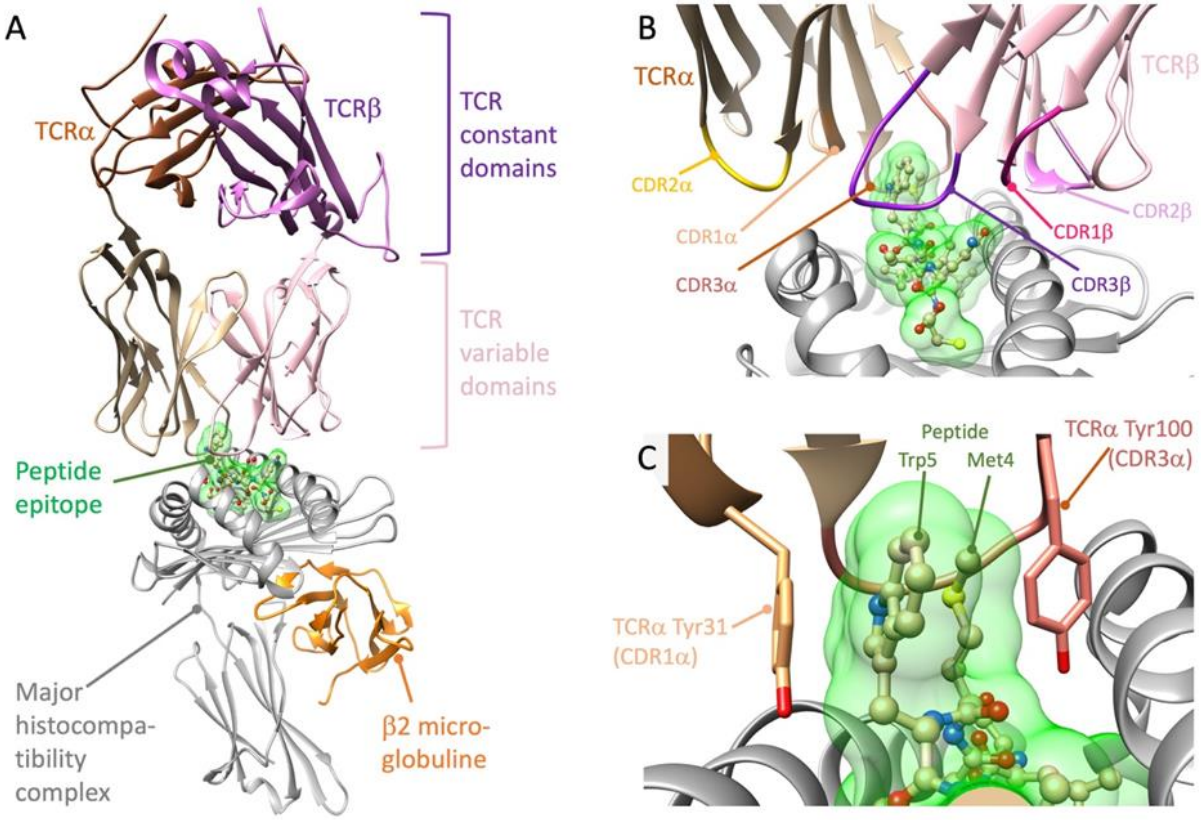


Figure 1. Structure of the TCR-pMHC complex with detailed interaction between the NY-ESO-1 peptide and its TCR. **A)** TCR-pMHC complex (PDB ID 2BNR, Chen *et al.*, 2005[1]), showing the recognition, by the 1G4 TCR, of the NY-ESO-1$_{(157\text{-}165)}$ -SLLMWITQC peptide presented in the context of HLA-A2. TCR, MHC and β2-microglobulin are shown in ribbon representation. TCRα constant and variable domains are colored in light and dark brown, respectively, while the TCRβ constant and variable domains are colored in light and dark pink. The MHC molecule is colored in grey and the β2-microglobulin in orange. The peptide epitope is shown in ball and stick representation, with a transparent green surface. **B)** Zoom on the TCR-pMHC interface. The representation and color coding are identical to A), with the exception of the CDR1, CDR2 and CDR3 loops of the TCR, which are colored differently. **C)** Zoom on the peptide Trp5. Tyr100 of CDR3α and Tyr31 of CDR1α are shown in stick representation.

