T-cell Receptor Diversity Estimates for Repertoires (TCRDivER) uses sequence similarity to find signatures of immune response

Milena Vujović[1], **Paolo Marcatili**[1], **Benny Chain**[3], **Joseph Kaplinsky**[1,2], **Thomas Lars Andresen**[1]

***For correspondence:**
milvu@dtu.dk (MV);
joseph.kaplinsky@ludwig.ox.ac.uk (JK)

[1]DTU HealthTech, Department of Health Technology, Technical University of Denmark, Denmark; [2]Ludwig Institute for Cancer Research Ltd, University of Oxford, Nuffield Department of Medicine, United Kingdom; [3]UCL Division of Infection and Immunity, University College London, United Kingdom

Abstract

We propose TCRDivER, a global approach to T-cell repertoire comparison using diversity profiles sensitive to both clone size and sequence similarity. As immunotherapies improve, the long standing biological interest in connecting outcome with T cell receptor (TCR) repertoire status has become more urgent. Here we show that new insights can be extracted from high throughput repertoire sequencing data. Most current efforts focus on identification of immunisation-specific sequence motifs or on monitoring changes in frequency of individual clones. Applying TCRDivER to murine spleen samples shows it characterises an additional dimension of repertoire variation, beyond conventional diversity estimates, allowing distinction between immunised and non-immunised samples. We further apply TCRDivER to repertoires from human blood. In both cases we show characteristic relationships between repertoire features. These reveal biologically interpretable relationships between sequence similarity and clonal expansions. We thereby demonstrate a new tool for investigation in clinical and research applications.

Introduction

The T cell compartment of adaptive immunity plays a crucial role in cancer immunity, auto-immune and infectious diseases. Adaptive immune responses as a whole draw on diverse T-cell receptors. Due to the phenomenon of epitope spreading, T cells diversify their antigen-specific response by reacting to non-dominant epitopes present on the antigen, in addition to the main dominant epitope driven response ***Didona and Di Zenzo*** (***2018***); ***Vanderlugt and Miller*** (***2002***). On the other hand, it has been shown that T cells responding to the same epitope share more sequence similarity ***Dash et al.*** (***2017***). In addition, TCRs often exhibit cross-reactivity in order to ensure broad epitope recognition responses, despite the limited number of unique TCRs within each repertoire ***Petrova et al.*** (***2012***); ***Antunes et al.*** (***2017***); ***Bentzen and Hadrup*** (***2019***).

T cells are generated through the imprecise stochastic process of V(D)J recombination giving rise to 10^{20} or more possible TCR combinations ***Miles et al.*** (***2011***); ***Mora and Walczak*** (***2018***). The number of possible generated TCRs is much larger than those estimated to be present within any individual T cell repertoire ***Laydon et al.*** (***2015***); ***Robins et al.*** (***2009***, 2010). The complexity and com-

position of TCR repertoires makes it difficult to compare and stratify individuals based on immune status or to even establish a healthy baseline. T cells activate and proliferate upon antigen-specific contact, creating a complex mix of receptors. Initial experimental approaches, such as spectratyping ***Choi et al.*** (***1989***); ***Gorski et al.*** (***1994***); ***Memon et al.*** (***2012***); ***Ochsenreither et al.*** (***2008***) and flow cytometry ***Ciupe et al.*** (***2013***); ***Muraro et al.*** (***2000***) aimed to reveal oligoclonal expansions of T cells by tracking clonal sizes of CDR3s with the same length. However, they provided no insight into TCR similarity. Recent advances in high throughput sequencing (HTS) now allows characterisation of adaptive immune receptors in increasing depth and with improved quantitation. HTS methods supply information on both clonal sizes and sequence relatedness. However, this development has given rise to the need for summary measures to interpret the data generated by such experiments. Several methods have emerged to fulfil the demand to stratify repertoires either by the TCR antigen specificities ***Glanville et al.*** (***2017***); ***Dash et al.*** (***2017***); ***Sidhom et al.*** (***2018***) or by finding characteristics of TCR sequences ***Thomas et al.*** (***2014a***); ***Sun et al.*** (***2017***); ***Cinelli et al.*** (***2017***). Still, many of the employed methods aim to uncover epitope similarity without simultaneously examining T cell clonal expansions, or vice versa. Measures that capture the global repertoire structure by incorporating both characteristics of the adaptive immune response, could potentially be used to stratify patients for disease outcome or therapy. Thereby, transcending the notion of "public TCRs" into "public repertoire structures" responsible for therapeutic outcome.

A popular approach in characterisation of repertoires has been through measures of diversity. They have been widely used in evaluation of therapy and disease effects ***Twyman-Saint Victor et al.*** (***2015***); ***Rudqvist et al.*** (***2018***); ***Sherwood et al.*** (***2013***); ***Robert et al.*** (***2014***); ***Warren et al.*** (***2011***) or attempts at repertoire classification and diagnosis ***Carey et al.*** (***2016***); ***Chang et al.*** (***2019***); ***Robins et al.*** (***2009***). However, there are a variety of ways in which the intuitive idea of diversity can be formalised giving rise to ambiguity. In the naive sense, diversity is estimated based on the number of and clonal expansion of unique TCRs in a repertoire. Commonly used diversity estimates are richness (number of different TCR clones), clonality (number of expanded clones) and diversity indices such as Shannon entropy ***Spellerberg and Fedor*** (***2003***), Simpson ***SIMPSON*** (***1949***), Gini-Simpson ***Jost*** (***2006***) and Berger-Parker ***Berger and Parker*** (***1970***) index. Different diversity indices will weight expanded clones differently, thereby imposing a threshold on the clonal frequencies within the repertoire. Thus counting unique clones with species richness will give rare clones the same weight as expanded clones. Entropy, will give give more weight to expanded clones than to rare clones. No single index will capture all information about the clone size distribution. Notably, this ambiguity has led to no clear consensus which diversity index should be applied in practical cases of interpreting immune diversity ***Izraelson et al.*** (***2018***); ***Chiffelle et al.*** (***2020***). The approach of using individual diversity indices provides no repertoire characteristics truly independent of sample size ***Laydon et al.*** (***2015***) and can lead to erroneous conclusions on ordering repertoires (Figure 1 A.). Additionally, measures of diversity should not only rely on clone counts but should also account for sequence similarity of receptors.

The first problem of debatable usage of individual indices, can be surmounted by estimating them simultaneously in a single expression of diversity: the diversity of order q, $D(q)$, which subsumes most of the commonly used indices ***Jost*** (***2006***, 2010). Accounting for the distribution of clone sizes, diversity can be estimated in the form of "diversity profiles" ***Greiff et al.*** (***2015***); ***Mora and Walczak*** (***2016***); ***Chiffelle et al.*** (***2020***). Such profiles define "effective numbers" of receptors when viewed at different resolutions, making use of a single parameter (q) to systematically shift focus from counting each unique clone to giving weight only to the largest clone in a repertoire (Figure 1 A, C). The use of diversity profiles gives insight into T cell clonal expansions, as the relationship between diversities calculated at different clonality weights q can be correlated to the ratio of common to rare clones ***Leinster and Cobbold*** (***2012***). This approach has been previously implemented for one B- and three T-cell repertoires in the work of ***Greiff et al.*** (***2015***). Keeping in mind that the study focused solely on clonal frequency, the authors report remarkable separation based on immunological status, in 3 out of 4 immune repertoire datasets. However, as naive diver-

sity estimates only take clone frequencies into account they are not sensitive to minor polyclonal expansions of TCRs reacting to the same antigen which are mounting a unified front of reacting similar T cells.

The second problem, incorporating sequence similarity in diversity estimates, has been less thoroughly explored. One approach is to count clusters of similar receptors ***Sidhom et al.*** (***2018***). Another approach is to use an effective number with sensitivity to sequence similarity ***Arora et al.*** (***2018***). These approaches suffer from a similar limitation as use of a single diversity index in that they adopt either a single arbitrary cutoff or a single sequence similarity distance in their definition of effective number e.g. a single similarity corrected diversity index. Here we make use of approach used by ***Leinster and Cobbold*** (***2012***) in ecology to explore 2 dimensional profiles of effective numbers. Our approach of using similarity scaled diversity $D(q, \lambda)$ allows for simultaneous characterisation of the clonal distributions and similarity of receptor repertoires (Figure 1 B). Instead of depending on a single parameter, q, our profiles depend on two parameters, q and λ. As in conventional diversity profiles, the q parameter probes the structure of the clone size distribution. The λ parameter plays an analogous role for sequence similarity (Figure 1 D.). As λ varies from infinity down to zero the effective diversity gradually merges together more and more similar sequences. Incorporating this additional aspect to the diversity estimation allows us not only to probe the clone size distribution, but also TCR similarity which may provide information on repertoire convergence through expansion of similar clones.

In this study, we showcase a new tool for estimating TCR repertoire diversity using similarity sensitive diversity estimates: TCRDivER. We apply TCRDivER to previously published murine TCRβ sequence data from CD4$^+$ T cells following immunization ***Sun et al.*** (***2017***). We show that TCRDivER, by simultaneously probing clonal expansion and sequence similarities reveals novel TCR repertoire traits. Using features of the similarity scaled diversity profile we detect differences in response to immunisation protocols at all sampling times, indicating unique features arise within repertoires can be detected as early as 5 days and persist for several months.

Notably we find strong nonlinear correlations between features of the similarity scaled diversity profiles, including feature which characterise the average distance between sequences or the balance between large and small clones. The strength of the correlation indicates biological constraints on repertoire development which couple together clone size with sequence similarity. The nonlinear shape of the correlations reveal that these features can exist in multiple discrete equilibria.

We validate our finding that TCR repertoires reside in a non-linear space on an independent dataset of human bulk TCR sequences extracted from non-small-cell lung cancer (NSCLC) patients following CTLA-4 blockade treatment ***Formenti et al.*** (***2018***). We show that by estimating repertoire diversity with TCRDivER we can unearth information which might allow us to understand more subtle differences between repertoires, stratify them and ultimately guide therapy regimes.

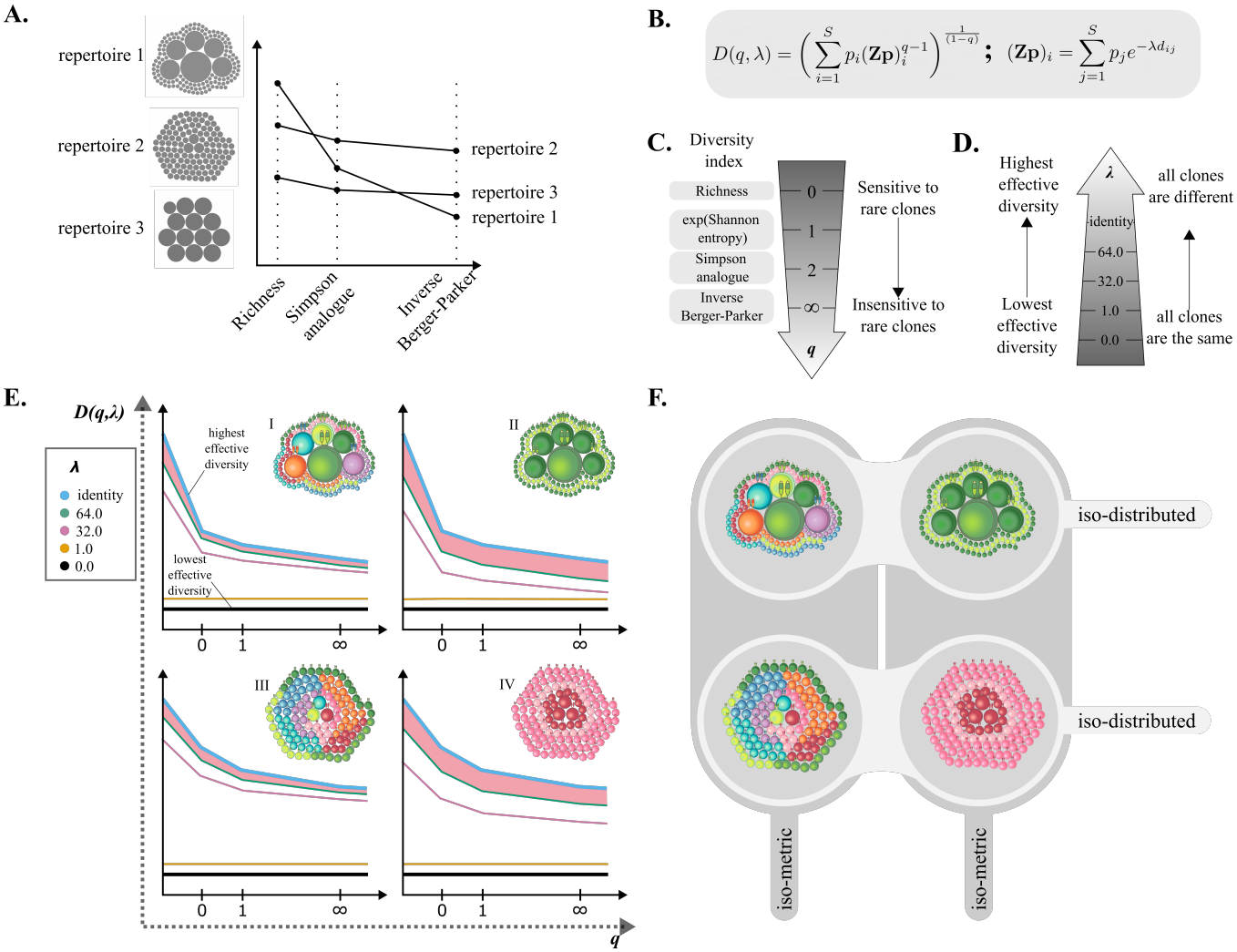


Figure 1. A. Naive diversity indices (richness, Simpson analogue and Inverse Berger-Parker) for three model repertoires. The repertoires contain the same total number of T cells, however they differ in the number of unique T cell clones and clone size distribution. Each T cell clone within the repertoire is presented as a grey circle and the size of each circle corresponds to its relative frequency. Repertoire 1 has the most uneven clonal distribution. Repertoires 2 and 3 have a more uniform distribution of clones. The number of unique clones falls from repertoire 1 through to repertoire 3. As different diversity indices are applied, the ordering of repertoires changes. **B.** Formula for calculating similarity scaled diversity $D(q, \lambda)$. Here p_i is the fraction of cells in clone i, **Z** is a similarity kernel between clones, q controls sensitivity to clone size, and λ controls sensitivity to sequence similarity. **C.** Relationship of some commonly used diversity indices to the naive diversity of order q. **D.** Effect of introducing the λ distance scaling into the diversity calculation. As λ increases the distance between clones increases until at $\lambda = \infty$ clones have no similarity i.e. the similarity kernel **Z** is the identity. **E.** Schematic representation of diversity profiles of four model repertoires are shown along the repertoires in the top right corner of each diversity profile. Repertoire composition is schematically represented with T cells of varying size and colour. Analogously to A. the size of individual cells corresponds to the clone frequency within the repertoire. More similar colouring indicates higher T cell sequence similarity in the repertoire (repertoires II and IV). Analogously, divergent colouring corresponds to higher sequence dissimilarity (repertoires I and III). **F.** Further explanation of structure differences between the four model repertoires in E. Repertoires sharing the same clone distributions are shown in rows (I = II and III = IV). Repertoires sharing same similarity relationships between T cell clones are shown in columns (I = III and II = IV).

Results

We analysed the frequency distribution and similarity of $CDR_\beta 3$ amino acid sequences in TCR repertoires previously published in ***Sun et al.*** (***2017***). Briefly, the dataset consists of $CD4^+$ T-cell repertoires harvested from murine spleens following immunisation with Complete Freund's Adjuvant (CFA) with or without the addition of Ovalbumin (OVA) antigen. The T cells were harvested post immunisation at three timepoints: early (days 5 and 14) and late (day 60). Additionally, we have analysed untreated mouse repertoires from the same study. We used TCRDivER to calculate diversities $D(q, \lambda)$, with varying orders of q and λ. From these we constructed diversity profiles (divPs), which we present as graphs of the natural logarithm of diversity versus the varying order of q for each lambda. Key features were extracted for analysis as shown bellow.

To validate some of our findings we analysed a human TCR repertoire data set previously published in ***Formenti et al.*** (***2018***). In short, T cells were isolated from blood samples taken from non-small-cell lung cancer patients prior and post treatment with CTLA-4 blockade (ipilimumab) in combination with radiation therapy (RT). The obtained T cells were sequenced in bulk. We analysed these repertoires as with the murine data, using TCRDivER to construct diversity profiles for analysis.

TCRDivER reveals unique TCR repertoire features

Examples of diversity profiles constructed for the murine samples are shown in figure 2. Each dataset was sampled for 50,000 sequences in order to eliminate effects of sequencing depth. For each sample a series of curves are plotted, corresponding to different values of λ. The constructed diversity profiles provide a graphically intuitive way to caputre the shape of a repertoire. Here we highlight some features of these plots in order to develop an understanding of how features of the plots map to structural and immunological characteristics of TCR repertoires.

In our framework the naive diversity profile corresponds to the case that the receptor of each T cell clone is considered totally distinct, with no consideration of similarity to other clones i.e. the highest effective diversity. In reality there will be some degree of functional overlap between clones, which will reduce the functional diversity below the naive value. The naive diversity ($\lambda = \infty$) is therefore a base case of maximal diversity. At the opposite extreme, $\lambda = 0$, all clones are considered considered functionally identical. Biologically, this would correspond to non-specific binding of TCRs to peptide-MHC complexes. In this case the functional diversity is therefore minimal and equals one. In each repertoire sample these two extreme cases can be seen bounding the profile (these curves are labelled in example figure 1 E.) The parameter λ interpolates between these two extreme cases, as the intermediate profiles in each sample correspond to intermediate values of λ.

We begin our account by highlighting features of the naive diversity, i.e. the upper bounding curves, in each sample. We plotted naive diversity profiles from all samples together showing that crossings in the range $0 \leqslant q \leqslant 2$ are common events (See *Supplementary information - Section 1.2 and 1.6*). This confirmed in our data set that the ranking of repertoires based on a single value of q would indeed depend strongly on the chosen index (similar to what is shown in example figure 1 A.). We concluded that the previously mentioned justification for analysing profiles across a range of orders q is not merely theoretical.

The highest value of the naive diversity at $q = 0$ gives the number of unique TCR sequences observed in the sample of 50,000 sequences. At $q = \infty$ we read off the effective number of clones in the repertoire if it consisted only of the largest clones. The rate of fall of naive diversity as q rises therefore encodes information about the balance between larger and smaller clones. To characterise this we derived an expression for the gradient of the naive diversity at $q = 1$ and found that it is proportional to the variance of the clone size distribution i.e. the ratio of rare to common T cell clones in a repertoire (see Appendix 1 Evaluating the slope at $q = 1$).

Notably, when examining the diversity profiles in Figure 2 and *Supplementary information - Section 1.1*, a sharper slope can be seen in the curves from repertoires that have been immunised

compared to the untreated ones, especially for later time points. This is quantified in figure 3 A by plotting the slope between $q = 0$ and $q = 1$ for each treatment group. The increasing value of the slope is indicative of an increased clonal expansion at later timepoints. The impact of clonal expansion in reducing diversity is seen to be mild at earlier time points after vaccination (5 and 14 days) and more marked at the day 60 time point.

In order to explore the effects of similarity scaling on measures of diversity, we investigate in depth the features of similarity sensitive profiles with values of $\lambda < \infty$. A first feature of the similarity sensitive profiles occurs near the flat curve for $\lambda = 0$ at the bottom of the profile. We observed that for small values of λ the curves are approximately flat and are evenly spaced on a log scale. This phenomenon can be seen in figure 2 by observing that the spacing between the curves for $\lambda = 0$ (blue) and $\lambda = 1$ (orange) is equal to the spacing between the curves $\lambda = 1$ (orange) and $\lambda = 2$ (red). To further investigate the biological significance of this we derived an expression for the diversity at small λ using perturbation theory (See Appendix 2 Evaluation $\Delta ln(D(q, \lambda))$ for small λ: Perturbation around $\lambda = 0$). Interestingly, this shows that the spacing is proportional to the mean distance between sequences in the repertoire. In particular, the spacing of profiles, which we denote $\Delta ln(D(q, \lambda))$ at small λ, is dependent solely on the distance and frequency of CDR3s, and not on the weight q. Notably, this measure naturally integrates increases in similarity from both expansion of particular clones and selection of clones with similar sequence. We are therefore able to use these spacings to gain biological insight in to repertoire structure following immunisation. The spacing of profiles, $\Delta ln(D(q, \lambda))$ at small λ, is presented for different treatment groups in figure 3 A. We concluded that while there may be a small rise in spacing at early time points after vaccination (5 and 14 days), there is a distinct decline of around 15% at day 60 indicating an increase of CDR3 similarity at later time points.

A second feature is the rate at which diversity falls as λ falls from $\lambda = \infty$ to lower values, where $\lambda = \infty$ characterises the naive profile. Unlike the case at small λ, the value of $\Delta ln(D(q, \lambda))$ at large λ is no longer independent of q. Remembering that $\lambda = \infty$ corresponds to no effective clustering of similar sequences, large values of λ correspond to just a small amount of effective clustering counting together only the most similar clones. If such clustering produces a large fall in effective diversity then the repertoire must contain many similar clones. Conversely, if such clustering produces only a small fall in diversity then the clones must be spaced further apart.

