ABSTRACT

Motivation: The effectiveness of vertebrate adaptive immunity depends crucially on the establishment and maintenance of extreme diversity in the antigen receptor repertoire. Spectratype analysis is a method used in clinical and basic immunological settings in which antigen receptor length diversity is assessed as a surrogate for functional diversity. The purpose of this paper is to describe the systematic derivation and application of statistical methods for the analysis of spectratype data.

Results: The basic probability model used for spectratype analysis is the multinomial model with \( n \), the total number of counts, indeterminate. We derive the appropriate statistics and statistical procedures for testing hypotheses regarding differences in antigen receptor distributions and variable repertoire diversity in different treatment groups.

We then apply these methods to spectratype data obtained from several healthy donors to examine the differences between normal CD4\(^+\) and CD8\(^+\) T cell repertoires, and to data from a thymus transplant patient to examine the development of repertoire diversity following the transplant.

Availability: http://www.duke.edu/~kepler/spa.html

Contact: kepler@duke.edu

1 INTRODUCTION

1.1 Biology of antigen receptors

The immune system of all jawed vertebrates has two major divisions, denoted as the innate and adaptive systems, which are distinguished from each other by the absence or presence, respectively, of randomly rearranged antigen receptors. Pathogenic microorganisms display enormous molecular variability compared with their vertebrate hosts, owing to their short generation times and large populations. The strategy of the adaptive immune system, therefore, is to counter this extraordinary variability by using somatic diversification, which occurs at time scales typical of individual somatic cell turnover, rather than the much longer time scales of host germline turnover, or generations. These antigen receptors are the T cell receptors (TCR) and immunoglobulins (Ig), borne by T and B cells, respectively. For an excellent introduction to the biology of the immune system, see Janeway et al. (2005).

Both T cell and B cell receptors are encoded by gene segments that must be rearranged on the chromosome to produce a complete productive gene. This process involves the stochastic selection of gene segments from each of two or three libraries (depending on the type of chain in question) and the further stochastic selection of specific recombination points given these segments, as well as diversity derived from additional non-templated nucleotides. The region containing these highly variable junctions is the third of three complementarity-determining regions (CDRs) that are seen crystallographically to contact antigen. The sources of TCR diversity are naturally broken down hierarchically into gene segment family (library), segment within family, CDR3 length and CDR3 nucleotide diversity.

The immune response to infection involves the manifold expansion of a small number of T cell clones, and consequently, a decrease in the TCR repertoire diversity. Severe immune disruptions, either caused by genetic lesions disrupting normal developmental processes or acquired during life as a result of lymphoma or leukemia, result in loss of antigen receptor diversity or of one or more antigen receptors altogether.

Transplantation of bone marrow or of primary lymphoid organs is performed to establish or re-establish the missing populations of T or B cells. These interventions are often followed up by spectratype analyses to monitor the progress of the diversification of the affected cellular populations. Spectratype analysis (Cochet et al., 1992; Pannetier et al., 1993, 1997) provides information on antigen receptor diversity at the level of CDR3 length. The point is not that length diversity itself is of particular relevance (though it might be), but that length heterogeneity is very likely to be representative of overall sequence heterogeneity.

Spectratyping has proven to be a valuable method for the monitoring of antigen receptor repertoire diversity subsequent to lymphoid transplantation and to infection (Cochet et al., 1992; Pannetier et al., 1993, 1997; Bousso et al., 2000; Sarzotti et al., 2003; Markert et al., 2003). The purpose of this paper is to present a set of statistical methods that bring further power and flexibility to the use of this assay. Previous work on bringing statistical methods to spectratype analysis include Collette and Siz (2002) and Collette et al. (2003), who have developed statistical methods and implemented them in an Excel package, and Gorochov et al. (1998). These methods, while clearly valuable, are based on \textit{ad hoc} measures of spectratype differences. Our aim in the present paper, in contrast, is to start from first principles and derive the statistical measures and tests most natural for the assay itself.

The methods developed in this paper can be applied to both TCR and Ig; for the sake of clarity and simplicity, we will specifically describe the procedures and background in terms of TCR.