[1] Chen J-L, Stewart-Jones G, Bossi G, Lissin NM, Wooldridge L, Choi EML, Held G, Dunbar PR, Esnouf RM, Sami M, et al.: Structural and kinetic basis for heightened immunogenicity of T cell vaccines. J Exp Med 2005, 201:1243–1255.

Figure 2

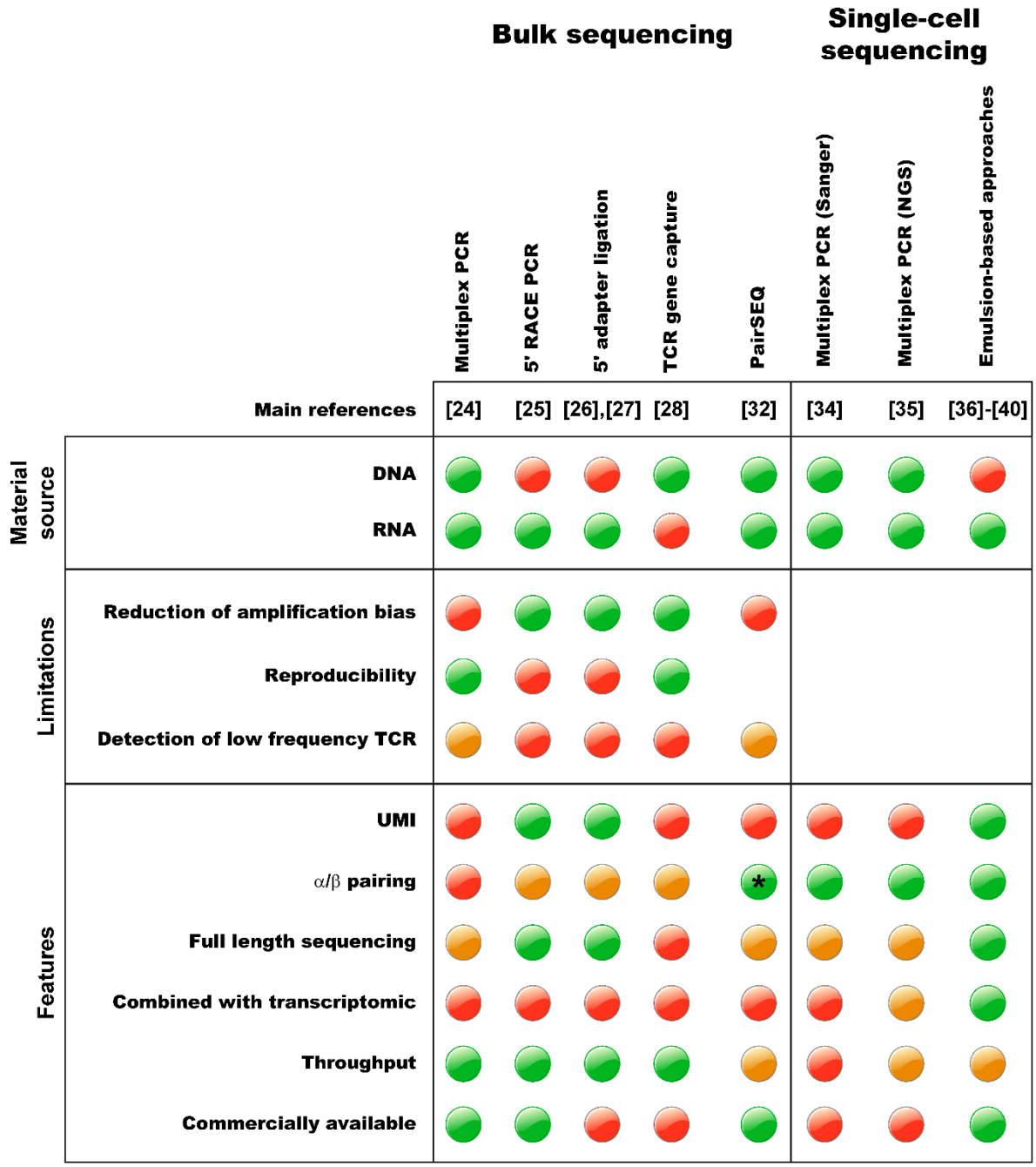


Figure 2. Comparison of bulk and single-cell technologies for TCR sequencing. Each of the presented methods is shown here with its corresponding reference publications. The material source is displayed as well as the features linked to the sequencing approach. The throughput ranges from bulk to single-cell methods thus showing the limitation of scTCR-Seq in terms of number of sequenced cells compared to bulk methods. The reduction of amplification bias, reproducibility and detection of low frequency TCRs is only applied to bulk sequencing, since these impact quantitatively the sequenced TCR chain. In case of single-cell sequencing, the frequency of TCR is directly linked to ratio of cells of a specific clonotype to the total number of cells and is not distorted by amplification bias or reproducibility on TCR transcripts. *****Compared to scTCR-Seq, PairSeq pairs α and β chains from bulk sequencing but the yield is much lower since many cells are needed for its combinatorial analyses allowing successful pairing.

