

SpA: Web-accessible Spectratype Analysis: data management, statistical analysis and visualization

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

¹ Department of Biostatistics & Bioinformatics and Center for Bioinformatics & Computational Biology, Institute for Genome Sciences & Policy, ² Department of Pediatrics, and ³ Department of Immunology Duke University Medical Center, Durham, NC 27708

ABSTRACT

Summary: SpA is a Web-accessible system for the management, visualization and statistical analysis of T-cell receptor spectratype data. Users upload data from their spectratype analyzers to SpA, which saves the raw data and user-defined supplementary covariates to a secure database. The statistical engine performs several data analyses and statistical summaries. The visualization engine displays spectratype histograms in a Java applet and in an image file suitable for download. All of these results are also saved on the database and remain accessible to the user. Additional statistical tools specific to the analysis of multiple spectratypes are also available through the SpA interface.

Availability: The service is freely accessible via the web at <http://www.duke.edu/~kepler/spa.html>. 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.

Contact: kepler@duke.edu

1. INTRODUCTION

The vertebrate immune system depends crucially on the generation and maintenance of an enormous diversity of specialized receptors, called T cell receptors (TCR) and B cell receptors, are used to sense the presence of microbial pathogens (see, eg, Janeway et al, 2005). Loss of this diversity can compromise the effectiveness of immune surveillance; it can also signal other underlying deficiencies. Spectratype analysis was developed for gauging TCR diversity by measuring the length-distribution of the third complementarity determining region (CDR3) in rearranged T-cell receptor β -chain (TCRB) genes using PCR amplification and size separation of the amplified products (Cochet et al, 1992; Pannetier et al., 1993; Pannetier, 1997).

Spectratype data is typically analyzed subjectively, using expert judgment to classify CDR3 length histograms into a small number of categories (see, eg, Sarzotti et al., 2003). Although such analyses have yielded much useful information in both basic biological and clinical settings, we developed an objective, statistically rigorous approach to quantitative spectratype analysis based on the hierarchical-relative multinomial model (Kepler et al., 2005). Here we describe *SpA* (Spectratype Analysis), a comprehensive data management system that integrates these statistical tools with tools for the management and visualization of spectratype and covariate data. A system designed with similar purposes in mind was developed by Collette and Six (2002) for use within spreadsheet programs and is available from those authors as well.

* To whom correspondence should be addressed. Tel.: +1-919-681-0620; fax: +1-919-668-2465.
E-mail address: kepler@duke.edu (T. B. Kepler)

2. IMPLEMENTATION AND FUNCTIONALITY

SpA is written in Fortran90, C++, and Java, and is interconnected to a relational database Oracle on a Linux platform.

Raw spectratype data, processed data, results of summary statistical analysis and graphics are stored in an Oracle database. These individual data are available for comparative analyses of multiple datasets, hypothesis testing and statistical modeling using the integrated statistical engine. The system architecture of SpA is briefly diagrammed in Figure 1.

2.1 Data Fetching and Preprocessing

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. The software typically produces 24 spectratype data files (one for each primer pair, or, roughly, each TCRBV family). In SpA, the user compresses these 24 files into a single ZIP-formatted archive (zipfile) which is then uploaded through the input interface. After the archive has

been uploaded, the decompression module opens the archive and stores the individual files in a temporary directory. The data are then preprocessed to convert the called peak locations and sizes into CDR3 lengths and relative abundances and stores these preprocessed data along with the raw peak calls and a preprocessing log detailing the data conversion in the database.

SpA registered users (registration is free of charge) can customize the flexible interface to facilitate the use of the software for his/her specific purpose, specifying any number of user-defined covariates (such as subject age, clinical status, etc) to be