Our measure is highlighted by the pink area in figure 2, defined as lying between the naive profile $\lambda = \infty$ and the profile for $\lambda = 16$. By using the area highlighted in the figure as the feature of interest we are effectively averaging over q. Notably, like $\Delta ln(D(q, \lambda))$ at small λ, the area between $\lambda = \infty$ and $\lambda = 16$ is a probe of distance. However, in this case the weighting is toward similar clones. i.e. larger spacings correspond to more similarity of sequences in the repertoire. We have shown that, in the case of a uniformly distributed CDR3s in a repertoire, with the increase of similarity between CDR3s the area between the λ curves increases (see Appendix 2 Evaluation $\Delta ln(D(q, \lambda))$ for larger λs and it's relationship to distance). In the case of natural repertoires the effect of clonal expansion will interplay with the similarity. The area between profiles for $\lambda = \infty$ and $\lambda = 16$, is presented for different treatment groups in figure 3 A. We concluded that there is a tendency to fall at the early time points after vaccination (5 and 14 days), with a further fall of comparable magnitude by day 60. As the area is influenced by q it is also closely connected to the slope of diversity curves and therefore clonal expansion. Therefore the fall of value of area cannot be attributed to a decrease in similarity at later time points. When other repertoire features are taken into account, such as the slope and the trend of $\Delta ln(D(q, \lambda))$ at small λ, the decrease in area can be explained by the driving effect of clonal expansions at later time points.

Comparing the $\Delta ln(D(q, \lambda))$ at small and large values of λ we were able to make some biological conclusions about the structure of the repertoires. At early time points there is a reduction in diversity of atypically similar (~ 1 amino acid difference) sequences. This may correspond to the expansion of responding clones with distinct sequences at the expense of background diversity. At early time points these expansions have little impact on the mean diversity, but by day 60 they

have reduced mean diversity across the whole repertoire.

TCRDivER features improve separation of biologically distinct repertoires

In order to test if the similarity scaled diversity profiles can be used to classify repertoires we used principal component analysis (PCA). We carried out this analysis first on the values for the naive diversity profiles alone and then for the all values in the complete similarity scaled diversity profile (Figure 3 B. and *Supplementary information - Section 1.3*.).

Both the naive and similarity scaled profiles show a strong PC1 which is driven by the expansion of large clones at the late day 60 time point. However, relative to the naive profile, PCA on the similarity scaled profiles shows more than twice the variance in PC2. In contrast to the naive profiles that give an effective 1 dimensional separation, the similarity scaled profiles are able to give a robust 2 dimensional separation. It can be seen that this allows for substantial separation of immunised vs. untreated controls in the second dimension (Figure 3 C.).

To validate the observation of improved PCA separation, we analysed an additional data set of human TCR repertoires. Briefly, this arose from blood samples collected prior and post immunotherapy from 40 patients diagnosed with stage 4 non-small-cell lung carcinoma (NSCLC). The therapy consisted of a regime of radiation and administering CTLA-4 blockade. After therapy completion each patient was categorised according to RECIST response criteria into four categories based on the therapy outcome (for further details see Section Data Acquisition and Description). We analysed the data as before by calculating the diversity $D(q, \lambda)$ and constructing diversity profiles (see *Supplementary information - Section 2.1*.).

To test if the features we have identified are useful in capturing key dimensions of variation in the similarity scaled diversity profiles we then extracted these features and carried out PCA on the features rather the raw values. These include all the areas between different λ curves, average values of $\Delta \ln(D(q, \lambda))$ for small λs and slopes for $q = 0 \to 1$, $q = 1 \to 2$ and $q = 0 \to 2$. To further mitigate the effect of repertoire size on the analysis we have log-transformed the values of $D(q, \lambda)$ prior to feature extraction. The analysed features are therefore ratios and relationships of the natural logarithm of $D(q, \lambda)$. The results of the PCA analysis on the human can be seen in *Supplementary information - Section 2.2*.

In both the mouse and human data sets we found that the PCA on features was qualitatively similar to that on the full diversity profile. However, the variance explained by PC1 was reduced while that explained by PC2 was increased, leading to improved 2 dimensional separations. In the case of the mouse data the effect was modest, while in the human data it was more substantial, leading to an increase in variance explained by PC2 from 11% to 20%.

The reduction of variance explained in PC1 indicates that these features are indeed acting as useful summaries of redundant (linearly correlated) information in the full profile. Therefore making use of such features, rather than the raw profile values may help reduce experimental noise and improve robustness.

Non-linear relationships between TCRDivER features are driven by the structure of repertoires

Because a significant proportion of the variance is not captured by PC1 alone, there must be some non-linear relationships between similarity sensitive diversity profile features. We therefore decided to look more closely at pairwise relationships between these features, as shown in figure 3 D.

This revealed two interesting characteristics of the relationships. Firstly, while many of the plots are clearly non-linear, they lie on surprisingly tight curves. This indicates that there is some constraint at work in the structure of the repertoire meaning that the value of one feature tightly constrains the value of other features. However, the second interesting characteristic is that these constraints are not unique.

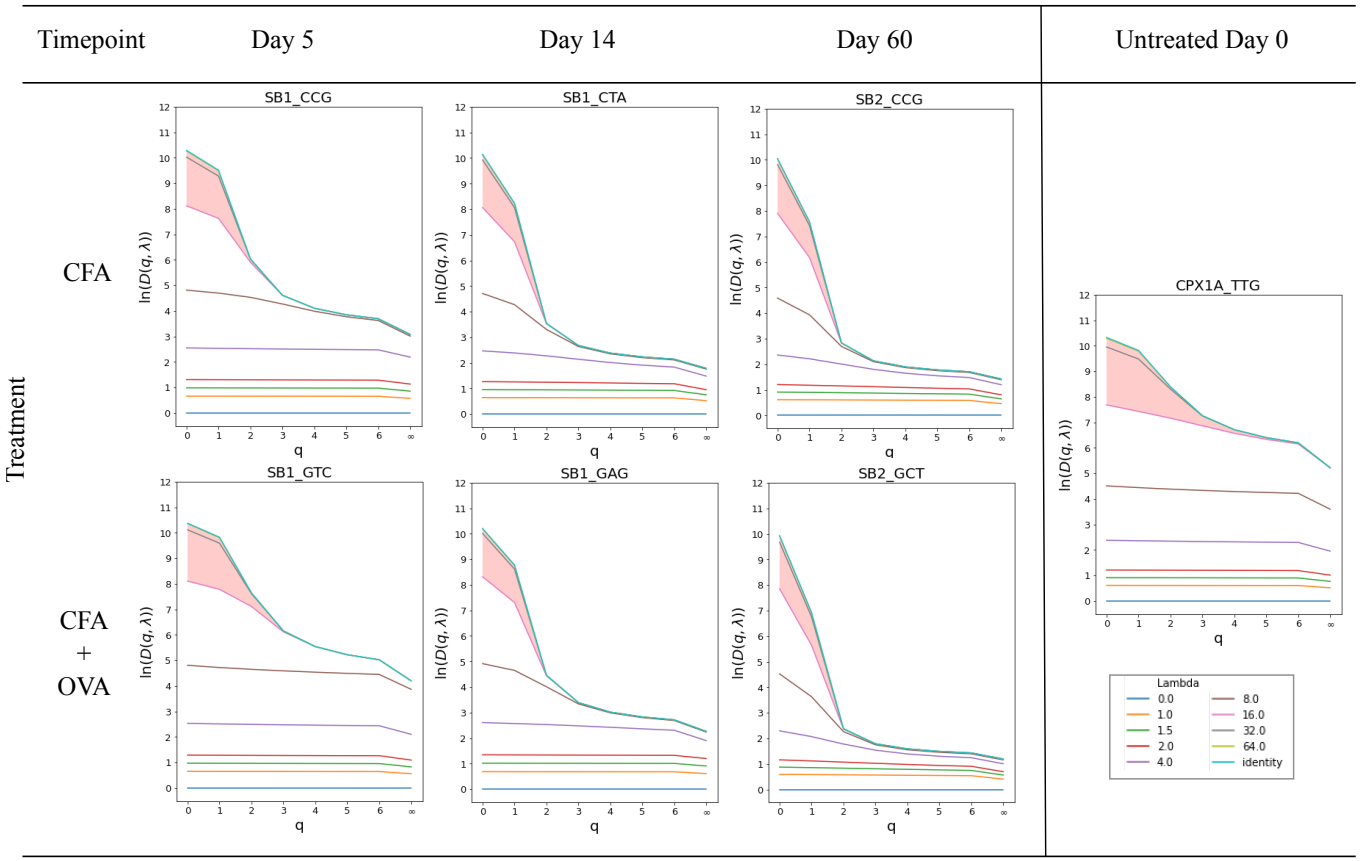


Figure 2. Diversity profiles (divPs) of calculated from CDR3 frequency within each repertoire and their similarity. DivPs of repertoires stemming from immunisation have been shown to the left, while the untreated is show on the right. Natural logarithm transformed values of diversity ${}^{q}D$ for each calculated $\lambda = 0.0, 1.0, 1.5, 2.0, 4.0, 8.0, 16.0, 32.0, 64.0$ and identity versus the increasing order of q. The legend for all diversity profiles is shown at the bottom right. The highlighted area represents the area between λ 16.0 and identity curves. It highlights the change in repertoire CDR3 similarity unification for repertoires of different origin. Diversity profiles of only one sample per group are shown, the rest can be found in *Supplementary information - Section 1.1*.

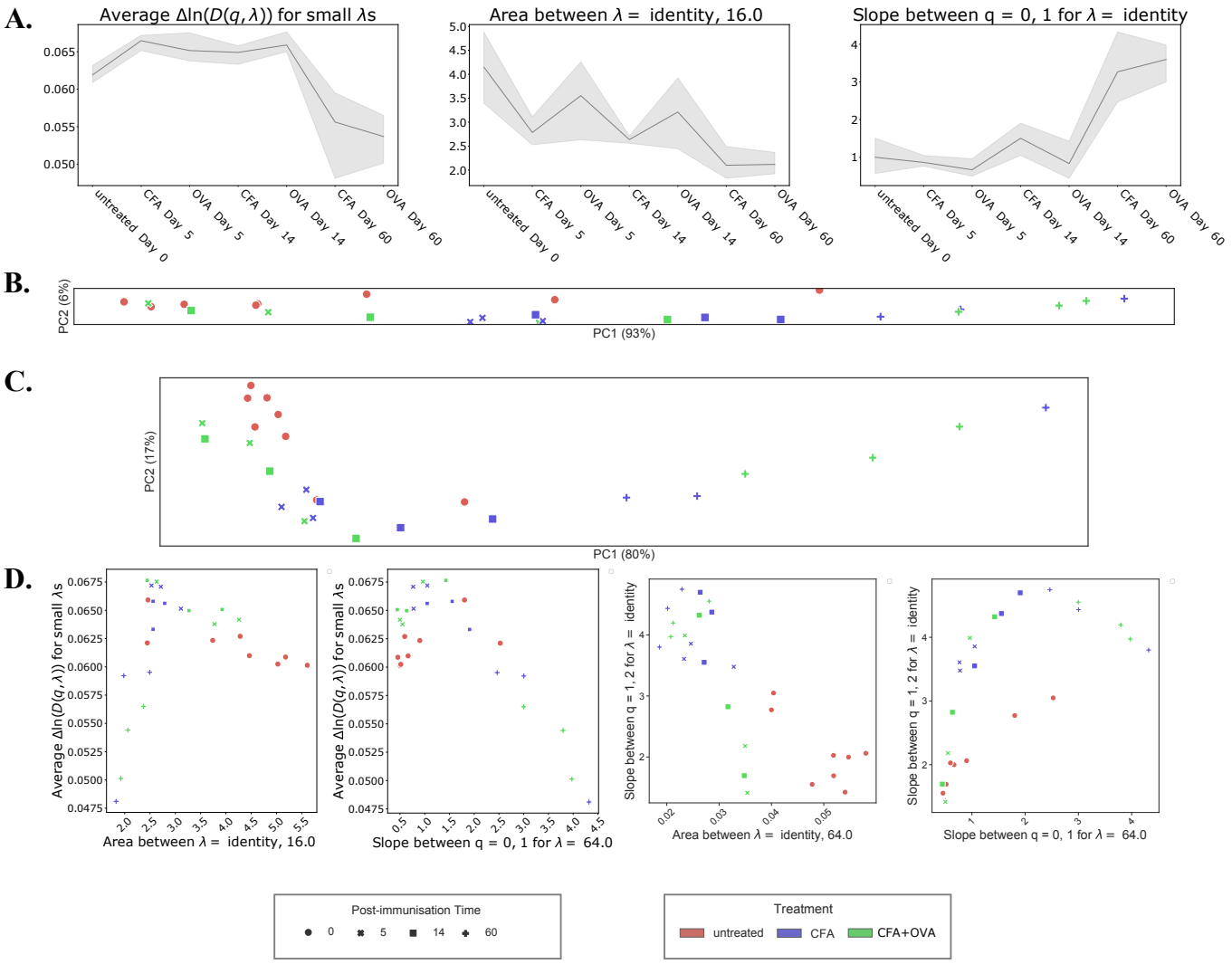


Figure 3. A. Trends of three features extracted from divPs are shown versus the treatment regime and timepoints ending with the latest timepoint. The features are, from left to right: average $\Delta \ln(D(q,\lambda))$ for small λs, between curves of $\lambda =$ identity and 16.0 and slope of $q = 0 \rightarrow 1$ for value of λ identity. The line connects the mean values of the features for all samples within a group and the shaded area represents the confidence interval. **B.** PCA on naive diversity values ${}^{q}D$, i.e. $\lambda =$ identity. The PCA plot aspect ratio has been adjusted and corresponds to variation explained by the first two principal components. **C.** PCA on features extracted from the diversity profiles constructed on the diversity values ${}^{q}D^{Z}$. These include areas between all lambda curves, average $\Delta \ln {}^{q}D^{Z}$ for small λs and slopes $q = 0 \rightarrow 1$, $q = 0 \rightarrow 2$ and $q = 1 \rightarrow 2$. As in C., the aspect ratio corresponds to variation found by PCA. **D.** Graphs showing relationships between some of the divP features. From left to right: average $\Delta \ln(D(q,\lambda))$ for small λs is shown versus the area between curves of $\lambda =$ identity and 16.0; average $\Delta \ln(D(q,\lambda))$ for small λs is shown versus the slope of $q = 0 \rightarrow 1$ for value of λ 64.0; slope of $q = 1 \rightarrow 2$ for value of λ identity (i.e. naive diversity) is shown versus the area between curves of $\lambda =$ identity and 64.0; slope of $q = 1 \rightarrow 2$ for value of λ identity (i.e. naive diversity) is shown versus the slope of $q = 0 \rightarrow 1$ for value of λ 64.0. The legend corresponds to figures B, C and D.

An example is the relationship between $\Delta ln(D(q,\lambda))$ at small λ and the area between profiles for $\lambda = \infty$ and $\lambda = 16$, as shown in the first panel of figure 3 D. As noted above, these both depend on the distances between sequences in the repertoire. Rising $\Delta ln(D(q,\lambda))$ at small λ indicates mean distances between sequences in the repertoire increasing. Falling values of the area indicate fewer clones at very close distances to another clone. In a simple ideal case where the clone size and distance distributions are uniform these two effects perfectly coincide. We illustrate this using model data in Appendix 2 Evaluation $\Delta ln(D(q,\lambda))$ for larger λs and it's relationship to distance. In real data with non-uniform clone size and distance distributions $D(q,\lambda)$ provides a measure where the differential effect of changes in the most similar sequences and changes in comparisons across the repertoire as a whole can be characterised.

For data from the mouse immunisation experiments, figure 3 D illustrates a non-linear change point in the relationship between $\Delta ln(D(q,\lambda))$ at small λ and area the between profiles for $\lambda = \infty$ and $\lambda = 16$. Above a threshold in area for larger λs of around 2.5 we see the expected behaviour for the simple ideal case. Below this threshold we see the more complex phenomenon where a smaller mean distance between sequences goes with fewer small inter-clone distances. This is most developed in the samples from late time points where expanded clones are present. Within these clones the distances between sequences will be zero, pushing down the mean distance between sequences. At the same time the presence of large expanded clones means that these clones are less likely to have close neighbours.

The example of the relationship between $\Delta ln(D(q,\lambda))$ at small λ and the area between profiles for $\lambda = \infty$ and $\lambda = 16$ illustrates a non-unique constraint. A value of 0.06 for $\Delta ln(D(q,\lambda))$ at small λ constrains the value of the area between profiles for $\lambda = \infty$ and $\lambda = 16$ but to one of two possible values - either around 2.0 or around 5.5. This again emphasises the importance of taking multiple features of the diversity profile to more fully characterise repertoire structure. Similar non-unique constraints can be seen in the other panels of figure 3 D.

The mathematical form of the diversity we have adopted does impose some restrictions on possible diversity profiles. For example, the effective diversity must always fall (or stay constant) as q rises, reflecting the down weighting of small clones. To test if these relationships might be an artefact imposed by the mathematical form of the diversity we replaced the clone size distribution with pseudo-random numbers while keeping the distance matrix fixed, with results shown in figure 4. This eliminated observed correlations, showing that the correlations are not mathematically necessary. To confirm that the correlations are not a product of the particular distance definition adopted we repeated the analysis using an alternative metric based on amino acid properties. This shows qualitatively similar correlations (See *Supplementary information - Section 1.9*).

Turning to the human data set we found that pairwise relationships between features were quantitatively quite different that the mouse data, but displayed the same characteristics of lying on curves and giving rise to non-unique constraints (Figure 5 A.). Given differences in species, tissue and treatment, it is unsurprising that the range of repertoire structures observed differs considerably. At least some of these differences are captured in features of the similarity scaled diversity profiles. Despite these differences, the human data corroborates the notion that regardless of the immunisation strategy and dataset (human or murine), the natural TCR repertoires reside in a subspace governed by a complex interplay of TCR clonality and similarity.

Discussion

The complex structure of immune repertoires makes them challenging to compare and classify. Previous work making use of sequence information to understand TCR repertoires has focused on determining the antigen specificity of particular sequences. In contrast, TCRDivER is able to make use of sequence information to reveal structural similarity between repertoires that have little or no sequence overlap. With 2 tunable parameters TCRDivER becomes an effective computational microscope, able to focus on different scales of structure in the immune repertoire. The resulting diversity profiles then provide highly interpretable summaries of global multiscale structure.

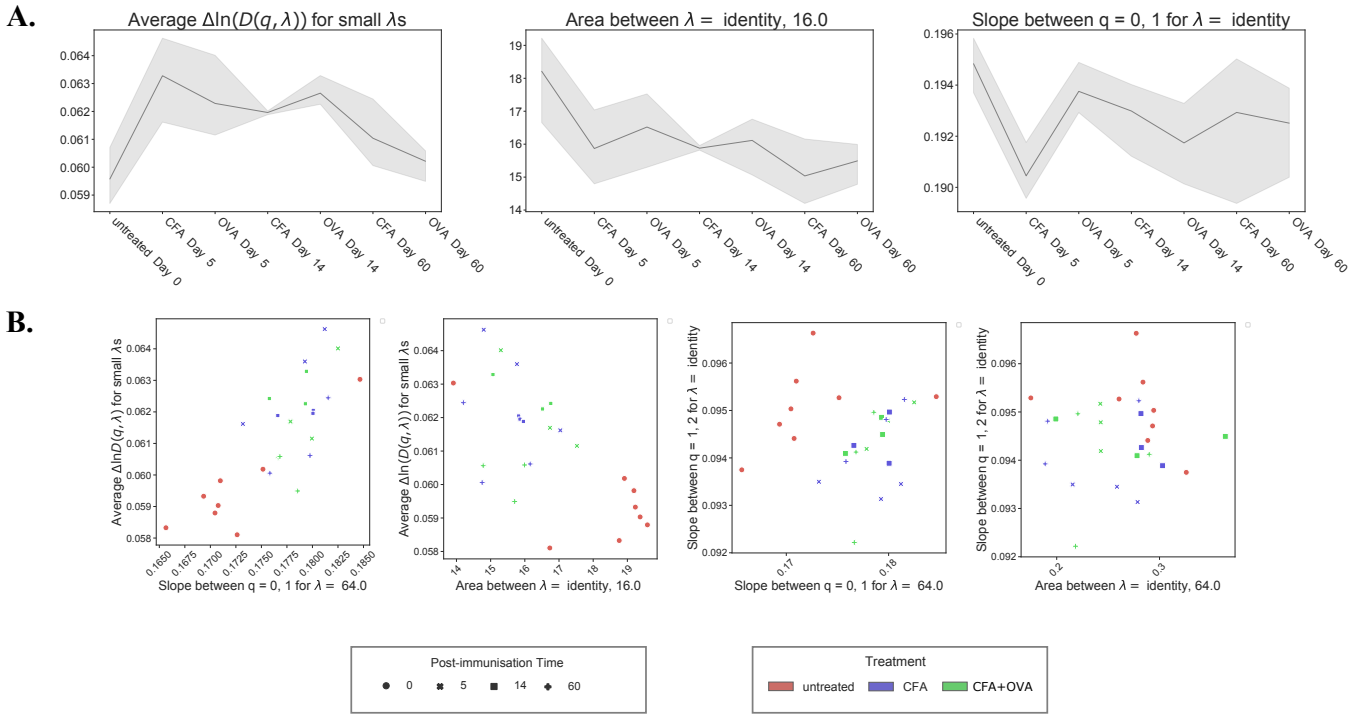


Figure 4. A. Trends of three features for the randomised murine dataset are shown versus the treatment regime and timepoints ending with the latest timepoint. The features are, from left to right: average $\Delta \ln(D(q,\lambda))$ for small λs, between curves of λ = identity and 16.0 and slope of $q = 0 \rightarrow 1$ for value of λ identity. The line connects the mean values of the features for all samples within a group and the shaded area represents the confidence interval. **B.** Graphs showing relationships between some of the divP features of the murine dataset with random frequencies. From left to right: average $\Delta \ln(D(q,\lambda))$ for small λs is shown versus the slope of $q = 0 \rightarrow 1$ for value of λ 64.0; average $\Delta \ln(D(q,\lambda))$ for small λs is shown versus the area between curves of λ = identity and 16.0; slope of $q = 1 \rightarrow 2$ for value of λ identity (i.e. naive diversity) is shown versus the slope of $q = 0 \rightarrow 1$ for value of λ 64.0; slope of $q = 1 \rightarrow 2$ for value of λ identity (i.e. naive diversity) is shown versus the area between curves of λ = identity and 64.0.