© The Author 2005. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org
Fig. 1. Schematic diagram illustrating the spectratyping assay. T cells are fractionated from a sample of peripheral blood, the mRNA is isolated and TCRB-family PCR primers are used to amplify part of the rearranged TCR containing the complete CDR3. This heterogeneous mixture of DNA molecules is then size-separated by electrophoresis. Finally, the quantity of material deposited in each band is quantified by densitometry, and the spectratype trace is produced.

The paper is structured as follows. First, we describe the molecular biological assay itself and present examples of the data generated. Next, we derive the probability density functions (pdfs) and compute their relevant characteristics. We next develop two fundamental statistical procedures and illustrate their use with data from healthy human subjects and from human thymus transplant patients. Finally, we discuss the mathematical relationship between the TCR repertoire diversity and the primary statistic, d, derived in our analysis.

1.2 Spectratype analysis

The spectratype assay begins with the collection of a peripheral blood sample from the subject, and the isolation of CD4 or CD8 T cells from it, although total CD3 or PBMC can be used. PCR is used to specifically replicate the variable-length region (CDR3) of the rearranged TCR variable region beta chain (TCRBV). Primers specific to individual TCRBV families (or family subsets) are used to provide independent spectratypes for each. The resulting mixture of CDR3 replicons is size-separated by electrophoresis, and quantified by densitometry (Fig. 1).

Spectratypes are presented as histograms of the number of T cells bearing receptors versus receptor length for each of the 24 TCR V beta families tested. While the utility of these histograms for the monitoring of TCR diversity is widely accepted, the theoretical relationship between these histograms and the underlying receptor diversity has not been described. Spectratype histograms have largely been analyzed by subjective classification. We have undertaken this study to allow the quantitative and objective analysis of spectratype data, and to render the relationship between spectratype and receptor diversity unambiguous.

2 METHODS

2.1 Experimental methods

All of the data discussed in this paper were collected from human subjects as follows.

The immunoscope analysis of PCR-amplified products is performed as described by Pannetier et al. (1997). Briefly, RNA is extracted from PBMC samples (2–10 × 10^6 cells/sample) using Triazol (Life Technology, Gaithersburg, MD) and reverse transcribed to single-stranded cDNA with AMV reverse transcriptase using an oligo (dT) primer according to the manufacturer’s protocol (Promega, Madison, WI). The newly synthesized cDNA is then used as a template for 24 PCRs. The PCRs are carried out in 20 μl volumes by standard procedures (Pannetier et al., 1997), using 24 Vbeta primers and one Cbeta 1 primer. The samples are subjected to 40 cycles of denaturation (25 s at 94°C), annealing (45 s at 60°C) and elongation (45 s at 72°C). After the last cycle, a final elongation step (5 min at 72°C) is performed. The PCR products are visualized on a 1.5% agarose gel by ethidium bromide staining before using 2 μl of the amplified products for a run-off elongation reaction with a fluorescent (G-FAM) C beta 2 primer, as described previously (Pannetier et al., 1997; Bouso et al., 2000; Sarzotti et al., 2003; Markert et al., 2003). The elongation products are then run on an ABI 3100 Genetic Analyzer. Fluorescence-labeled size markers (Applied Biosystems, CA), are loaded with the run-off products. After analysis on an automated sequencer (Applied Biosystems) size determination of the run-off products and the analysis of the CDR3 region products are performed using the GeneScan software.

2.2 Computational methods

Software for the numerical computations was written in Fortran90. Web-based implementation is described by M. He, J.K. Tomfohr, B.H. Devlin, M. Sarzotti, M.L. Markert and T.B. Kepler (submitted for publication) Additional routine statistical analyses were performed using Splus 6.1 (Insightful, Inc.).

3 APPROACH

In this section, we derive the pdfs required to devise the statistical methods for data analysis, identify the statistics that will most directly address the issues of interest in these analyses, and compute the expected values and sampling distributions of these statistics.