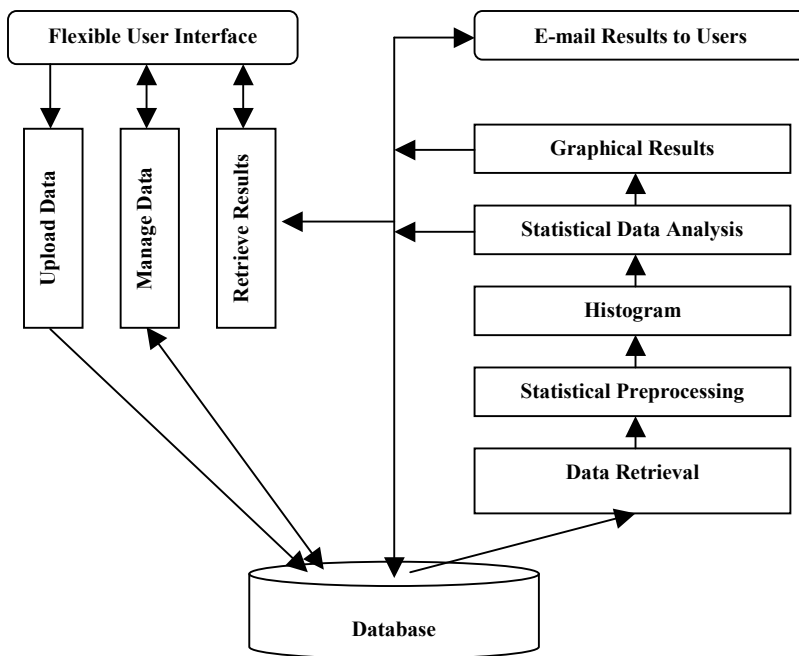


Fig. 1. SpA system architecture

stored with the spectratype data and used in subsequent data analyses. In subsequent sessions, the interface will display the previously-entered covariate descriptors. Additional covariates can be added at any time. The interface for adding covariates is shown in Figure 2A. To comparative or integrative analyses, users can add or modify covariates at any time, even after the spectratype data is uploaded to the SpA system.

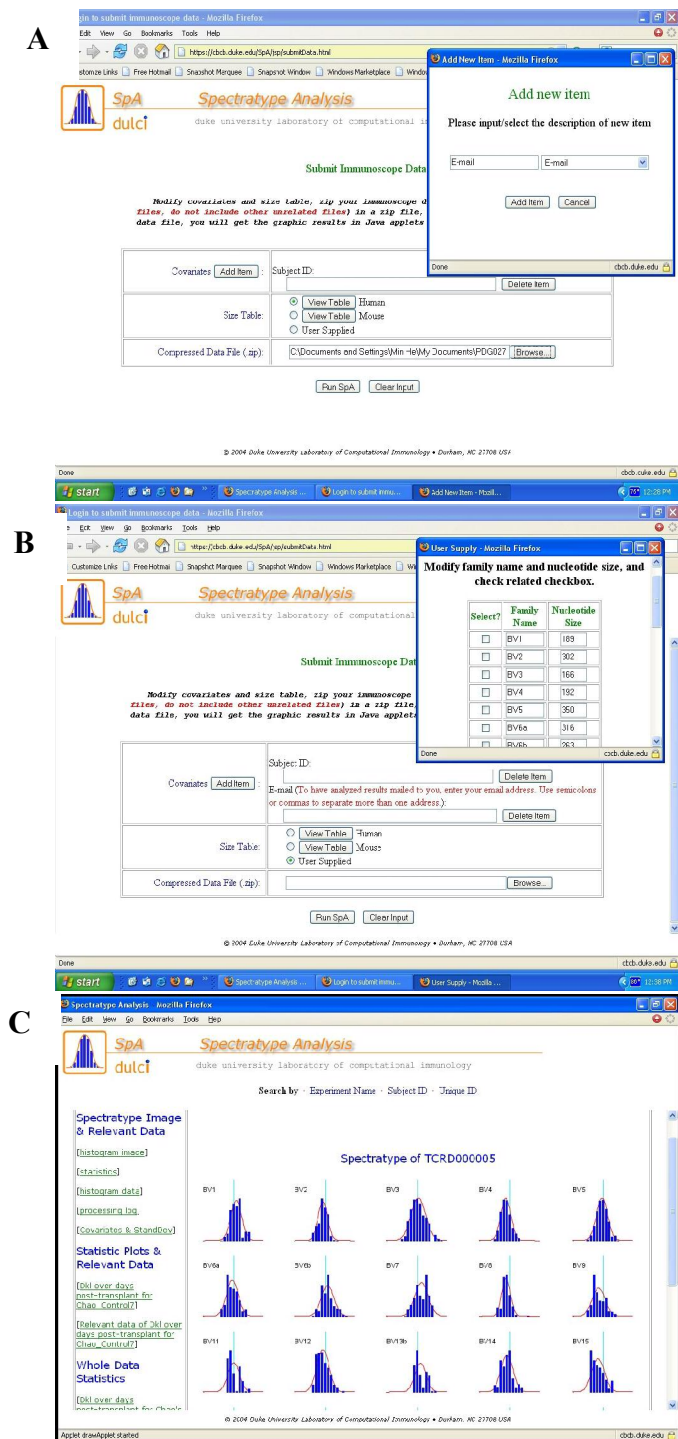


Fig. 2 SpA screenshots. A: Interface for adding covariates; B: Interface for modifying the size conversion table; C: Interface for spectratype visualization.