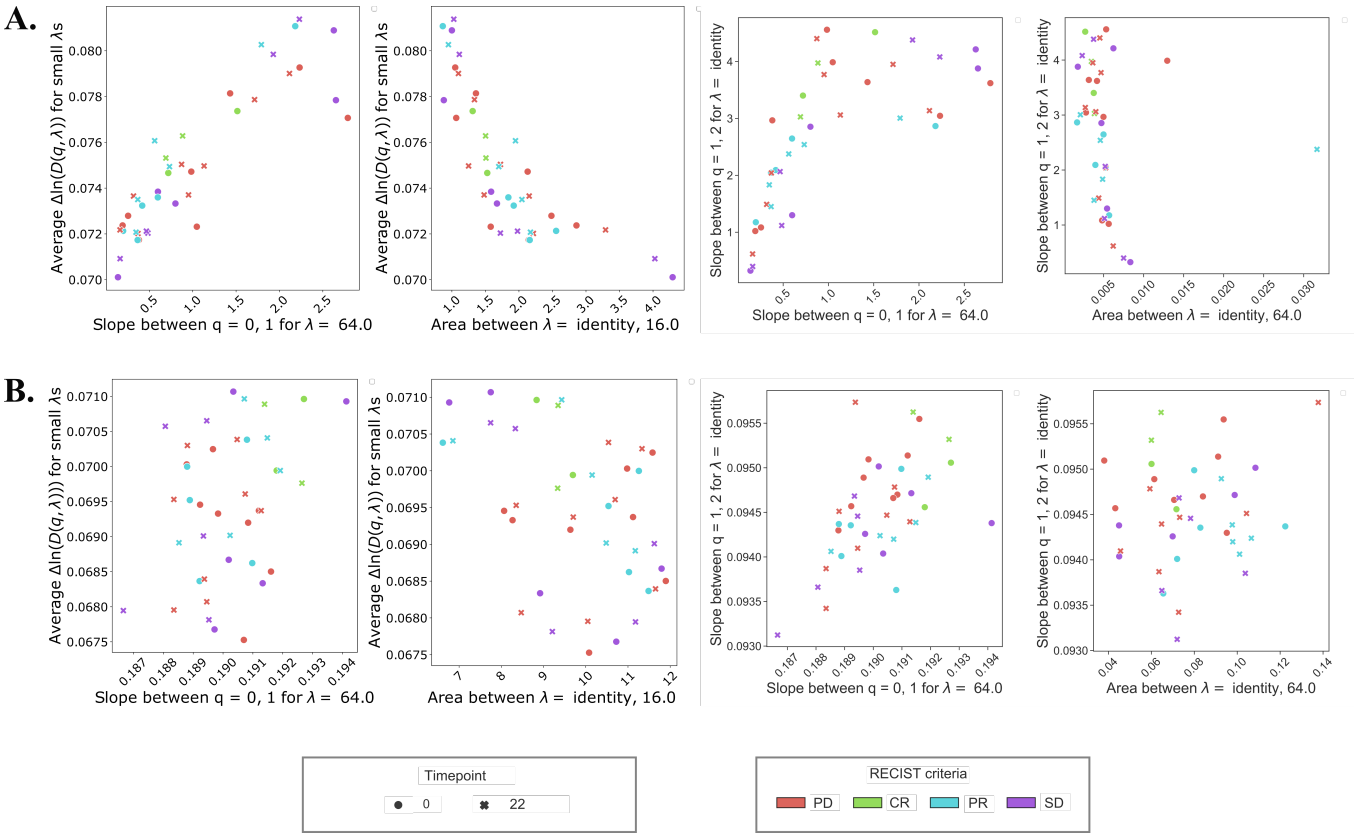


Figure 5. A. Graphs showing relationships between some of the divP features extracted from diPs of the human dataset. From left to right: average $\Delta \ln(D(q,\lambda))$ for small λs is shown versus the slope of $q = 0 \rightarrow 1$ for value of λ 64.0; average $\Delta \ln(D(q,\lambda))$ for small λs is shown versus the area between curves of $\lambda =$ identity and 16.0; slope of $q = 1 \rightarrow 2$ for value of λ identity (i.e. naive diversity) is shown versus the slope of $q = 0 \rightarrow 1$ for value of λ 64.0; slope of $q = 1 \rightarrow 2$ for value of λ identity (i.e. naive diversity) is shown versus the area between curves of $\lambda =$ identity and 64.0. **B.** As A., but comparing features from the human dataset with randomized frequencies. The legend corresponds to both A. and B.

Here we provide a proof-of-concept study in application of similarity scaled diversity estimates to TCR repertoire analysis. We have applied principal components analysis as a means of qualitative repertoire stratification and shown it is able to capture variation in the diversity profiles which characterise immunisation history. The similarity scaled diversity profiles are themselves quite rich summaries and in the future we anticipate that they may be subject to more sophisticated machine learning techniques. At present this possibility is limited by the number of available samples for which comparable data is available. The need to define a similarity scaled diversity index with a parameter λ arises naturally from a desire to generalise the idea of naive diversity profiles. While there may be a variety of ways to define such a generalisation, it should be noted that the specific form of equation 7 (also shown in figure 1 B.) that we use possesses important abstract mathematical properties that will enable further investigation ***Tom Leinster*** (***2013***).

As has previously been shown ***Greiff et al.*** (***2015***) the use of a single diversity index to rank or classify repertoires is not robust, since the ranking will depend on the index selected. The use of naive diversity profiles is a step forward in so far as they reflect the contribution of both large and small clones. However, as applied to real TCR data naive diversity profiles typically give a single effective dimension of separation. This is reflected in our results that show 93% of variance explained by PC1 carried out on the naive diversity profile. The common choice to model clone size distributions using power laws ***Altan-Bonnet et al.*** (***2020***) that have a single tunable parameter further supports the idea that range of possible biological distributions is essentially 1 dimensional. In this case, any practical classification of repertoires based on naive diversity will be based on a 1 dimensional separation.

The novel features identified in our similarity scaled diversity profiles provide a genuine second dimension of variation in the structure of the repertoire based on sequence similarity. As shown in our PCA analysis this opens the practical possibility of 2 dimensional separations of repertoires which are inherently more powerful. Notably, this approach can make use of sequence data to classify repertoires together even when they share no similar sequences.

The striking relationships between similarity profile features appear to reflect biological structure in the TCR repertoires. We hypothesise that the non-linear relationships we observe reflect the range of possible biological variation and are thus analogous to the way in which clone size distributions are well approximated by power laws. Our observation motivates investigations of extending of power law distributions to include description of sequence similarity.

Any calculation involving an all-against-all comparison will inevitably scale with the square of the number sequences. Because TCRDivER is parallelisable, with the 50,000 sequences per sample analysed here it is very practical to run on commonly available computer clusters. When distributed over 8 cores of an Intel Xeon Gold 6126 2.60 GHz processor, each repertoire computation took under 6 hours. There is scope for several-fold speed up, including optimisation of the distance function. A possible gain would come from replacement of the exact all-against-all comparison at high lambda with comparison against approximate k-nearest neighbours using a ball tree algorithm. However, at low lambda values the all-against-all comparison cannot be avoided reflecting the way in which the similarity scaled diversity incorporates genuinely global information about the repertoire.

The form of equation 7 is motivated by rather general mathematical considerations, but these still leave the metric used to compare sequences undetermined. There is no 'true' metric, in the sense that a assigning a single number to the distance between two sequences cannot fully capture all the ways in which binding affinities vary. In our work we used two metrics which plausibly reflect biological functional similarity (through use of evolutionary data in BLOSUM45 matrix) and biochemical similarity (thorough the Atchley factors). While these gave qualitatively similar results, the best metric to use for a given question is an open question in the field of TCR analysis as a whole.

While in this study we have applied similarity scaled diversity profiles to TCR repertoires, we believe that the same concept should also be applicable for understanding antibody repertoires.

The features of similarity scaled diversity profiles can easily be translated in to properties of the repertoire. The functional biological significance of the similarity scaled diversity (as indeed the naive diversity) is likely to be more variable and subject to experimental investigation. TCRDivER provides an important tool to enable those investigations.

Materials and methods

Data Acquisition and Description

The murine dataset consits of previously published data that has been analysed as part of a larger dataset in the work by ***Sun et al.*** (***2017***). Briefly, CD4$^+$ T cells were isolated from spleens of 18 C57BL/6 mice immunised with Complete Freund's adjuvant (CFA) with or without an addition of Ovalbumin antigen (OVA). The samples were collected at different times post-immunisation: at day 5 and 14 (early timepoints) and day 60 (late timepoint). In addition, CD4$^+$ T cells were collected from 8 healthy unimmunised mice prior to study start. An overview of the dataset is given in Table 1. We have received the dataset already analysed with Decombinator, which described in depth in ***Thomas et al.*** (***2013***, 2014b). The data we have analysed consisted of a list of CDR3 sequences present in each sample. The raw fastq files are available at http://www.ncbi.nlm.nih.gov/sra/?term=SRP075893.

Additionally, we have analysed a human TCR dataset previously analysed by ***Formenti et al.*** (***2018***). The participants of the study were 39 patients diagnosed with metastatic non-small-cell lung cancer (NSCLC). They were treated with daily radiation therapy regimen in two phases of the trial (phase I - 6Gy $\times$ 5 and phase II 9.5 Gy $\times$) and intravenous ipilimumab (CTLA-4 blockade) following the first radiation treatment and subsequently repeated every 3 weeks for four cycles. The assessment of patient treatment response was performed with PET/CT scans at day 88 and evaluated using Response Criteria In Solid Tumors (RECIST). The patients were then classified, according to RECIST, into complete responders (CR), partial responders (PR) with tumour decrease in size $\leqslant 30\%$, stable disease (SD) with insufficient shrinkage to qualify for any of the other criteria , and progressive disease (PD) with increase in size $> 20\%$ or appearance of new lesions. Out of 39 patients 20 were evaluable at day 88. Serial blood samples for peripheral blood mononuclear cells (PBMCs) were collected at baseline (day 0), and on days 22, 43, 64, and 88. The isolated PBMC were subjected to amplification and sequencing of bulk TCRβ CDR3 regions by Adaptive Biotechnologies. We have obtained the data from the Adaptive Bioctechnologies ImmunoSEQ database ***Imm*** (***????***). Since the samples collected at later timepoints (day 43 and onward) were not available for all of the 20 evaluable patients, we have restricted our analysis to samples collected at baseline and day 22 of treatment. An overview of samples included in our analysis is given in Table 2.

Subsampling to reduce computational load

We have only considered "In" frame reads of CDR3s in our analysis. In order to reduce the computational load of calculating pairwise similarity between a large number of CDR3 regions (order of magnitude $\approx 10^5$), we have performed subsampling prior to analysis. We will refer to individual

Table 1. Murine Dataset overview

Treatment	Sample collection time (days)	Number of mice
CFA	5	3
CFA	14	3
CFA	60	3
CFA+OVA	5	3
CFA+OVA	14	3
CFA+OVA	60	3
Non-immunised	0	8

Table 1–source data 1. Overview of samples available in the analysed dataset, a part of a previously published dataset by ***Sun et al.*** (***2017***). In total 26 CD4$^+$ murine spleen samples were analysed. Sample collection time is given as the number of days post immunisation. Note that mice culled at day 60 received an additional booster shot of immunising agent (CFA or CFA+OVA).

CDR3 sequences as sequences, and a collection of identical CDR3 sequences as a clone. Each repertoire was down sampled to contain 50000 CDR3 sequences. The sampling was random, based on the original CDR3 frequency distribution. After sampling the number of CDR3 clones was $\leqslant 50000$. An overview of number of unique clones for the murine and human dataset is given in Table 3 and 4, respectively. For each sample, the CDR3 clone count (number of identical CDR3s within a clone) was transformed into frequencies, so that the all the CDR3 clone frequencies sum up to 1:

$$\sum_{i=1}^{S} p_i = 1 \tag{1}$$

,where p_i is the frequency of the *i*-eth CDR3 clone in the repertoire. The clone CDR3 amino acid sequences with their respective frequencies were used as input further downstream.

Calculating the Distance Matrix

In order to reduce computational time and memory usage of calculating pairwise comparison, we have divided the distance matrix into smaller portions ("chunks") which are separately calculated. The algorithm takes in the list of *S* CDR3 clones, and splits it up into *n* sub lists of predefined length, default value is 100. For each sublist pairwise comparisons are made between the CDR3s in the sublist versus all the CDR3s in the original list. An example of the distance matrix calculation can be seen in Figure 6. Global alignment between two CDR3 clone sequences was performed with a gap penalty of 10 and scored using the BLOSUM45 ***Henikoff and Henikoff*** (***1992***) substitution matrix. The alignments were created using the PairwiseAligner function in the Bio.Align package within Python3 ***Van Rossum and Drake*** (***2009***). The distance between two CDR3s, $d(CDR3_i, CDR3_j)$, was calculated based on the alignment scores:

$$d(CDR3_i, CDR3_j) = 1 - \frac{BLOSUM45score(CDR3_i, CDR3_j)}{max(BLOSUM45score(CDR3_i, CDR3_i), BLOSUM45score(CDR3_j, CDR3_j))} \tag{2}$$

Alternatively, we have also employed a biochemical scoring based on the Atchley factors ***Atchley et al.*** (***2005***). The factors are based on biochemical properties of amino acid residues that are grouped and transformed into five Atchley factors. For each CDR3 an average value for individual Atchley factors was computed. The distance between two CDR3 sequences is then calculated as an Euclidean distance between the averaged five Atchley factors.

The obtained distance matrix is mathematically connected to the similarity kernel Z as shown

Table 2. Human Dataset overview

RECIST criteria	Sample collection time (days)	Number of patients
PD	0	8
PD	22	8
SD	0	5
SD	22	5
PR	0	5
PR	22	5
CR	0	2
CR	22	2

Table 2–source data 1.

Overview of samples used from the dataset previously published by ***Formenti et al.*** (***2018***). In total 40 human PMBC samples were sequenced for TCRβ CDR3. Sample collection time has been given as the number of days after start of therapy, with 0 being baseline prior to first treatment.

in equation 3

$$Z_{ij} = e^{-\lambda d_{ij}}, \tag{3}$$

where the λ provides scalling.

Calculating Naive Diversity

Naive diversity does not take similarity into account and is calculated based on the frequencies of CDR3 clones within the repertoire ***Jost*** (***2006***, 2010):

$$D(q) \equiv \left(\sum_{i=1}^{S} p_i^q \right)^{\frac{1}{(1-q)}} \tag{4}$$

, where p_i is the frequency of the i-eth CDR3 clone and q is the diversity order, as diversity indices are a functions of $\sum_{i=1}^{S} p_i^q$. We have chosen a list of qs which subsume the use of some common diversity indices such as 0, 1, 2 and ∞, which correspond to the richness, exponent of the Shannon diversity ***Spellerberg and Fedor*** (***2003***), Simpson ***SIMPSON*** (***1949***); ***Jost*** (***2006***) and Berger-Parker ***Berger and Parker*** (***1970***) index, respectively. To extend our surveying the clone size distribution space we have also added 3, 4, 5 and 6th order of diversity. For values of $q = \{1, \infty\}$ the diversity was calculated as the limit of q approaching the values of 0 and ∞.

$$^{\infty}D = \frac{1}{p_{max}} \tag{5}$$

$$^{1}D = e^{(\ln {}^{1}D)} = e^{(-\sum_{i=1}^{S} p_i \ln p_i)} \tag{6}$$

A detailed derivation of the equations is provided in Appendix 1.

Calculating Similarity scaled Diversity

Similarity scaled diversity of order q takes CDR3 clone sequence distances $d(CDR3_i, CDR3_j)$ along with their respective frequencies. We have adapted the method of calculating similarity-sensitive diversity measures, as proposed by ***Leinster and Cobbold*** (***2012***). We have again chosen a list of qs, $q = 0, 1, 2, 3, 4, 5, 6, \infty$, in the same manner as when calculating the naive diversity. Furthering the method, ***Leinster and Cobbold*** (***2012***), propose the use of similarity-sensitive diversity measures $^{q}D^{Z}$ (Equation 7), dependent on relative abundances and species similarity data as distance d_{ij}. We introduced an alteration of the original approach is introducing the similarity scaling factor λ, which allows us to weight TCR distances in much the same way as we do clone sizes (Figure 1 D.). Here we choose a list of values $\lambda = \{0.0, 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.75, 1.0, 1.5, 2.0, 4.0, 8.0, 16.0, 32.0, 64.0\}$. The value of λ = identity corresponds to the naive diversity calculation, $D(q)$.

$$^{q}D^{Z} = \left(\sum_{i=1}^{S} p_i (\mathbf{Zp})_i^{q-1} \right)^{\frac{1}{(1-q)}} \qquad (\mathbf{Zp})_i = \sum_{j=1}^{S} Z_{ij} p_j \qquad Z_{ij} = e^{-\lambda d_{ij}} \tag{7}$$

Algorithm overview

An overall scheme of the TCRDivER algorithm is given in Figure 7. The algorithm has been implemented within Python (v3.6) ***Van Rossum and Drake*** (***2009***) and it's freely available at https://github.com/sciencisto/TCRDivER.

Downstream analysis of TCRDivER output

The final output of TCRDivER is a table containing all the values of diversity calculated with a range of qs and λs. The full downstream analysis is summarised in a Python3.6 jupyter notebook ***Kluyver et al.*** (***2016***) available alongside the main TCRDivER algorithm at https://github.com/sciencisto/TCRDivER. The diversity profiles were constructed using the seaborn package ***Waskom***

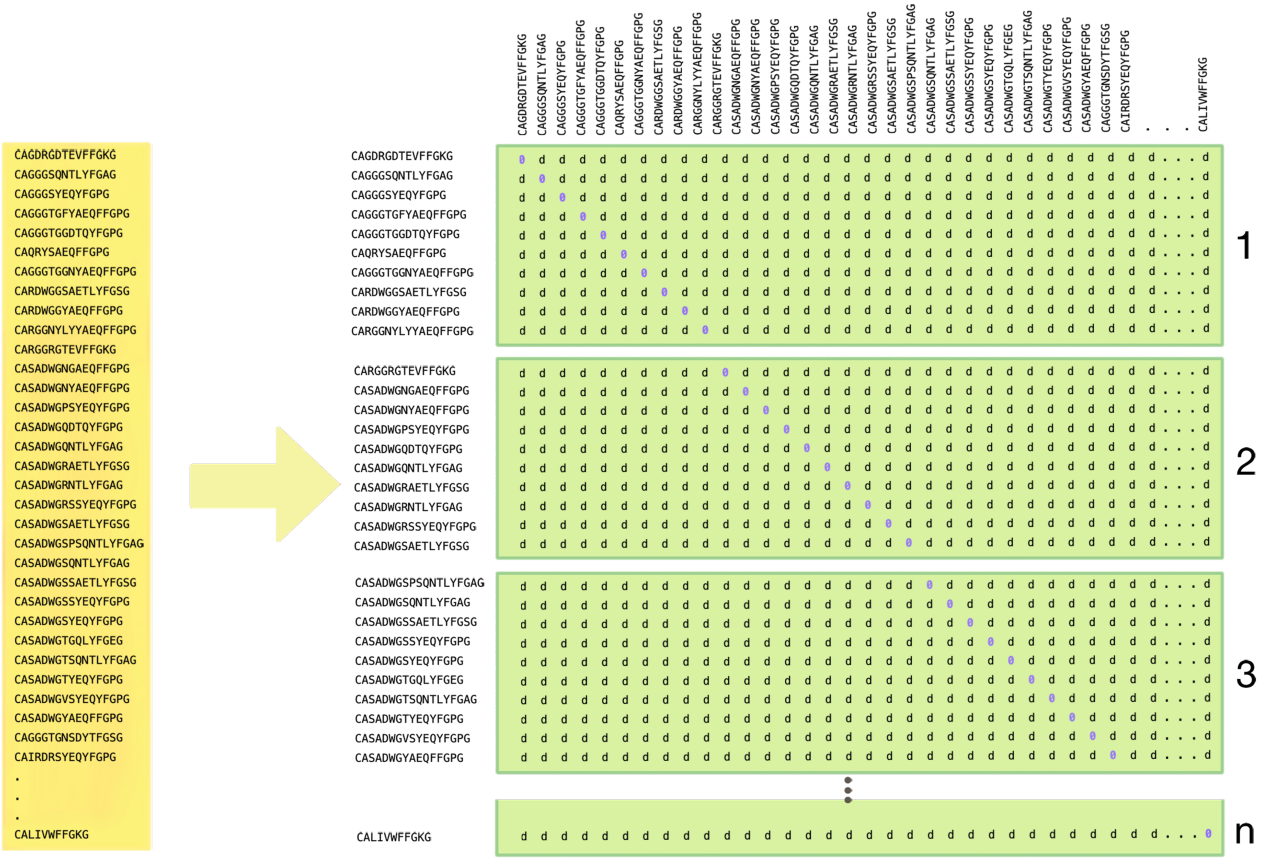


Figure 6. An overview of the calculation of the distance matrix. The list of CDR3 sequences is divided into lists of equal length, here 10 sequences, the default value is 100. These 10 CDR3s are then pairwise compared with all the other CDR3s in the total list of CDR3 sequences. Each portion of the distance matrix i.e. chunk has 10 rows and S columns, where S corresponds to the total number of CDR3 clone sequences. In the end there are n chunks, where n is equal to the floored division of total number of sequences by the length of chunk $n = \frac{S}{length\ of\ chunk}$. Distances *d* are $d(CDR3_i, CDR3_j)$ calculated based on the BLOSUM45 alignment score. The diagonal of the combined distance matrix is 0.