Figure 3

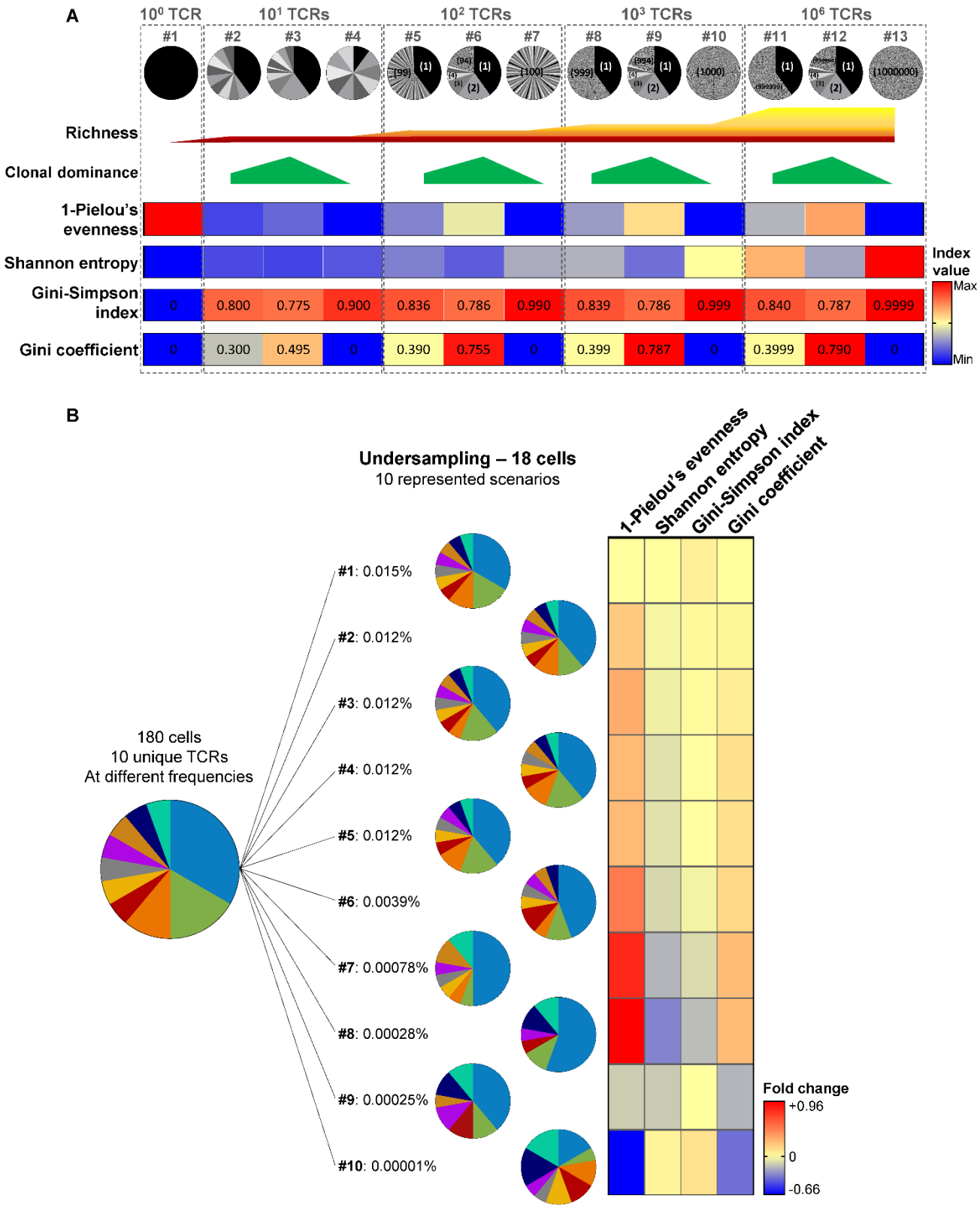


Figure 3. Diversity metrics variation upon richness and evenness changes. A) Different scenarios of clonal distribution in a repertoire of ten million cells sequenced at the same depth for normalization are represented here. The changes in clonality (1-Pielou's evenness), Shannon entropy, Gini-Simpson index and Gini coefficient are shown in a heatmap, whose axis variables depend on the index used