3.1 Derivation of probability distributions

In order to analyze spectratype data appropriately, we need to determine the relevant probability distributions. The densitometric intensity of the peak at any given electrophoretic displacement is, ideally, proportional to the number of TCR with CDR3 of the corresponding length. If we could simply count these TCR, a multinomial model would be appropriate. In our case, however, we do not measure the absolute number of T cells in the sample. In this subsection, we derive a family of density functions obtained from the multinomial, but in which n, corresponding to the total number of T cells sampled, is uncertain.

3.1.1 Multinomial distribution with n uncertain

To establish notational conventions, suppose m is a L-dimensional multinomial random variable with population parameter vector q. Then, the probability mass function (pmf) for m is given by

\[ f_m(m|n, q) = C_n(m) \prod_{i=1}^{L} q_{i}^{m_i}, \]

where \( C_n(m) \) is the multinomial coefficient

\[ C_n(m) = \frac{\Gamma(n+1)}{\prod_{i=1}^{L} \Gamma(m_i+1)} \]

for \( \sum_{i=1}^{L} m_i = n \) and zero otherwise.
3.1.2 Density functions for relative frequencies

The absolute number $m_i$ of T cells with CDR3 length $i$ is not measured; only the relative frequencies $r_i = m_i/n$ are. For $n$ sufficiently large, the relative frequencies may be treated as continuous random variables with pdf determined by a transformation of variables (Gardiner, 2004)

$$f_r(r|n, q) = n^{L-1}C_n(nr) \prod_{i=1}^{L} q_i r_i^{n_i-1}$$

where $n^{L-1}$ is the value of the determinant of the Jacobian transformation matrix.

We assume that for each $i$, $nr_i$ is large enough for Stirling’s approximation (Weisstein, 2004a, http://mathworld.wolfram.com/StirlingsApproximation.html) to provide an adequate representation of the gamma function, and use it to simplify the multinomial coefficient

$$\log C_n(nr) \equiv \log \Gamma(n+1) - \sum_{i=1}^{L} \log \Gamma(r_i n + 1)$$

$$= \sum_{i=1}^{L} \left\{ -nr_i \log r_i - \frac{1}{2} \log r_i \right\} - \frac{L-1}{2} \log 2\pi n + O\left(\frac{1}{n}\right).$$

Using this expression in Equation (1) we obtain

$$f_r(r|n, q) = \frac{n^{L-1}e^{-nD(q|q)}}{(2\pi)^{L-1} \prod_{i=1}^{L} r_i} \delta(r-1),$$

where $\delta$ is the Dirac delta and we use the dotted index convention: $\equiv \sum r_i$, and $D(q|q)$ is the Kullback–Leibler divergence or relative entropy (Kullback and Leibler, 1951),

$$D(q|q) = \sum_{i=1}^{L} q_i \log \frac{q_i}{r_i}.$$  

We determine the pdf for $D(q|q)$ by first computing its cumulant generating function (cgf) (Weisstein, 2004b, http://mathworld.wolfram.com/Cumulant-GeneratingFunction.html)

$$h(s) \equiv \log \int d^L r e^{sD(q|q)} f_r(r|n, q)$$

$$= \lambda (\log n - \log 2\pi) + \log \int d^L r \frac{e^{-(n-1)D(q|q)}}{\prod_{i=1}^{L} r_i} \delta(r-1)$$

$$= -\lambda \log \left(1 - \frac{s}{n}\right).$$

We recognize this cgf as that of a gamma random variable with shape parameter $\lambda \equiv (L-1)/2$ and scale parameter $1/\lambda$.

$$f_D(D|n) = \frac{n^\lambda}{\Gamma(\lambda)} D^{\lambda-1} e^{-nD}. \quad (2)$$

When $n$ is uncertain but can be described as a gamma random variable with shape parameter $\gamma$ and scale parameter $1/\xi$, we marginalize Equation (2) over $n$ to get

$$f_D(D|\gamma, \xi) = \frac{\Gamma(\gamma + \lambda)}{\Gamma(\lambda)} \xi D^{\lambda-1}. \quad \frac{1}{\lambda + \gamma + 1} D^{\lambda+\gamma-1}.$$