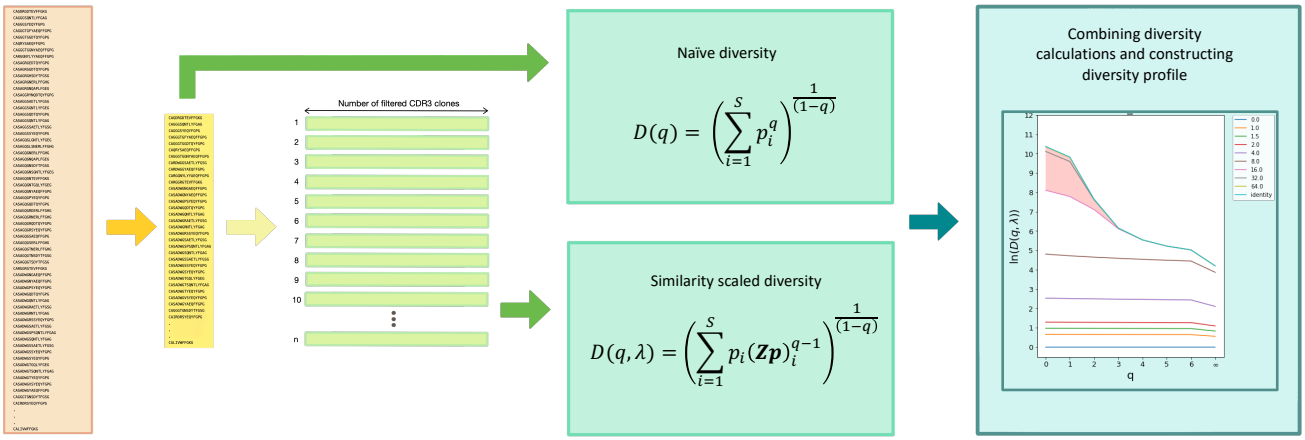


Figure 7. An overview of the TCRDivER algorithm. **I.***Input*. The algorithm takes as input of CDR3 clone sequences with their clone sizes expressed as count or frequency. **II.***Filtering*. The input sample is filtered to contain only "In" frame CDR3 regions. Afterwards, the repertoire is sampled for CDR3s randomly, based on the frequency (count) distribution in the original repertoire. The default subsampling size is 50000 CDR3 sequences, therefore the filtered and subsampled repertoires contain ≤ 50000 unique CDR3s. The CDR3 sequence counts in the subsampled repertoire are transformed into frequencies, so that the final output is a list of unique CDR3s with their respective frequencies summing up to 1 within the repertoire. **III.***Calculating distance matrix* The filtered and downsampled repertoire is then provided as input for calculating the distance matrix. This step is split up so the original list of unique CDR3 sequences is divided into sublists of equal length which are in turn pairwise compared against the whole list of CDR3s using the BLOSUM45 alignment score (see Figure 6). The output is a set of files which contain portions of the distance matrix. If concatenated they would form the complete distance matrix. However, manipulating such a large file would computationally expensive. **IV.***Calculating similarity scaled diversity* This step is split into two parts: calculating naive and similarity scaled diversity (see sections *Calculating Naive Diversity* and *Calculating Similarity scaled Diversity*. The first script takes in only the filtered and subsampled list of CDR3s and their respective frequencies, since $D(q)$ is not dependent on CDR3 sequence distances. The output of this calculation is a .tsv file containing values of diversity at different values of q. The second script takes in both the list of CDR3s with their frequencies and the distance matrix "chunk" files. The calculation is done in a parallel fashion to reduce computational time. As output a .tsv file is given containing the values of calculated diversity at different values of q and λ. **V.** *Combining diversity calculations* The final step is joining the two diversity calculations into a complete overview of the diversity. Two previously obtained .tsv files with calculated diversity values are combined into one file containing all calculated values of diversity. Downstream this file is used for constructing the diversity profiles and further statistical analysis.

et al. (***2017***). Areas between λ curves of $\ln(D(q,\lambda))$ was calculated within the numpy framework ***Oliphant*** (***2006***) as an integration using the composite trapezoidal rule. Average $\Delta \ln(D(q,\lambda))$ for small λs was calculated for $\lambda = \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ as the mean of the average difference between $\ln(D(q,\lambda))$ at each calculated q. Slopes of diversity when $q = 0 \rightarrow 1$, $q = 1 \rightarrow 2$ and $q = 0 \rightarrow 2$ were calculated as differences between the diversities when $q = 0, 1 and 2$. We are aware that this is not the slope of diversity at the specific values of q, as we also provide the mathematical evaluation of the slope at these timepoints (See Appendix 1 Evaluating the slope at $q = 1$). However, due to time and memory considerations we have opted for the simplified calculation as we feel that it represents the slope sufficiently. For future implementations, we will update the algorithm to include the analytical slope evaluation. Principal components analysis was performed as implemented in the scikit-learn package ***Pedregosa FABIANPEDREGOSA et al.*** (***2011***).

Table 3. Murine Dataset Subsampling

Sample name	Treatment	Sample collection time (days)	Number clones prior to sampling	Number of clones after sampling
SB1_AAA	CFA	5	655436	30400
SB1_CCG	CFA	5	250334	29493
SB1_TTG	CFA	5	1091977	32263
SB1_ACC	CFA	14	411320	27156
SB1_CTA	CFA	14	443454	25445
SB1_GCT	CFA	14	357390	25435
SB2_ATT	CFA	60	252947	19626
SB2_CCG	CFA	60	253711	23146
SB2_CGT	CFA	60	154234	15539
SB1_ATT	CFA+OVA	5	805062	31315
SB1_CAC	CFA+OVA	5	470031	30520
SB1_GTC	CFA+OVA	5	428108	32077
SB1_AGG	CFA+OVA	14	572343	33190
SB1_GAG	CFA+OVA	14	437636	27132
SB1_TAT	CFA+OVA	14	581210	32121
SB2_CAC	CFA+OVA	60	153275	16251
SB2_GCT	CFA+OVA	60	197386	20608
SB2_GTC	CFA+OVA	60	193514	17966
CPX1A_GGA	Non-immunised	0	413430	32295
CPX1A_TTG	Non-immunised	0	283449	30541
CPX1B_CAC	Non-immunised	0	171819	30131
EAE1A_GGA	Non-immunised	0	201127	28170
EAE1A_TTG	Non-immunised	0	155225	28699
EAE1B_CCG	Non-immunised	0	80632	14866
EAE1B_TTG	Non-immunised	0	89643	21506
SB1_TGC	Non-immunised	0	284924	21727

Table 3–source data

1. Overview of number CDR3 clones prior and post subsampling in $CD4^+$ TCR repertoires of the murine dataset

Table 4. Human Dataset Subsampling

Sample name	RECIST criteria	Sample collection time (days)	Number clones prior to sampling	Number of clones after sampling
Pt10_PD_PBMC_Day0	PD	0	261326	39555
Pt16_PD_PBMC_Day0	PD	0	170417	36110
Pt27_PD_PBMC_Day0	PD	0	102012	25848
Pt28_PD_PBMC_Day0	PD	0	197523	38151
Pt36_PD_PBMC_Day0	PD	0	42953	21211
Pt38_PD_PBMC_Day0	PD	0	69248	18942
Pt40_PD_PBMC_Day0	PD	0	119828	20905
Pt43_PD_PBMC_Day0	PD	0	144548	32998
Pt10_PD_PBMC_Day22	PD	22	261955	39417
Pt16_PD_PBMC_Day22	PD	22	182238	35294
Pt27_PD_PBMC_Day22	PD	22	30562	17699
Pt28_PD_PBMC_Day22	PD	22	162481	36018
Pt36_PD_PBMC_Day22	PD	22	52416	22467
Pt38_PD_PBMC_Day22	PD	22	83106	20022
Pt40_PD_PBMC_Day22	PD	22	153595	26737
Pt43_PD_PBMC_Day22	PD	22	149218	32983
Pt5_SD_PBMC_Day0	SD	0	124049	36074
Pt9_SD_PBMC_Day0	SD	0	56702	21371
Pt22_SD_PBMC_Day0	SD	0	72589	26421
Pt30_SD_PBMC_Day0	SD	0	74019	17423
Pt32_SD_PBMC_Day0	SD	0	75631	18287
Pt5_SD_PBMC_Day22	SD	22	150203	36579
Pt9_SD_PBMC_Day22	SD	22	43496	20669
Pt22_SD_PBMC_Day22	SD	22	97833	30146
Pt30_SD_PBMC_Day22	SD	22	82704	20291
Pt32_SD_PBMC_Day22	SD	22	59017	19849
Pt1_PR_PBMC_Day0	PR	0	174555	33446
Pt17_PR_PBMC_Day0	PR	0	169671	37698
Pt23_PR_PBMC_Day0	PR	0	91260	16753
Pt37_PR_PBMC_Day0	PR	0	130788	33824
Pt44_PR_PBMC_Day0	PR	0	108678	29786
Pt1_PR_PBMC_Day22	PR	22	205316	33568
Pt17_PR_PBMC_Day22	PR	22	229810	32628
Pt23_PR_PBMC_Day22	PR	22	94146	17549
Pt37_PR_PBMC_Day22	PR	22	195069	35062
Pt44_PR_PBMC_Day22	PR	22	120658	28538
Pt3_CR_PBMC_Day0	CR	0	128860	29075
Pt4_CR_PBMC_Day0	CR	0	119253	27209
Pt3_CR_PBMC_Day22	CR	22	92123	27116
Pt4_CR_PBMC_Day22	CR	22	108383	29732

Table 4–source data 1.

Overview of number CDR3 clones prior and post subsampling in bulk TCR repertoires of the human dataset.

Acknowledgements

We would like to thank Prof. Nir Friedman, Dr. Shlomit Reich-Zeliger and Dr. Erik Shifrut from the Weizmann Institute, Rehovot, Israel for generating and sharing the TCR sequence data analysed in this paper.

Appendix 1

Evaluation of naive diversity of the first order

We start with:

$$\ln\left(D(q)\right) = \ln\left(\sum_{i=1}^{S} p_i^q\right)^{\frac{1}{(1-q)}}$$

Exploring the limit as q approaches 1 allows us to apply L'Hopitals rule:

$$\lim_{q\to 1} \ln\left(D(q)\right) = \lim_{q\to 1} \frac{\ln \sum_{i=1}^{S} p_i^q}{1-q} = \lim_{q\to 1} \frac{\left(\ln \sum_{i=1}^{S} p_i^q\right)'}{(1-q)'}$$

The solution is:

$$\ln\left(D(1)\right) = -\sum_{i=1}^{S} p_i \ln p_i$$

This is equivalent to:

$$D(1) = \frac{1}{p_1^{p_1} p_2^{p_2} \cdots p_i^{p_i}}$$

Evaluating the slope at $q = 1$

We start with the definition of ${}^q D$:

$$D(q) = \left(\sum_{i=1}^{N} p_i^q\right)^{\frac{1}{1-q}}$$

define S as the internal sum

$$S = \sum_{i=1}^{N} p_i^q$$

Now we can write

$$\ln(D(q)) = \frac{1}{1-q} \ln S$$

Let $'$ denote differentiation with respect to q. Now evaluate

$$\begin{aligned}\ln(D(q))' &= \frac{1}{(1-q)^2} \ln S + \frac{1}{1-q}(\ln S)' \\ &= \frac{\ln S + (1-q)(\ln S)'}{(1-q)^2}\end{aligned}$$

To evaluate the limit as q goes to 1 we need to apply l'hopital's rule twice. Calling the numerator t

$$\begin{aligned}t &= \ln S + (1-q)(\ln S)' \\ t' &= (1-q)(\ln S)'' \\ t'' &= (1-q)(\ln S)''' - (\ln S)''\end{aligned}$$

Since $\ln S$ and all its derivatives are finite as q goes to 1

$$\lim_{q\to 1} t'' = -\lim_{q\to 1} (\ln S)''$$

Call the denominator b

$$b'' = 2$$

and

$$\lim_{q\to 1} b'' = 2$$

Putting this together we have

$$\begin{aligned}\lim_{q\to 1} \ln(D(q))' &= \frac{\lim_{q\to 1} t''}{\lim_{q\to 1} b''} \\ &= -\frac{1}{2}(\ln S)''\Big|_{q=1}\end{aligned}$$

We need

$$(\ln S)' = \frac{\sum_{i=1}^{N} p_i^q \ln p_i}{S}$$

$$(\ln S)'' = \frac{\sum_{i=1}^{N} p_i^q (\ln p_i)^2}{S} - \frac{(\sum_{i=1}^{N} p_i^q \ln p_i)^2}{S^2}$$

Because p_i is a probability distribution

$$\lim_{q\to 1} S = 1$$

which means (since the limit of quotient is the quotient of the limits)

$$\lim_{q\to 1} S'' = \sum_{i=1}^{N} p_i^q (\ln p_i)^2 - (\sum_{i=1}^{N} p_i^q \ln p_i)^2$$

Finally, we want to evaluate $D(q)'$ and then take the limit as q goes to 1.

$$D(q)' = D(q)\,(\ln(D(q)))'$$

and

$$\begin{aligned}\lim_{q\to 1} D(q)' &= \lim_{q\to 1} D(q) \lim_{q\to 1} -\frac{1}{2}(\ln S)'' \\ &= -\frac{1}{2}\, D(1) \left(\sum_{i=1}^{N} p_i^q (\ln p_i)^2 - (\sum_{i=1}^{N} p_i^q \ln p_i)^2 \right)\end{aligned}$$

To generalise to the case $\mathbf{Z} \neq \mathbf{I}$ we simply have to replace S with

$$S = \sum_{i=1}^{N} p_i (\mathbf{Zp})_i^{q-1}.$$

In this case

$$(\ln S)' = \frac{\sum_{i=1}^{N} p_i (\mathbf{Zp})_i^{q-1} \ln(\mathbf{Zp})_i}{S}$$

$$(\ln S)'' = \frac{\sum_{i=1}^{N} p_i (\mathbf{Zp})_i^{q-1} (\ln \mathbf{Zp})_i)^2}{S} - \frac{(\sum_{i=1}^{N} p_i (\mathbf{Zp})_i^{q-1} \ln(\mathbf{Zp})_i)^2}{S^2}$$

Evaluation of similarity scaled diversity of the first order

We start with:

$$\ln\ D(q,\lambda) = \ln\left(\sum_{i=1}^{S} p_i(\mathbf{Zp})_i^{q-1}\right)^{\frac{1}{(1-q)}}$$

Rewritting the equation, calculating the limit as q approaches 1 and applying L'Hopitals rule:

$$\lim_{q\to 1}\ln\ D(q,\lambda) = \lim_{q\to 1}\frac{\ln\sum_{i=1}^{S} p_i(\mathbf{Zp})_i^{q-1}}{1-q} = \lim_{q\to 1}\frac{\left(\ln\sum_{i=1}^{S} p_i(\mathbf{Zp})_i^{q-1}\right)'}{(1-q)'}$$

The result is:

$$\ln\left(D(1,\lambda) = -\sum_{i=1}^{S} p_i \ln\left(\mathbf{Zp}\right)_i\right.$$

, which is equivalent to:

$$D(q,\lambda) = \frac{1}{(\mathbf{Zp})_1^{p_1}(\mathbf{Zp})_2^{p_2}\cdots(\mathbf{Zp})_i^{p_i}}$$

Evaluation of naive diversity of the infinity order

We start with the formula for naive diversity and extract the largest clone frequency p_{max}:

$$D(q) = \left(\sum_{i=1}^{S} p_i^q\right)^{\frac{1}{1-q}} = (p_{max})^{\frac{q}{1-q}}\left(1+\sum_{j=1}^{S} p_j'^q\right)^{\frac{1}{1-q}}$$

,where $p_j' = \frac{p_j}{p_{max}}$ for $j \neq max$, and p_{max} is represented in the first term of the sum. Since a limit of products is a product of limits, it follows:

$$\lim_{q\to\infty} D(q) = \lim_{q\to\infty}(p_{max})^{\frac{q}{1-q}}\lim_{q\to\infty}(1+\sum_{j=1}^{S} p_j'^q)^{\frac{1}{1-q}}$$

The first limit is evaluated as:

$$\lim_{q\to\infty}(p_{max})^{\frac{q}{1-q}} = \frac{1}{p_{max}}$$

The second limit is evaluated by taking the logarithm:

$$\log\left(\lim_{q\to\infty}(1+\sum_{j=1}^{S} p_j'^q)^{\frac{1}{1-q}}\right) = \lim_{q\to\infty}\log\left((1+\sum_{j=1}^{S} p_j'^q)^{\frac{1}{1-q}}\right) = \lim_{q\to\infty}\frac{1}{(1-q)}\log(1+\sum_{j=1}^{S} p_j'^q)$$

Since $0 < \sum_j^S p_j'^q < 1$, the bounds of logarithm are:

$$0 < \log(1+\sum_{j=1}^{S} p_j'^q) < \log 2$$

, which gives:

$$\lim_{q\to\infty}\frac{1}{(1-q)}\log(1+\sum_{j=1}^{S} p_j'^q) = 0$$

$$\implies \log\left(\lim_{q\to\infty}(1+\sum_{j=1}^{S} p_j'^q)^{\frac{1}{1-q}}\right) = 0$$

$$\implies \lim_{q\to\infty}(1+\sum_{j=1}^{S} {p_j'}^{q})^{\frac{1}{1-q}} = 1$$

$$\implies \lim_{q\to\infty} D(q) = \frac{1}{p_{max}}$$

Evaluation of similarity scaled diversity of the infinity order

We start with:

$$D(q,\lambda) = \left(\sum_{i=1}^{S} p_i(\mathbf{Zp})_i^{q-1}\right)^{\frac{1}{(1-q)}} = ((\mathbf{Zp})_{max})^{\frac{q-1}{1-q}}\left(p_{max}(1+\sum_j p_j'(\mathbf{Zp})_j'^{q-1}\right)^{\frac{1}{1-q}}$$

where the term that has been pulled out is the one for which $(\mathbf{Zp})_i$ is maximum. The p_{max} is the corresponding p_i. As before, the p_j' are defined as $\frac{p_j}{p_{max}}$ for $j \neq max$ and $(\mathbf{Zp})_j'$ is defined as $\frac{(\mathbf{Zp})_j}{(\mathbf{Zp})_{max}}$. Again, the limit splits in to two factors:

$$\lim_{q\to\infty}(\mathbf{Zp})_{max})^{\frac{q-1}{1-q}} = \frac{1}{(\mathbf{Zp})_{max})}$$

Taking the log of the second term gives:

$$\lim_{x\to\infty}\frac{1}{1-q}\log\left(p_{max}(1+\sum_j p_j'(\mathbf{Zp})_j'^{q-1}\right)$$

and now the log is bounded by:

$$\log p_{max} < \log\left(p_{max}(1+\sum_j p_j'(\mathbf{Zp})_j'^{q-1}\right) < \log 2$$

so again the limit of the log second factor in (*) is 0, and limit of the factor itself is 1. The end result is:

$$\lim_{q\to\infty} D(q,\lambda) = \frac{1}{(\mathbf{Zp})_{max}}$$

which reduces to the correct limit when **Z**=**I** which is the naive diversity.

Appendix 2

Evaluation $\Delta ln(D(q,\lambda))$ for small λ: Perturbation around $\lambda = 0$

Conjecture: gradient of $\left.\frac{D(q,\lambda)}{dq}\right|_{q=0}$ is a decreasing function of λ. N.B. $D(q,\lambda)$ is an increasing function of λ for all q.

