

SpA: web-accessible spectratype analysis: application to investigate the development of TCR diversity in a patient with complete DiGeorge syndrome

Min He,¹ Blythe H. Devlin,² M. Louise Markert,^{2,3} Marcella Sarzotti,³ and Thomas B. Kepler^{1,3,*}

¹Department of Biostatistics & Bioinformatics, ²Department of Pediatrics, ³Department of Immunology, Duke University Medical Center, Durham, NC 27710

Min He:

Box 90090 DUMC, Durham, NC 27708

Phone: +1 919 684-6055; Fax: +1 919 668-2465; Email: min.he@duke.edu

Blythe H. Devlin:

Box 3068 DUMC, Durham, NC 27710

Phone: +1 919 668-1546/ 6001; Fax: +1 919 681-8676; Email: devli002@mc.duke.edu

M. Louise Markert:

Box 3068 DUMC, Durham, NC 27710

Phone: +1 919 684-6263; Fax: +1 919 681-8676; Email: marke001@mc.duke.edu

Marcella Sarzotti:

Box 2926 DUMC, Durham, NC 27710

Phone: +1 919 684-6373; Fax: +1 919 684-4288; Email: msarzott@duke.edu

Thomas B. Kepler:

Box 90090 DUMC, Durham, NC 27708

Phone: +1 919 681-0620; Fax: +1 919 668-2465; Email: kepler@duke.edu

The authors who will be presenting the paper if it is accepted: Min He and Blythe H. Devlin.

Keywords: SpA (spectratype analysis); D_{KL} (Kullback-Leibler divergence); TCR (T cell receptor); CDR3 (third complementary-determining region); DGS (DiGeorge syndrome)

Abstract

T-cell receptors (TCR) and immunoglobulins (Ig) are somatically diversified through an active process; the status of the repertoire diversity in these antigen receptors provides an insight into the immune health of the subject and a window onto the dynamical workings of lymphocyte population dynamics. One measure of this diversity is the distribution of lengths of the third complementary-determining region (CDR3) of TCR and Ig. Antigen-receptor spectratype analysis is the technique commonly used to measure this distribution, and thereby the receptor repertoire diversity for basic immunological experiments and medical procedures, notably hematopoietic stem-cell and lymphoid tissue transplantations.

A web-accessible spectratype analysis (SpA) has been developed for data management, statistical analysis, and visualization of TCR and Ig spectratype data of human and mouse subjects. Users upload spectratype data from their analyzer to SpA, which saves the raw data and user-defined and user-supplied supplementary covariates to a secure database. The analysis engine performs several data analyses, providing estimated relative frequencies, and summary statistics. The visualization engine presents analyzed histogram results in a Java applet and a PNG (Portable Network Graphics) image. Relevant statistics, histogram data, and processing data are also provided for downloading or online browsing. One of the statistics provided, the Kullback-Leibler divergence (D_{KL}), is an objective measure of the level of TCR diversity. Given spectratypes obtained from different dates, SpA can produce a plot of D_{KL} vs. time giving a convenient visual representation of the progression of TCR diversity over the course of a treatment study. This paper presents an analysis of the development of TCR diversity in a patient with complete DiGeorge syndrome (lacking all thymic function) who has undergone thymus tissue transplantation for T cell reconstitution.

Availability: The SpA service is freely accessible at <https://spa.dulci.org/>. Additional technical support and specialized statistical analysis and consultation are available by arrangement with the authors and, depending on the service requested, may be subject to fee.

1. Introduction

T-cell receptors (TCR) are membrane-bound protein complexes, expressed on T lymphocytes that bind antigenic peptides presented in the context of molecules of the Major Histocompatibility Complex (MHC) by other host cells. TCR β -chain variable (TCRBV) repertoire diversity results from the largely random recombination of gene segments known as variable (V), diversity (D), and joining (J) gene segments (Tonegawa, 1983). In addition to randomness in the choice of segments used in any particular recombination, there is randomness in the nature of the resulting junctions: the locations of the precise recombination sites have a random component and there are often non-templated nucleotides added to the junctions. These latter processes lead to a distribution in the lengths of the third complementarity determining region (CDR3) of the TCR, which encompasses both the VD and the DJ junctions. The diversity of the TCR repertoire is essential for protection against the broad array of pathogens that might be encountered, and is, furthermore, an important indicator of immune health.

