BMC Immunology

METHODOLOGY ARTICLE Open Access

Quantification of total T-cell receptor diversity by flow cytometry and spectratyping

Stanca M Ciupe[1*], Blythe H Devlin[2], Mary Louise Markert[2,3] and Thomas B Kepler[4]

Abstract

Background: T-cell receptor diversity correlates with immune competency and is of particular interest in patients undergoing immune reconstitution. Spectratyping generates data about T-cell receptor CDR3 length distribution for each BV gene but is technically complex. Flow cytometry can also be used to generate data about T-cell receptor BV gene usage, but its utility has not been compared to or tested in combination with spectratyping.

Results: Using flow cytometry and spectratype data, we have defined a divergence metric that quantifies the deviation from normal of T-cell receptor repertoire. We have shown that the sample size is a sensitive parameter in the predicted flow divergence values, but not in the spectratype divergence values. We have derived two ways to correct for the measurement bias using mathematical and statistical approaches and have predicted a lower bound in the number of lymphocytes needed when using the divergence as a substitute for diversity.

Conclusions: Using both flow cytometry and spectratyping of T-cells, we have defined the divergence measure as an indirect measure of T-cell receptor diversity. We have shown the dependence of the divergence measure on the sample size before it can be used to make predictions regarding the diversity of the T-cell receptor repertoire.

Background

The immune system's ability to fight a large array of foreign particles is facilitated by the diversity of the T-cell receptor (TCR) repertoire [1]. This diversity is generated during thymocyte development by a process of somatic recombination. Inside the thymus, the constant (C) and variable (V) domains of the α and β chains of the TCR are assembled via random genetic rearrangements of the variable (V), diversity (D) and joining (J) gene segments [2]. Additional diversity is added through imprecise joining of the V and J regions along with random nucleotide additions and deletions at the V(D)J junctions [2,3]. Consequently, most of the variability lies in the third complementary determining region (CDR3) which is encoded by the V(D)J junction and comes in contact with the antigenic peptide on the surface of peptide/major histocompatibility complex (pMHC) molecules [4,5]. While the total number of lymphocytes in the blood can be directly measured, assessment of the diversity of the TCR repertoire requires more complex and indirect assays in a research setting. Such assays include flow cytometry, spectratyping and nucleotide sequencing.

Different T-cell clones use different V gene families in the rearrangement of their β chains. Through the use of commercially available monoclonal antibodies (named TCR Vβ), one can use standard flow cytometry on whole blood samples to determine the percentage of CD4 T-cells that use a given TCR BV family in subjects or controls. Measures of heterogeneity of TCR BV family usage in these CD4 T-cells can be used as a substitute for TCR repertoire diversity [6]. Flow cytometry is not only faster, cheaper, and technically simpler to use; the data reflects real population percentages.

Spectratyping uses messenger RNA (mRNA) from T-cells to amplify, by PCR, the complementary DNA (cDNA) across the CDR3 region. This generates information about the heterogeneity of the relative frequencies of different CDR3 length products within a functional TCR BV family. Because different T-cell clones have different sequences or lengths of CDR3, analysis of the CDR3 length distributions can be used to determine the overall TCR repertoire diversity [7-11]. Spectratyping has the advantage of providing a finer level of resolution than

*Correspondence: stanca@vt.edu
[1]Department of Mathematics, Virginia Tech, 460 McBryde Hall, Blacksburg, VA 24060, USA
Full list of author information is available at the end of the article

just analyzing BV gene family expression on the T-cells of flow cytometry. Although spectratyping provides the total number of CDR3 sizes and their pattern of distribution, the investigator cannot determine the frequency of cells used by a particular BV family. Amplifications of variations from a background distribution of each individual BV family may lead to over-representation of immunodominant clonotypes and therefore yield results that are not representative of the contribution of those cells in the entire T-cell repertoire.

TCR diversity can also be assessed by nucleotide sequencing of DNA CDR3 regions, but this is labor-intensive and generates an even lower level of resolution of the whole T-cell repertoire compared to spectratyping [12].

This paper focuses on the role of flow cytometry in measuring T-cell population diversity and compares it to T-cell population diversity as given by spectratyping. Traditionally, spectratyping data is quantified using a wide range of methods from visual [13,14] to quantitative scoring [15-17]. Our group previously described the use of a likelihood method for measuring deviation from a normal TCR repertoire [9,11]. For each observed CDR3 length distribution by spectratyping, we compute the Kullback-Leibler divergences between the patient CDR3 length distribution and a known reference distribution [9,11]. We have modified the Kullback-Leibler divergence to measure the deviation of T-cell receptor diversity from normal. This was done by accounting for both the TCR BV family usage as measured by flow cytometry and by comparing the utility of this method to CDR3 length distribution as measured by spectratyping [11].

Estimator bias is a concern when using this method of divergence scoring. In particular, it is desirable to determine how much deviation in the computation of the divergence occurs when the initial number of lymphocytes used in generating the data is varied. We have addressed this question in the context of divergence measures generated individually by flow cytometry and spectratyping. The results are especially useful when using the techniques for limited numbers of cells.

Results

We used the Kullback-Leibler divergence to quantify similarities between different frequency distributions in the T-cell repertoire diversity when measured by either flow cytometry or spectratyping. We started with two assumptions: 1) the reference distribution corresponds to a polyclonal TCR repertoire and 2) in individual subjects, a positive divergence determines the deviation from the normal TCR repertoire. The flow divergence, D_f, is the distance between the individual and the perfectly sampled reference control distributions of all TCR BV family usage measured by flow cytometry. The spectratype divergence, D_s, is the distance between the individual and the perfectly sampled reference control distributions of the CDR3 lengths of each TCR BV family and averaged over all TCR BV families as measured by spectratyping (see section Kullback-Leibler divergence and [9]).