We start with the assumption that for λ around 0:

$$\ln\left(D(q,\lambda)\right) \propto \lambda$$

Where $\ln\left(D(q,\lambda)\right)$ is:

$$\begin{aligned}
\ln(D(q,\lambda)) &= \ln\left(\sum_{i=1}^{S} p_i(\mathbf{Zp})_i^{q-1}\right)^{\frac{1}{(1-q)}} \\
&= \frac{1}{1-q}\ln\left(\sum_{i=1}^{S} p_i(\mathbf{Zp})_i^{q-1}\right) \\
&= \frac{1}{1-q}\ln\left(\sum_{i=1}^{S} p_i\Big(\sum_{j=1}^{S} p_j e^{-\lambda d_{ij}}\Big)_i^{q-1}\right)
\end{aligned}$$

For $\lambda \to 0$ by applying Taylor expansion $e^{-\lambda d_{ij}}$ reduces to $1-\lambda d_{ij}$ which gives:

$$\ln(D(q,\lambda)) \approx \frac{1}{1-q}\ln\left(\sum_{i=1}^{S} p_i\Big(\sum_{j=1}^{S} p_j(1-\lambda d_{ij})\Big)_i^{q-1}\right)$$

We can then rewrite:

$$\begin{aligned}
\Big(\sum_{j=1}^{S} p_j(1-\lambda d_{ij})\Big)_i^{q-1} &\approx \Big(p_0(1-\lambda d_{i0}) + p_1(1-\lambda d_{i1}) + \cdots + p_j(1-\lambda d_{ij})\Big)^{q-1} \\
&\approx \left(1-\lambda\Big(\sum_{j=1}^{S} p_j d_{ij}\Big)\right)^{q-1}
\end{aligned}$$

By applying the binomial expansion we arrive at:

$$\Big(\sum_{j=1}^{S} p_j(1-\lambda d_{ij})\Big)_i^{q-1} \approx \left(1-(q-1)\lambda\Big(\sum_{j=1}^{S} p_j d_{ij}\Big)\right)$$

By substituting the derived expressions in the formula for $\ln\left(D(q,\lambda)\right)$ and keeping in mind that $\sum_{i=1}^{S} p_i = 1$, we can write:

$$\begin{aligned}
\ln\left(D(q,\lambda)\right) &= \frac{1}{1-q}\ln\left(\sum_{i=1}^{S} p_i\Big(\sum_{j=1}^{S} p_j e^{-\lambda d_{ij}}\Big)_i^{q-1}\right) \\
&\approx \frac{1}{1-q}\ln\left(\sum_{i=1}^{S} p_i\Big(1-(q-1)\lambda\Big(\sum_{j=1}^{S} p_j d_{ij}\Big)\Big)\right) \\
&\approx \frac{1}{1-q}\ln\left(1-\sum_{i=1}^{S} p_i(q-1)\lambda\Big(\sum_{j=1}^{S} p_j d_{ij}\Big)\right) \\
&\approx \frac{1}{1-q}\ln\left(1-(q-1)\lambda\sum_{i=1}^{S} p_i\Big(\sum_{j=1}^{S} p_j d_{ij}\Big)\right)
\end{aligned}$$

By applying the linear approximation $ln(1-x) \approx x$, we finally arrive:

$$\ln(D(q,\lambda)) \approx \frac{1}{1-q}\Big(-(q-1)\lambda \sum_{i=1}^{S} p_i\big(\sum_{j=1}^{S} p_j d_{ij}\big)\Big)$$
$$\approx \lambda \sum_{i=1}^{S} p_i\big(\sum_{j=1}^{S} p_j d_{ij}\big)$$

Note that the final form of the evaluation of $D(q,\lambda)$ for $\lambda \to 0$ is independent of the order of diversity q. It is solely dependent on the distance between CDR3 sequences weighted by their respective frequencies.

Evaluation $\Delta ln(D(q,\lambda))$ for small λ and it's relationship to distance

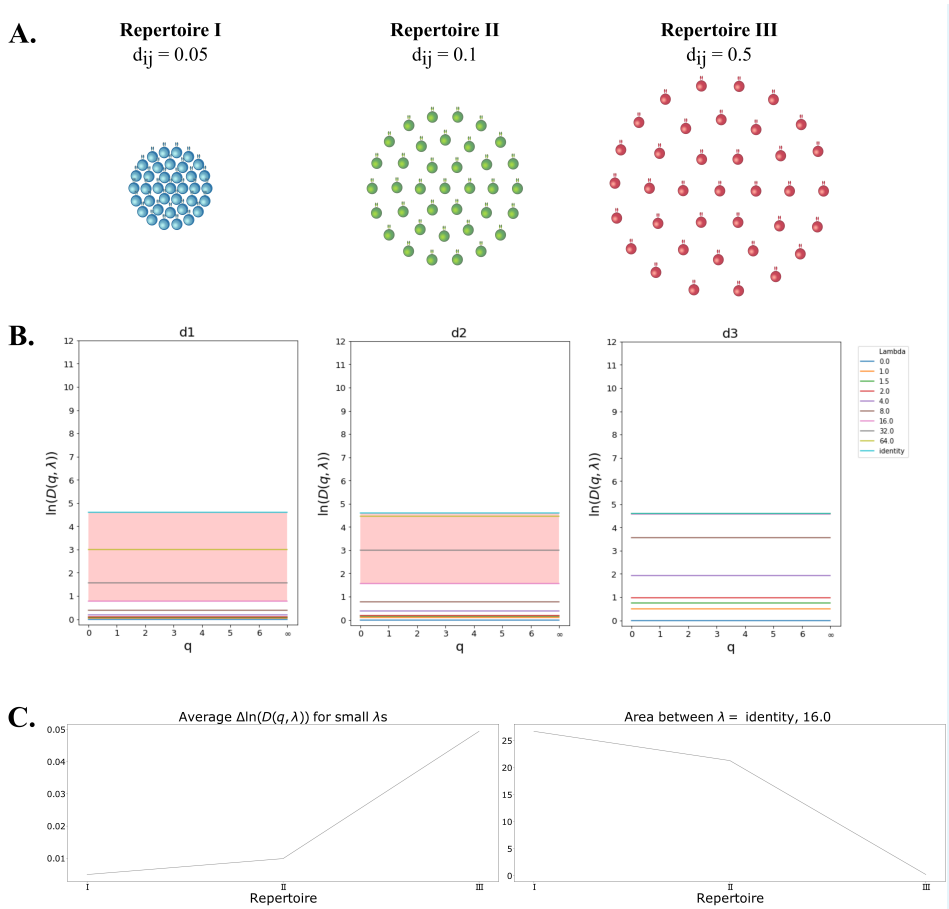
By evaluating $\Delta ln(D(q,\lambda))$ for two values of small λ, where $\lambda' > \lambda''$ we arrive at:

$$\Delta ln(D(q,\lambda)) \approx D(q,\lambda') - D(q,\lambda)''$$
$$\approx \lambda' \sum_{i=1}^{S} p_i\big(\sum_{j=1}^{S} p_j d_{ij}\big) - \lambda' \sum_{i=1}^{S} p_i\big(\sum_{j=1}^{S} p_j d_{ij}\big)$$
$$\approx (\lambda' - \lambda'') \sum_{i=1}^{S} p_i\big(\sum_{j=1}^{S} p_j d_{ij}\big)$$

It is evident that $\Delta ln(D(q,\lambda))$ is linearly dependent on the distances between CDR3s and their probabilities. In the case of two hypothetical repertoires, **I** and **II**, which have a uniform distribution of CDR3 frequencies within the repertoire $p_i^I = p_i^{II} = p$ and distances between CDR3s $d_{ij}^I > d_{ij}^{II}$, $\Delta ln(D(q,\lambda))$ for repertoire **I** is larger than $\Delta ln(D(q,\lambda))$ for repertoire **II**. That is with the increase of similarity between CDR3s, the area between the curves for small λs decreases (Appendix 2 Figure 1 C.). Alternatively, if the distances between CDR3s of the two repertoires are the same $d_{ij}^I = d_{ij}^{II} = d$, and the distribution is still uniform, but the number of clones differs so that repertoire **I** has less clones than **II** i.e. $p_i^I > p_i^{II}$, then $\Delta ln(D(q,\lambda))$ is larger than $\Delta ln(D(q,\lambda))$. Meaning that repertoires with more abundant clones have a larger $\Delta ln(D(q,\lambda))$ for small λs.

Evaluation $\Delta ln(D(q,\lambda))$ for larger λs and it's relationship to distance

In order to evaluate the relationship between CDR3 clone distance and the area between the curves of larger λs we have constructed three mock repertoires. The reperotires constist of 100 CDR3s that are uniformly distributed in the repertoire, i.e. $p_i = \frac{1}{S} = \frac{1}{100} = 0.01$. For each mock repertoire a mock distance matrix was calculated so that the distance between the CDR3s within the repertoire were equal, but that they differ between the repertoires. The distances were $d_{i,j}^I = 0.05$, $d_{i,j}^{II} = 0.1$ and $d_{i,j}^{III} = 0.5$, for repertoire I, II and III respectively when $i \neq j$, else $d_{i,j} = 0$ for $i = j$. Individual λ curves of the diversity profiles straight lines - a remnant of uniform distribution of CDR3 frequencies in the repertoire (Appendix 2 Figure 1).



Appendix 2 Figure 1. Effect of CDR3 distance shown in three mock repertoires with a uniform distribution of 100 CDR3 clones in in the repertoire. **A.** Schematic representation of the three mock repertoires with the distances d_{ij} between CDR3s increasing from repertoire I to III. **B.** Diversity profiles calculated based on the probability distribution and d_{ij} for CDR3s in the mock repertoires. The frequency of seeing each CDR3 clone in all the repertoires, since they consits of 100 uniformly distributed CDR3s, is $p_i = \frac{1}{100} = 0.01$ **C.** Calculated values of average $\Delta \ln(D(q, \lambda))$ for small λs and calculated area between λ identity and 16 curves for the three repertoires, shown left to right respectively.

Acknowledgements

References

immunoSEQ – Adaptive Biotechnologies;. https://www.adaptivebiotech.com/products-services/immunoseq/.

Altan-Bonnet G, Mora T, Walczak AM. Quantitative immunology for physicists. Physics Reports. 2020 3; 849:1–83. doi: 10.1016/j.physrep.2020.01.001.

Antunes DA, Rigo MM, Freitas MV, Mendes MFA, Sinigaglia M, Lizée G, Kavraki LE, Selin LK, Cornberg M, Vieira GF. Interpreting T-Cell cross-reactivity through structure: Implications for TCR-based cancer immunotherapy. Frontiers in Immunology. 2017 10; 8(OCT):1210. doi: 10.3389/fimmu.2017.01210.

Arora R, Burke H, Arnaout R. Immunological Diversity with Similarity. bioRxiv. 2018 12; p. 483131. https://doi.org/10.1101/483131, doi: 10.1101/483131.

Atchley WR, Zhao J, Fernandes AD, Drüke T. Solving the protein sequence metric problem. Proceedings of the National Academy of Sciences of the United States of America. 2005 5; 102(18):6395–6400. www.genome., doi: 10.1073/pnas.0408677102.

Bentzen AK, Hadrup SR. T-cell-receptor cross-recognition and strategies to select safe T-cell receptors for clinical translation. Immuno-Oncology Technology. 2019 9; 2:1–10. doi: 10.1016/j.iotech.2019.06.003.

Berger WH, Parker FL. Diversity of planktonic foraminifera in deep-sea sediments. Science. 1970 6; 168(3937):1345–1347. doi: 10.1126/science.168.3937.1345.

Carey AJ, Gracias DT, Thayer JL, Boesteanu AC, Kumova OK, Mueller YM, Hope JL, Fraietta JA, Zessen DBHv, Katsikis PD. Rapid Evolution of the CD8+ TCR Repertoire in Neonatal Mice. The Journal of Immunology. 2016 3; 196(6):2602–2613. http://www.ncbi.nlm.nih.gov/pubmed/26873987http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4779665, doi: 10.4049/JIMMUNOL.1502126.

Chang LW, Doan LT, Fields P, Vignali M, Akilov OE. The Utility of T-Cell Clonality in Differential Diagnostics of Acute Graft-versus-Host Disease from Drug Hypersensitivity Reaction. Journal of Investigative Dermatology. 2019; doi: 10.1016/j.jid.2019.11.009.

Chiffelle J, Genolet R, Perez MA, Coukos G, Zoete V, Harari A, T-cell repertoire analysis and metrics of diversity and clonality. Elsevier Ltd; 2020. doi: 10.1016/j.copbio.2020.07.010.

Choi Y, Kotzin B, Herron L, Callahan J, Marrack P, Kappler J. Interaction of Staphylococcus aureus toxin 'superantigens' with human T cells. Proceedings of the National Academy of Sciences of the United States of America. 1989 11; 86(22):8941–8945. https://www.pnas.org/content/86/22/8941https://www.pnas.org/content/86/22/8941.abstract, doi: 10.1073/pnas.86.22.8941.

Cinelli M, Sun Yuxin, Best K, Heather JM, Reich-Zeliger S, Shifrut E, Friedman N, Shawe-Taylor J, Chain B. Feature selection using a one dimensional naïve Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires. Bioinformatics. 2017 1; 33(7):951–955. https://doi.org/10.1093/bioinformatics/btw771, doi: 10.1093/bioinformatics/btw771.

Ciupe SM, Devlin BH, Markert ML, Kepler TB. Quantification of total T-cell receptor diversity by flow cytometry and spectratyping. BMC Immunology. 2013 8; 14(1):35. https://bmcimmunol.biomedcentral.com/articles/10.1186/1471-2172-14-35, doi: 10.1186/1471-2172-14-35.

Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB, Nguyen THO, Kedzierska K, La Gruta NL, Bradley P, Thomas PG. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature. 2017 6; 547(7661):89–93. http://www.nature.com/doifinder/10.1038/nature22383, doi: 10.1038/nature22383.

Didona D, Di Zenzo G, Humoral epitope spreading in autoimmune bullous diseases. Frontiers Media S.A.; 2018. www.frontiersin.org, doi: 10.3389/fimmu.2018.00779.

Formenti SC, Rudqvist NP, Golden E, Cooper B, Wennerberg E, Lhuillier C, Vanpouille-Box C, Friedman K, Ferrari de Andrade L, Wucherpfennig KW, Heguy A, Imai N, Gnjatic S, Emerson RO, Zhou XK, Zhang T, Chachoua A, Demaria S. Radiotherapy induces responses of lung cancer to CTLA-4 blockade. Nature Medicine. 2018 12; 24(12):1845–1851. https://doi.org/10.1038/s41591-018-0232-2, doi: 10.1038/s41591-018-0232-2.

Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, Ji X, Han A, Krams SM, Pettus C, Haas N, Arlehamn CSL, Sette A, Boyd SD, Scriba TJ, Martinez OM, Davis MM. Identifying specificity groups in the T cell receptor repertoire. Nature. 2017 7; 547(7661):94–98. doi: 10.1038/nature22976.

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. The Journal of Immunology. 1994; 152(10).

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 Medicine. 2015 5; 7(1):49. http://genomemedicine.com/content/7/1/49, doi: 10.1186/s13073-015-0169-8.

Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America. 1992 11; 89(22):10915–10919. https://www-pnas-org.proxy.findit.dtu.dk/content/89/22/10915https://www-pnas-org.proxy.findit.dtu.dk/content/89/22/10915.abstract, doi: 10.1073/pnas.89.22.10915.

Izraelson M, Nakonechnaya TO, Moltedo B, Egorov ES, Kasatskaya SA, Putintseva EV, Mamedov IZ, Staroverov DB, Shemiakina II, Zakharova MY, Davydov AN, Bolotin DA, Shugay M, Chudakov DM, Rudensky AY, Britanova OV. Comparative analysis of murine T-cell receptor repertoires. Immunology. 2018 2; 153(2):133–144. http://doi.wiley.com/10.1111/imm.12857, doi: 10.1111/imm.12857.

Jost L. Entropy and diversity. Oikos. 2006 5; 113(2):363–375. http://doi.wiley.com/10.1111/j.2006.0030-1299.14714.x, doi: 10.1111/j.2006.0030-1299.14714.x.

Jost L. The Relation between Evenness and Diversity. Diversity. 2010 3; 2. doi: 10.3390/d2020207.

Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C. Jupyter Notebooks—a publishing format for reproducible computational workflows. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016* IOS Press BV; 2016. p. 87–90. doi: 10.3233/978-1-61499-649-1-87.

Laydon DJ, Bangham CRM, Asquith B. Estimating T-cell repertoire diversity: Limitations of classical estimators and a new approach. Philosophical Transactions of the Royal Society B: Biological Sciences. 2015 8; 370(1675). doi: 10.1098/rstb.2014.0291.

Leinster T, Cobbold CA. Measuring diversity: the importance of species similarity. Ecology. 2012 3; 93(3):477–489. http://doi.wiley.com/10.1890/10-2402.1, doi: 10.1890/10-2402.1.

Memon SA, Sportès C, Flomerfelt FA, Gress RE, Hakim FT. Quantitative analysis of T cell receptor diversity in clinical samples of human peripheral blood. Journal of Immunological Methods. 2012 1; 375(1-2):84–92. /pmc/articles/PMC3253939/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3253939/, doi: 10.1016/j.jim.2011.09.012.

Miles JJ, Douek DC, Price DA, Bias in the αB T-cell repertoire: Implications for disease pathogenesis and vaccination. John Wiley & Sons, Ltd; 2011. doi: 10.1038/icb.2010.139.

Mora T, Walczak AM. Rényi entropy, abundance distribution, and the equivalence of ensembles. Physical Review E. 2016 5; 93(5):052418. https://journals.aps.org/pre/abstract/10.1103/PhysRevE.93.052418, doi: 10.1103/PhysRevE.93.052418.

Mora T, Walczak AM. Quantifying lymphocyte receptor diversity. In: *Systems Immunology* CRC Press; 2018.p. 183–198. doi: 10.1201/9781315119847-11.

Muraro PA, Jacobsen M, Necker A, Nagle JW, Gaber R, Sommer N, Oertel WH, Martin R, Hemmer B. Rapid identification of local T cell expansion in inflammatory organ diseases by flow cytometric T cell receptor Vβ analysis. Journal of Immunological Methods. 2000 12; 246(1-2):131–143. https://pubmed.ncbi.nlm.nih.gov/11121554/, doi: 10.1016/S0022-1759(00)00309-4.

Ochsenreither S, Fusi A, Busse A, Nagorsen D, Schrama D, Becker J, Thiel E, Keilholz U. Relative quantification of TCR Vbeta-chain families by real time PCR for identification of clonal T-cell populations. Journal of Translational Medicine. 2008 7; 6. https://pubmed.ncbi.nlm.nih.gov/18593466/, doi: 10.1186/1479-5876-6-34.

Oliphant TE. A guide to NumPy, vol. 1. Trelgol Publishing USA; 2006.

Pedregosa FABIANPEDREGOSA F, Michel V, Grisel OLIVIERGRISEL O, Blondel M, Prettenhofer P, Weiss R, Vanderplas J, Cournapeau D, Pedregosa F, Varoquaux G, Gramfort A, Thirion B, Grisel O, Dubourg V, Passos A, Brucher M, Perrot andÉdouardand M, Duchesnay a, Duchesnay EDOUARDDUCHESNAY F. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot; 2011.

Petrova G, Ferrante A, Gorski J, Cross-reactivity of T cells and its role in the immune system. Begell House Inc.; 2012. doi: 10.1615/CritRevImmunol.v32.i4.50.

Robert L, Tsoi J, Wang X, Emerson R, Homet B, Chodon T, Mok S, Huang RR, Cochran AJ, Comin-Anduix B, Koya RC, Graeber TG, Robins H, Ribas A. CTLA4 blockade broadens the peripheral T-cell receptor repertoire. Clinical Cancer Research. 2014 5; 20(9):2424–2432. doi: 10.1158/1078-0432.CCR-13-2648.

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 $\alpha\beta$ T cells. Blood. 2009 11; 114(19):4099–4107. doi: 10.1182/blood-2009-04-217604.

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. Science Translational Medicine. 2010 9; 2(47):64–47. doi: 10.1126/scitranslmed.3001442.

Rudqvist NP, Pilones KA, Lhuillier C, Wennerberg E, Sidhom JW, Emerson RO, Robins HS, Schneck J, Formenti SC, Demaria S. Radiotherapy and CTLA-4 blockade shape the tcr repertoire of tumor-infiltrating t cells. Cancer Immunology Research. 2018 2; 6(2):139–150. doi: 10.1158/2326-6066.CIR-17-0134.

Sherwood AM, Emerson RO, Scherer D, Habermann N, Buck K, Staffa J, Desmarais C, Halama N, Jaeger D, Schirmacher P, Herpel E, Kloor M, Ulrich A, Schneider M, Ulrich CM, Robins H. Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue. Cancer Immunology, Immunotherapy. 2013 9; 62(9):1453–1461. doi: 10.1007/s00262-013-1446-2.

Sidhom JW, Bessell CA, Havel JJ, Kosmides A, Chan TA, Schneck JP. ImmunoMap: A bioinformatics tool for T-cell repertoire analysis. Cancer Immunology Research. 2018 2; 6(2):151–162. doi: 10.1158/2326-6066.CIR-17-0114.

SIMPSON EH. Measurement of Diversity. Nature. 1949 4; 163(4148):688–688. http://www.nature.com/articles/163688a0, doi: 10.1038/163688a0.

Spellerberg IF, Fedor PJ. A tribute to Claude Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon-Wiener' Index; 2003.

Sun Y, Best K, Cinelli M, Heather JM, Reich-Zeliger S, Shifrut E, Friedman N, Shawe-Taylor J, Chain B. Specificity, privacy, and degeneracy in the CD4 T cell receptor repertoire following immunization. Frontiers in Immunology. 2017 4; 8(APR). doi: 10.3389/fimmu.2017.00430.

Thomas N, Best K, Cinelli M, Reich-Zeliger S, Gal H, Shifrut E, Madi A, Friedman N, Shawe-Taylor J, Chain B. Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. Bioinformatics. 2014 11; 30(22):3181–3188. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu523, doi: 10.1093/bioinformatics/btu523.