Limited diversity of the TCR repertoire is observed in a number of clinical states related to immune deficiencies and therapies of immunoreconstitution. Thus, to estimate the diversity of the TCR repertoire T cells can be isolated and the distribution of CDR3 lengths of the various TCR beta gene families can be assessed by spectratyping. Subjects with limited repertoire diversity will exhibit only limited CDR3 lengths for the various TCRBV families.

Spectratype analysis estimates the CDR3 length distribution for each functional TCRBV subfamily of the T cell mRNA by reverse transcription of the RNA followed by PCR amplification and size separation of the amplified products (Cochet et al., 1992; Pannetier et al., 1993, Pannetier et al., 1997). Software developed for DNA sequence analysis, such as GeneScan[®] (Applied Biosystems, Foster City) and Genotester[®] (Amersham, Uppsala), is typically used for peak detection and intensity quantification. Spectratype data is often analyzed qualitatively, using expert judgment to classify CDR3 length histograms into a small number of categories, such as *Gaussian*, *polyclonal skewed* and *oligoclonal* (Markert et al., 2003, Sarzotti et al., 2003). While this subjective method has yielded much useful information in both basic biological and clinical settings, there is a real need for an objective, statistically rigorous approach to spectratype analysis, as well as a data management system that integrates seamlessly the spectratype data, the relevant covariate data, visualization and statistical analysis.

We have developed an objective, a statistically rigorous approach to quantitative spectratype analysis based on the hierarchical-relative multinomial model (Kepler et al., 2005) and a web-accessible spectratype analysis (SpA) system that integrates statistical approach with tools for investigators to quantitatively analyze spectratype data, which allows users to submit spectratype data and view existing analysis on the web. After an investigator submits his/her spectratype data, the histogram image and relevant statistical data of the analyzed results can be navigated on the web interactively. If the investigator provides valid email addresses when the spectratype data is submitted, the analyzed results are automatically returned by email when the whole analysis is done (He et al., 2005). Here we describe an application in assessing the development of TCR diversity in a patient with complete DiGeorge syndrome (DGS) who has undergone thymus tissue

transplantation for T cell reconstitution is given.

2. Methods and implementation

TCR diversity is essential to the effective functioning of the immune system; careful measurement of the TCR diversity can provide valuable information on the state of the immune system. We have started from first principles and derived statistical methods that allow spectratype data to be quantified and used for hypothesis tests and parameter estimation. The primary statistic that arises, the Kullback-Leibler divergence (D_{KL}), is generally a measure of difference between two probability functions, and in the appropriate context is a natural measure of deviation from maximum diversity (Kepler et al., 2005).

SpA has been implemented in Java programming languages, and interconnected to a relational database Oracle 10g on Linux. Analyzed statistical results and its relevant input data are stored in Oracle database for investigator's retrieving and viewing the analyzed results at any time, as well as extending the system to do other kinds of statistics in the future (He et al., 2005).

3. Application

DGS is a rare congenital (i.e. present at birth) disease whose symptoms vary greatly between individuals but commonly include a history of recurrent infection due to defects in the thymus, heart defects, hypoparathyroidism, and characteristic facial features. DGS may be associated with a deletion in chromosome 22. Some effects, for example the cardiac problems and the hypoparathyroidism, can be treated either surgically or therapeutically. Less than 1% of infants with DGS is born with no detectable thymus function and, thus, lacks T cells. This is called complete DGS and is a fatal condition because the lack of T cells leads to opportunistic infections. Thymic transplantation is a promising, well tolerated therapy for infants with complete DGS. (Markert et al., 1998, 1999, 2003). To illustrate one of SpA's applications, we assess the development of TCR diversity in a research subject with complete DGS who has undergone thymus tissue transplantation. The subject in the example developed clonal T cells athymically prior to treatment (Markert et al., 2004A, Markert et al., 2004B).

