*Bioinformatics*, 31(23), 2015, 3844–3846
doi: 10.1093/bioinformatics/btv456
Advance Access Publication Date: 6 August 2015
Applications Note

OXFORD

Sequence analysis

# ARResT/AssignSubsets: a novel application for robust subclassification of chronic lymphocytic leukemia based on B cell receptor IG stereotypy

**Vojtech Bystry[1,†], Andreas Agathangelidis[2,†], Vasilis Bikos[1,†], Lesley Ann Sutton[3], Panagiotis Baliakas[3], Anastasia Hadzidimitriou[4,3], Kostas Stamatopoulos[4,3] and Nikos Darzentas[1,*], also on behalf of ERIC, the European Research Initiative on CLL**

[1]CEITEC—Central European Institute of Technology, Masaryk University, Brno, Czech Republic, [2]Division of Molecular Oncology and Department of Onco-Hematology, IRCCS San Raffaele Scientific Institute and Università Vita-Salute San Raffaele, Milan, Italy, [3]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden and [4]Institute of Applied Biosciences, Center for Research and Technology Hellas, Thessaloniki, Greece

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first three should be regarded as Joint First Authors.

Associate Editor: John Hancock

Received on April 7, 2015; revised on July 13, 2015; accepted on July 30, 2015

## Abstract

**Motivation:** An ever-increasing body of evidence supports the importance of B cell receptor immunoglobulin (BcR IG) sequence restriction, alias stereotypy, in chronic lymphocytic leukemia (CLL). This phenomenon accounts for ∼30% of studied cases, one in eight of which belong to major subsets, and extends beyond restricted sequence patterns to shared biologic and clinical characteristics and, generally, outcome. Thus, the robust assignment of new cases to major CLL subsets is a critical, and yet unmet, requirement.

**Results:** We introduce a novel application, ARResT/AssignSubsets, which enables the robust assignment of BcR IG sequences from CLL patients to major stereotyped subsets. ARResT/AssignSubsets uniquely combines expert immunogenetic sequence annotation from IMGT/V-QUEST with curation to safeguard quality, statistical modeling of sequence features from more than 7500 CLL patients, and results from multiple perspectives to allow for both objective and subjective assessment. We validated our approach on the learning set, and evaluated its real-world applicability on a new representative dataset comprising 459 sequences from a single institution.

**Availability and implementation:** ARResT/AssignSubsets is freely available on the web at http://bat.infspire.org/arrest/assignsubsets/

**Contact:** nikos.darzentas@gmail.com

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

## 1 Introduction

Research into the immune pathogenesis of lymphomas and leukemias over the last three decades has highlighted the important role of antigen recognition by lymphocyte receptors, by uncovering restrictions in their immunogenetic makeup. Such research in large cohorts of cases of chronic lymphocytic leukemia (CLL), the most common leukemia in an ageing Western population, revealed that subsets of patients with stereotyped, or quasi-identical, B cell receptor immunoglobulin (BcR IG, or antibody) collectively account for ~30% of cases (Agathangelidis *et al.*, 2012; Darzentas *et al.*, 2010), and represent disease sub-entities with shared biologic and clinical profiles, including outcome (Baliakas *et al.*, 2014; Rossi *et al.*, 2009; Stamatopoulos *et al.*, 2007). Intriguingly, in the latest and largest such study of over 7500 patients, 19 major subsets, each with at least 20 cases, accounted for 12% of the cohort and 41% of all stereotyped cases (Agathangelidis *et al.*, 2012). This is remarkable given the negligible chance, in the range of 1:$10^{12}$, of finding two B cell clones with stereotyped BcR IG. Thus, major stereotyped subsets have attracted great interest.

To date, assignment of new cases to these major subsets has been based on the *ad hoc* application of published criteria (Darzentas and Stamatopoulos, 2013), mostly by groups with advanced knowledge in immunogenetics, and thus with limited applicability, and arguably reliability, in routine practice. We therefore introduce a novel application, ARResT/AssignSubsets, which enables the robust assignment of submitted BcR IG sequences from CLL patients to the existing 19 major stereotyped subsets.

## 2 Methods

All sequence datasets used in this work are available through IMGT/CLL-DB [imgt.org/CLLDBInterface/] in accordance with its bylaws.