We find that the mean and variance are

$$E[D] = \frac{\lambda \xi}{\gamma - 1} = \frac{1}{n_0} \left[1 + c_0 + O(c_0^2)\right], \quad (3)$$

where $E$ is the expectation operator, and

$$\text{var}(D) = \frac{\lambda \xi^2 (2\gamma - 3)}{(\gamma - 1)^2 (\gamma - 2)} = \frac{1}{n_0^2} \left[2 + 5c_0 + O(c_0^2)\right].$$

for $n_0$ and $c_0$ are the prior mean and coefficient of variation, respectively.

3.1.3 Hierarchical relative multinomial distribution

The distributions we have just derived are useful for pairwise spectratype comparisons. There are situations of great interest for clinical applications, however, where the quantities of interest are the divergences of multiple observed spectratypes from some control or asymptotic spectratype. In these cases, the patient TCR repertoire can be characterized in terms of their divergence from the control repertoire, while the measured spectratype is additionally divergent from true patient spectratype due to sampling variability.

We, therefore, need to consider a hierarchical model in which there are two sampling stages. We start with a relative multinomial with population parameter vector $p$ corresponding to the ideal, perfectly sampled healthy CDR3 length distribution. We imagine that the actual (but not yet observed) CDR3 length distributions in individual subjects are samples $q$ from this ideal distribution. The parameter $d^{-1}$ describes the completeness of this stage one sampling; the larger $d^{-1}$, the more similar will be $p$ and $q$. The observed spectratypes $r$ are samples or size $n$ (stage 2) from these individuals. The pdf for this model is

$$f_H(r|p, n, d^{-1}) = \int d^L q f_r(r|q, n) f_q(q, d^{-1}). \quad (4)$$

We perform the integration by Laplace’s method (Erdélyi, 1956). The critical point is given by

$$0 = q_i \frac{\partial \log f_H}{\partial q_i} = nr_i - d^{-1} q_i \left[1 + \log \frac{q_i}{p_i}\right] - \frac{1}{2} - \mu q_i,$$

where $\mu$ is a Lagrange multiplier enforcing the normalization of $q$. We solve for $\mu$ by summing on $i$, getting

$$\mu = n - 1 + D_0 d - \frac{L}{2},$$

where $D_0 \equiv D(q|p)$. Then

$$0 = n(r_i - q_i) + d^{-1} \left(D_0 - \log \frac{q_i}{p_i}\right) + \frac{1}{2} (Lq_i - 1). \quad (5)$$

The Hessian matrix, whose determinant is required for the application of Laplace’s method, is

$$H_{ij} \equiv \frac{\partial^2 \log f_H}{\partial q_i \partial q_j} = -\delta_{ij} \frac{1}{q_i} \left(n r_i + d^{-1} - 1\right),$$

where $\delta_{ij} = 1$ for $i = j$ and equals zero otherwise.
With these results and using appropriate care in applying Laplace’s method to integrals with integrands with constraints, we are able to write the pdf,

\[ \log f_H(r|p, n, d^{-1}) = -nD_1 - d^{-1}D_0 - \lambda \log \left( \frac{1}{n} + d \right) \]

\[ - \lambda \log 2\pi - \frac{1}{2} \sum_{i=1}^{L} \log r_i, \quad (6) \]

where \( D_1 \equiv D(r, q) \).

The first hierarchical stage is the one of biomedical interest, and the parameter \( d \), measuring the divergence of the true subject spectratype \( q \) from the control spectratype is the quantity we will be trying to estimate. The second stage arises in sampling. The inverse \( n^{-1} \) of the sample size, according to Equation (3), gives the mean sampling divergence. We will hereafter assume adequate sampling, \( nd \gg 1 \). Where the sampling is inadequate, the measurements can provide no more than an approximate lower bound on \( d \).