3.1 Spectratype analysis

Fig. 1A, 1B, 1C, and 1D show the spectratype analysis for patient DIG102 at 1 month prior to transplantation (Fig. 1A), and at 3.5 months (Fig. 1B), 9.5 months (Fig. 1C) and 21 months (Fig. 1D) post-transplantation. In Figure 1, the repertoire observed prior to transplantation was very limited, indicating clonal populations of T cells. By 9.5 months post transplantation, the repertoire had normalized.

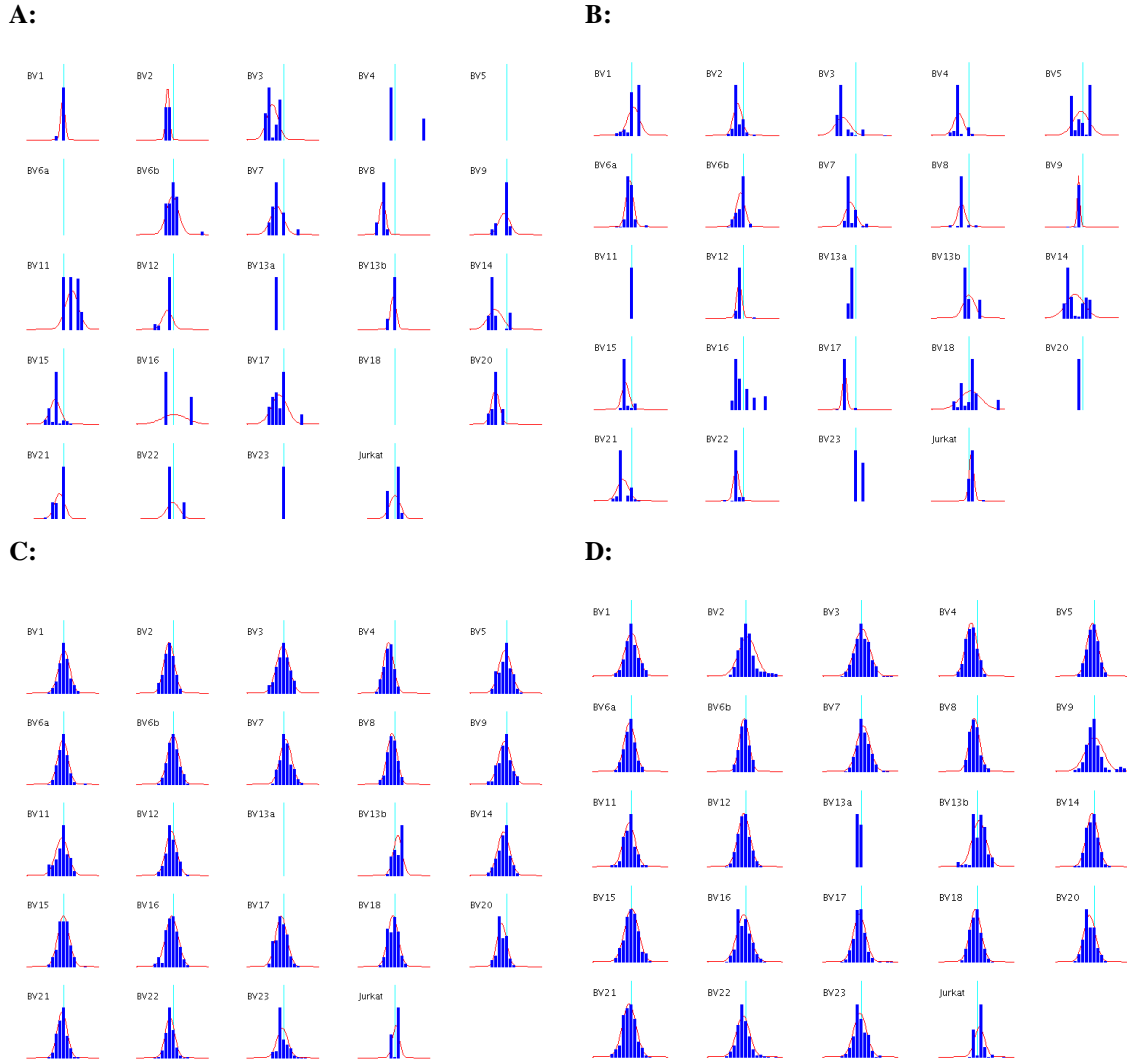


Fig. 1 Spectratypes from patient DIG102 A) pre-transplantation; B) at 3.5 months post-transplantation; C) at 9.5 months post-transplantation; D) at 21 months post-transplantation. All testing was on peripheral blood mononuclear cells.

In a normal individual's CD4 T cells, every TCRBV family is usually represented in the TCRBV repertoire, and the distribution of CDR3 lengths within each family is quasi-Gaussian. We do occasionally observe technical problems and a family may fail to amplify, such as panel C, family BV13a. The repertoire in CD8 T cells usually has some skewed patterns e.g. BV13b in panel C. These represent T cell responses to infection.

3.2 D_{KL} over days post-transplant for all TCRBV families

The D_{KL} measures the degree of discrepancy between the observed data histogram and the controls' averaged data histogram to which that histogram is most similar. D_{KL} is

given by $D_{KL} = \sum_{i=1}^{n_{\text{peak}}} f_i \log \frac{f_i}{g_i}$, where f_i is the observed proportion of total peak area corresponding to CDR3 length i and g_i is the controls' averaged proportion of total peak area corresponding to CDR3 length i . Therefore, a D_{KL} of smaller indicates that the data histogram is closer to controls' averaged one; larger D_{KL} values indicate greater deviation from the controls' averaged one. Fig. 2 shows, for subject DIG102, the D_{KL} (Y-axis) trend for each TCRBV family as well as the average of all the TCRBV families at different time points post-transplantation (X-axis). In Fig. 2, symbol "■" means that there is no D_{KL} value (NA) of next time point in the related family, while "✕" means that there is NA of previous time point in the related family. While computing the average of D_{KL} for all families at one time point, the "NA" value is ignored.