([0:1] for 1-Pielou's evenness, [0:19.93] for Shannon entropy, [0:0.9999991] for Gini-Simpson index and [0:0.79] for Gini coefficient]). **B)** Clonal distribution of ten unique TCRs in a repertoire composed of 180 cells is presented here. A simulation of extraction of 18 cells was repeated ten million times to obtain the frequency of manifestation of each subsampled scenario. The five most frequent situations of subsampling are shown and five others were randomly chosen according to the Monte Carlo approach. The four diversity indexes presented in A) are here represented as fold change from the initial scenario of 180 cells. In A) and B), the Gini-Simpson index was calculated based on the formula[1] without replacement since the repertoires in B) are composed of only 180 or 18 entities. This formula tends to the one with replacement (5) in case of large datasets such as in A).

[1] $1 - \sum_{i=1}^{N} \left(\frac{n_i(n_i-1)}{n(n-1)}\right)$ with n_i is the clone size of the i^{th} clonotype (*i.e.* number of entities weighting a specific sequence) and n the total number of entities found in the overall repertoire.

Funding and acknowledgements

This work was supported by FNS grant 310030_182384.
We thank Fabrizio Benedetti for assistance in the mathematical and statistical interpretation of diversity indices.

Conflict of interest statement

Nothing declared.

Credit authorship contribution statement

Johanna Chiffelle: Writing – Original Draft Writing – Review & Editing

Raphael Genolet: Writing – Original Draft Writing – Review & Editing

Marta A. S. Perez: Writing – Original Draft Writing – Review & Editing

George Coukos: Writing – Review & Editing

Vincent Zoete: Writing – Review & Editing

Alexandre Harari: Funding Acquisition, Supervision, Writing – Review and Editing

References

1. Rosati E, Dowds CM, Liaskou E, Henriksen EKK, Karlsen TH, Franke A: **Overview of methodologies for T-cell receptor repertoire analysis**. *BMC Biotechnol* 2017, **17**:61.

2. De Simone M, Rossetti G, Pagani M: **Single Cell T Cell Receptor Sequencing: Techniques and Future Challenges**. *Front Immunol* 2018, **9**:1638.

3. Davis MM, Boyd SD: **Recent progress in the analysis of αβ T cell and B cell receptor repertoires**. *Curr Opin Immunol* 2019, **59**:109–114.

- 4. Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V: **Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires**. *Front Immunol* 2018, **9**:224.

Review on the currently available computational tools used for immune repertoire characterization (B and T-cells) with a focus on diversity, clustering based on similarity architecture, antibody evolutionary relationships and overlap of repertoires.

- 5. Laydon DJ, Bangham CRM, Asquith B: **Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach**. *Philos Trans R Soc Lond B Biol Sci* 2015, **370**:1675.

Study presenting the undersampled experimental limitation in capturing total diversity of T-cell repertoires and comparing the different available estimators used to correct this undesired effect. The authors developed a rarefaction-based diversity estimator shown to be superior to commonly used estimators of TCR sequences richness.

6. Arnaud J, Huchenq A, Vernhes MC, Caspar-Bauguil S, Lenfant F, Sancho J, Terhorst C, Rubin B: **The interchain disulfide bond between TCR alpha beta heterodimers on human T cells is not required for TCR-CD3 membrane expression and signal transduction.** *Int Immunol* 1997, **9**:615–626.

7. Bassing CH, Swat W, Alt FW: **The Mechanism and Regulation of Chromosomal V(D)J Recombination**. *Cell* 2002, **109**:S45–S55.

8. Davis MM, Bjorkman PJ: **T-cell antigen receptor genes and T-cell recognition**. *Nature* 1988, **334**:395–402.

9. Mora T, Walczak AM: **Quantifying lymphocyte receptor diversity**. *bioRxiv* 2016, doi:10.1101/046870.

10. Borrman T, Cimons J, Cosiano M, Purcaro M, Pierce BG, Baker BM, Weng Z: **ATLAS: A database linking binding affinities with structures for wild-type and mutant TCR-pMHC complexes**. *Proteins* 2017, **85**:908–916.