Two quantities that will prove to be of value in what follows are

Two quantities that will prove to be of value in what follows are:

\[ D_s \equiv \sum r_i \log \frac{r_i}{p_i} \]

and

\[ S_d \equiv \sum r_i \left( \log \frac{r_i}{p_i} - D_s \right)^2. \]

Expanding \( q_i \) in \( \epsilon \equiv (dn)^{-1} \) using Equation (5) gives

\[ q_i = r_i + \epsilon r_i \left( D_s - \log \frac{r_i}{p_i} \right) + O(\epsilon^2). \]

Substituting this expression into Equation (6) gives the pdf valid under these assumptions,

\[ \log f_H(r|p, n, d^{-1}) = -m \left( D_s - \frac{sD}{2dn} \right) - \lambda \log \left( \frac{d}{n} + 1 \right) \]

\[ - \lambda \log 2\pi - \frac{1}{2} \sum_{i=1}^{L} \log r_i + O(\epsilon^2). \]

Define the generating function

\[ h_H(s, t) \equiv \lambda \log \left( \frac{nm}{2\pi (n + d^{-1})} \right) \]

\[ + \log \int d^L r \exp \left( \frac{(s - d^{-1})}{2} \left( D_s - \frac{d^{-1} - s}{2(n - t)} \right) \right) \]

\[ = \lambda \left[ \log \left( \frac{1}{d^{-1} - s} + \frac{1}{n - t} - d \right) + O(\epsilon^2) \right]. \]

By definition, we have

\[ E \left[ D_s - \frac{sD}{2dn} \right] = \frac{\partial h_H}{\partial s}(0, 0) = \lambda \left( d - \frac{1}{n} \right) + O(d\epsilon^2) \]

and

\[ E \left[ S_d \right] = \frac{2}{dt^2} \frac{\partial^2 h_H}{\partial t^2}(0, 0) = 2\lambda \left( d - \frac{1}{n} \right) + O(d^2\epsilon^2). \]

Combining these results, we find that

\[ E[D_s] = \lambda \left( d + \frac{1}{n} \right). \]

Finally, we can write down estimators for \( d \) and \( 1/n \)

\[ \frac{1}{2\lambda} E \left[ D_s + \frac{1}{2} S_d \right] = d \]

and

\[ \frac{1}{2\lambda} E \left[ D_s + \frac{1}{2} S_d \right] = \frac{1}{n}. \]

4 DATA ANALYSIS

The pdfs derived above serve as the point of departure for the development of statistical methods for the analysis of spectratype data. In this section, we introduce these tests, and demonstrate their use with spectratype data collected from healthy volunteers, as well as data from thymus transplant patients.

The one-sample test is of limited utility, so in the interest of economy, we start with the tests for the comparison of parameter vectors in two or more treatment groups. Then, we describe the technique of linear modeling for the analysis of variable deviations from a given parameter vector. This latter is based on the hierarchical model.

4.1 Comparison of parameter vectors

The first scenario to explore is one in which there are two or more treatment groups, and the issue to be addressed is whether the distribution of CDR3 lengths is the same in all groups. The null hypothesis is that the population parameter vector \( q \) is identical in all groups. This will look very familiar, since, as we will see, the Kullback–Leibler divergence supports an additive decomposition of variability analogous to the partitioning of variance in linear models.

Denote by \( r_{ijk} \) the relative frequency of counts with CDR3 length \( i \) in the \( k \)-th member of group \( j \), which has \( n_j \) members. Define the sample mean parameter vector for group \( j \):

\[ \tilde{r}_j \equiv \frac{1}{n_j} \sum_{k=1}^{n_j} r_{ijk}. \]

If the number of groups is \( G \), the grand sample mean is

\[ \bar{r}_j \equiv \frac{1}{n} \sum_{j=1}^{G} n_j r_{ij}. \]

Denote by \( r_{jk} \) the vector with components \( r_{ijk} \).

Then the total divergence is partitioned as

\[ D_{tot} \equiv \sum_{jk} D(r_{jk}:q) \]

\[ = \sum_{jk} \left[ D(r_{jk}:\bar{r}_k) + D(\bar{r}_k:\bar{r}) + D(\bar{r}:q) \right]. \]

Under the null hypothesis that the groups have identical population parameter vectors, the expected values for each of these partial
The null hypothesis is rejected for values of \( f \) larger than the appropriate critical value.