Thomas N, Best K, Cinelli M, Reich-Zeliger S, Gal H, Shifrut E, Madi A, Friedman N, Shawe-Taylor J, Chain B. Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. Bioinformatics. 2014 8; 30(22):3181–3188. https://doi.org/10.1093/bioinformatics/btu523, doi: 10.1093/bioinformatics/btu523.

Thomas N, Heather J, Ndifon W, Shawe-Taylor J, Chain B. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. Bioinformatics. 2013 1; 29(5):542–550. https://doi.org/10.1093/bioinformatics/btt004, doi: 10.1093/bioinformatics/btt004.

Tom Leinster. The Magnitude of Metric Spaces. Documenta Mathematica. 2013; 18:857–905. https://www.math.uni-bielefeld.de/documenta/vol-18/27.html.

Twyman-Saint Victor C, Rech AJ, Maity A, Rengan R, Pauken KE, Stelekati E, Benci JL, Xu B, Dada H, Odorizzi PM, Herati RS, Mansfield KD, Patsch D, Amaravadi RK, Schuchter LM, Ishwaran H, Mick R, Pryma DA, Xu X, Feldman MD, et al. Radiation and dual checkpoint blockade activate non-redundant immune mechanisms in cancer. Nature. 2015 4; 520(7547):373–377. http://www.nature.com/articles/nature14292, doi: 10.1038/nature14292.

Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.

Vanderlugt CL, Miller SD. Epitope spreading in immune-mediated diseases: implications for immunotherapy. Nature Reviews Immunology. 2002 2; 2(2):85–95. http://www.nature.com/articles/nri724, doi: 10.1038/nri724.

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 Research. 2011 5; 21(5):790–797. doi: 10.1101/gr.115428.110.

Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S, Gemperline DC, Augspurger T, Halchenko Y, Cole JB, Warmenhoven J, de Ruiter J, Pye C, Hoyer S, Vanderplas J, Villalba S, Kunter G, Quintero E, Bachant P, Martin M, Meyer K, et al., mwaskom/seaborn: v0.8.1 (September 2017). Zenodo; 2017. https://doi.org/10.5281/zenodo.883859, doi: 10.5281/zenodo.883859.

Supplementary information TCRDivER: T cell Receptor Diversity Estimates for Repertoires

Milena Vujović Paolo Marcatili Benny Chain Joseph Kaplinsky
Thomas Lars Andresen

1 Murine Dataset

Results concerning the murine CD4$^+$ TCR repertoire dataset following immunisation with Complete Freund's Adjuvant (CFA) with or without the additon of Ovalbumin (OVA) antigen. The dataset also contains unimminused mice. The samples were collected at 3 timepoints (5, 14 and 60 day) for immunised mice, and day 0 for unimmunised mice. Two TCR distance metrics were used a BLOSUM45 and Atchley factor based score. Each subsection will therefore be marked with the distance metric

1.1 Diversity Profiles-BLOSUM45

Diversity profiles calculated for the murine dataset with 50000 subsample size. The profiles are organised into tables according to post-immunisation sample collection time with untreated mice, day 5, day 14 and day 60 in Table 1, 2, 3 and 4, respectively.

Table 1: Untreated day 0

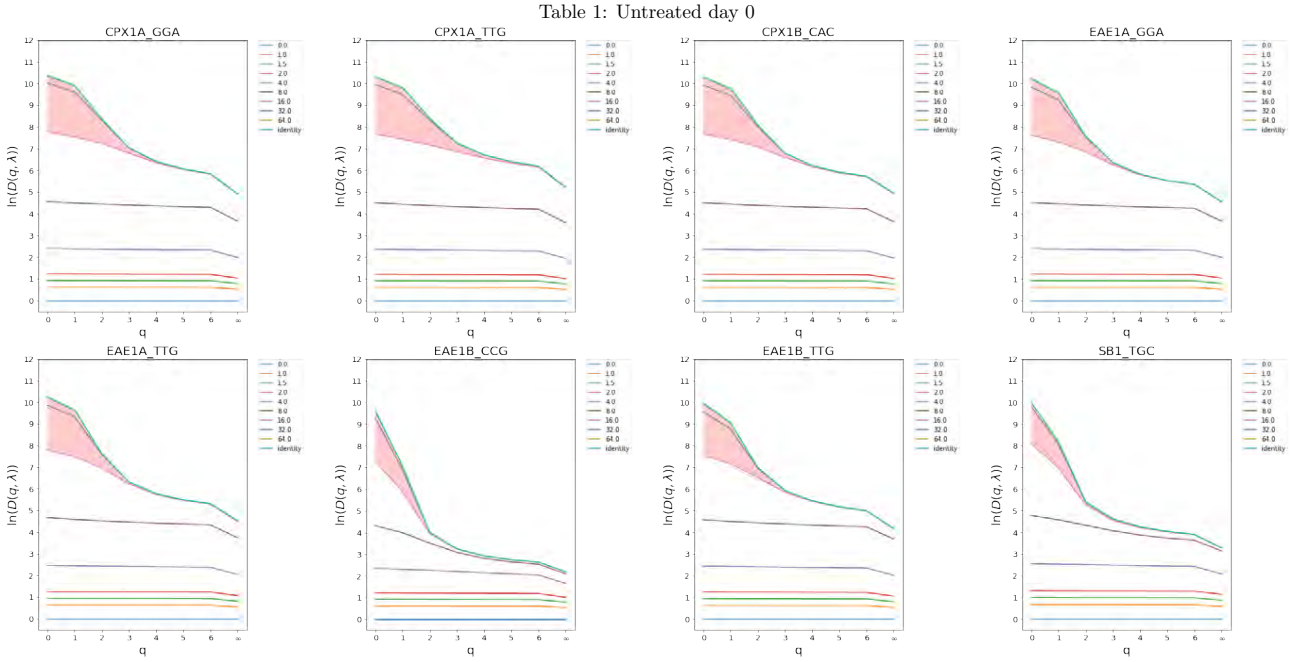


Table 2: Immunised day 5

CFA

CFA + OVA

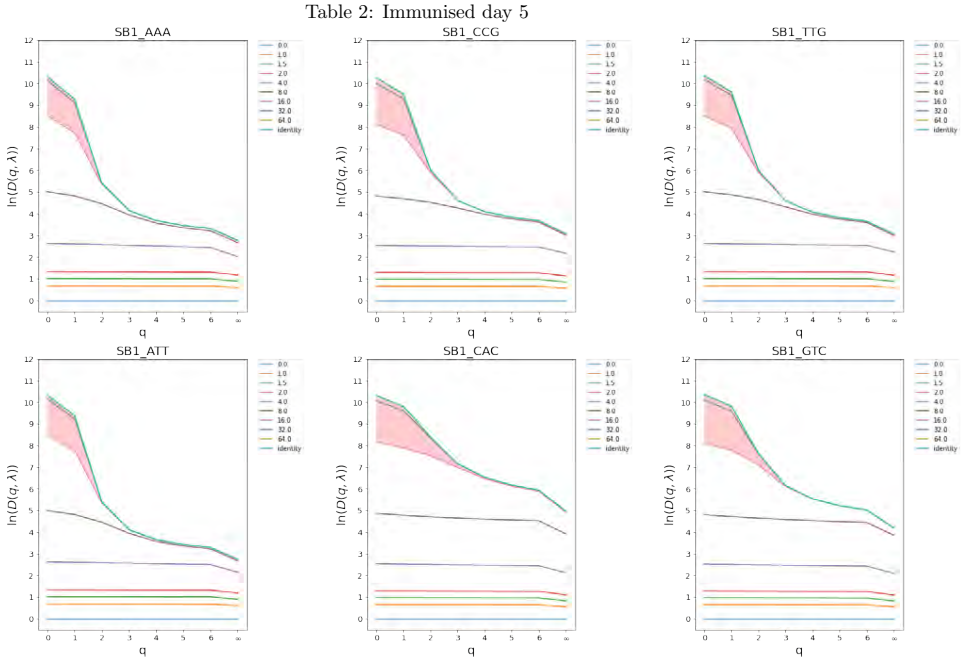


Table 3: Immunised day 14

CFA

CFA + OVA

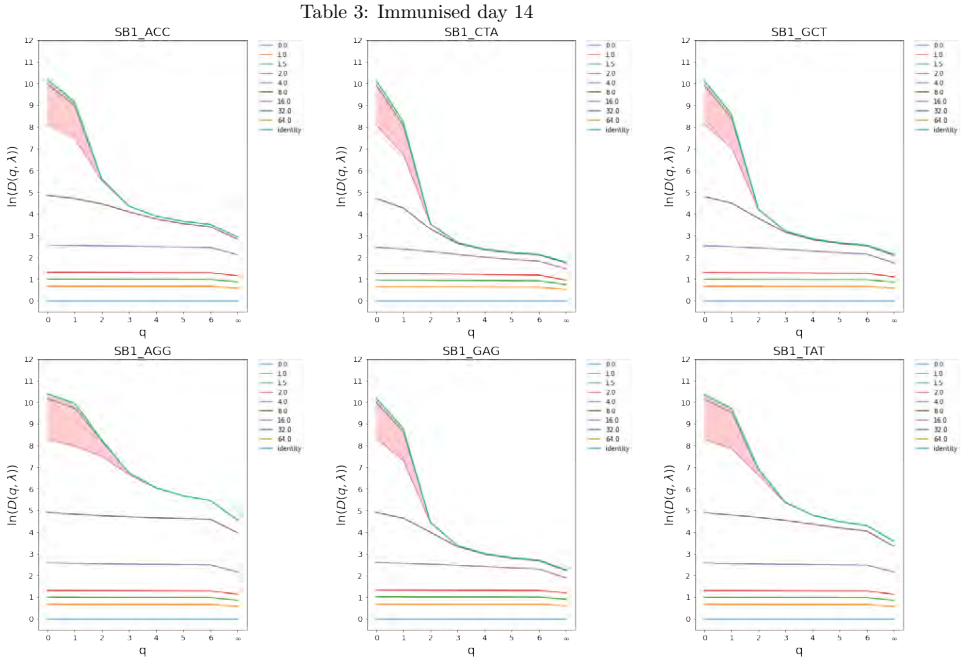
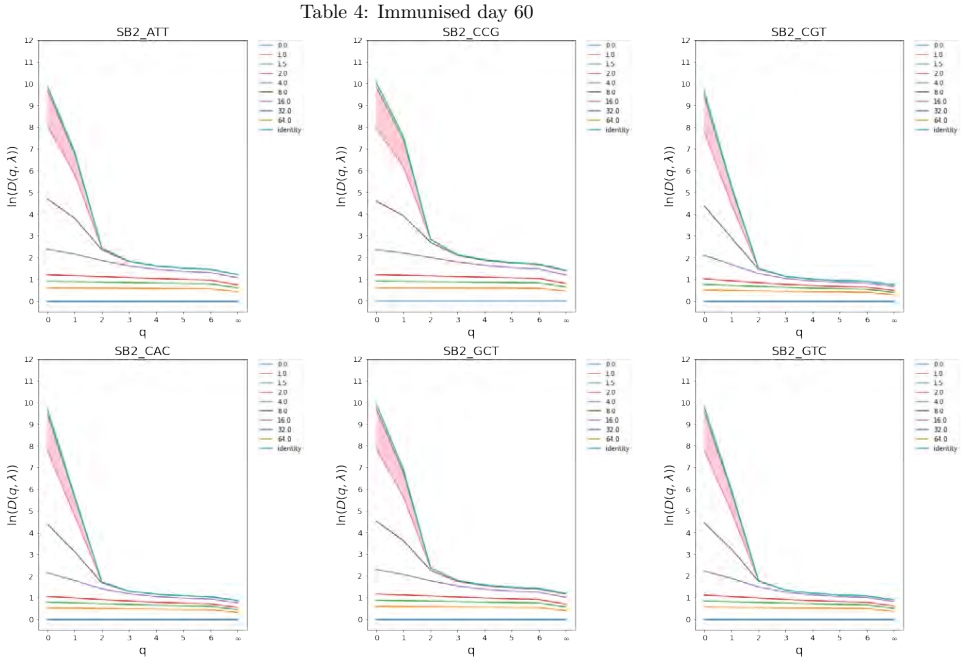


Table 4: Immunised day 60

CFA

CFA + OVA



1.2 Naive ($q = 0$) diversity profiles-BLOSUM45

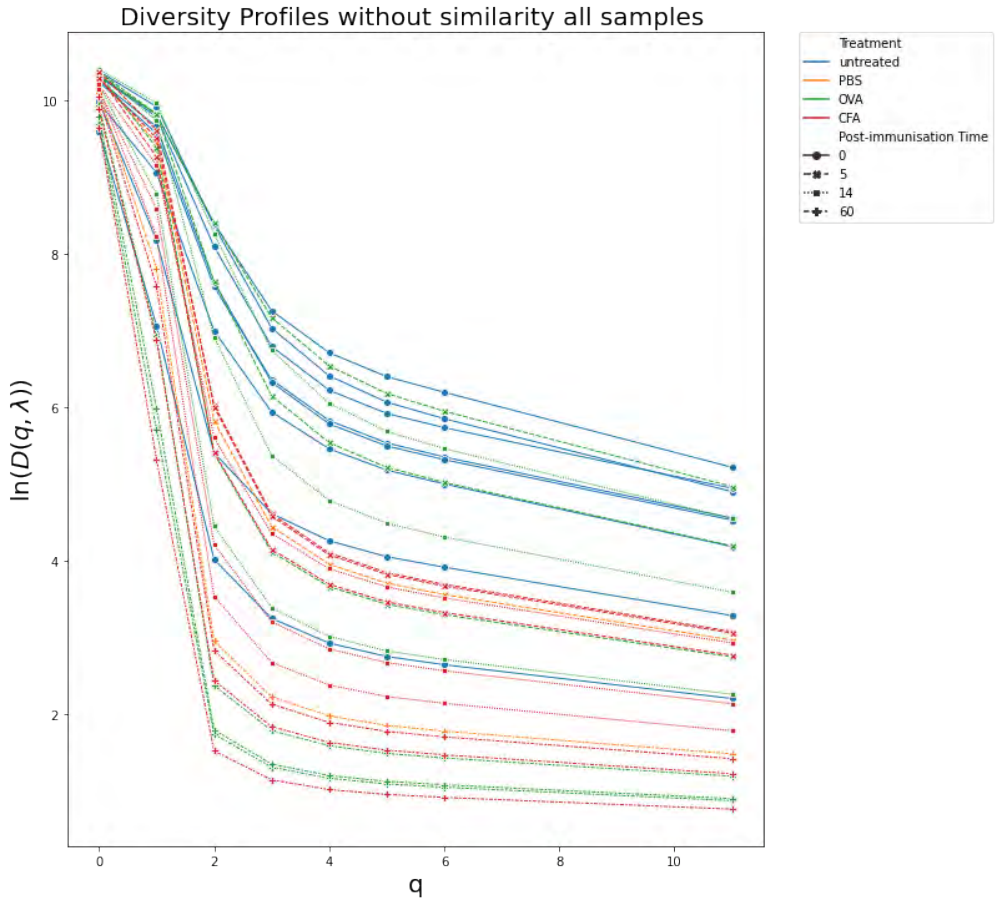


Figure 1: Naive ($q = 0$) diversity profiles plotted for all murine samples. Frequent crossings of the curves can be observed illustrating that the rank order of samples depends on the specific choice of index.

1.3 PCA on natural logarithm transformed values of true diversity-BLOSUM45

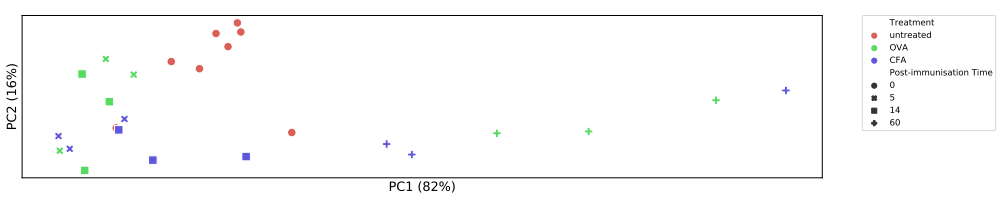
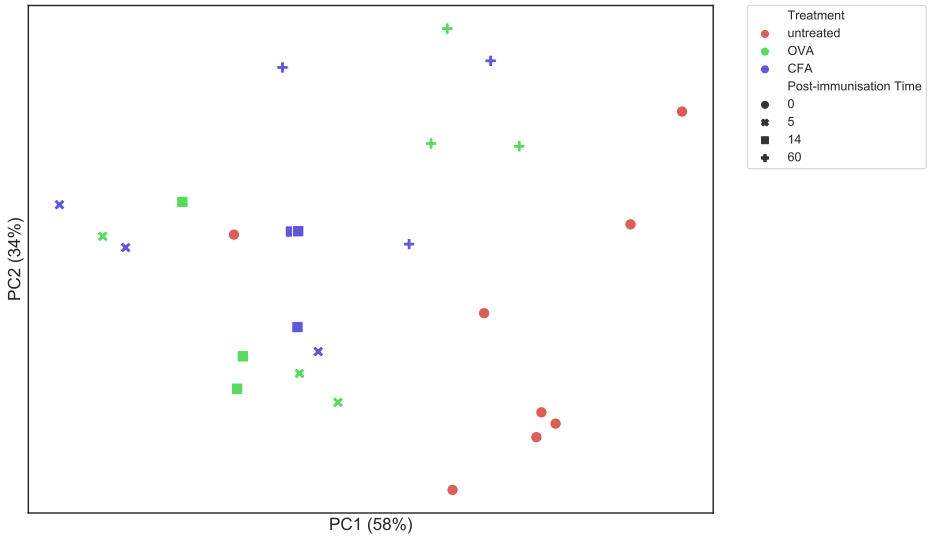
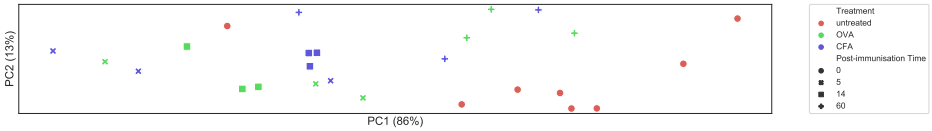


Figure 2: PCA on values of true diversity $D(q, \lambda)$. for the murine dataset. The aspect ratio corresponds to variation found by PCA.

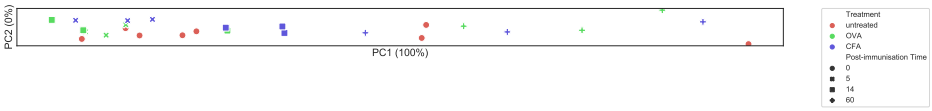
1.4 PCA on diversity values from the randomised murine dataset with random frequencies-BLOSUM45



(a)



(b)



(c)

Figure 3: Principal Components Analysis on diversity calculated for the randomised murine dataset. The aspect ratio corresponds to variation found by PCA.**a.** PCA on features extracted from the diversity profiles constructed from the true diversity $D(q, \lambda)$. **b.** PCA on values of true diversity $D(q, \lambda)$. **c.** PCA on naive diversity values $D(q)$, i.e. $\lambda =$ identity.

1.5 Diversity Profiles-Atchley factor distance

Diversity profiles calculated for the murine dataset with 50000 subsample size. The profiles are organised into tables according to post-immunisation sample collection time with untreated mice, day 5, day 14 and day 60 in Table 1, 2, 3 and 4, respectively.