We specifically wanted to assess the performance of the divergences D_f and D_s in predicting the diversity of the T-cell receptor repertoire under stressful, i.e. cell limited, circumstances. While D_f and D_s account for deviations from normal of distributions of TCR BV family usage and CDR3 lengths within each TCR BV family, additional variability is added due to the dependence on the number of measured events, n, for every individual patient/control (see Figures 1 and 2). Knowing the sample size n and the dimensions of the measured space, L_i, we derived the corrected divergence value, $D_{i,\text{corr}}$ (see section 'Sampling bias - theoretical derivation') to be given by

$$D_{i,\text{corr}} = D_i - \frac{L_i - 1}{2n}, \tag{1}$$

where $i = f, s$ for flow cytometry and spectratyping, respectively. L_f is the number of BV families used in the flow cytometry assay (in our case 18) and L_s is the number of CDR3 lengths used in the spectratype assay (in our case 14).

Therefore, only the number of measured events, n, and the dimension of the measured space, L_i are needed to correct the divergence measures. We used this formula to assess the performance of D_f and D_s measures in an athymic DiGeorge subject (Figure 1) during a period of limited numbers of peripheral blood T-cells as the patient underwent immune reconstitution following thymus transplantation.

Flow cytometry results

Flow divergence measurements, D_f, were determined at seven time points following thymus transplantation in DiGeorge subject 5 (Table 1). For each time point, the number of CD4 T-cell was known (Table 1). The corrected divergence $D_{f,\text{corr}}$ is found by subtracting $(L_f - 1)/2n$, where $L_f = 18$, from the measured divergence D_f at each time point (Table 1). The measured and corrected divergences as a function of $1/n$ are plotted in Figure 1(a). When we use samples with low event numbers, we noted an overestimate in the measured D_f compared to D_f estimates from samples with high event numbers, for which the correction is not significant. Formula (1) helped address the effect of event number on the D_f prediction.

To further test the dependence of D_f on the sample size we assumed that D_f is a function of the decreasing event numbers in the CD4 T-cell gate used for TCR BV analysis. For this analysis we used a single blood sample collection from each of four complete DiGeorge subjects after thymus transplantation and from each of four healthy

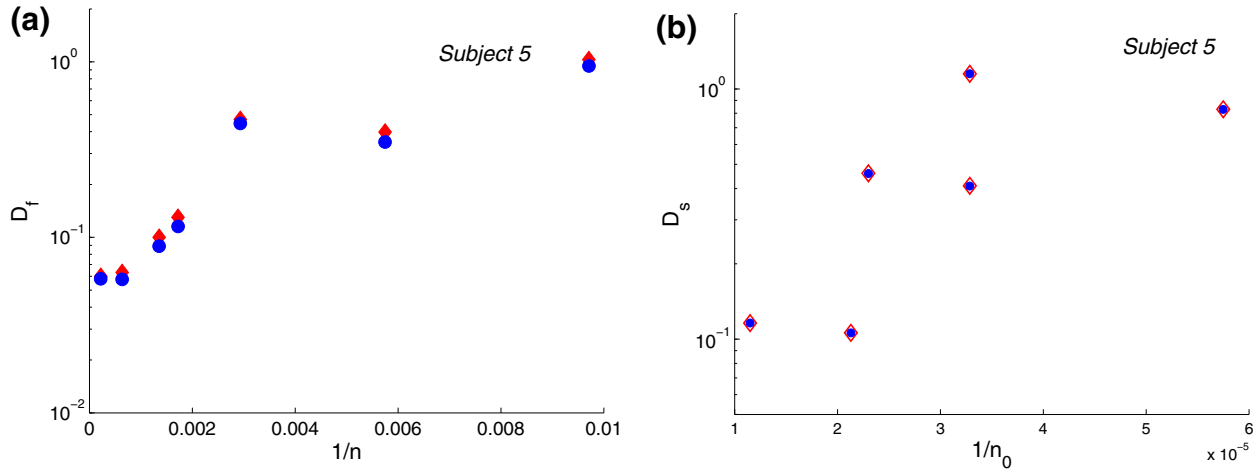

Figure 1 Measured and corrected divergence measures as function of inverted sample number. (a) Measured flow divergence, D_f, (red solid diamonds) and corrected flow divergence, $D_{f,\text{corr}}$, (blue circles) as functions of the inverted sample number $1/n$; **(b)** Measured spectratype divergence, D_s, (red empty diamonds) and corrected spectratype divergence, $D_{s,\text{corr}}$, (blue circles) as a function of the inverted sample number $1/n_0$ in one DiGeorge patient.

controls. Each blood sample was serially diluted, followed by flow cytometry. The results are presented in Table 2 and the plot of D_f as a function of n is presented in Figure 2.

For each of these eight cases, we wanted to predict the corrected divergence value, $D_{f,\text{corr}}$, using the measured D_fs and determine their dependence on the sample size n. We define a three parameter linear model given by

$$y(n) = \alpha + C/n + \varepsilon, \tag{2}$$

where, $y(n)$ is the observed D_f and n is the number of CD4 T-cells in the sample. The intercept α is the true divergence, $D_{f,\text{corr}}$, and the slope C quantifies the rate at which the diversity is dependent on the sample size. In equation (1), slope C corresponds to the $(L_f - 1)/2$ value, which for an assay that uses 18 BV families, reduces to 8.5. The errors, ε, are independent and normally distributed.

We derived estimates and 95% confidence intervals for parameters α and C for each of eight individuals by fitting

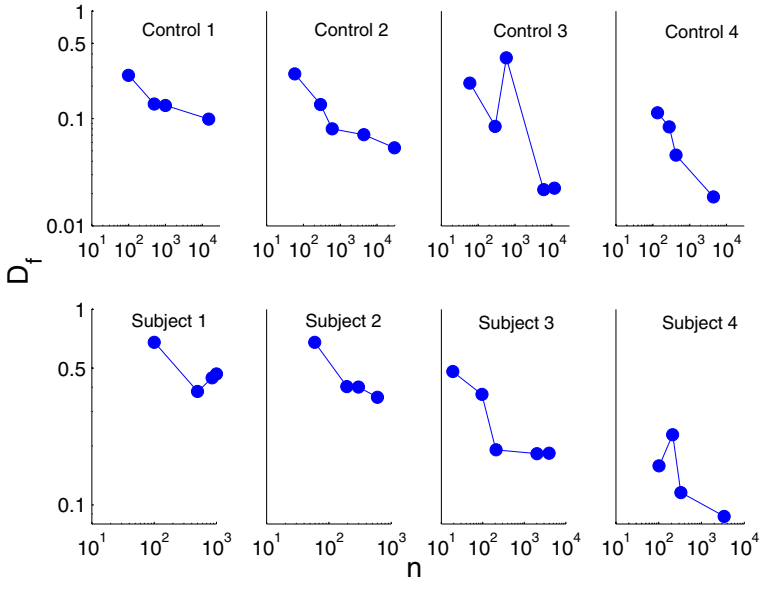

Figure 2 Flow divergence, D_f, as a function of sample size n (●), presented on a log-log scale.