**Example.** Comparison of CD4\(^+\) and CD8\(^+\) T cell repertoires in healthy volunteers. We collected peripheral blood samples from healthy volunteers and fractionated them into CD4\(^+\) and CD8\(^+\) T cells as described in Section 2.1. We compared spectratypes between these two subsets in each of the TCRBV families for which complete spectratype could be obtained using the \( f \) statistic computed as given in Equation (7). The null hypothesis tested is that the spectratype population parameter vectors in the two subsets are identical, and thus the divergence between two sample mean parameter vectors is attributable entirely to the same sources as the within-subset spectratype variability. Figure 2 shows the histograms corresponding to two of the TCRBV families studied, TCRBV1 and TCRBV5. TCRBV1 is judged to exhibit differences between the subsets, and TCRBV5 is not (Table 1). Nevertheless, the observed difference between the CD4 and CD8 spectratypes is consistent between the two families, with CD8 cells favoring shorter CDR3 lengths in both.

### 4.2 Variable divergence from a given population parameter vector

A second relevant scenario is one in which we have two or more spectratypes, and our interest is in describing how much they each differ from a given population parameter vector \( \mathbf{q} \). This acquires additional salience when \( \mathbf{q} \) is the parameter vector corresponding to the maximally diverse TCR repertoire. In that case, as we show in the Discussion below, the divergence from \( \mathbf{q} \) is a direct measure of the corresponding repertoire diversity.

The measure of the departure from the population parameter vector is the parameter \( d \) defined in Equation (4). The analysis following that equation shows that log \( d \) is distributed approximately normally, with the log transformation regularizing the variance. The approach we take here, then, is to estimate \( d \) as appropriate to the model under investigation and use the logs of these estimates in parameter fitting and hypothesis testing. All the machinery of normal-model statistics including analysis of variance, regression and linear modeling, more generally, are then at our disposal.

**Example.** Development of a diverse TCR repertoire following thymus transplantation. The immune response to infectious agents involves the selective expansion of particular T cell specificities, a perturbation that transiently reduces the diversity of the TCR repertoire. Autoimmune disease is similarly accompanied by decreased TCR diversity, as is the uncontrolled expansion of T cells that defines leukemias. In each of these cases, the return to health and to a steady-state, implies a return to the prior state of diversity. This process can be described in terms of continuous-time statistical models.

In this example, we examine the establishment of diversity over time in a Di George syndrome patient (DIG 102; Markert et al., 2004) following thymus transplantation. We use a three-parameter piecewise linear model given by

\[
\log d_i = \alpha + H(t_i - \tau)\beta(t_i - \tau) + \epsilon_i.
\]

In this model, the divergence is constant until time \( \tau \), when healthy T cells passing through the transplanted thymus begin to reach effective levels. \( \tau \) is thus the regression breakpoint and is to be estimated using the data. \( H \) is the Heaviside function, having value one for non-negative argument and zero otherwise. The intercept \( \alpha \) is treated
The rate at which the TCR repertoire diversifies. The errors, \( \epsilon \), represent the maximum size. Let the number of functional clones in family \( i \) be \( n_i \).

Each functional T cell clone, defined as the set of T cells that respond to the same antigenic peptide–MHC ligands, is restricted in size by intraclonal competition for TCR-specific growth signals. Assume, as a simplification, that each functional clone has the same maximum size. Let the number of functional clones in family \( i \) with CDR3-length class \( j \) be \( n_{ij} \).

