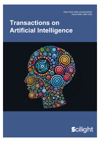
Scilight

Transactions on Artificial Intelligence
https://www.sciltp.com/journals/tai



Supplementary Materials

T-Cell Receptor Repertoire in Autoimmune Diseases and Their Machine Learning-Based Prediction Analysis

Tongfei Shen, Miaozhe Huo and Shuaicheng Li *

Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, China
* Correspondence: shuaicli@cityu.edu.hk

1. Methods

1.1. Data Sources

Publicly available TCR β repertoire sequencing datasets were utilized for this exploratory analysis. Twenty samples were randomly selected from each disease group: multiple sclerosis (MS), rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), type 1 diabetes (T1D), and healthy controls (CK). The datasets were drawn from previously published studies [44,74,91–93]. In particular, the study by Martinez et al. [93] provided the healthy control cohort, while disease-specific repertoires were obtained from the other cited works. Exact accession identifiers and sample-level metadata were not uniformly available across all datasets; therefore, this analysis was restricted to the subset of samples explicitly reported in the source publications.

1.2. Sample Selection

Within each dataset, twenty samples were randomly selected to ensure balanced representation across groups. No additional inclusion or exclusion criteria beyond those defined in the original publications were applied. Sample types (e.g., peripheral blood, cerebrospinal fluid, or tissue) and sequencing platforms (e.g., Illumina MiSeq/HiSeq) were consistent with the specifications of the respective source studies [44,74,91–93]. The datasets are available in the immuneACCESS database under the following accession IDs: mitchell-2022-jci, mustjoki-2017-natcomms, gold-2019-cr, and martinez-2025-s. Given that experimental details varied across studies, these factors were treated as equivalent for exploratory purposes in the current analysis.

1.3. Preprocessing and Annotation

Raw repertoire data were processed according to the pipelines described in the original publications. VDJ gene segment assignment was performed using the methods reported in each study (e.g., MiXCR, IgBlast, or equivalent annotation tools). Filtering thresholds, duplicate collapsing strategies, and preprocessing steps were not modified beyond those already applied in the source datasets. As such, the present work represents a harmonized re-analysis of published repertoires rather than a de novo pipeline.

1.4. Normalization

To account for differences in sequencing depth across datasets, repertoires were downsampled to a fixed number of clonotypes per sample prior to comparative analysis. This normalization strategy ensured that diversity metrics were not confounded by unequal read depths.

1.5. Feature Extraction

The following repertoire features were extracted:

- V and J gene usage frequencies, computed at the clonotype level and averaged across samples within each group.
- Shannon diversity indices of V and J gene usage, calculated per sample and then summarized by group means.
- CDR3 amino acid length distributions, including mean length and variance per group.
- Shannon diversity indices of CDR3 length distributions, computed per sample and averaged across groups.
- Proportions of conserved G/C nucleotide insertions, derived from the annotated junctional sequences.

Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1.6. Statistical Analysis

Comparisons between autoimmune disease groups and healthy controls were performed using two-tailed Student's *t*-tests. Reported p-values reflect uncorrected significance testing; no multiple-comparison adjustments were applied given the exploratory nature of the study. All statistical analyses were conducted at the per-sample level, with group-level summaries reported as mean ± standard deviation.

1.7. Limitations

This analysis is preliminary and subject to several limitations: heterogeneous sample types and sequencing platforms across datasets, lack of uniform accession identifiers, and reliance on preprocessing pipelines defined by the original studies. Consequently, the results should be interpreted as exploratory trends rather than definitive conclusions.

2. Supplementary Results for a Preliminary Small-Scale Case Study on Feature Extraction Methods for Alteration

CDR3 amino acid length distributions were broadly comparable across cohorts; autoimmune disease groups exhibited slightly longer mean CDR3 lengths than controls, with MS showing the largest mean (14.703 ± 0.237) versus controls (14.595 ± 0.182). These differences did not reach conventional statistical thresholds (all $p > 0.05$; t-test), and variability measures were similar across groups (mean SD $\approx$ 1.85–1.88), suggesting limited influence of autoimmune disease status on overall CDR3 length architecture (Figure S1A, main text Figure 4D).