Table 1 Average CD4 T-cell sample size, measured flow divergence D_f, and corrected flow divergence $D_{f,\text{corr}}$ in a DiGeorge subject

Days after transplant	Average CD4 nr in gate (n)	Measured flow D_f value	Corrected flow $D_{f,\text{corr}}$ value
70	341	0.47	0.44
88	103	1.02	0.94
117	174	0.39	0.34
145	581	0.129	0.11
181	737	0.103	0.091
398	1569	0.063	0.057
868	4514	0.06	0.058

Values are measured over time following thymic transplantation.

$y(n)$, as given by (2), to the measured D_f values in Table 1 for CD4 T-cell numbers n. For the fitting routine we used a descent method for univariate functions [18]. The parameter values and their confidence intervals are presented in Table 3. The regression curves and data are presented in Figure 3.

Moreover, if we consider the slope C to be equal among the subjects we can simultaneously fit the following model to the data from all subjects.

$$y_i(n) = \alpha_i + C/n + \varepsilon_i, \tag{3}$$

where α_i are the corrected divergence values for the patient i, with $i = 1, ..., 8$. The rate at which the diversity is dependent on the sample size, C, is considered constant among the subjects. The errors for each of the subjects, ε_i, are independent and normally distributed.

The fitting procedure was done using a quasi-Newton method for finding the minimum of a multivariate function [18]. The predicted parameter values and their confidence intervals are presented in Table 4. The regression curves and data are presented in Figure 4.

Events estimation

From the flow cytometry analysis we can estimate the minimum number of CD4 T-cells needed in a sample for an accurate $D_{f,\text{corr}}$ estimate. If we want our estimates to be 90% accurate, *i.e.*, err $= 0.1$, then the ratio between the corrected and measured divergence has to be less than err,

$$\frac{1/nC}{\alpha + 1/nC} < \text{err}. \tag{4}$$

This translates into the following condition

$$n > \frac{C(1-\text{err})}{\text{err} \times \alpha}. \tag{5}$$

Table 2 Summary of T-cell sample size and the corresponding flow divergence values D_f

Subject	Average CD4 T-cell nr in gate n	Measured flow divergence D_f
Control 1	66	0.252
	340	0.135
	675	0.132
	10051	0.098
Control 2	58	0.260
	290	0.135
	603	0.079
	4438	0.070
	29438	0.053
Control 3	60	0.214
	290	0.084
	585	0.366
	5965	0.021
	11889	0.022
Control 4	136	0.112
	282	0.083
	425	0.045
	4354	0.018
Subject 1	89	0.679
	445	0.379
	756	0.445
	887	0.466
Subject 2	59	0.678
	194	0.403
	299	0.399
	605	0.355
Subject 3	19	0.479
	95	0.366
	207	0.191
	2013	0.182
	3946	0.183
Subject 4	103	0.158
	213	0.229
	329	0.115
	3367	0.087

From our estimates $C = 7.705$ and $\alpha = 0.19$ (median 0.12). This implies the sample size, n, must be larger than 364 (median 577) cells for an accurate $D_{f,\text{corr}}$ estimate. In our case, we gated the flow cytometry on CD4 T-cells, so more than 364 CD4 T-cells, or events, must be captured in the flow analysis.

Table 3 Parameter values and confidence intervals for model (2)

Subject		Value	CI
Control 1	α	0.107	[0.079,0.135]
	C	9.7	[6.1, 13.4]
Control 2	α	0.07	[0.02,0.129]
	C	10.9	[4.7, 17.2]
Control 3	α	0.111	[-0.17,0.373]
	C	6.9	[-29, 43]
Control 4	α	0.02	[-0.027,0.067]
	C	13	[2, 24]
Subject 1	α	0.39	[0.214, 0.574]
	C	25	[-6.3, 56]
Subject 2	α	0.32	[0.253, 0.377]
	C	21.3	[14.5, 28.1]
Subject 3	α	0.205	[0.087, 0.322]
	C	5.5	[0.7, 10.4]
Subject 4	α	0.113	[-0.116, 0.342]
	C	7.9	[-33, 49]

Spectratype results

Spectratype divergence measurements, D_s, were determined in five patients for three to seven time points following thymic transplantation (Table 5). For each time point, the number of CD3 T-cell used to isolate RNA, n_0, is known (Table 5). Starting with a fixed amount of RNA, complementary DNA (cDNA) is generated in a reverse transcriptase reaction and used with each of $\pi = 23$ different primers to amplify the CDR3 region from each BV gene.

The corrected $D_{s,\mathrm{corr}}$ is found by subtracting $(L_s-1)/2n$, where $n = n_0/\pi$, from the measured divergence at each time point, where $L_s = 14$ (Table 5). The measured and corrected divergences as a function of $1/n_0$ are plotted in Figure 1(b). We note that there is no correction in the measured spectratype divergence, D_s, since the number n_0 of CD3 T-cells that we are starting with is always high.

Total divergence

By combining the individual contributions of flow and spectratype divergence, we defined the total divergence, D (see section 'Kullback-Leibler divergence'). D measures the divergence of the individual from the perfectly sampled reference control and accounts for differences in distributions of CDR3 lengths within each TCR BV family by spectratyping as well as differences in distributions of overall TCR BV families by flow cytometry. Corrections in the flow and spectratype divergences are sufficient to ensure that the total divergence is independent of the sample size.

