Bioinformatics Advance Access published October 13, 2016

Sequence analysis

ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data

Vojtech Bystry[1#], Tomas Reigl[1#], Adam Krejci[1,2#], Martin Demko[1], Barbora Hanakova[1], Andrea Grioni[1,3], Henrik Knecht[4], Max Schlitt[4], Peter Dreger[5], Leopold Sellner[5], Dietrich Herrmann[4], Marine Pingeon[6], Myriam Boudjoghra[6], Jos Rijntjes[7], Christiane Pott[4], Anton W. Langerak[8], Patricia J. T.A. Groenen[7], Frederic Davi[6], Monika Brüggemann[4] and Nikos Darzentas[1,*], also on behalf of EuroClonality-NGS

[1]CEITEC - Central European Institute of Technology, Masaryk University, Brno, Czech Republic, [2]RECAMO, Masaryk Memorial Cancer Institute, Zluty kopec 7, 65653, Brno, Czech Republic, [3]Centro Ricerca Tettamanti, Clinica Pediatrica, Università di Milano-Bicocca, Ospedale San Gerardo/Fondazione MBBM, Monza, Italy, [4]Department of Hematology, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany, [5]Department of Medicine V, University Hospital Heidelberg, Im Neuenheimer Feld 410, 69120 Heidelberg, Germany, [6]Department of Hematology, Hopital Pitié-Salpêtrière and Pierre et Marie Curie University, Paris, France, [7]Department of Pathology, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands, [8]Department of Immunology, Erasmus MC, University Medical Center, Rotterdam, The Netherlands

*To whom correspondence should be addressed.

#Authors contributed equally.

Associate Editor: Dr. Inanc Birol

Abstract

Motivation: The study of immunoglobulins and T cell receptors using next-generation sequencing has finally allowed exploring immune repertoires and responses in their immense variability and complexity. Unsurprisingly, their analysis and interpretation is a highly convoluted task.

Results: We thus implemented ARResT/Interrogate, a web-based, interactive application. It can organize and filter large amounts of immunogenetic data by numerous criteria, calculate several relevant statistics, and present results in the form of multiple interconnected visualizations.

Availability and implementation: ARResT/Interrogate is implemented primarily in R, and is freely available at http://bat.infspire.org/arrest/interrogate/

Contact: nikos.darzentas@gmail.com

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

1 Introduction

Immunoglobulins (IG) and T cell receptors (TR) are highly adaptive molecular receptors responsible for antigen recognition in immunological responses. Fundamental to their adaptiveness is their enormous inherent variability, achieved through stochastic processes during B and T cell maturation. The advent of high-throughput profiling of IG and TR repertoires (Benichou et al., 2012) has been instrumental for understanding normal and pathologic immune responses, which include a wide range of diseases with an underlying immune cause. This unprecedented capability has also brought along novel and unique challenges.

The first task of immunoprofiling is sequence annotation, such as which variable (V), diversity (D), and joining (J) genes have been rearranged, or what is the sequence of the hypervariable complementarity-determining region 3 (CDR3). IMGT® (Lefranc et al., 2015) is the global reference in the field of antigen receptor sequence analysis and immunogenetic annotation.

Mining these inherently complex immunogenetic annotations of usually millions of reads and tens to hundreds of samples for biologically relevant information is a non-trivial task. There is an increasing number of published software applications to tackle this challenge, all with their unique features and advantages, but also limitations like limited interactivity (Alamyar et al., 2012; Shugay, 2015) or scope restricted to repertoire studies (Moorhouse et al., 2014) or minimal residual disease (MRD) monitoring (Giraud et al., 2014).

In this work, we put together in one application features and functionalities we believe are needed for wide-ranging *in silico* immunoprofiling. These insights are a result of collaborative efforts within the EuroClonality-NGS consortium, which strives to develop, standardize, and validate *in vitro* assays and bioinformatics for IG/TR NGS analysis.

2 Methods

ARResT/Interrogate is primarily based on R and Shiny, a framework for user interactivity and web-based accessibility. The analytical core relies on the 'data.table' R package for efficient data handling based on advanced indexing techniques. Therefore, ARResT/Interrogate is able to maintain sufficient responsiveness even with datasets of tens of thousands of clonotypes from millions of reads and dozens of samples.

ARResT/Interrogate has four step-wise functions: input processing, data selection and filtering, comparative calculations, and visualization.

Input processing. An integrated parser processes multiple IMGT/HighV-QUEST runs and their major immunogenetic annotations. Of these, the V, D, and J genes and alleles are combined with the amino acid sequence of the junction (which encompasses the CDR3) to construct IMGT-like clonotypes (Li et al., 2013). These annotations are referred to as 'feature types' and their corresponding individual values as 'features'; for example, feature type "V gene" contains feature "TRBV20-1" (Figure 1).

Data selection and filtering. Users can annotate samples with arbitrary metadata (e.g. patient data, sampling dates) and use these to select and group samples of interest. The next necessary step is to select feature types to focus on. This creates a table of abundance per feature per sample, with abundance expressed as relative or absolute count of reads or clonotypes. Individual features can be filtered in or out using a combination of four filters: abundance, variation across samples, correlation of abundance profiles across samples, and text regular expression (see supplementary section S2.1).