11. Gálvez J, Gálvez JJ, García-Peñarrubia P: **Is TCR/pMHC Affinity a Good Estimate of the T-cell Response? An Answer Based on Predictions From 12 Phenotypic Models**. *Front Immunol* 2019, **10**:349.

12. Gowthaman R, Pierce BG: **TCR3d: The T cell receptor structural repertoire database**. *Bioinformatics* 2019, **35**:5323–5325.

13. Bagaev DV, Vroomans RMA, Samir J, Stervbo U, Rius C, Dolton G, Greenshields-Watson A, Attaf M, Egorov ES, Zvyagin IV, et al.: **VDJdb in 2019: database extension, new analysis**

infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res* 2020, **48**:D1057–D1062.

14. Sharon E, Sibener LV, Battle A, Fraser HB, Garcia KC, Pritchard JK: **Genetic variation in MHC proteins is associated with T cell receptor expression biases**. *Nat Genet* 2016, **48**:995–1002.

15. Blevins SJ, Pierce BG, Singh NK, Riley TP, Wang Y, Spear TT, Nishimura MI, Weng Z, Baker BM: **How structural adaptability exists alongside HLA-A2 bias in the human αβ TCR repertoire**. *PNAS* 2016, **113**:E1276–E1285.

16. Cole DK, Miles KM, Madura F, Holland CJ, Schauenburg AJA, Godkin AJ, Bulek AM, Fuller A, Akpovwa HJE, Pymm PG, et al.: **T-cell Receptor (TCR)-Peptide Specificity Overrides Affinity-enhancing TCR-Major Histocompatibility Complex Interactions**. *J Biol Chem* 2014, **289**:628–638.

●● 17. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, Ji X, Han A, Krams SM, Pettus C, et al.: **Identifying specificity groups in the T cell receptor repertoire**. *Nature* 2017, **547**:94–98.

The authors developed GLIPH (grouping of lymphocyte interactions by paratope hotspots), an algorithm that clusters TCRs with a high probability of shared specificity.

18. Sela-Culang I, Kunik V, Ofran Y: **The Structural Basis of Antibody-Antigen Recognition**. *Front Immunol* 2013, **4**:302.

19. Lanzarotti E, Marcatili P, Nielsen M: **T-Cell Receptor Cognate Target Prediction Based on Paired α and β Chain Sequence and Structural CDR Loop Similarities**. *Front Immunol* 2019, **10**:2080.

20. Jensen KK, Rantos V, Jappe EC, Olsen TH, Jespersen MC, Jurtz V, Jessen LE, Lanzarotti E, Mahajan S, Peters B, et al.: **TCRpMHCmodels: Structural modelling of TCR-pMHC class I complexes**. *Sci Rep* 2019, **9**:14530.

21. Zvyagin IV, Tsvetkov VO, Chudakov DM, Shugay M: **An overview of immunoinformatics approaches and databases linking T cell receptor repertoires to their antigen specificity**. *Immunogenetics* 2020, **72**:77–84.

22. Lanzarotti E, Marcatili P, Nielsen M: **Identification of the cognate peptide-MHC target of T cell receptors using molecular modeling and force field scoring**. *Mol Immunol* 2018, **94**:91–97.

23. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, Bolotin DA, Staroverov DB, Putintseva EV, Plevova K, et al.: **Towards error-free profiling of immune repertoires**. *Nat Methods* 2014, **11**:653–655.

24. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS: **Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells**. *Blood* 2009, **114**:4099–4107.

25. 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.

26. Kitaura K, Shini T, Matsutani T, Suzuki R: **A new high-throughput sequencing method for determining diversity and similarity of T cell receptor (TCR) α and β repertoires and identifying potential new invariant TCR α chains**. *BMC Immunol* 2016, **17**:38.

27. Ruggiero E, Nicolay JP, Fronza R, Arens A, Paruzynski A, Nowrouzi A, Ürenden G, Lulay C, Schneider S, Goerdt S, et al.: **High-resolution analysis of the human T-cell receptor repertoire**. *Nat Commun* 2015, **6**:8081.

28. Linnemann C, Heemskerk B, Kvistborg P, Kluin RJC, Bolotin DA, Chen X, Bresser K, Nieuwland M, Schotte R, Michels S, et al.: **High-throughput identification of antigen-specific TCRs by TCR gene capture**. *Nat Med* 2013, **19**:1534–1541.

29. Polz MF, Cavanaugh CM: **Bias in Template-to-Product Ratios in Multitemplate PCR**. *Appl Environ Microbiol* 1998, **64**:3724–3730.