Discussion

The data used in our study came from flow cytometry and spectratype assays in both DiGeorge subjects after thymus transplantation and healthy adult volunteers. This study presents significant information regarding the utility of

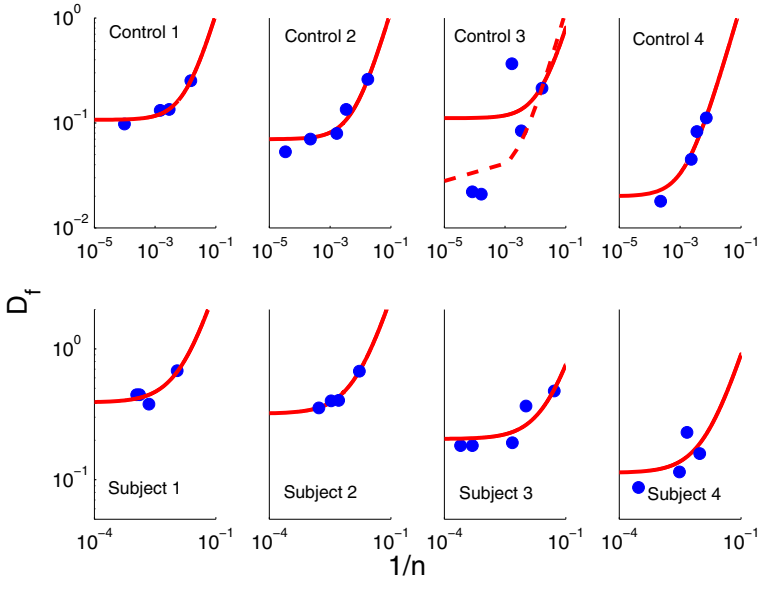

Figure 3 Flow divergence D_f as a function of the inverted sample number $1/n$ in eight subjects. The solid line represents the fit of the three parameter linear model (2) to the data (•). Results are presented on a log-log scale. The same model was fitted to a data set that excluded point (0.0017, 0.366) for control 3 (dashed line). The best parameter estimates and their 90% confidence intervals are presented in Table 3.

Table 4 Parameter values and confidence intervals for model (3)

Subject	α	CI
Control 1	0.117	[0.033,0.202]
Control 2	0.085	[0.009,0.161]
Control 3	0.107	[0.032,0.184]
Control 4	0.039	[-0.045,0.123]
Subject 1	0.46	[0.38, 0.55]
Subject 2	0.41	[0.32, 0.49]
Subject 3	0.175	[0.089, 0.261]
Subject 4	0.113	[0.029, 0.2]
Subject	*C*	CI
All	7.705	[4.55, 10.85]

flow cytometry, as well as spectratyping, to assess the diversity of the antigen receptor repertoire. Importantly, these data identify a bias in measurement errors which must be corrected. The paper presents the relationships between the number of gated events in the flow cytometry assay, as well as the number of CD3 T-cells in the spectratype assay, and the information-theory measures, D_f and D_s, used as surrogates of TCR diversity.

We addressed a critical issue of estimator bias. Starting with the assumption that such a bias exists, we have derived ways to account for the error in the measured divergences. We show that D_f and D_s can be corrected by substracting a number inversely proportional to the sample size.

For the flow cytometry data, the constant of proportionality can either be deduced theoretically as a function of the total number of BV TCR families used in the flow cytometry assay, or derived from a statistical model applied to individual data. Both methods predict similar results, with the constant equal to 8.5 in the theoretical approach and 7.7 in the statistical approach. It is important to note that we found a direct correlation between the measured D_f and the sample size in five out of eight subjects (see Table 6).

Our study allows us to predict a lower bound for the number of CD4 T-cells needed in the flow cytometry gated events. We have shown that at least 364 CD4 T-cells have to be counted as gated events for a 90% confidence in the D_f measures. With fewer gated events, the D_f measurement cannot be used as a substitute for diversity. This is particularly important to keep in mind when assessing patients with limited numbers of T-cells, such as those undergoing immune reconstitution following thymus, stem cell or bone marrow transplantation. Each of these is a clinical situation in which the development of the T-cell repertoire correlates to immune competency. Thus, these data provide a quantitative basis by which T-cell repertoire diversity can be assessed by flow cytometry.

For the spectratype data, the results are quite different. Although, using the same theoretical approach, we derive a constant, $C = 6.5$, that accounts for measurement bias;

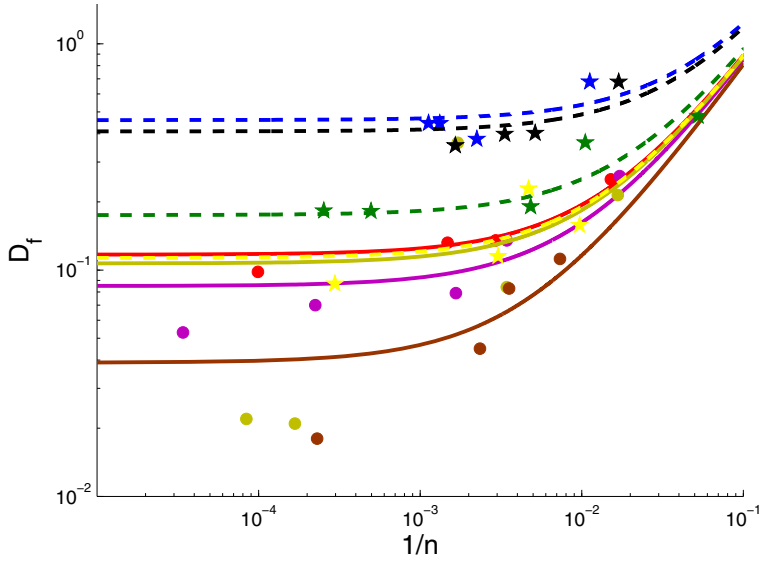

Figure 4 Flow divergence D_f as a function of the inverted sample number $1/n$ for the same slope C. The solid and dashed lines shows the fit of a three parameter linear model (3) to the data (•). The results are presented on a log-log scale. The best parameter estimates and their 90% confidence intervals are presented in Table 4.

