Bioinformatics, 36(6), 2020, 1930–1932
doi: 10.1093/bioinformatics/btz804
Advance Access Publication Date: 30 October 2019
Applications Note

Sequence analysis

ImSpectR: R package to quantify immune repertoire diversity in spectratype and repertoire sequencing data

Martijn Cordes[1,2,*], Karin Pike-Overzet[1], Marja van Eggermond[1], Sandra Vloemans[1], Miranda R. Baert[1], Laura Garcia-Perez[1], Frank J. T. Staal[1], Marcel J. T. Reinders[2,3] and Erik B. van den Akker[2,3,4]

[1]Department of Immunohematology & Blood Transfusion and [2]Leiden Computational Biology Center, Leiden University Medical Center, Leiden, The Netherlands, [3]The Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands and [4]Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on April 9, 2019; revised on September 29, 2019; editorial decision on October 24, 2019; accepted on October 25, 2019

Abstract

Summary: An effective immune system is characterized by a diverse immune repertoire. There is a strong demand for accurate and quantitative methods to assess the diversity of the immune repertoire for various (pre-)clinical applications, including the diagnosis and prognosis of primary immune deficiencies, or to assess the response to therapy. Current strategies for immune diversity assessment generally comprise the visual inspection of the length distribution of rearranged T- and B-cell receptors. Visual inspections, however, are prone to subjective assessments and thus lead to biases. Here, we introduce *ImSpectR*, a unified approach to quantify immunodiversity using either spectratype, repertoire sequencing or single cell RNA sequencing data. *ImSpectR* scores various types of deviations from the expected length distribution and integrates these into one measure, allowing for robust quantitative comparisons of immune diversity across individuals or conditions.

Availability and implementation: R-package is available for download on GitHub at https://github.com/martijncordes/ImSpectR.

Contact: m.cordes@lumc.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The immune repertoire refers to the diversity of T-cell receptors (TCR) and B cell receptors (BCR) on T- and B-lymphocytes each expressing a unique antigen-specific receptor. The most diverse part of the receptor is the third complementarity-determining region 3 (CDR3) that binds a specific antigen. The diversity of the immune repertoire is key in combatting viral infections and other pathogens. Indeed, rare inherited variations that hamper immune cell development generally lead to the development of severe primary immunodeficiencies, as no diverse immune repertoire can be formed. Patients with such primary immunodeficiencies often require hematopoietic stem cell transplantation or gene therapy (Wiekmeijer *et al.*, 2016). To monitor immune reconstitution upon treatment, accurate quantification of the immune repertoire diversity is required.

While deep-sequencing methods have been developed for CDR3 length distribution analyses, spectratyping, i.e. measuring CDR3 lengths of all possible TCR Vβ-Cβ combinations by capillary electrophoresis (Pannetier *et al.*, 1993), is still considered to be the gold standard for clinical applications. Existing methods (Gorochov *et al.*, 1998; Long *et al.*, 2006) for CDR3 length distribution analysis are typically developed to detect gross clonal abnormalities and are therefore generally not accurate enough to detect subtle changes in immunodiversity. Moreover, classical methods typically ignore sequences that result in out-of-frame rearrangements, which lead to non-functional T-cell receptors, and thus miss important information embedded in the complexity of the actual distribution patterns. As a result, CDR3 length distributions are often visually inspected, which is prone to introduce subjective variation and thus may lead to biases.

Here we present *ImSpectR*, an R package for robust scoring of immunodiversity of TCRs. Our package is able to assess the complexity of the distribution of CDR3 lengths from spectratype, ImmunoSeq repertoire sequencing or single cell RNA sequencing data from either human or mouse by scoring various types of deviations from the expected distribution and integrating these into one score.

2 Materials and methods

Briefly, amplification of CDR3 regions, either by spectratyping or sequencing, will produce fragments of different sizes

(See Supplementary Chapter S0). The distribution of these CDR3 fragment sizes represents the junctional diversity of TCRs and can thus be used to assess the diversity of the T cell repertoire. A density plot of CDR3 fragment sizes (Fig. 1A), can be used to display the junctional diversity. Peaks in this plot represent CDR3 regions of specific sizes expressed by particular rearranged Vβ segments. Fragments typically display a 3 bp size difference in length, as these constitute the rearranged receptor sequences that are transcribed *in-frame*. The rearrangements leading to CDR3 regions are stochastic, causing the overall CDR3 length distribution in naïve T cells to adhere to a Gaussian distribution. Consequently, *ImSpectR* models the different fragment lengths of *in-frame* CDR3 regions with a mixture of Gaussians. Deviations from this model reflect a skewed or otherwise suboptimal immunodiversity and is used to score CDR3 length patterns (described in Supplementary Materials). *ImSpectR* is organized in three different modules (Fig. 1A and Supplementary Materials).