30. Wulf MG, Maguire S, Humbert P, Dai N, Bei Y, Nichols NM, Corrêa IR, Guan S: **Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other**. *J Biol Chem* 2019, **294**:18220–18231.

31. Liu X, Zhang W, Zeng X, Zhang R, Du Y, Hong X, Cao H, Su Z, Wang C, Wu J, et al.: **Systematic Comparative Evaluation of Methods for Investigating the TCRβ Repertoire**. *PLoS One* 2016, **11**:e0152464.

32. Howie B, Sherwood AM, Berkebile AD, Berka J, Emerson RO, Williamson DW, Kirsch I, Vignali M, Rieder MJ, Carlson CS, et al.: **High-throughput pairing of T cell receptor α and β sequences**. *Sci Transl Med* 2015, **7**:301ra131.

● 33. Friedensohn S, Khan TA, Reddy ST: **Advanced Methodologies in High-Throughput Sequencing of Immune Repertoires**. *Trends Biotechnol* 2017, **35**:203–214.

Detailed review presenting the different methodologies of immune repertoires sequencing and highlighting the emergence of single-cell technologies combining TCR sequencing with T-cell transcriptomic profiling.

34. Kim S-M, Bhonsle L, Besgen P, Nickel J, Backes A, Held K, Vollmer S, Dornmair K, Prinz JC: **Analysis of the Paired TCR α- and β-chains of Single Human T Cells**. *PLOS ONE* 2012, **7**:e37338.

35. Han A, Glanville J, Hansmann L, Davis MM: **Linking T-cell receptor sequence to functional phenotype at the single-cell level**. *Nat Biotechnol* 2014, **32**:684–692.

36. Turchaninova MA, Britanova OV, Bolotin DA, Shugay M, Putintseva EV, Staroverov DB, Sharonov G, Shcherbo D, Zvyagin IV, Mamedov IZ, et al.: **Pairing of T-cell receptor chains via emulsion PCR**. *Eur J Immunol* 2013, **43**:2507–2515.

37. McDaniel JR, DeKosky BJ, Tanno H, Ellington AD, Georgiou G: **Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes**. *Nat Protoc* 2016, **11**:429–442.

38. Spindler MJ, Nelson AL, Wagner EK, Oppermans N, Bridgeman JS, Heather JM, Adler AS, Asensio MA, Edgar RC, Lim YW, et al.: **Massively parallel interrogation and mining of natively paired human TCRαβ repertoires**. *Nat Biotechnol* 2020, **38**:609–619.

●● 39. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, Nainys J, Wu K, Kiseliovas V, Setty M, et al.: **Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment**. *Cell* 2018, **174**:1293-1308.e36.

One of the first study using the single-cell technology from 10x Genomics allowing TCR sequencing in parallel to transcriptomic signature of immune cells. They profiled the tumor microenvironment from

breast cancer patients and discovered continuous T-cell differentiation stated and T-cell phenotypes associated with specific TCR sequences.

40. Yost KE, Satpathy AT, Wells DK, Qi Y, Wang C, Kageyama R, McNamara KL, Granja JM, Sarin KY, Brown RA, et al.: **Clonal replacement of tumor-specific T cells following PD-1 blockade**. *Nat Med* 2019, **25**:1251–1259.

●● 41. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB, Nguyen THO, Kedzierska K, et al.: **Quantifiable predictive features define epitope-specific T cell receptor repertoires**. *Nature* 2017, **547**:89–93.

Dash and colleagues defined a novel distance measure on the space of TCRs, TCRdist, allowing for clustering and visualization of repertoire diversity.

●● 42. Ostmeyer J, Christley S, Toby IT, Cowell LG: **Biophysicochemical Motifs in T cell Receptor Sequences Distinguish Repertoires from Tumor Infiltrating Lymphocyte and Adjacent Healthy Tissue**. *Cancer Res* 2019, **79**:1671–1680.

Ostemeyer and colleagues analyzed immune repertoires of patients using biophysicochemical descriptors of the TCR interface in order to cluster and identify disease-associated TCRs.

●● 43. Reuben A, Zhang J, Chiou S-H, Gittelman RM, Li J, Lee W-C, Fujimoto J, Behrens C, Liu X, Wang F, et al.: **Comprehensive T cell repertoire characterization of non-small cell lung cancer**. *Nat Commun* 2020, **11**:1–13.