Table 5 CD3 T-cell sample size, measured spectratype divergence D_s, and corrected spectratype divergence $D_{s,,\text{corr}}$ in a DiGeorge subject

Subject	Days after transplant	CD3 T-cells n_0	Measured D_s value	Corrected $D_{s,\text{corr}}$ value
Subject 1	9	420,000	0.91	0.9096
	34	12,220,000	0.61	0.61
	70	550,000	0.97	0.9697
Subject 4	540	670,000	0.039	0.0388
	1540	1,260,000	0.073	0.0729
	2017	1,140,000	0.076	0.0759
Subject 5	70	700,000	1.15	1.1498
	88	400,000	0.83	0.8296
	117	700,000	0.41	0.4098
	145	1,000,000	0.46	0.4599
	181	1,080,000	0.106	0.1059
	398	2,000,000	0.116	0.1159
Subject 6	175	1,440,000	0.107	0.1069
	209	800,000	0.168	0.1678
	286	1,480,000	0.086	0.0859
	730	1,200,000	0.12	0.1199
Subject 7	102	380,000	0.43	0.4296
	130	460,000	0.23	0.2297
	166	500,000	0.08	0.0797
	372	1,250,000	0.14	0.1399

Values are measured over time following thymic transplantation.

thus, the corrected spectratype divergence is identical to the observed divergence. Moreover, we find no correlation between the measured spectratype divergence, D_s, and the sample size in four out of five patients (Table 7).

The total divergence actively incorporates the flow divergence. Correction in the flow divergence, D_f, guarantees independence of the total divergence, D, from the sample size.

Conclusions

In conclusion, sample size is a sensitive parameter in the predicted flow divergence values, but not in the spectratype divergence values. Although using flow cytometry to assess T-cell repertoire diversity is a valuable tool, one must have sufficient cells, or events, in the flow cytometry gate before using either the flow or the total divergence as a prediction for the TCR repertoire diversity.

Methods

Human subjects

Blood samples used in our study come from healthy adult controls and from infants with complete DiGeorge anomaly after thymus transplantation [19]. Blood was obtained under protocols approved by Duke University Medical Center Internal Review Board (IRB). T-cell repertoire evaluation was done by flow cytometry. Whole blood samples were evaluated using 22 monoclonal antibodies directed against CD4 and a total of 18 TCR BV families (Beckman Coulter and BD Biosciences - see Tables 8 and 9).

Table 6 Correlation coefficient and p-values as given by a Pearson comparison test, between the inverse average number of CD4 T-cell used in flow cytometry assays and the flow divergence

Subject	Correlation coefficient	p-value
Control 1	0.99	0.0076
Control 2	0.98	0.0031
Control 3	0.32	0.58
Control 4	0.96	0.035
Subject 1	0.92	0.075
Subject 2	0.99	0.005
Subject 3	0.9	0.036
Subject 4	0.5	0.49

Table 7 Correlation coefficient and p-values as given by a Pearson comparison test, between the inverse total number of CD3 T-cell used in spectratype assays and the spectratye divergence

Subject	Correlation coefficient	p-value
Subject 1	0.92	0.25
Subject 4	-0.98	0.11
Subject 5	0.66	0.15
Subject 6	0.97	0.03
Subject 7	0.64	0.35

Human subjects

Subjects were enrolled in protocols that were approved by the Duke University Health System Institutional Review Board and were reviewed by the Food and Drug Administration under an Investigational New Drug application. All subjects were children. The parent(s) of each subject provided written informed consent.

Table 8 List of TCR BV families and antibodies used in the flow cytometry assay

Antibody names	Clone	Family name *
Vβ1	BL37.2	TRBV9
Vβ2	MPB2D5	TRBV20
Vβ3	CH92	TRBV28
Vβ4	WJF24	TRBV29
Vβ5.1	IMMU157	TRBV5
Vβ5.3	3D11	TRBV5
Vβ5.2	36213	TRBV5
Vβ7.1	ZOE	TRBV4
Vβ7.2	Zizou4	TRBV4
Vβ8.1 & Vβ8.2	56C5	TRBV12
Vβ9	FIN9	TRBV3
Vβ11	C21	TRBV25
Vβ12	VER2.32.1	TRBV10
Vβ13.2	H132	TRBV6
Vβ13.6	JU-74	TRBV6
Vβ14	CAS1.1.3	TRBV27
Vβ16	TAMAYA 1.2	TRBV14
Vβ17	E17.5F3	TRBV19
Vβ18	BA62	TRBV18
Vβ20	ELL 1.4	TRBV30
Vβ22	IMMU 546	TRBV2
Vβ23	AF23	TRBV13

The antibodies were purchased from Immunotech (Beckman Coulter) and used for the analysis. A kit IOTest Beta Mark became available during the study and was used in place of individually purchased antibodies.
*Nomenclature of the IMGT, the international ImMunoGeneTics information system http://www.imgt.org.

Table 9 List of TCR VB families and antibodies excluded from the flow cytometry studies

Antibody names	Clone	Family name*
Vβ13.1 & 13.4 & 13.6	IMMU 222	TRBV6-5 & 6-6 & 6-9
Vβ21.3	IG125	TRBV11-2

These antibodies are included in the kit but were not included in the analysis.
*Nomenclature of the IMGT, the international ImMunoGeneTics information system http://imgt.cines.fr.

Flow cytometry

Reference distributions of TCR BV family usage determined by flow cytometry were generated from peripheral blood samples of fifty healthy individuals (see Table 10). Similar distributions of TCR BV usage were derived from four additional controls and four DiGeorge subjects [19] who underwent thymus transplantation.

Spectratyping

CD3 T-cells from the peripheral blood of patients were isolated. RNA was prepared and used for cDNA synthesis.

Table 10 Mean % of CD4 T-cells that use a TCR BV family as predicted by the flow cytometry assay

Antibody names	Mean % of CD4 T-cells
Vβ1	3.21
Vβ2	9.79
Vβ3	4.80
Vβ4	2.58
Vβ5.1	6.78
Vβ5.3	0.97
Vβ5.2	0.70
Vβ7.1	1.89
Vβ7.2	1.12
Vβ8.1 & Vβ8.2	4.71
Vβ9	3.48
Vβ11	0.73
Vβ12	1.85
Vβ13.2	2.66
Vβ13.6	1.84
Vβ14	3.03
Vβ16	0.91
Vβ17	5.79
Vβ18	1.96
Vβ20	2.35
Vβ22	4.12
Vβ23	0.45

Note that the antibody used in flow cytometry assay covers approximately 70% of CD4 T-cells.
The values are averaged across 50 normal volunteers.