Preprocessing: For spectratyping raw intensity FSA files are imported, and the fluorescent peaks are matched to the expected DNA sizes of the provided size-standard ladder. For sequencing data, tables containing CDR3 sequencing lengths annotated to a particular V family are used (see Supplementary Chapter S1).

Peak pattern detection: Dynamic time warping (Giorgino, 2009) is employed to scan for a pattern in the raw spectratypes best matching the expected Gaussian model. From the best match fragment sizes and fragment abundances are extracted and used for scoring (see Supplementary Chapter S2).

Model fitting: For peak pattern mixture modelling, a mixture of Gaussians is fitted to the matched peak pattern, i.e. for every peak a three basepair-wide Gaussian matching the height of the expected model is fitted to the matched peak pattern. A peak score is calculated by comparing properties (peak heights, peak areas and residual data) of an observed CDR3 length distribution to the expected model resulting in a score between 0 and 100 (Fig. 1A) (see Supplementary Chapter S3).

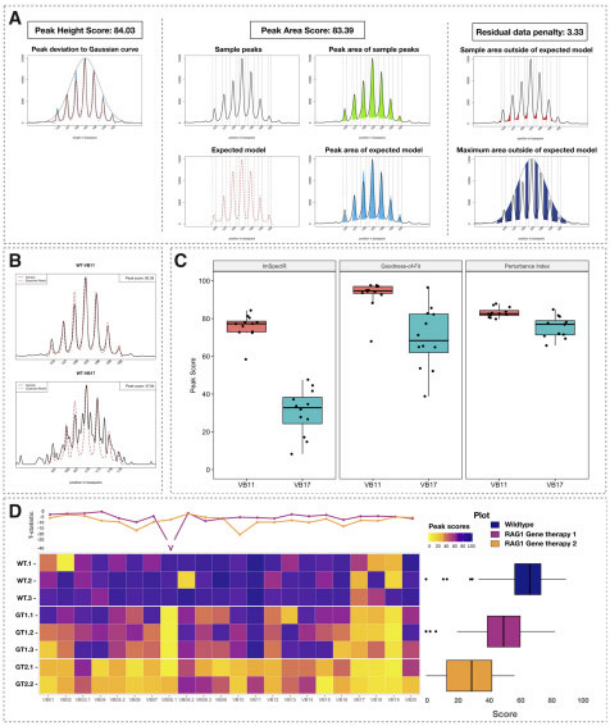

Fig. 1. (**A**) Subscores contributing to the peak score. Peak Height Score: Deviation of sample peaks from the overall Gaussian curve. Peak Area Score: Area of the model (light blue) compared to the area of the sample (light green). Residual data penalty: Area subjected to penalties (dark blue) and actual signal outside of the expected model (red). (**B**) CDR3 length distribution plots of Vβ11 and Vβ17 from a WT C57/Bl6 mouse and scored with *ImSpectR*. (**C**) Vβ11 and Vβ17 peak patterns of 12 WT mice scored with *ImSpectR*, χ^2 Goodness-of-Fit (Gorochov *et al.*, 1998) and REPERTOIRE (Long *et al.*, 2006). (**D**) Heatmap of *ImSpectR* scores per sample (vertical axis) per Vβ family (horizontal axis). Top: Differences in t-tests between WT versus GT1 and WT versus GT2 calculated per Vβ family. Right: Boxplots of scores across all Vβ families, illustrating that GT1 reconstitutes immunodiversity better than GT2 (see also Supplementary Fig. S11)