**Sequence annotation and curation**: Immunogenetic annotations are obtained from IMGT/V-QUEST (Giudicelli *et al.*, 2011), the widely accepted reference for antigen receptor sequence analysis (Lefranc, 2014). These annotations are used to extract sequence features (e.g. the variable heavy complementarity determining region 3, or VH CDR3) for the learning and assignment phases below, but also for the sequences to be validated by ARResT/SeqCure [bat.infspire.org/arrest/seqcure/], developed to systematically report compromising issues based on expert in-house rules applied on the annotations. Such issues include non-ACGT characters or short sequence length, both of which compromise sequence annotation; out-of-frame VH CDR3 with or without stop codons, which implies an unproductive chain; or absence of a gene rearrangement altogether.

At the center of ARResT/AssignSubsets lies a set of rules captured in a probabilistic model, a Bayes classifier implemented in R [r-project.org].

**Learning phase**: Learning of the model to be later used for assignment was based on a positive set of 929 major subset members, and a negative set ('pool' cohort) of 6 667 sequences belonging to minor subsets as well as non-subset, heterogeneous cases (Agathangelidis *et al.*, 2012).

The learning phase evaluates sequence features of these two sets divided into two groups: core and secondary. Core features are based on the latest accepted stringent criteria for subset discovery (Agathangelidis *et al.*, 2012), and if the value of a core feature of a new sequence is not shared among the members of a subset, then the sequence cannot be assigned to that subset. Currently, the three core features are: (i) VH CDR3 length, a critical determinant of the structure of the antigen recognition loop (Barrios *et al.*, 2004); (ii) immunoglobulin heavy variable (IGHV) gene phylogenetic clan, implying meaningful sequence similarity through common ancestry (Kirkham *et al.*, 1992) and (iii) mutational status of the rearrangement, with 'mutated' for <98% nucleotide identity to germline and 'unmutated' otherwise, an important prognostic indicator in patients with CLL (Damle *et al.*, 1999; Hamblin *et al.*, 1999). Secondary features are less strictly controlled for, i.e. unobserved values are accepted (but scored negatively)—these are: (i) rearranged IGHV and immunoglobulin heavy joining genes; and (ii) the VH CDR3 amino acid sequence through amino acid frequencies at any given sequence position.

To keep the model robust and avoid overfitting, the Bayesian network topology is kept as simple as possible, with only obvious dependencies captured: VH CDR3 amino acid frequencies at any position, and relative frequencies of rearranged IGHV genes of the same phylogenetic clan. To avoid creating unrealistically stringent criteria, we include a 'relaxation' coefficient to the probability calculation of the observed frequency of each feature in all available sequences. Finally, to reduce the 'noise' created by relatively less frequent values, we apply the power function on the probability distribution of each feature.

**Assignment phase**: Assignment is based on evaluating submitted sequence features with the learned Bayesian model. Each sequence first acquires an absolute score for each major subset of the positive set and for the negative set or 'pool' cohort. The absolute score is the minus logarithm of the exact probability of assigning to a set, with '−Inf' (minus infinity) meaning that the sequence core features did not match the subset core features. The difference between the absolute score of each major subset and that of the 'pool' cohort is then calculated as a relative score, with positive numbers for submitted sequences closer to subsets than to the 'pool' cohort. Assignment to the best-scoring subset uses a per-subset threshold, based on the range of scores achieved by existing members of that subset. The difference between the relative score and the subset threshold is 'translated' to confidence, ranging from 'borderline' to 'extreme', to further assist the user.

**Output**: Real-time output consists of progress reports, links, information and help, results and tables. These include the detailed ARResT/SeqCure report on the 'health' of the submitted sequences; absolute and relative frequencies of assignment to each of the 19 major CLL subsets; and an assignment report for each submitted sequence, including its 'health', the confidence of the assignment, and, when possible, heat maps of core and secondary features with their significance with respect to the submitted sequence and the best-scoring, but not necessarily assigned, subset. Additional information can be found in the Supplementary data and on the home page of ARResT/AssignSubsets.

## 3 Results

To validate ARResT/AssignSubsets we performed 100 learning and assignment runs, each using randomly selected 80% of the 7 596 sequences of our full cohort for learning, and the remaining 20% as 'new' sequences for assigning. False-positives (i.e. of the 'pool' cohort but assigned to a subset, or falsely assigned) and false-negatives (i.e. of a subset but assigned to the 'pool' cohort, or falsely unassigned) were closely inspected. We confirm high levels of average specificity (99.7%), sensitivity (95.2%) and overall accuracy (99.2%). Favoring specificity is by design, due to the potentially important clinical implications of the assignment results.