The cDNA was used as a template for 23 TCR BV specific primer pairs to amplify the complete CDR3 region by PCR [10]. Each PCR product, representing a different TCR BV family, was size separated by electrophoresis and the product lengths were identified using the GeneScan software (Applied Biosciences). An example of spectratype data in a healthy adult is presented in Figure 5, which shows the histograms of the number of CD4 T-cells versus CDR3 length for each TCR BV family.

Kullback-Leibler divergence

Flow Kullback-Leibler divergence

Let $P = \{P_i, i = 1, ..., n_F\}$ be the relative frequencies corresponding to the ideal, perfectly sampled reference distribution of BV family i usage, where n_F is the number of BV families (in our case 18). Let $p = \{p_i, i = 1, ..., n_F\}$ be the relative frequency of cells that use BV family i in individual control/subjects. The null hypothesis is that a normal polyclonal TCR repertoire has a distribution identical with that of the reference distribution. Deviation from the normal repertoire seen in subjects can be quantified by the flow Kullback-Leibler divergence [9]

$$D_f = \sum_{i=1}^{n_F} p_i \log \frac{p_i}{P_i}. \tag{6}$$

The flow Kullback-Leibler divergence is a measure of the distance between the two frequency distributions or,

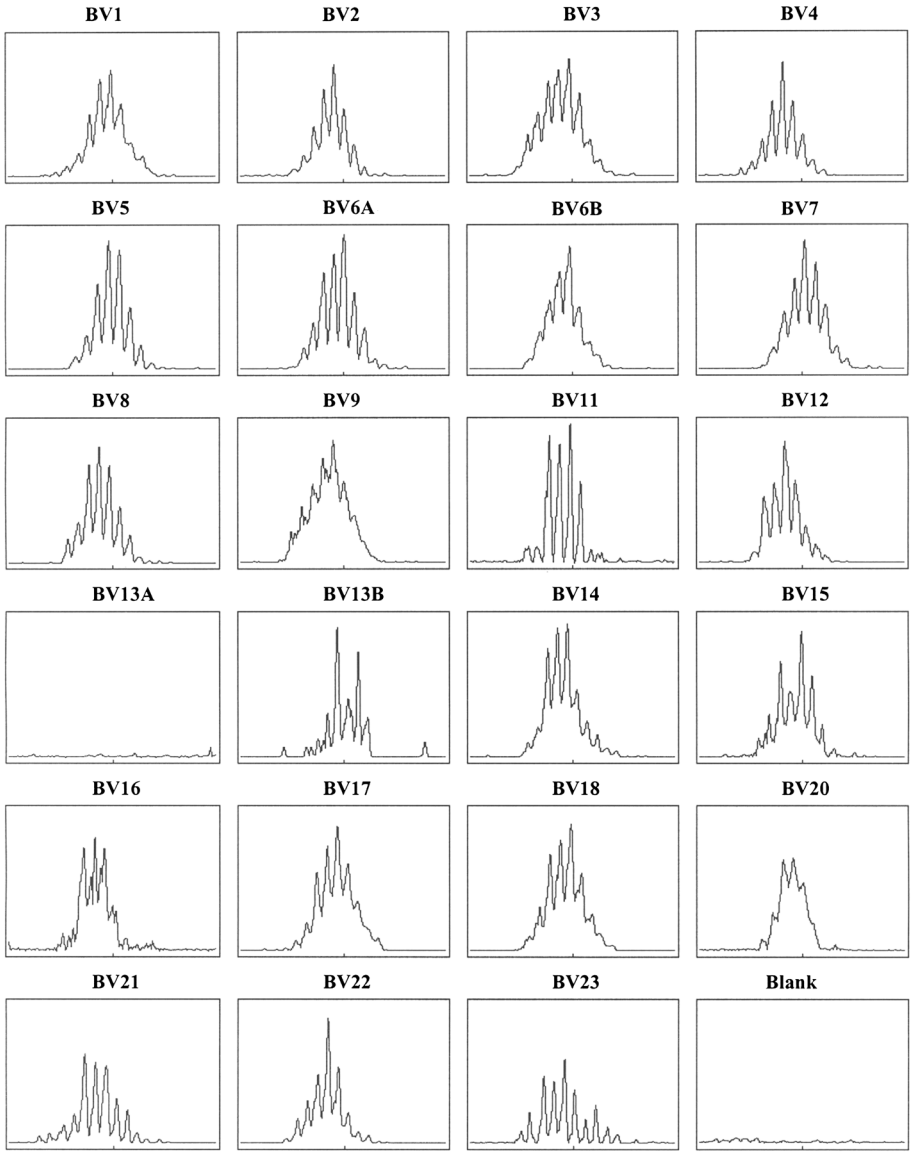

Figure 5 CD4 T-cell spectratype data. Spectratype histograms show the number of CD4 T-cells bearing receptors versus CDR3 length for each TCR BV families tested.

equivalently, it is the inefficiency of assuming that the distribution of BV family usage is p_i, $i = 1, ..., n_F$, when the true frequency usage is $P_i, i = 1, ..., n_F$.

Spectratype Kullback-Leibler divergence

Similarly, let $p = \{p_{ij} = q_i r_{j/i}, i = 1, ..., n_F$ and $j = 1, ..., n_C\}$, and $P = \{P_{ij} = Q_i R_{j/i}, i = 1, ...n_F$ and $j = 1, ...n_C\}$, respectively, be the relative numbers of T-cells of CDR3 lengths j, given that the BV family i is used in individual patient/controls and reference controls as determined by spectratype. Here n_C is the number of CDR3 lengths (in our case 14), $(q, Q)_i$ are the relative frequencies of cells which use the BV family i and $(r, R)_{j/i}$ the relative frequencies of of cells that have CDR3 length j, given that they use the BV family i. The null hypothesis is that a normal polyclonal TCR repertoire has a distribution of CDR3 lengths identical with that of the reference distribution. Deviation of from the normal repertoire, as seen in patients, can be quantified by the spectratype divergence for each TCR BV family i as follows

$$D_{s/i} = \sum_{i=1}^{n_C} r_{j/i} \log \frac{r_{j/i}}{R_{j/i}}, \tag{7}$$

and the total spectratype divergence, which is the average of spectratype divergences of TCR BV families i, $i \in \{1, ..., n_F\}$ is given by