922

Table 5: Untreated day 0

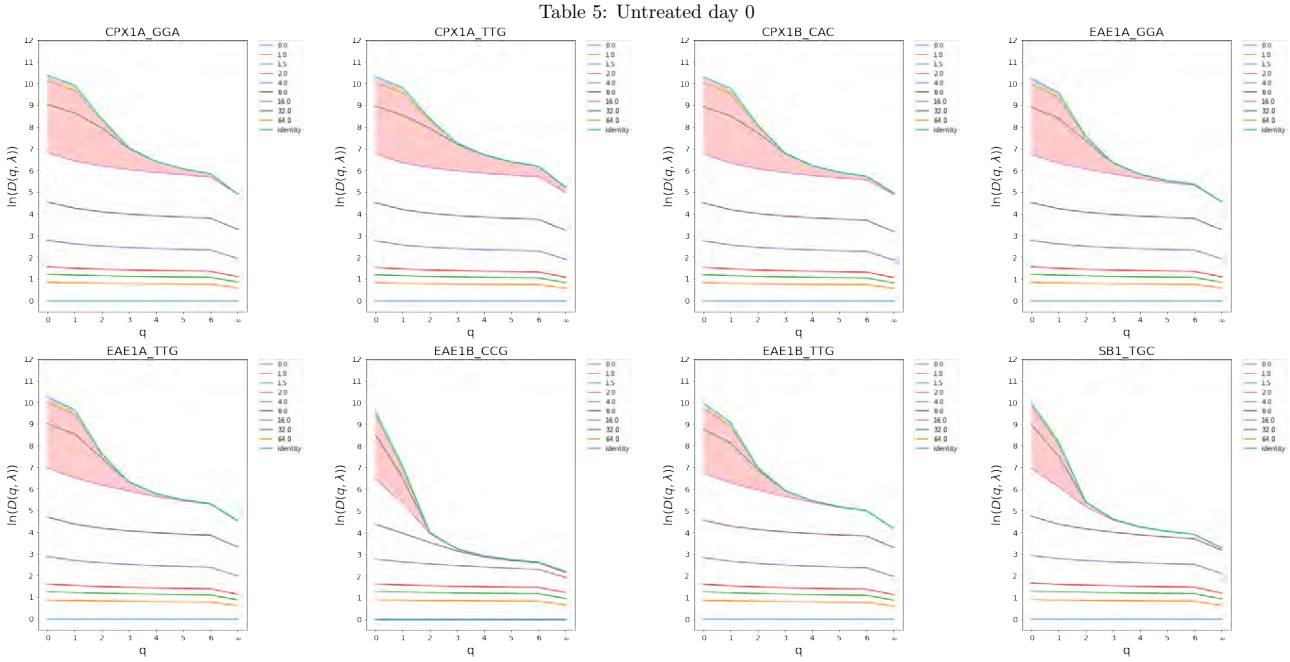


Table 6: Immunised day 5

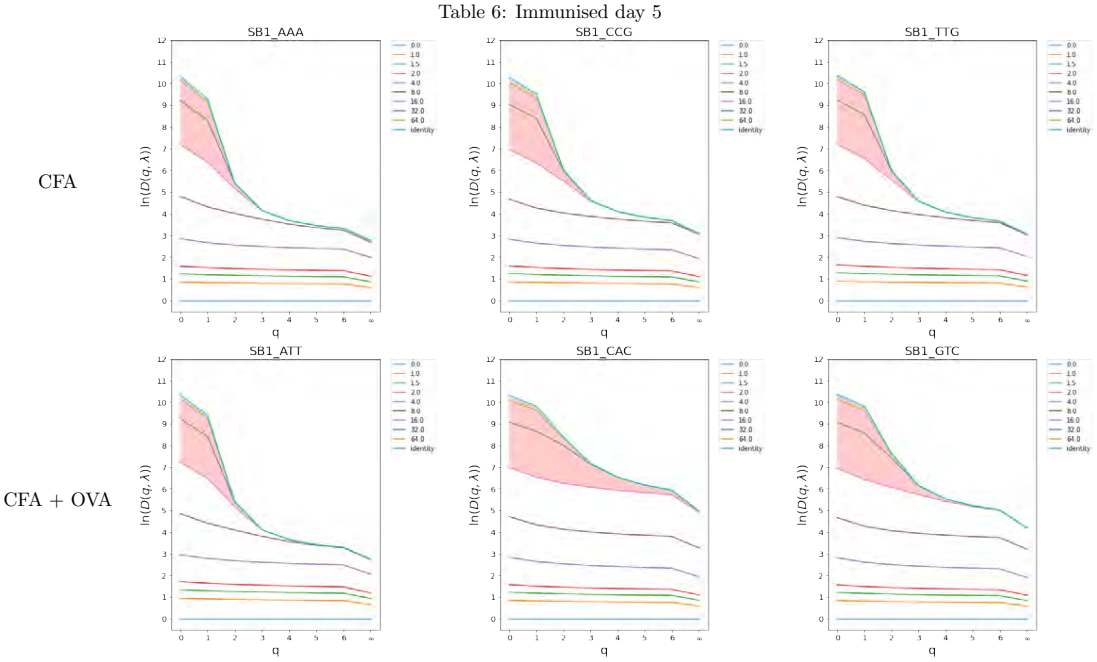


Table 7: Immunised day 14

CFA

CFA + OVA

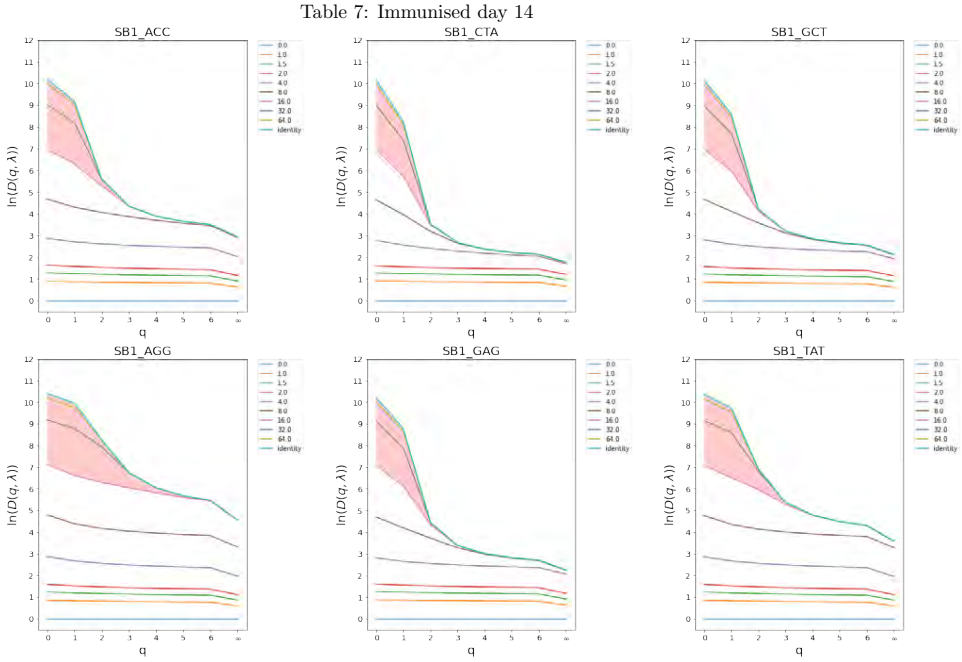
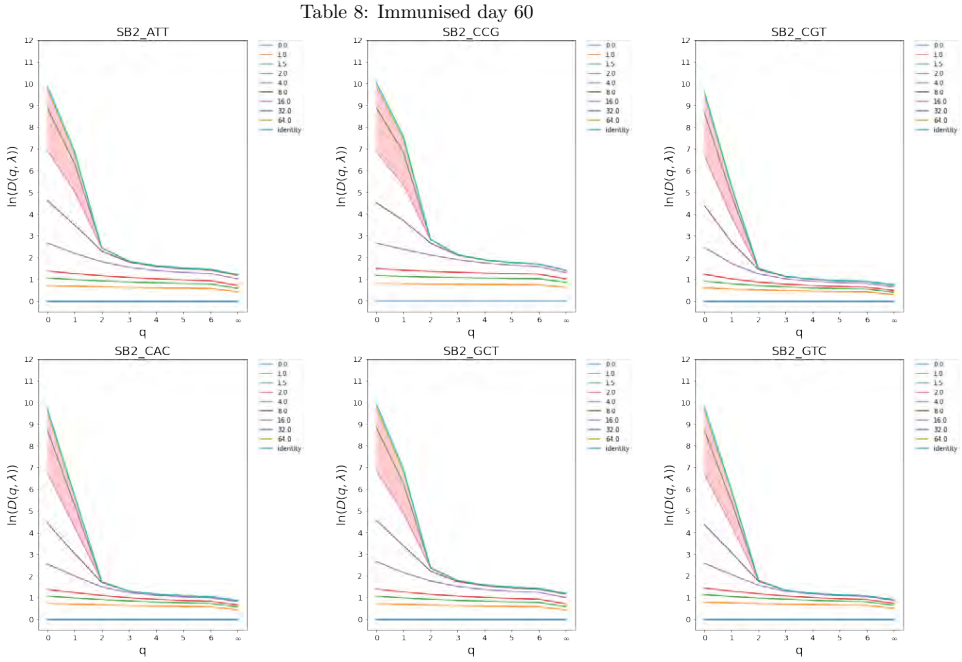


Table 8: Immunised day 60

CFA

CFA + OVA



1.6 Naive ($q = 0$) diversity profiles–Atchley factor distance

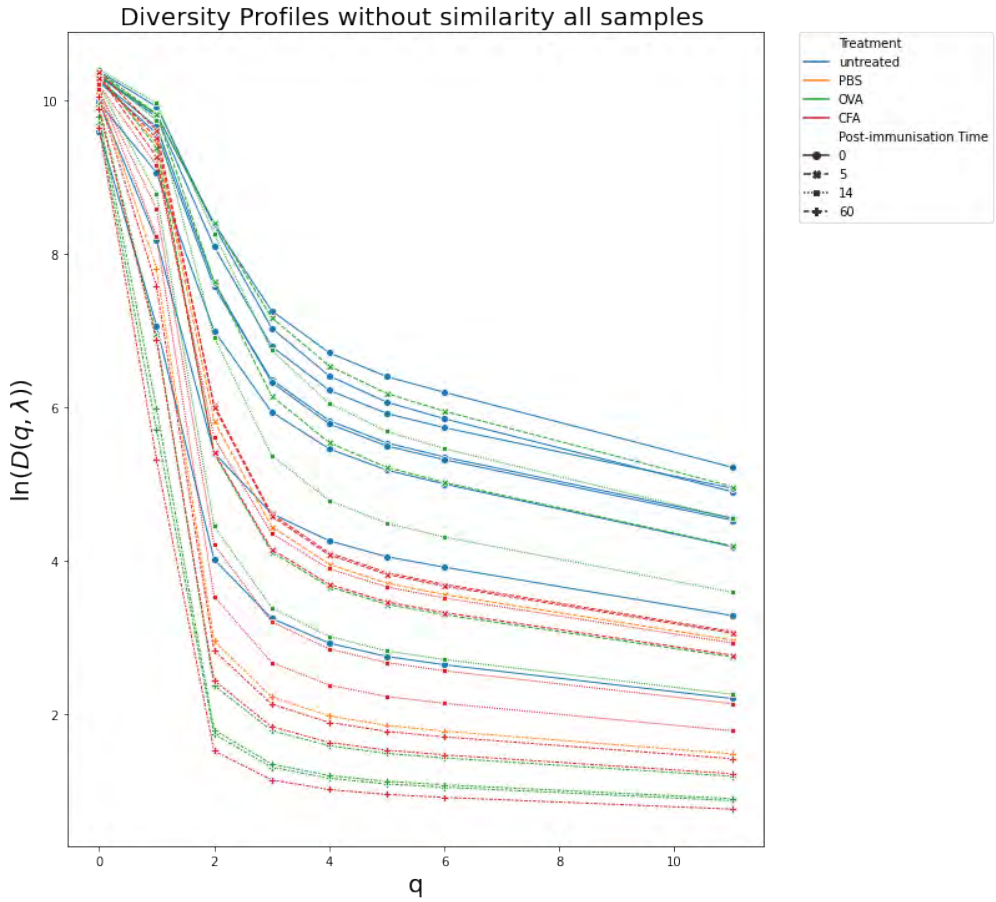


Figure 4: Naive ($q = 0$) diversity profiles plotted for all murine samples. Frequent crossings of the curves can be observed illustrating that the rank order of samples depends on the specific choice of index.

1.7 PCA on natural logarithm transformed values of diversity–Atchley factor distance

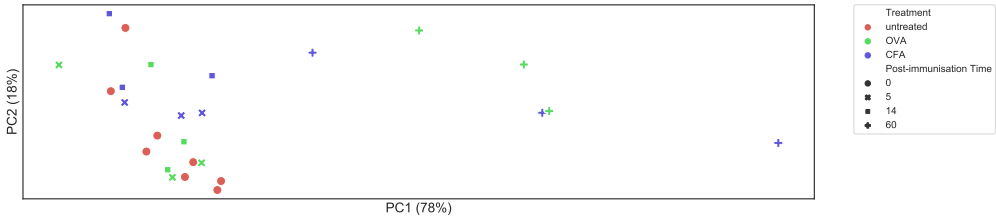
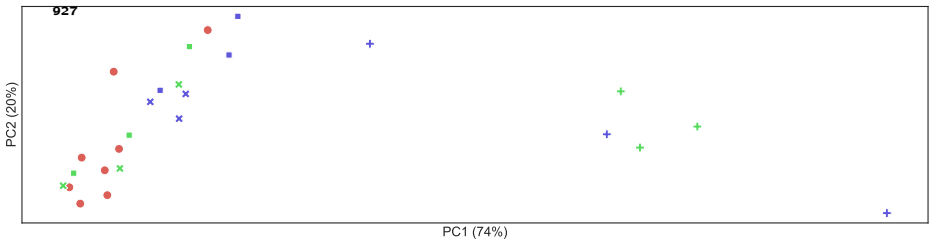
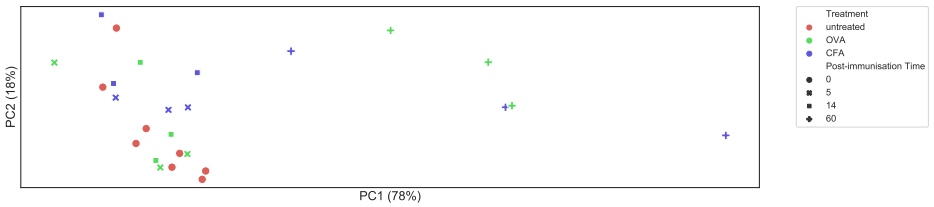


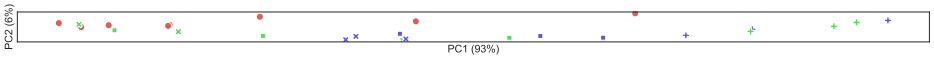
Figure 6: PCA on values of true diversity $D(q, \lambda)$. for the murine dataset. The aspect ratio corresponds to variation found by PCA.



(a)



(b)



(c)

Figure 5: Principal Components Analysis on diversity calculated for the murine dataset using the Atchley Factor distance for TCRs. The aspect ratio corresponds to variation found by PCA.**a.** PCA on features extracted from the diversity profiles constructed from the true diversity $D(q, \lambda)$. **b.** PCA on values of true diversity $D(q, \lambda)$. **c.** PCA on naive diversity values $D(q)$, i.e. $\lambda =$ identity.

1.8 Trends of three features extracted from divPs versus timepoint and treatment regime-Atchley factor distance

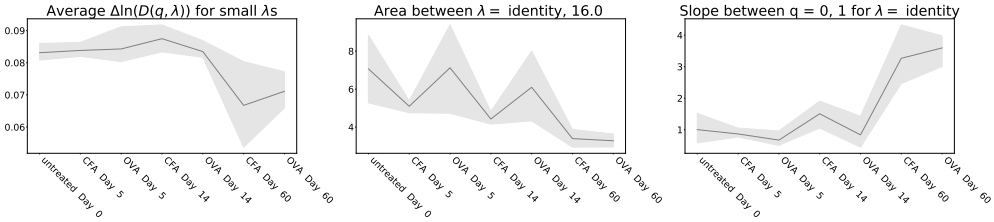
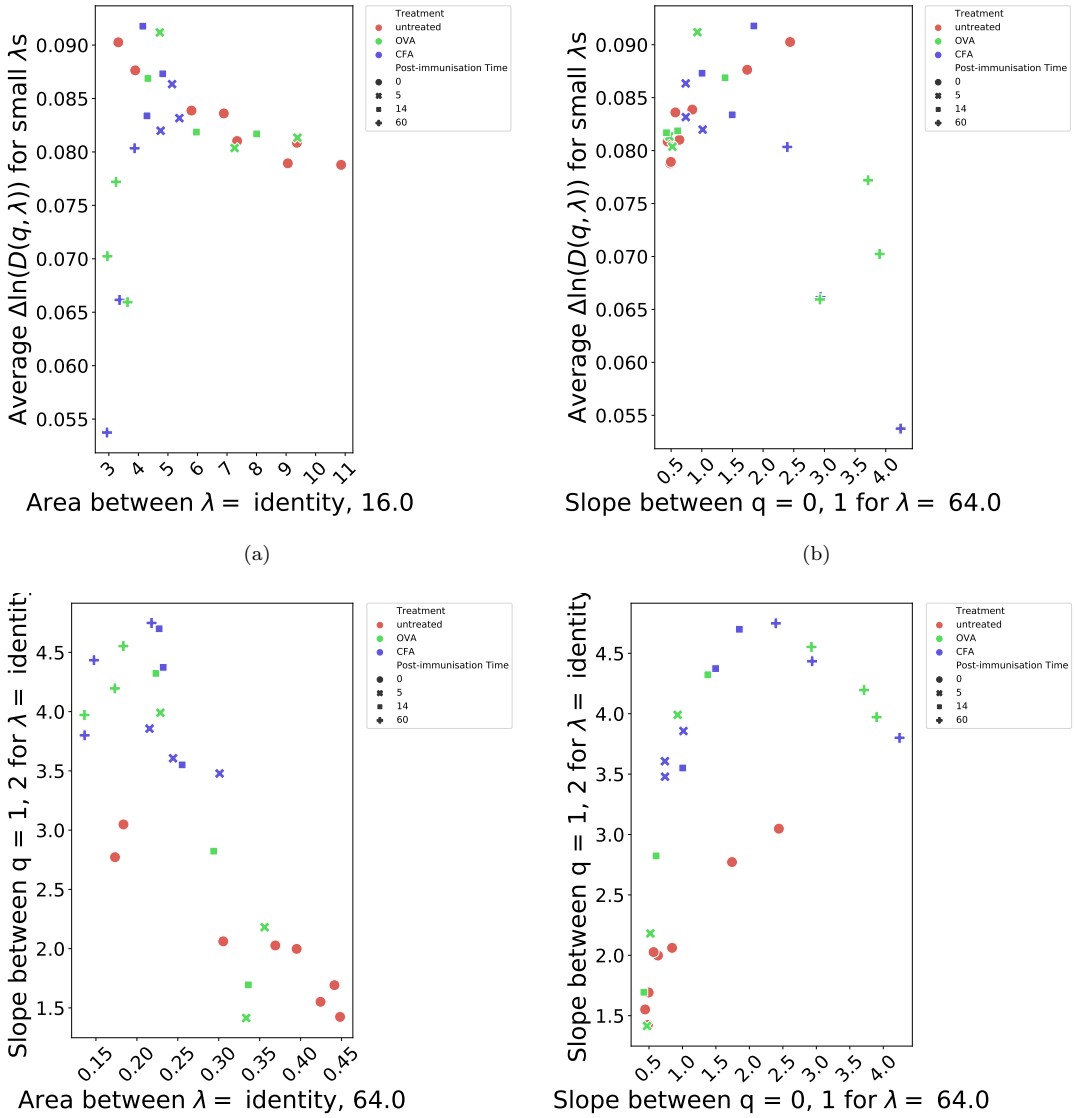


Figure 7: Trends of three features extracted from divPs are shown versus the treatment regime and timepoints ending with the latest timepoint. The features are, from left to right: average $\Delta \ln D(q, \lambda)$ for small λs, between curves of $\lambda =$ identity and 16.0 and slope of $q = 0 \rightarrow 1$ for value of λ identity. The line connects the mean values of the features for all samples within a group and the shaded area represents the confidence interval.

1.9 DivP features relationships -Atchley factor distance



(a) (b) (c) (d)

Figure 8: Graphs showing relationships between some of the divP features. **a.** average $\Delta \ln D(q, \lambda)$ for small λs is shown versus the area between curves of $\lambda =$ identity and 16.0; **b.** average $\Delta \ln D(q, \lambda)$ for small λs is shown versus the slope of $q = 0 \rightarrow 1$ for value of λ 64.0; **c.** slope of $q = 1 \rightarrow 2$ for value of λ identity (i.e. naive diversity) is shown versus the area between curves of $\lambda =$ identity and **d.** 64.0; slope of $q = 1 \rightarrow 2$ for value of λ identity (i.e. naive diversity) is shown versus the slope of $q = 0 \rightarrow 1$ for value of λ 64.0.

2 Human Dataset

Results concerning the human TCR repertoire dataset following combination therapy (RT and anti-CTLA4 blockade ipilimumab. The patients were stratified into RECIST response criteria and by time of sample collection. We show the results in the same order.

2.1 Diversity Profiles

Diversity profiles calculated for the human dataset with 50000 subsample size. The profiles are organised according to RECIST criteria and timepoint in Tables 5 to 12 for progressive disease (PD) day 0 and 22, stable disease (SD) day 0 and 22, partial responders (PR) day 0 and 22 and complete responders (CR) day 0 and 22, respectively. The distance metric used in estimating diversity was based on the BLOSUM45 alignment.