To evaluate the applicability and robustness of ARResT/AssignSubsets to a real-world situation, we analyzed 459 new sequences from the University of Athens, Greece (now also in IMGT/CLL-DB), a cohort representative of our larger cohort in terms of IGHV gene repertoire and mutational status. Our results confirmed the published incidence of major subsets (~12%), with 48/459 (10.5%) sequences assigned to 15/19 major subsets. Subsets 1 and 4 acquired almost half of the new assignments (13 and 10 cases, respectively), as expected, while subset 2 was under-represented in this Mediterranean population, confirming geographic biases (Ghia *et al.*, 2005).

## 4 Conclusions

CLL is clinically and biologically heterogeneous, and still incurable, and could benefit from accurate prognostic markers and classifiers, in an effort to implement rationally designed treatment(s). Indeed, studies focusing on major CLL subsets, which, remarkably, describe one in eight CLL patients, have revealed subset-biased biologic and clinical behaviors (Baliakas *et al.*, 2014; Rossi *et al.*, 2009; Stamatopoulos *et al.*, 2007). ARResT/AssignSubsets is an important step towards enabling scientists to tap into this expanding knowledge in a robust and standardized way.

## Acknowledgements

We thank the IgCLL group [igcll.org], the IMGT® [imgt.org], the IMGT/CLL-DB initiative, and Prof. P. Panagiotidis, University of Athens.

## Funding

V.Bystry, V. Bikos and N.D. by CEITEC MU, SuPReMMe, SYLICA, ESHLO::EuroClonality, FP7 NGS-PTL/2012-2015/no.306242, Czech Ministry of Education, Youth and Sports (2013–2015, no.7E13008); A.A. by Associazione Italiana per la Ricerca sul Cancro AIRC (Triennial fellowship 'Guglielmina Lucatello é Gino Mazzega'); A.H. and K.S. by ENosAI (code 09SYN-13-880, co-funded by the EU and the Hellenic GSRT); P.B. by The Swedish Cancer Society, the Swedish Research Council, Uppsala University Hospital, and Lion's Cancer Research Foundation, Uppsala.

*Conflict of Interest*: none declared.

## References

Agathangelidis,A. *et al.* (2012) Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. *Blood*, **119**, 4467–4475.

Baliakas,P. *et al.* (2014) Clinical effect of stereotyped B-cell receptor immunoglobulins in chronic lymphocytic leukaemia: a retrospective multicentre study. *Lancet Haematol.*, **1**, e74–e84.

Barrios,Y. *et al.* (2004) Length of the antibody heavy chain complementarity determining region 3 as a specificity-determining factor. *J. Mol. Recogn.*, **17**, 332–338.

Damle,R.N. *et al.* (1999) Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood*, **94**, 1840–1847.

Darzentas,N. and Stamatopoulos,K. (2013) Stereotyped B cell receptors in B cell leukemias and lymphomas. *Methods Mol. Biol.*, **971**, 135–148.

Darzentas,N. *et al.* (2010) A different ontogenesis for chronic lymphocytic leukemia cases carrying stereotyped antigen receptors: molecular and computational evidence. *Leukemia*, **24**, 125–132.

Ghia,P. *et al.* (2005) Geographic patterns and pathogenetic implications of IGHV gene usage in chronic lymphocytic leukemia: the lesson of the IGHV3-21 gene. *Blood*, **105**, 1678–1685.

Giudicelli,V. *et al.* (2011) IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harbor Protoc.*, **2011**, 695–715.

Hamblin,T.J. *et al.* (1999) Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood*, **94**, 1848–1854.

Kirkham,P.M. *et al.* (1992) Immunoglobulin VH clan and family identity predicts variable domain structure and may influence antigen binding. *EMBO J.*, **11**, 603–609.

Lefranc,M.P. (2014) Immunoglobulin and T cell receptor genes: IMGT((R)) and the birth and rise of immunoinformatics. *Front. Immunol.*, **5**, 22.

Rossi,D. *et al.* (2009) Stereotyped B-cell receptor is an independent risk factor of chronic lymphocytic leukemia transformation to Richter syndrome. *Clin. Cancer Res.*, **15**, 4415–4422.

Stamatopoulos,K. *et al.* (2007) Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: pathogenetic implications and clinical correlations. *Blood*, **109**, 259–270.