$$D_s = \frac{1}{n_F} \sum_{i=1}^{n_F} D_{KL,\text{spec}/i}. \tag{8}$$

Total Kullback-Leibler divergence

We can combine these two measures to obtain a total divergence measure from normal repertoire, derived as follows

$$\begin{aligned} D &= \sum_{i=1}^{n_F} \sum_{j=1}^{n_C} p_{ij} \log \frac{p_{ij}}{P_{ij}} = \sum_{i=1}^{n_F} \sum_{j=1}^{n_C} q_i r_{j/i} \log \frac{q_i r_{j/i}}{Q_i R_{j/i}} \\ &= \sum_{i=1}^{n_F} q_i \log \frac{q_i}{Q_i} + \sum_{i=1}^{n_F} q_i \sum_{j=1}^{n_C} r_{j/i} \log \frac{r_{j/i}}{R_{j/i}} \\ &= D_f + \sum_{i=1}^{n_F} q_i D_{s/i}, \end{aligned} \tag{9}$$

Sampling bias - theoretical derivation

The distribution of BV family usage (CDR3 length within a BV family) of a perfectly sampled reference control can be described by a L_f (L_s)-dimensional multinomial distribution with the parameter vector P, where P_i is the relative numbers of T-cells that use the BV family (CDR3 length) i. The distribution of the actual, but not yet observed, BV family (CDR3 length) usage in individual patient/controls are subsamples q of the ideal distribution, where q_i are the relative numbers of T-cells that use the BV family (CDR3 length) i. The distance between these two distributions is given by the parameter d^{-1}, with a large d accounting for a closer similarity between P and q. Finally, the observed distribution of BV family usage (CDR3 length), p, are samples of n measured events for every individual patient/control, where p_i are the relative numbers of T-cells that use the BV family (CDR3 length) i. Here L_f (L_s) is the dimension of the measured space, *i.e.* the number of BV families used in the flow cytometry assay, in our case 18 (the number of CDR3 lengths used in spectratyping assay, in our case 14).

For a large sampling number, n, we can consider the relative frequencies P, q and p to be continuous variables and define their probability distribution functions, pdf, as

$$f(p|P, n, d^{-1}) = \int f(p|q, n) f(q|P, d^{-1}) d^{L_i} q \tag{10}$$

where $i = f, s$. The pdf of p, for np_i large enough, can be approximated using Stirling's formula (see [9] for a complete computation). Therefore,

$$\begin{aligned} f(p|q, n) &= n^{L_i - 1} \frac{\Gamma(n+1)}{\Gamma(np_i + 1)} \prod_{i=1}^{L_i} q_i^{p_i n} \\ &\approx \frac{n^{(L_i-1)/2} e^{-nD(p|q)}}{\sqrt{(2\pi)^{L_i - 1} \prod_{i=1}^{L_i} p_i}} \delta(\sum_i p_i - 1), \end{aligned} \tag{11}$$

where δ is the Dirac delta function and

$$D(p|q) = \sum_{i=1}^{L_i} p_i \log \frac{p_i}{q_i}, \tag{12}$$

is the Kullback-Leibler divergence between p and q.

As shown in Kepler et al. [9] Laplace's integration method with constraints [20] can be used to asymptotically approximate the integral (10) as follows

$$\begin{aligned} f(p|P, n, d^{-1}) = &\left\{ 2\pi \left(\frac{1}{n} + d \right) \right\}^{-(L_i - 1)/2} \\ &\times \prod_{i=1}^{L_i} \frac{1}{\sqrt{p_i}} e^{-nD(p|q) - d^{-1} D(q|P)} \end{aligned} \tag{13}$$

and

$$\begin{aligned} \log f(p|P, n, d^{-1}) = &- nD(p|q) - d^{-1} D(q|P) \\ &- \frac{L_i - 1}{2} \log \left(\frac{1}{n} + d \right) \\ &- \frac{L_i - 1}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{L_i} \log p_i, \end{aligned} \tag{14}$$

Moreover, as shown in Kepler et al. [9], a Taylor expansion in $\epsilon = (nd)^{-1}$ of q_i around p_i, leads to the following expression for (14)

$$\begin{aligned}\log f(p|P,n,d^{-1}) = &- d^{-1}\left(D(p|P) - \frac{s_D}{2nd}\right)\\ &- \frac{L_i-1}{2}\log\left(\frac{1}{n}+d\right)\\ &- \frac{L_i-1}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{L_i}\log p_i + O(\epsilon^2)\end{aligned} \tag{15}$$

where

$$D(p|P) = \sum p_i \log\frac{p_i}{P_i}, \tag{16}$$

and

$$s_D = \sum p_i\left(\log\frac{p_i}{P_i} - D(p|P)\right)^2. \tag{17}$$

From this, one can derive the expected values, E, of $D(p|P)$ and s_D up to order ϵ to be (for a complete derivation refer to [9])

$$\begin{aligned}E\left[D(p|P)\right] &= \frac{L_i-1}{2}\left(\frac{1}{n}+d\right), \\ E\left[s_D\right] &= (L_i-1)\left(d-\frac{1}{n}\right) + O(d\epsilon^2).\end{aligned} \tag{18}$$

From here we can derive the corrected individual divergence,

$$D_{i,\mathrm{corr}} = D_f - \frac{L_i-1}{2n}, \tag{19}$$

which relaxes the concern of variability due to sampling error.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Conceived the study: BHD and TBK. Developed mathematical components: SMC and TBK. Developed empirical components: BHD and MLM. Interpreted results and wrote the manuscript: SMC, BHD, MLM and TBK. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by National Institute of Health grants R01 AI 54843, R01 AI 47040, M03 RR60 (Duke General Clinical Research Center, National Center for Research Resources, National Institute of Health), and Office of Orphan Products Development, Food and Drug Administration, grant FD-R-002606. MLM and TBK are members of the Duke Comprehensive Cancer Center. We acknowledge the technical assistance of Marilyn Alexieff, Jie Li, Chia-San Hsieh, Jennifer Lonon and Julie E. Smith, the clinical research assistance of Stephanie Gupton and Alice Jackson, and the regulatory affairs assistance of Elizabeth McCarthy and Michele Cox are appreciated as is the clinical care by the faculty and fellows of the Duke Pediatric Allergy and Immunology Division. We acknowledge the collaboration of surgeons James Jaggers, Andrew Lodge, Henry Rice, Micheal Skinner, and Jeffrey Hoehner. We appreciate the assistance of Drs. Michael Cook and Scott Langdon in the Duke Comprehensive Cancer Center flow cytometry and sequencing facilities.