Then the expected proportion of TCR in the \( i \)-th family is
\[
E[r_i] = q_i = \frac{n_i}{n},
\]
the expected proportion of TCR in CDR3-length class \( j \) conditional on being in family \( i \) is
\[
E[r_{ij}] = q_{ij} = \frac{n_{ij}}{n_i},
\]
and the expected proportion of TCR with sequence \( k \) conditional on CDR3 length \( j \) and family \( i \), is
\[
E[r_{ijk}] = q_{ijk} = \frac{1}{n_{ij}}.
\]

The total TCR repertoire diversity can be quantified by the entropy,
\[
S = -\sum_{ijk} r_{ijk} \log r_{ijk},
\]
which, using Equations (8)–(10), becomes
\[
S = -\sum_{ijk} r_{ijk} \log \left( \frac{r_{ijk}q_{ijk}}{n_qq_{ij}q_{kj}} \right)
\]
\[
= \log n - D - \sum_i r_i D_i - \sum_j \sum_i r_{ij} D_{ij}
\]
where the \( D \)'s are the Kullback–Leibler divergences at different levels:
\[
D = \sum_i r_i \log \frac{r_i}{q_i},
\]
\[
D_i = \sum_j r_{ij} \log \frac{r_{ij}}{q_{ij}},
\]
and
\[
D_{ij} = \sum_k r_{ijk} \log \frac{r_{ijk}}{q_{ijk}}.
\]

Spectratype divergences correspond to \( D \), for each TCRBV family \( i \). The first term in Equation (11), referring to the total number of (potential) T cell specificities is invariant. The variable part of the TCR repertoire entropy is given by the negative-weighted sum of the hierarchical Kullback–Leibler divergences.

Thus, when a good approximation of the ‘true’ population parameter vector is used, the analysis of deviation from this parameter vector, as described above, yields valid estimates of one component of the TCR repertoire diversity.

5 DISCUSSION

We have provided a first-principles method for the comparison of antigen receptor spectratypes under two distinct sets of circumstances. Yet spectratype data are typically regarded as providing information about the diversity of the antigen receptor repertoire. Here, we comment on the relationship between our methods and diversity per se.

5.1 Kullback–Leibler divergence and total TCR diversity

TCR diversity can be decomposed hierarchically, from distribution of TCRBV family usage, to distributions of CDR3 lengths within TCRBV family, to specific DNA sequence within CDR3 length. Alternative decompositions are possible, but these are the levels that correspond to conveniently available biological assays: family identification by flow cytometry, CDR3 length by spectratype, and specific DNA sequence by nucleotide sequencing. The data obtained in each of these assays represents the relative frequency of TCR counts in a subclass conditional on the parent class. We denote these relative frequencies \( r_i, r_{ij} \), and \( r_{ijk} \), respectively.

Each functional T cell clone, defined as the set of T cells that respond to the same antigenic peptide–MHC ligands, is restricted in size by intraclonal competition for TCR-specific growth signals. Assume, as a simplification, that each functional clone has the same maximum size. Let the number of functional clones in family \( i \) with CDR3-length class \( j \) be \( n_{ij} \).

Fig. 3. Data from a thymus transplantation patient, showing log \( m \) as a function of days post-transplantation. Different symbols distinguish different TCRBV families. The solid line shows the fit to a three-parameter piecewise linear model. Note that the first spectratype was performed before the transplant and recall that lower \( d \) corresponds to larger diversity and better prognosis.

The parameter values \( \pm \) their standard errors, estimated using data from DIG102 are, \( \hat{\alpha} = -0.159 \pm 0.150, \hat{\tau} = 111 \pm 5.4, \) and \( \hat{\beta} = -4.21 \times 10^{-2} \pm 3.86 \times 10^{-3}; \) the regression curve and data are shown in Figure 3.
The techniques we have described provide the objectivity and statistical power to further open new avenues for the application of spectratype analysis in both clinical and basic research areas.

We have made these methods publicly available via the World Wide Web at cbcb.duke.edu/SpA as described in greater detail by M.He, J.K.Tomfohr, B.H.Devlin, M.Sarzotti, M.L.Markert and T.B.Kepler (submitted for publication).

ACKNOWLEDGEMENTS

The authors gratefully acknowledge financial support from the National Institutes of Health through grants R01 AI 47040, R01 AI 54843, and M01-RR-00030, 5 P30 AI051445-03, and U54 AI057157-02, and thank the Statistical and Applied Mathematical Sciences Institute (SAMSI) for hospitality during part of this research effort.

Conflict of Interest: none declared.

REFERENCES