Study reporting the bulk TCRβ sequencing of peripheral blood, uninvolved tumor-adjacent lung and tumor repertoire of 236 NSCLC patients. This large number of samples allowed investigation into TCR repertoire diversity and overlap between tissue types. Some diversity and convergence metrics were used for this purpose, showing the limitations in their biological interpretation.

44. van de Sandt CE, Clemens EB, Grant EJ, Rowntree LC, Sant S, Halim H, Crowe J, Cheng AC, Kotsimbos TC, Richards M, et al.: **Challenging immunodominance of influenza-specific CD8 + T cell responses restricted by the risk-associated HLA-A*68:01 allomorph**. *Nat Commun* 2019, **10**:5579.

45. Greiff V, Miho E, Menzel U, Reddy ST: **Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires**. *Trends Immunol* 2015, **36**:738–749.

46. Alves Sousa A de P, Johnson KR, Ohayon J, Zhu J, Muraro PA, Jacobson S: **Comprehensive Analysis of TCR-β Repertoire in Patients with Neurological Immune-mediated Disorders**. *Sci Rep* 2019, **9**:1–10.

47. Chang C-M, Hsu Y-W, Wong HS-C, Wei JC-C, Liu X, Liao H-T, Chang W-C: **Characterization of T-Cell Receptor Repertoire in Patients with Rheumatoid Arthritis Receiving Biologic Therapies**. *Dis Markers* 2019, **2019**:2364943.

48. Krummey SM, Morris AB, Jacobs JR, McGuire DJ, Ando S, Tong KP, Zhang W, Robertson J, Guasch SA, Araki K, et al.: **CD45RB Status of CD8+ T Cell Memory Defines T Cell Receptor Affinity and Persistence**. *Cell Rep* 2020, **30**:1282-1291.e5.

49. Schober K, Voit F, Grassmann S, Müller TR, Eggert J, Jarosch S, Weißbrich B, Hoffmann P, Borkner L, Nio E, et al.: **Reverse TCR repertoire evolution toward dominant low-affinity clones during chronic CMV infection**. *Nat Immunol* 2020, **21**:434–441.

50. Cui J-H, Lin K-R, Yuan S-H, Jin Y-B, Chen X-P, Su X-K, Jiang J, Pan Y-M, Mao S-L, Mao X-F, et al.: **TCR Repertoire as a Novel Indicator for Immune Monitoring and Prognosis Assessment of Patients With Cervical Cancer**. *Front Immunol* 2018, **9**:2729.

● 51. Pruessmann W, Rytlewski J, Wilmott J, Mihm MC, Attrill GH, Dyring-Andersen B, Fields P, Zhan Q, Colebatch AJ, Ferguson PM, et al.: **Molecular analysis of primary melanoma T cells identifies patients at risk for metastatic recurrence**. *Nat Cancer* 2020, **1**:197–209.

The authors assessed the clonality of T-cell repertoire as a prognostic marker of recurrence in melanoma patients. Clonality was assessed by T-cell clone frequency and a single l diversity metric, the square-root of the Simpson index. However, repertoire clonality could not differ between progressors and nonprogressors patients.

52. Wu TD, Madireddi S, de Almeida PE, Banchereau R, Chen Y-JJ, Chitre AS, Chiang EY, Iftikhar H, O'Gorman WE, Au-Yeung A, et al.: **Peripheral T cell expansion predicts tumour infiltration and clinical response**. *Nature* 2020, **579**:274–278.

53. Khunger A, Rytlewski JA, Fields P, Yusko EC, Tarhini AA: **The impact of CTLA-4 blockade and interferon-α on clonality of T-cell repertoire in the tumor microenvironment and peripheral blood of metastatic melanoma patients**. *Oncoimmunology* 2019, **8**:e1652538.

54. Sheih A, Voillet V, Hanafi L-A, DeBerg HA, Yajima M, Hawkins R, Gersuk V, Riddell SR, Maloney DG, Wohlfahrt ME, et al.: **Clonal kinetics and single-cell transcriptional profiling of CAR-T cells in patients undergoing CD19 CAR-T immunotherapy**. *Nat Commun* 2020, **11**:1–13.

●● 55. Valpione S, Galvani E, Tweedy J, Mundra PA, Banyard A, Middlehurst P, Barry J, Mills S, Salih Z, Weightman J, et al.: **Immune-awakening revealed by peripheral T cell dynamics after one cycle of immunotherapy**. *Nat Cancer* 2020, **1**:210–221.