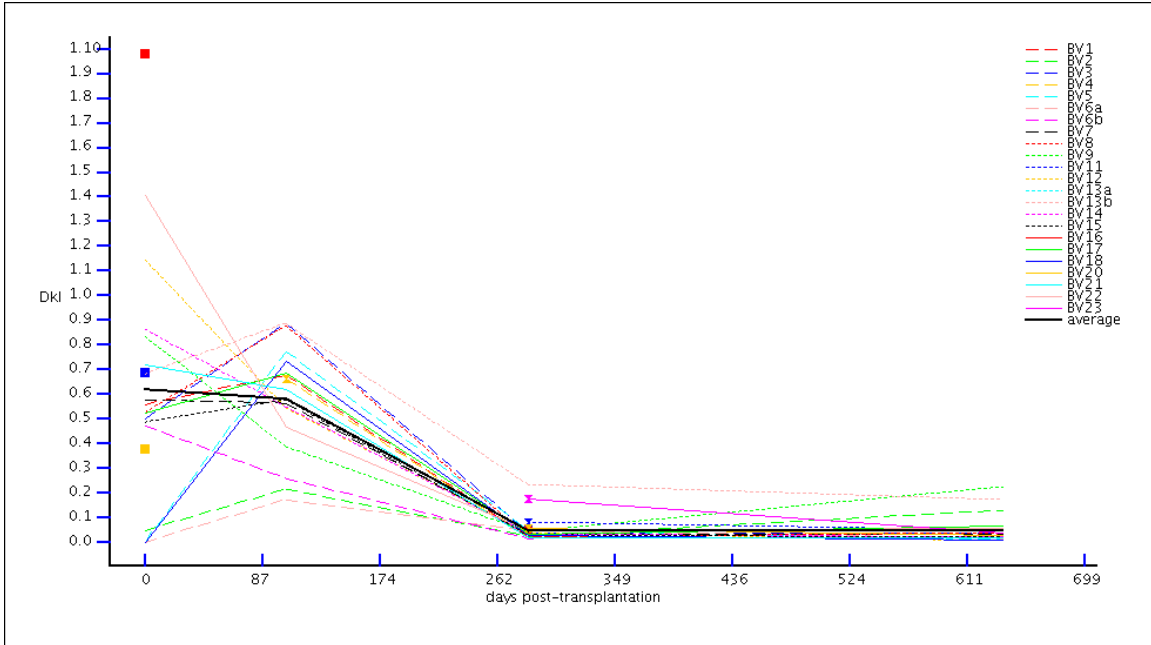


Fig. 2 Subject DIG102 D_{KL} versus days post-transplantation for each TCRBV family and the average of all families

The dark line in Fig. 2 is the average D_{KL} post-transplantation. From Fig. 2, we found that at 3.5 months post-transplantation, when the spectratype profile shows only a few distinct CDR3 lengths for each family, all the D_{KL} statistics are non-zero and range in value from 0.1 to 1.0. At 9.5 months post-transplantation, the D_{KL} statistic for every TCRBV family in patient DIG012 is < 0.2 indicating little divergence from the controls' average for most TCRBV families. Therefore by using a statistical analysis, we could measure the normalization of the TCRBV families in its increase in diversity in the T cell population around 9.5 months post-transplantation.

4. Conclusions

We have found that D_{KL} provides a summary statistic for the normality of the spectratype. Where D_{KL} is inordinately large, one can then turn to the skewness and kurtosis to further characterize the specific type of deviation from normality. The statistic result has been proven to be repeatable, reliable, highly sensitive to small changes in repertoire and

remarkably indicative of the stage of peripheral T cell development post-transplantation as well as other kinds of D_{KL} trend over post-transplantation.

SpA is a web-based spectratype analysis tool for investigators' quantitatively analyzing spectratype data, which allows users to submit spectratype data and view existing analysis interactively on the web. After investigator submits his/her spectratype data, the analyzed histogram results can be presented in a Java applet or as a PNG (Portable Network Graphics) image. The relevant statistics, histogram data, and processing data are also provided for investigator downloading and online browsing. And the analyzed results can also be sent to the investigator by email when the whole analysis is done. An approach of the statistics provided, the Kullback-Leibler divergence, is an objective measure of the level of TCR diversity. Given spectratypes obtained from different dates, SpA can produce a plot of D_{KL} vs. time giving a convenient visual representation of the progression of TCR diversity over the course of a treatment study. This paper briefly describes the methods and implementation of SpA. As an illustrative application, we present an analysis of the development of TCR diversity in a research subject with complete DGS who has undergone thymus tissue transplantation. To maintain the highest ethical standards in all research involving human subjects, the SpA system was set up a secure web system to prevent non-relevant user from browsing other investigators' analysis results. Any investigator can freely submit his/her spectratype data to SpA system and the analyzed results can be navigated on the web. It is available at <https://spa.dulci.org/>.

Acknowledgements

The authors thank Drs Nelson Chao and Congxiao Liu for granting us access to their spectratype data during the development of this project. We thank Lindsay Cowell, Shaza Fadel, Jun Lu, and Faheem Mitha for helpful discussions, and Bill Zeggert, Dan Ozaki and Jie Li for technical assistance. This work was supported financially by the Duke University Center for Translational Research NIH 5 P30 AI051445-03, grants R01 AI47040, R01 AE54843, and the Southeast Regional Center for Biodefense and Emerging Infections NIH U54 AI057157-02.