Comparative calculations. ARResT/Interrogate can calculate and visualize differences between samples and features. Samples are compared on the basis of the abundance of a single feature (e.g. TRBV20-1), or an entire feature type (e.g. V gene). Features are compared on the basis of their abundance distributions across samples. Groups of samples can also be statistically compared, for example, to assess immunogenetic differences before and after therapy (S2.3). ARResT/Interrogate can also perform principal component analyses (PCA) of samples and features.

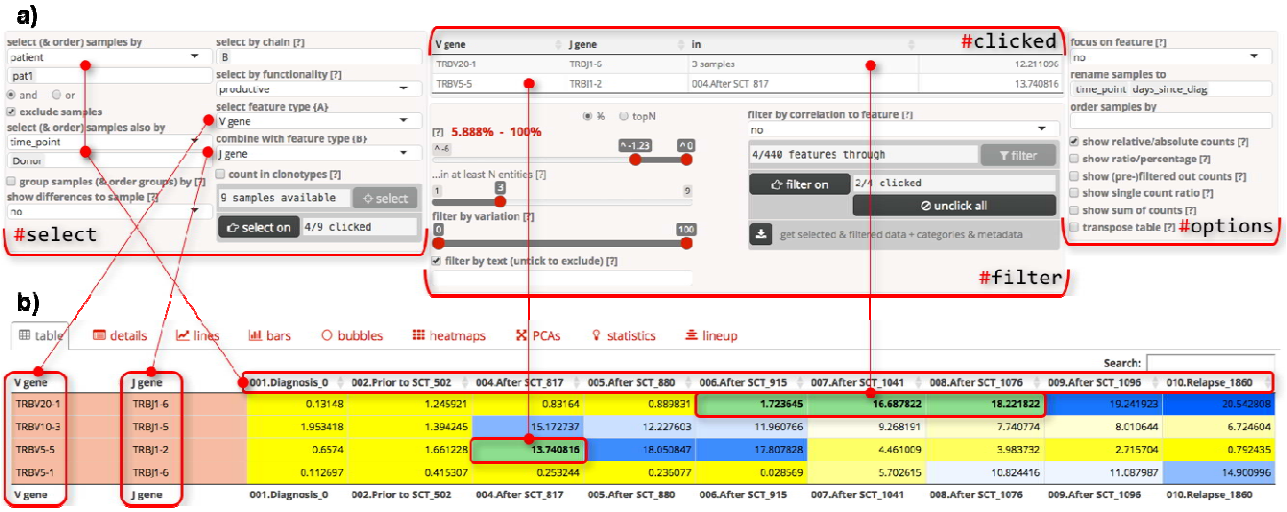

Figure 1: The controls panel (a) with the table view (b). **a)** The controls panel is divided into 3 parts: select, filter, and options. The former two are common for all the visualizations, while options change depending on which visualization view is currently selected. The table ('#clicked') above the filters shows features and samples currently highlighted in the visualizations and it is updated on the fly as the user clicks in the visualizations. **b)** In the table, for each feature in a row, abundance values are shown in columns of samples. Cells with features are colored in groups (in this case by receptor and chain i.e. "TRB"), cells with abundance values are colored in a heatmap-like fashion.

Visualization. Interactive views include tables; line charts, suitable for time-series analyses of clonal kinetics including MRD monitoring; bar charts, popular in clonality testing for lymphoma diagnostics; bubble charts; heatmaps, for sample-sample distance and sample-feature distributions; PCA scatterplots; statistical plots; and multiple sequence alignments. Customizing the visualizations (Figure 1a, #options) includes changing axis properties like values, labels, scales, orientation; and using extra virtual features such as sums of abundances. Interactivity includes zooming, feature highlighting or hiding, and tooltips with detailed information on any data point. Finally, visualizations are interconnected, with features selected in one automatically highlighted in others.

3 Results

Results from the validation and the expert evaluation of ARResT/Interrogate based on actual research data, as well as a running example, are available in the Supplementary Material.

4 Conclusions

We presented ARResT/Interrogate, an interactive data manipulation and visualization application for NGS-based immunoprofiling. It offers a wide variety of options and aims to serve as a user-friendly platform with flexible and powerful analytical capabilities.

Acknowledgements

Computational resources in CEITEC MU were provided by MetaCentrum (LM2010005), and CERIT-SC (CERIT Scientific Cloud, Operational Program Research and Development for Innovations, Reg. no. CZ.1.05/3.2.00/08.0144).

Funding

Authors from CEITEC MU were supported by research grant AZV-MZ-CR 16-34272A-4/2016, project CEITEC 2020 (LQ1601), and ESLHO::EuroClonality; A.K. was additionally supported by project MEYS-NPS I-LO1413.

Conflict of Interest: none declared.

References

Alamyar, E. *et al.* IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome research* 2012;8(1):26.

Benichou, J. *et al.* Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* 2012;135(3):183-191.

Giraud, M. *et al.* Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. *BMC Genomics* 2014;15:409.

Lefranc, M.P. *et al.* IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res* 2015;43(Database issue):D413-422.

Li, S. *et al.* IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat Commun* 2013;4:2333.

Moorhouse, M. *et al.* ImmunoGlobulin galaxy (IGGalaxy) for simple determination and quantitation of immunoglobulin heavy chain rearrangements from NGS. *Bmc Immunology* 2014;15.

Shugay, M. 2015. VDJtools: a framework for post-analysis of repertoire sequencing data. Release 1.0.4