The study analyzed the dynamics of peripheral immune repertoire in melanoma patients receiving immunotherapy. Early changes in T-cell clonality and diversity measured by Gini coefficient and Shannon entropy could anticipate patient clinical response.

56. Magurran A: *Measuring Biological Diversity*. Blackwell Publishing; 2004.

57. Hill MO: **Diversity and Evenness: A Unifying Notation and Its Consequences**. *Ecology* 1973, **54**:427–432.

58. Chao A, Jost L: **Estimating diversity and entropy profiles via discovery rates of new species**. *Methods Ecol Evol* 2015, **6**:873–882.

59. Rényi A: **On Measures of Entropy and Information**. In *Proceeding of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. . University of California Press; 1961:547–561.

60. Shannon CE: **A Mathematical Theory of Communication**. *Bell Syst Tech J* 1948, **27**:379–423.

61. Hosoi A, Takeda K, Nagaoka K, Iino T, Matsushita H, Ueha S, Aoki S, Matsushima K, Kubo M, Morikawa T, et al.: **Increased diversity with reduced "diversity evenness" of tumor infiltrating T-cells for the successful cancer immunotherapy**. *Sci Rep* 2018, **8**:1–12.

62. Liu Y-Y, Yang Q-F, Yang J-S, Cao R-B, Liang J-Y, Liu Y-T, Zeng Y-L, Chen S, Xia X-F, Zhang K, et al.: **Characteristics and prognostic significance of profiling the peripheral blood T-cell receptor repertoire in patients with advanced lung cancer**. *Int J Cancer* 2019, **145**:1423–1431.

63. Simpson EH: **Measurement of Diversity**. *Nature* 1949, **163**:688–688.

64. Farmanbar A, Kneller R, Firouzi S: **RNA sequencing identifies clonal structure of T-cell repertoires in patients with adult T-cell leukemia/lymphoma**. *NPJ Genom Med* 2019, **4**:1–9.

65. Jost L: **Partitioning Diversity into Independent Alpha and Beta Components**. *Ecology* 2007, **88**:2427–2439.

● 66. Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST: **A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status**. *Genome Med* 2015, **7**:49.

This study introduces a bioinformatics framework presenting the importance of characterizing immune repertoires clonality with diversity profiles composed of simultaneous analysis of different entropy-based diversity metrics to predict patient immunological status.

67. Postow MA, Manuel M, Wong P, Yuan J, Dong Z, Liu C, Perez S, Tanneau I, Noel M, Courtier A, et al.: **Peripheral T cell receptor diversity is associated with clinical outcomes following ipilimumab treatment in metastatic melanoma**. *J Immunother Cancer* 2015, **3**:23.

68. Pielou EC: **The measurement of diversity in different types of biological collections**. *J Theor Biol* 1966, **13**:131–144.

69. Rousseau R, Van Hecke P, Nijseen D, Bogaert J: **The relationship between diversity profiles, evenness and species richness based on partial ordering**. *Environ Ecol Stat* 1999, **6**:211–223.

70. Leydesdorff L: **Diversity and interdisciplinarity: how can one distinguish and recombine disparity, variety, and balance?** *Scientometrics* 2018, **116**:2113–2121.

71. Jaccard P: **The Distribution of the Flora in the Alpine Zone**. *New Phytologist* 1912, **11**:37–50.

72. Dice LR: **Measures of the Amount of Ecologic Association Between Species**. *Ecology* 1945, **26**:297–302.

73. Sorensen T: **A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons**. *Kongelige Danske Videnskabernes Selskab* 1948, **5**:1–34.

74. Bray JR, Curtis JT: **An Ordination of the Upland Forest Communities of Southern Wisconsin**. *Ecol Monogr* 1957, **27**:325–349.

75. Horn HS: **Measurement of "Overlap" in Comparative Ecological Studies**. *Am Nat* 1966, **100**:419–424.

76. 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.

77. Rempala GA, Seweryn M, Ignatowicz L: **Model for comparative analysis of antigen receptor repertoires**. *J Theor Biol* 2011, **269**:1–15.

78. Chao A, Jost L: **Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size**. *Ecology* 2012, **93**:2533–2547.

79. Koch H, Starenki D, Cooper SJ, Myers RM, Li Q: **powerTCR: A model-based approach to comparative analysis of the clone size distribution of the T cell receptor repertoire**. *PLoS Comput Biol* 2018, **14**:e1006571.