Table 9: Progressive Disease (PD) Day 0

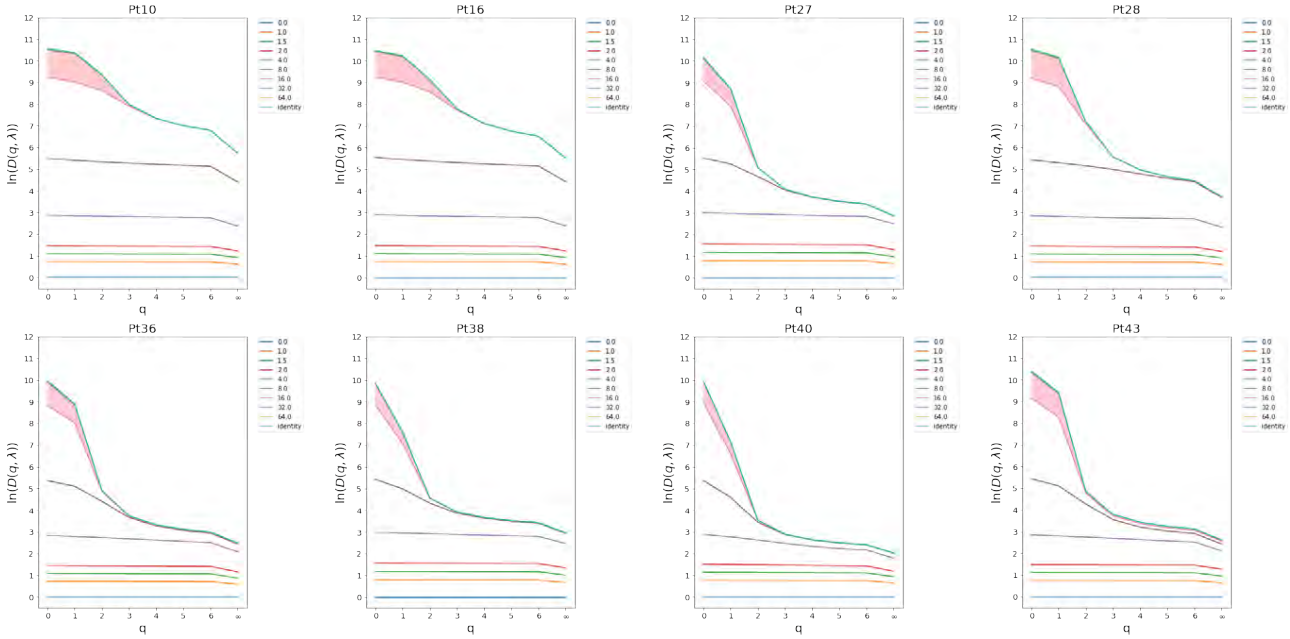
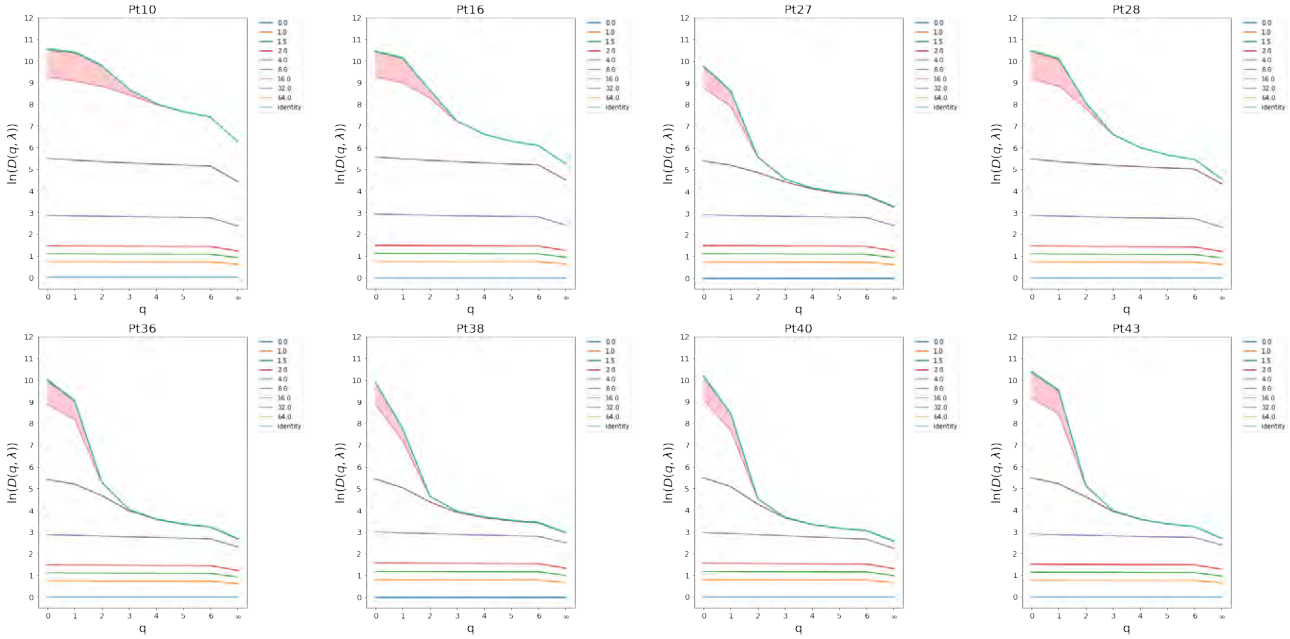
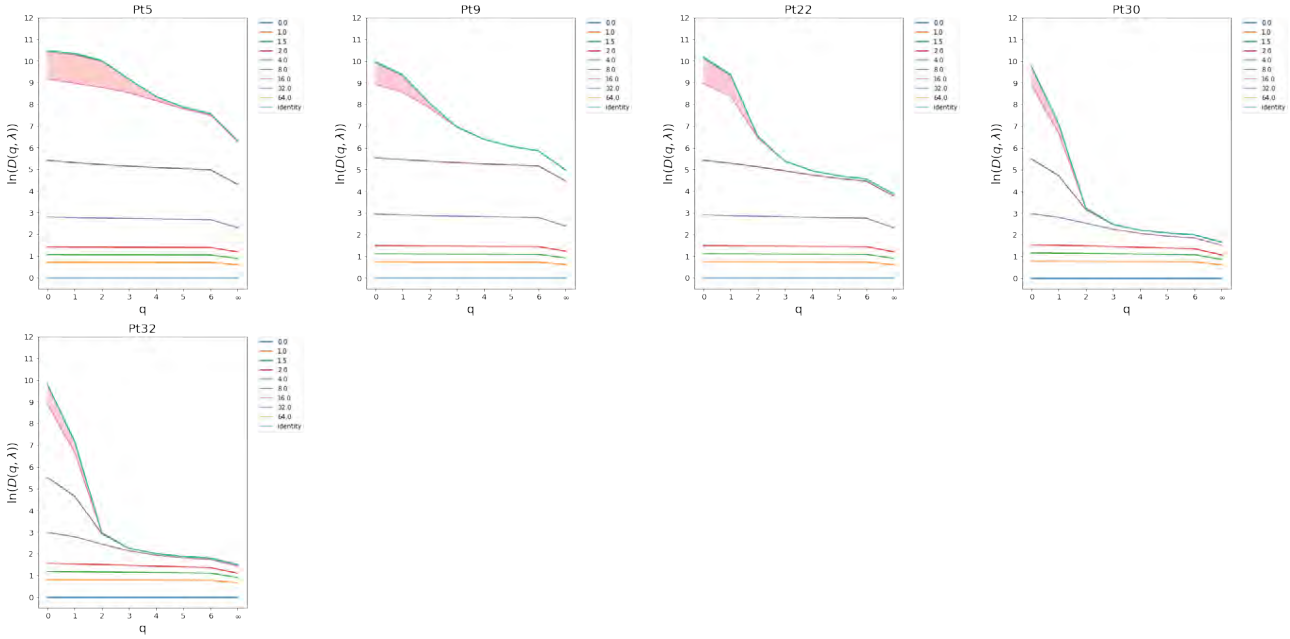


Table 10: Progressive Disease (PD) Day 22



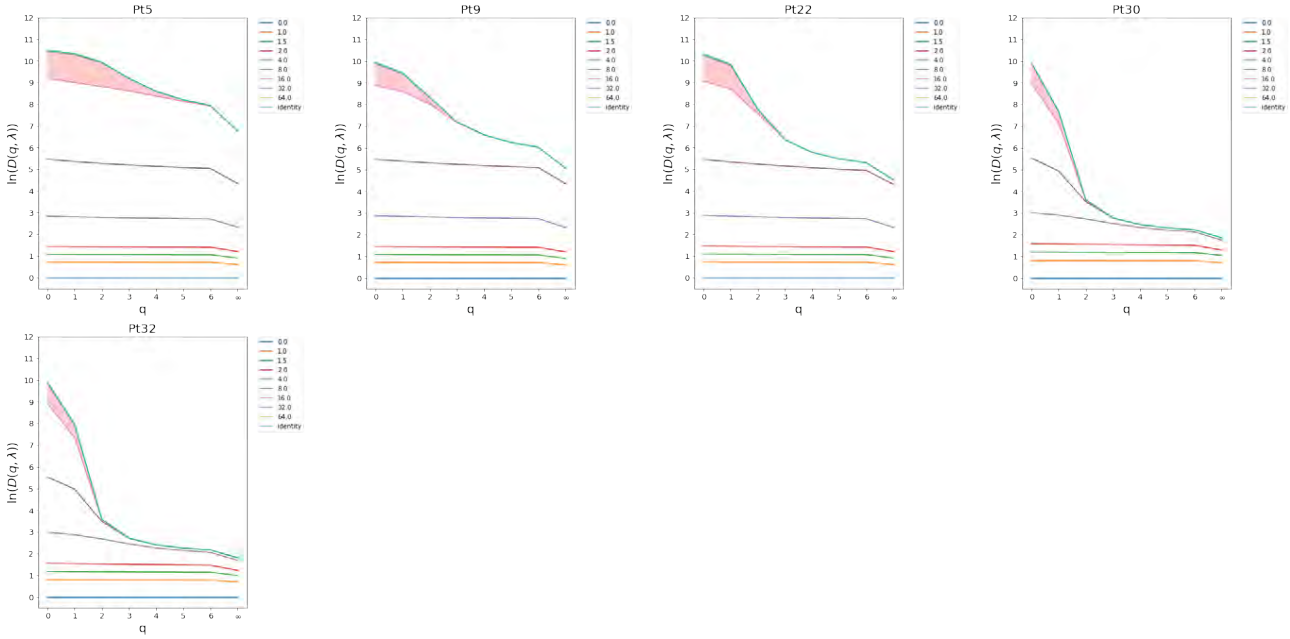
18

Table 11: Stable Disease (SD) Day 0



933

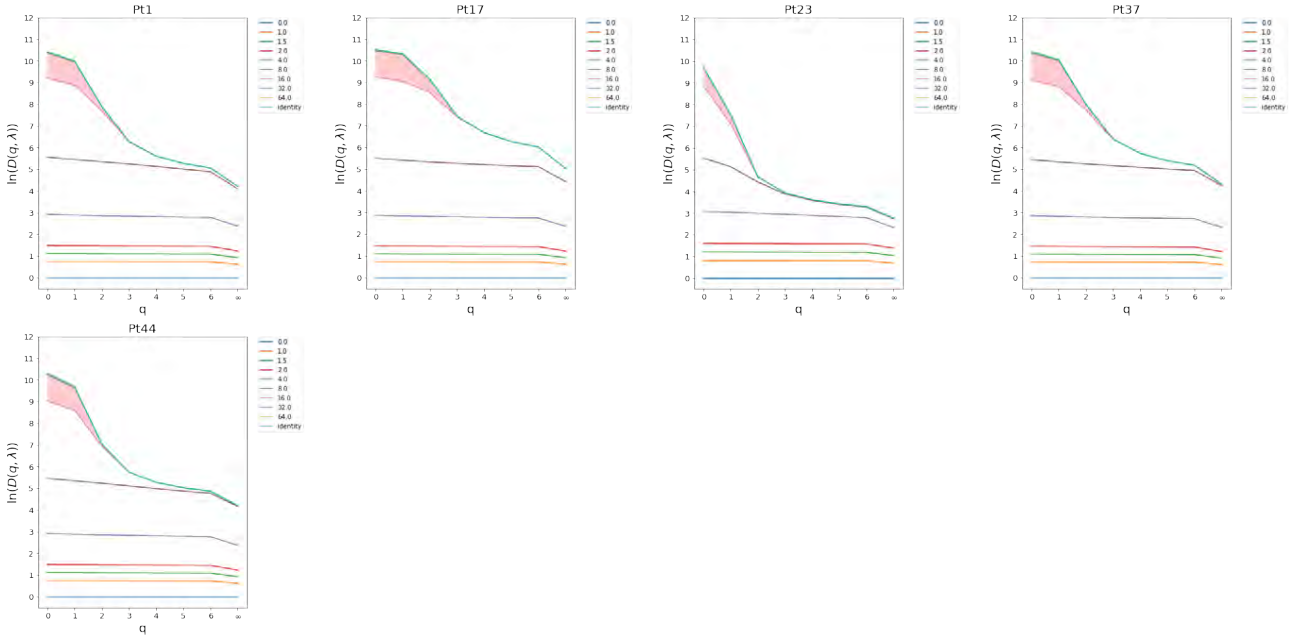
Table 12: Stable Disease (SD) Day 22



20

934

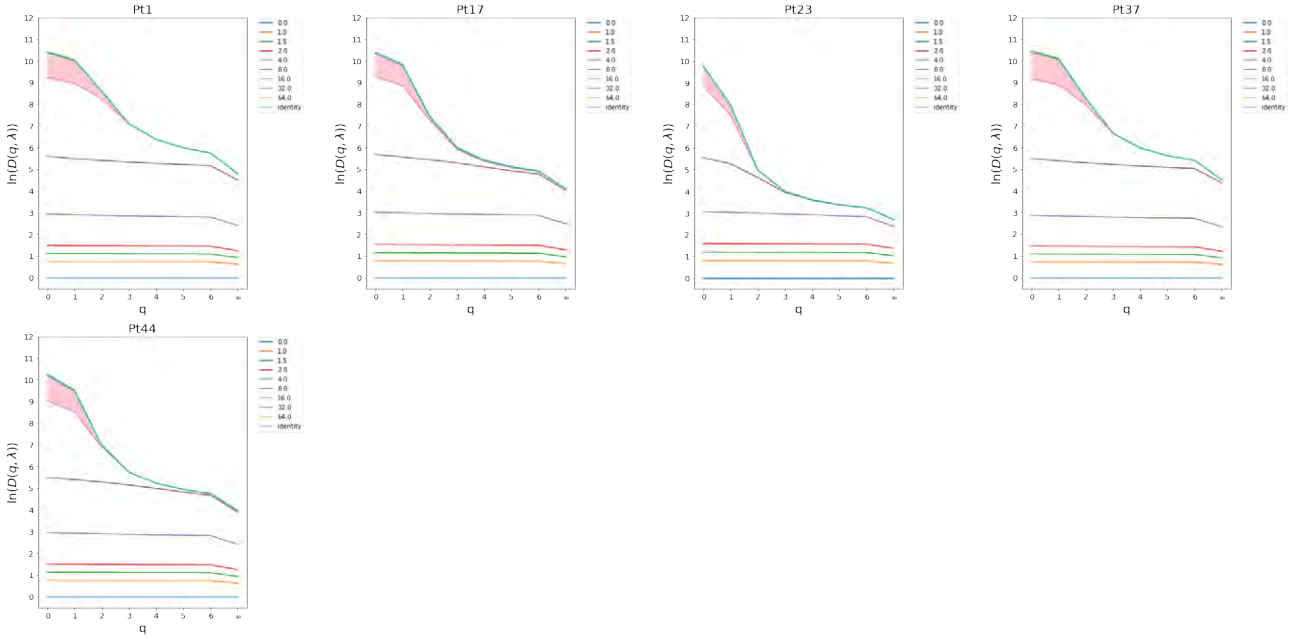
Table 13: Partial Responders (PR) Day 0



21

935

Table 14: Partial Responders (PR) Day 22



22

936

Table 15: Complete Responders (CR) Day 0

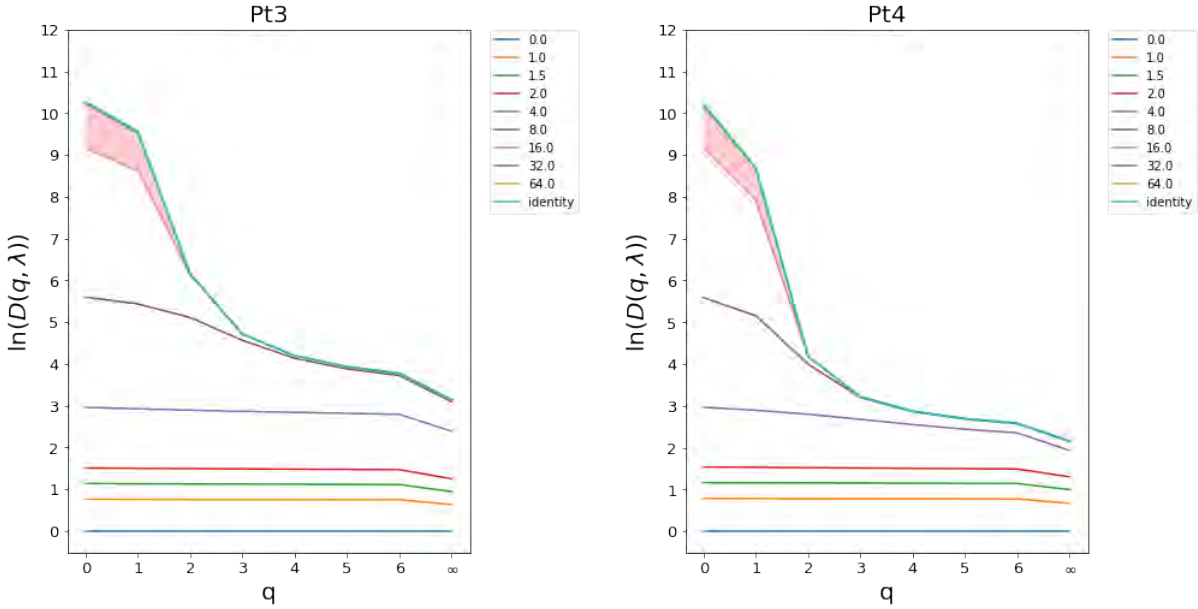
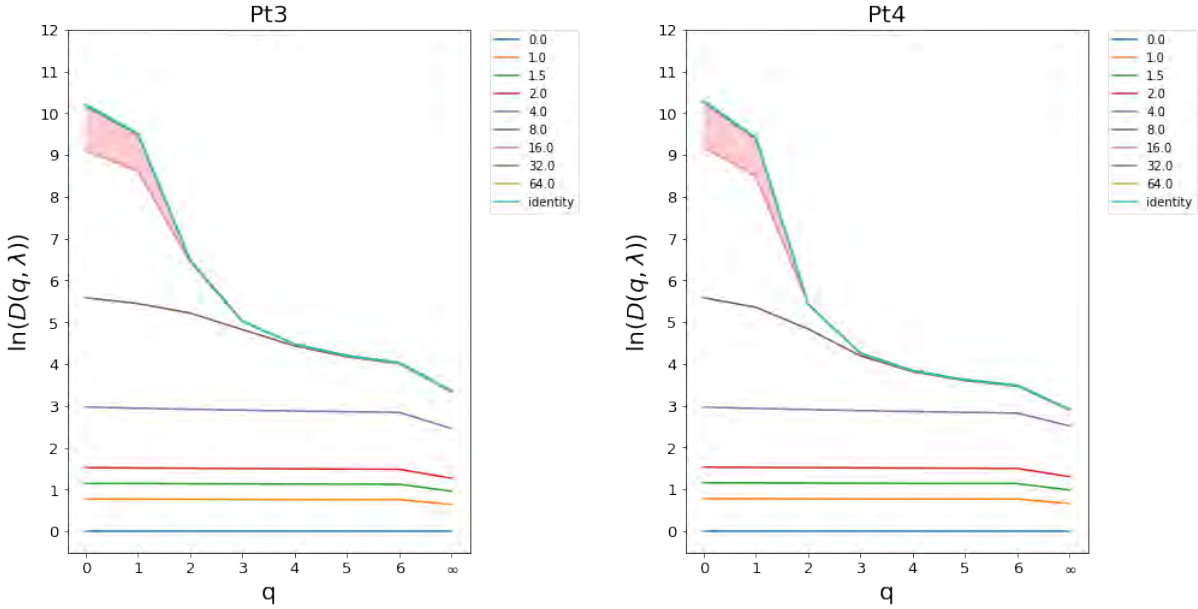
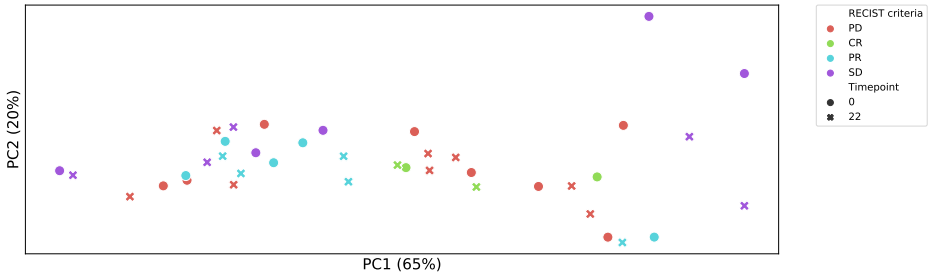


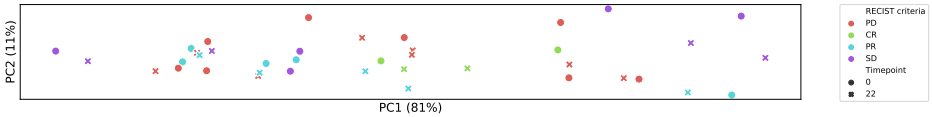
Table 16: Complete Responders (CR) Day 22



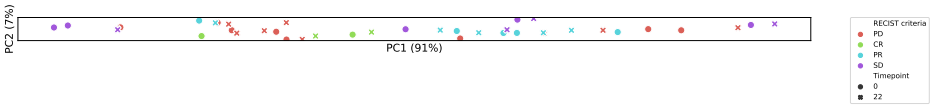
2.2 PCA on diversity profiles



(a)



(b)



(c)

Figure 9: Principal Components Analysis on diversity calculated for the human dataset.The aspect ratio corresponds to variation found by PCA. **a.** PCA on features extracted from the diversity profiles constructed from the true diversity $D(q, \lambda)$. **b.** PCA on values of true diversity $D(q, \lambda)$. **c.** PCA on naive diversity values $D(q)$, i.e. $\lambda =$ identity.