References:

Cochet, M., Pannetier, C., Regnault, A., Darche, S., Leclerc, C. & Kourilsky, P., 1992. Molecular detection and in vivo analysis of the specific T cell response to a protein antigen. *Eur. J. Immunol.* 22, 2639-2647.

He, M., Tomfohr, J.K., Devlin, B.H., Markert, M.L., Sarzotti, M., and Kepler, T.B., 2005. SpA: web-accessible spectratype analysis: data management, statistical analysis and visualization. *Bioinformatics.* 21, 3697-3699.

Kepler, T.B., He, M., Tomfohr, J.K., Devlin, B.H., Sarzotti, M., and Markert, M.L., 2005. Statistical Analysis of Antigen Receptor Spectratype Data. *Bioinformatics.* 21, 3394-3400.

Markert, M.L., Hummell, D.S., Rosenblatt, H.M., Schiff, S.E., Harville, T.O., Williams, L.W., Schiff, R.I., Buckley, R.H., 1998. Complete DiGeorge syndrome: persistence of profound immunodeficiency. *J. Pediatr.* 132(1), 15-21.

Markert, M.L., Boeck, A., Hale, L.P., Kloster, A.L., McLaughlin, T.M., Batchvarova, M.N., Douek, D.C., Koup, R.A., Kostyu, D.D., Ward, F.E., Rice, H.E., Mahaffey, S.M., Schiff, S.E., Buckley, R.H., Haynes, B.F., 1999. Transplantation of thymus tissue in complete DiGeorge syndrome. *N. Engl. J. Med.* 341(16), 1180-1189.

Markert, M.L., Sarzotti, M., Ozaki, D.A., Sempowski, G.D., Rhein, M.E., Hale, L.P., Le Deist, F., Alexieff, M.J., Li, J., Hauser, E.R., Haynes, B.F., Rice, H.E., Skinner, M.A., Mahaffey, S.M., Jagers, J., Stein, L.D., Mill, M.R., 2003. Thymus transplantation in complete DiGeorge syndrome: immunologic and safety evaluations in 12 patients. *Blood* 102(3), 1121-1130.

Markert, M.L., Alexieff M.J., Li, J., Sarzotti, M., Ozaki, D.A., Devlin, B.H., Sedlak, D.A., Sempowski, G.D., Hale, L.P., Rice, H.E., Mahaffey, S.M., and Skinner, M.A., 2004A. Postnatal thymus transplantation with immunosuppression as treatment for DiGeorge syndrome. *Blood* 104, 2574-2581.

Markert, M.L., Alexieff, M.J., Li, J., Sarzotti, M., Ozaki, D.A., Devlin, B.H., Sempowski, G.D., Rhein, M.E., Szabolcs, P., Hale, L.P., Buckley, R.H., Coyne, K.E., Rice, H.E., Mahaffey, S.M., Skinner, M.A., 2004B. Complete DiGeorge syndrome: development of rash, lymphadenopathy, and oligoclonal T cells in 5 cases. *J. Allergy Clin. Immunol.*, 113(4), 734-41.

Pannetier, C., Cochet, M., Darche, S., Casrouge, A., Zoller, M., Kourilsky, P., 1993. The size of the CDR3 hypervariable regions of the murine T-cell receptor B chains vary as a function of the recombined germ-line segments. *Proc. Natl. Acad. Sci. U.S.A.* 90, 4319.

Pannetier, C., Levraud, J.P., Lim, A., Even, J., Kourilsky, P., 1997. The spectratype approach for the analysis of T-cell repertoires. *The Human Antigen T Cell Receptor : selected protocols and applications.* J. R. Oksenberg, Austin, TX, p. 287.

Sarzotti, M. et al., 2003. T cell repertoire development in humans with SCID after nonablative allogeneic marrow transplantation. *J. Immunol.*, 170, 2771-2718.

Tonegawa, S., 1983. Somatic generation of antibody diversity. *Nature*, 302(5909), 575-581.