Author details
[1]Department of Mathematics, Virginia Tech, 460 McBryde Hall, Blacksburg, VA 24060, USA. [2]Department of Pediatrics, Duke University Medical Center, Durham, NC 27710, USA. [3]Department of Immunology, Duke University Medical Center, Durham, NC 27710, USA. [4]Department of Microbiology, Boston University School of Medicine, Boston MA 02118, USA.

Received: 19 April 2013 Accepted: 26 July 2013
Published: 6 August 2013

References
1. Nikolich-Zugich J, Slifka M, Messaoudi I: **The many important facets of T-cell repertoire diversity.** *Nat Rev Immunol* 2004, **4:**123.
2. Davis M, Bjorkman P: **T-cell antigen receptor genes and T- cell recognition.** *Nature* 1988, **334:**395–401.
3. Alt F, Oltz E, Young F, Gorman J, Taccioli J, Chen J: **VDJ recombination.** *Immunol Today* 1992, **13:**306–314.
4. Garcia K, Degano M, Stanfield R, Brunmark A, Jackson M, Peterson P, Teyton L, Wilson I: **An alphabeta T cell receptor structure at 2.5 A and its orientation in the TCR-MHC complex.** *Science* 1996, **274:**209.
5. Davis M, Boniface J, Reich Z, Lyons D, Hampl J, Arden B, Chien Y: **Ligand recognition by $\alpha\beta$ T-cell receptors.** *Annu Rev Immunol* 1998, **16:**523–544.
6. Markert ML, Alexieff MJ, Li J, Sarzotti M, Ozaki DA, Devlin BH, Sempowski GD, Hale LP, Buckley R, Rice HE, Mahaffey SM, Skinner MA: **Postnatal thymus transplantation with immunosuppression as treatment for DiGeorge syndrome.** *Blood* 2004, **104:**2574-2581.
7. Cochet M, Pannetier C, Regnault A, Darche S, Leclerc C, Kourilsky P: **Molecular detection and in vivo analysis of the specific T cell response to a protein antigen.** *Eur J Immunol* 1992, **22**(10):2639–2647.
8. Gorski J, Yassai M, Zhu X, Kissela B, Keever C, Flomberg N: **Circulating T cell repertoire complexity in normal individuals and bone marrow recipients analyzed by CDR3 size spectratyping. Correlation with immune status.** *J Immunol* 1994, **152:**5109–5119.
9. Kepler T, He M, Tomfohr J, Devlin B, Sarzotti M, Markert M: **Statistical analysis of antigen receptor spectratype data.** *Bioinformatics* 2005, **21**(16):3394–3400.
10. Pannetier C, Even J, Kourilsky P: **T-cell repertoire diversity and clonal expansions in normal and clinical samples.** *Immunol Today* 1995, **16:**176.
11. Ciupe S, Markert M, Devlin B, Kepler T: **The dynamics of T-cell receptor repertoire diversity following thymus transplantation in Digeorge anomaly.** *PLoS Comp Biol* 2009, **5:**1–13.
12. Pannetier C, Levraud J, Lim A, Even J, Kourilsky P: *The Immunoscope Approach for the Analysis of T Cell Repertoires.*(Oksenberg JR, ed.): The Antigen T Cell Receptor: Selected Protocols and Applications, 1998, Chapman and Hall,New York.
13. Ferrand C, Robinet E, Contassot E, Certoux J, Lim A, Hervé P, Tiberghien P: **Retrovirus-mediated gene transfer in primary T lymphocytes: influence of the transduction/selection process and of ex vivo expansion on the T cell receptor beta chain hypervariable region repertoire.** *Hum Gene Ther* 2000, **11:**1151–1164.
14. Kook H, Risitano A, Zeng W, Wlodarski M, Lottemann C, Nakamura R, Barrett J, Young N, Maciejewski J: **Changes in T-cell receptor VB repertoire in aplastic anemia: effects of different immunosuppressive regimens.** *Blood* 2002, **99:**3668–3675.
15. Bomberger C, Singh-Jairam M, Rodey G, Guerriero A, Yeager A, Fleming W, Holland HK, Waller E: **Lymphoid reconstitution after autologous PBSC transplantation with FACS-sorted CD34+ hematopoietic progenitors.** *Blood* 1998, **91:**2588–2600.
16. Peggs K, Verfuerth S, D'Sa S, Yong K, Mackinnon S: **Assesing diversity: immune reconstitution and T-cell receptor BV spectratype analysis following stem cell transplantation.** *J Haematol* 2003, **120:**154–165.
17. Wu C, Chillemi A, Alyea E, Orsini E, Neuberg D, Soiffer R, Ritz J: **Reconstitution of T-cell receptor repertoire diversity following T-cell depleted allogeneic bone marrow transplantation is related to hematopoietic chimerism.** *Blood* 2000, **95:**352–359.
18. Press W, Teukolsky S, Vetterling W, Flannery B: *Numerical Recipes with Source Code CD-ROM: The Art of Scientific Computing, 3rd edition*. Cambridge University Press, New York, NY; 2007.

19. Markert M, Sarzotti M, Ozaki D, Sempowski G, Rhein M, Hale L, Deist FL, Alexieff M, Li J, Hauser E, Haynes B, Rice H, Skinner M, Mahaffey S, Jaggers J, Stein L, Mill M: **Thymic transplantation in complete DiGeorge syndrome: immunologic and safety evaluations in 12 patients.** *Blood* 2003, **102:**1121–1130.
20. Erdélyi A: *Asymptotic Expansions*. New York: Dover Publications; 1956.

doi:10.1186/1471-2172-14-35
Cite this article as: Ciupe *et al.*: Quantification of total T-cell receptor diversity by flow cytometry and spectratyping. *BMC Immunology* 2013 **14**:35.