Examination of conserved G/C insertion proportions suggested a modest reduction in MS relative to controls, with RA, SLE, and T1D displaying slight increases; however, none of these differences reached conventional statistical thresholds. The term "conserved" here refers to the relative stability of G-C base pairing, as G-C pairs form three hydrogen bonds compared with two in A-T pairs. This enhanced stability is often observed in functionally important regions, such as highly conserved motifs or regulatory sequences. In the present dataset, G/C insertion patterns did not show robust alterations across cohorts and will require larger studies for confirmation (Figure S1B).

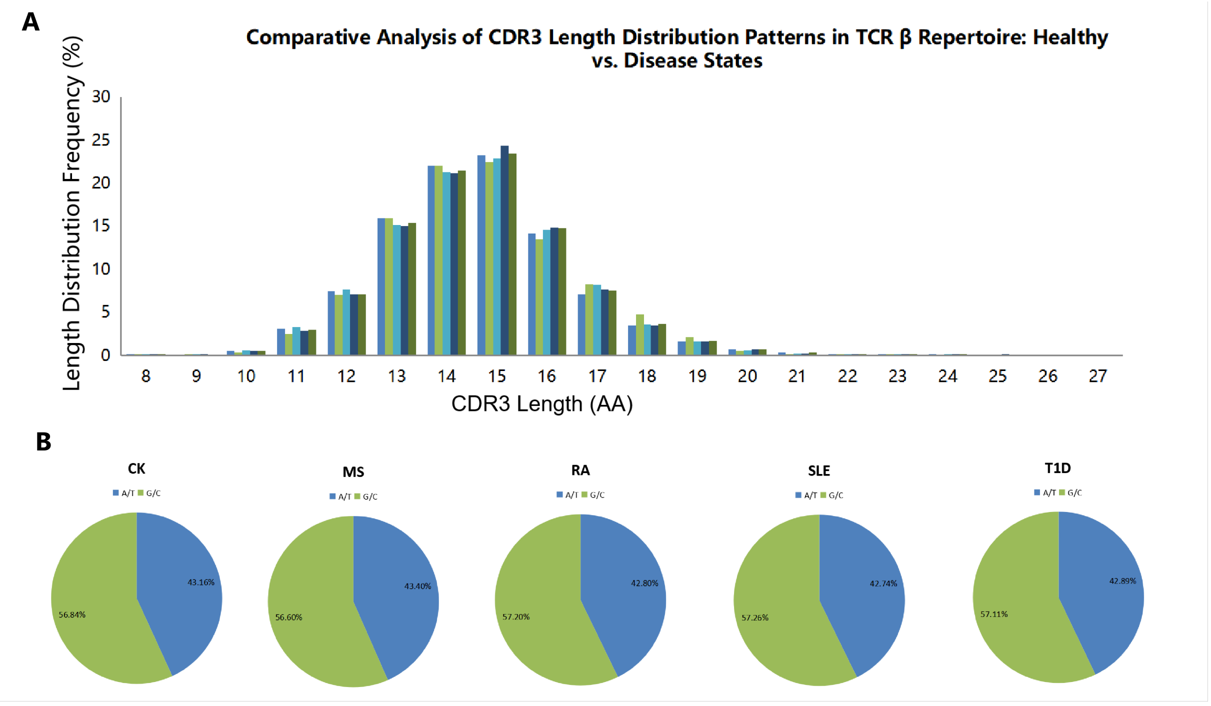


Figure S1. Preliminary small-scale case study of TCR β repertoire features across autoimmune disease groups and healthy control groups. (**A**) CDR3 amino acid length distribution patterns exhibit generally similar profiles across groups. (**B**) Proportions of conserved G/C nucleotide insertions show no substantial differences between autoimmune diseases and controls.