3 Results

We demonstrate the improved accuracy of *ImSpectR* over other existing methods for spectratype analysis, by analyzing Vβ17, a known dysfunctional family in the C57/Bl6 mouse strain (Louie *et al.*, 1989) (see Supplementary Chapter S4). While the Vβ17a gene in C57/Bl6 mice is rearranged and expressed, Vβ17b will never lead to a full-length and functional Vβ17 cell-surface protein, due to the presence of a premature stop codon in the germ line sequence. Consequently, Vβ17 patterns do not display a typical polyclonal pattern (Fig. 1B). Since the overall pattern of Vβ17 remains Gaussian shaped (Gapin *et al.*, 1998), competing methods for spectratype analysis score spectratypes of the non-functional Vβ17 almost as good as the functional Vβ11 (Fig. 1C, panel 2 and 3). In contrast, *ImSpectR* penalizes deviations in each individual peak, thus accurately capturing the difference between the non-functional Vβ17 and the functional Vβ11 (Fig. 1C first panel and Supplementary Figs S1 and S2) for full comparison on all Vβ families.

Next, we illustrate the use of *ImSpectR* by analyzing the diversity of immune reconstitution in Rag-1 deficient mice treated with two different experimental RAG-1 gene therapies (Pike-Overzet *et al.*, 2011). We scored 22 Vβ families measured in 3 Rag-1 deficient mice receiving either one of the two types of gene therapy ($N_{GT1} = 3$, $N_{GT2} = 2$), or transplanted wild-type stem cells ($N_{WT} = 3$) as a positive control (Supplementary Figs S3–S10). Our method comprehensively indicates a broader immune diversity in mice receiving GT1 over GT2, indicating an improved immune reconstitution by GT1 over GT2 (Fig. 1D), which is also supported by flow cytometric analysis and total T cell count (Supplementary Fig. S12). In addition, we demonstrate the accuracy of ImSpectR by scoring replicate measurements of C57/Bl6 WT control mice, which showed that the overall score distribution across the different Vβ genes are reproducible (Supplementary Chapter S3).

Finally, we demonstrate the broad applicability of *ImSpectR* by providing workflows and applying it to ImmunoSeq repertoire sequencing or single cell RNA sequencing data in either mouse or human (Supplementary Chapter S5).

4 Discussion

We present *ImSpectR*, a versatile, novel, quantitative and accurate method tailored to assess immunodiversity using CDR3 lengths from spectratype, or (single cell) RNA sequencing data from human or mouse. *ImSpectR* shows improved accuracy compared to existing methods and we demonstrate its merit by re-analyzing data from pre-clinical gene therapy studies aimed to reconstitute a diverse immune repertoire.

Acknowledgements

The authors thank Allan Thompson for additional lab work as well as, Thies Gehrmann, Amiet Chhatta and Harald Mikkers for fruitful discussions.

Funding

This work was funded by ZonMw E-RARE project [113302002 to M.C. and F.S.].

Conflict of Interest: none declared.

References

Gapin,L. *et al*. (1998) Quantitative analysis of the T cell repertoire selected by a single peptide–major histocompatibility complex. *J. Exp. Med*., **187**, 1871–1883.

Giorgino,T. (2009) Computing and visualizing dynamic time warping alignments in R: the DTW package. *J. Stat. Softw*., **31**, 1–24.

Gorochov,G. *et al*. (1998) Perturbation of CD4+ and CD8+ T-cell repertoires during progression to AIDS and regulation of the CD4+ repertoire during antiviral therapy. *Nat. Med*., **4**, 215–221.

Long,S.A. *et al*. (2006) Standardized analysis for the quantification of Vbeta CDR3 T-cell receptor diversity. *J. Immunol. Methods*, **317**, 100–113.

Louie,M.C. *et al*. (1989) Identification and characterization of new murine T cell receptor β chain variable region (Vβ) genes. *J. Exp. Med*., **170**, 1987–1998.

Pannetier,C. *et al*. (1993) The sizes of the CDR3 hypervariable regions of the murine T-cell receptor beta chains vary as a function of the recombined germ-line segments. *Proc. Natl. Acad. Sci. USA*, **90**, 4319–4323.

Pike-Overzet,K. *et al*. (2011) Correction of murine Rag1 deficiency by self-inactivating lentiviral vector-mediated gene transfer. *Leukemia*, **25**, 1471–1483.

Wiekmeijer,A.S. *et al*. (2016) Identification of checkpoints in human T-cell development using severe combined immunodeficiency stem cells. *J. Allergy Clin. Immunol*., **137**, 517–526.