Each TCRBV family produces a CDR3 length histogram which is rendered along with a curve showing the population mean curve appropriate to that family (Figure 2C). The image file is

Because different users may use different PCR primers to amplify CDR3, spectratype analysis requires the specification of a *Size Conversion Table* that provides the correspondence between amplicon absolute length and CDR3 length for each receptor family. Using SpA, the user can select from available standard Size Conversion Tables or supply his/her custom Size Conversion Tables, which will be stored for subsequent use by that user. The interface for modifying the Size Conversion Table is shown in Figure 2B.

2.2 Statistical Data Analysis

After preprocessing, the statistical data analysis modules are invoked to perform summary analyses of the histograms for each TCRBV family and for the assay as a whole. This information is made available for immediate viewing and stored for later retrieval.

Additional statistical analyses and visualizations are available for comparison and integration of multiple assays. For example, if the user has supplied categorical covariate data, such as presence or absence of a given drug therapy, statistical comparisons of repertoire diversity based on that covariate can be performed. If continuous covariate data are supplied, such as date-post-intervention, regressions of TCR diversity can be computed, relevant hypothesis tests performed, and plots of these fits produced. In addition to the specific procedures provided by SpA, we have developed an interface that integrates the powerful, general purpose data analysis package R (<http://www.r-project.org/>) to the system for more specialized procedures. These procedures can be designed and carried out by the user (It is free for academic users), or, through special arrangement, by the authors' research team.

2.3 Visualization

produced using the Java 2D image package, and can be saved on the user's machine as a PNG (Portable Network Graphics) file as well as being stored into the SpA database.

2.4 Security and Access

It is essential to maintain strict data confidentiality for both ethical and scientific reasons. Our system is developed on a secure server with secure login-based access. Academic users may register free of charge and receive the benefits of the customizable interfaces discussed above, in addition to data security. Unregistered users can gain access through the public interface, but in this case, leave their data open for public access. Technical support and assistance with more complex data analyses are available by arrangement with the authors. SpA is publicly available at <https://www.duke.edu/~kepler/spa.html>.

3. CONCLUSION

SpA is a Web-accessible spectratype statistical analysis and data management system. It allows users to submit spectratype and general covariate data for processing, analysis and visualization. Existing analyses can be interactively retrieved and used in comparative and integrative analyses. A detailed tutorial is available on the SpA site.

ACKNOWLEDGEMENTS

The authors 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 and the Southeast Regional Center for Biodefense and Emerging Infections NIH U54 AI057157-02 as well as grants R01 AI 47040 and R01 AI 54843 from the NIH.

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.
- Collette, A., Six, A. (2002) ISEApeaks: an Excel platform for GeneScan and Immunoscope data retrieval, management and analysis. *Bioinformatics* **18**:329-30.
- Janeway, C.A., Travers, P., Walport, M., Shlomchik, M.J. (2004) *Immunobiology*, 6th edition, Garland Publishing.
- Kepler, T.B., He, M., Tomfohr, J.K., Devlin, B.H., Sarzotti M., Markert, M.L. (2005) Statistical Analysis of Antigen Receptor Spectratype Data, *Bioinformatics*, in press.
- 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.*, **90**, 4319.
- Pannetier C, Levraud J-P, Lim A, Even J, Kourilsky P (1997) The Immunoscope Approach for the Analysis of T Cell Repertoires. In *The Antigen T Cell Receptor: Selected Protocols and Applications*. Edited by Oksenberg JR. Austin, TX: Landes; 1997: 287-325.
- Sarzotti M, Patel DD, Li X, Ozaki DA, Cao S, Langdon S, Parrott RE, Coyne K, Buckley RH. (2003) T cell repertoire development in humans with SCID after nonablative allogeneic marrow transplantation. *J Immunol.*, **170** (5), 2711-8.