http://dx.doi.org/10.7124/bc.000AFC

A comprehensive benchmarking study of adaptive immune receptor repertoire sequencing data aligners

F. Iordachi[1], A. Adamyan[2], S. Bostan[3], U. Krektun[4], L. Calujac[5], C. Dumitrescu[5]

[1] Stanford University's OHS
415, Broadway, Redwood City, CA 94063, USA
[2] Zaven and Sonia Akian College of Science and Engineering, American University of Armenia
40, Marshal Baghramyan Ave., Yerevan, Armenia, 0019
[3] New School International School of Georgia
35, Tskneti Hwy., Tbilisi, Georgia, 0162
[4] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
37, Beresteiskyi Ave., Kyiv, Ukraine, 03056
[5] Stefan cel Mare University of Suceava
13, University Str., Suceava, Romania, 720229
iordachifelicia19@gmail.com

Aim. The reliability of Adaptive Immune Receptor Repertoire sequencing (AIRR-Seq) data analysis hinges on the performance of alignment tools, but due to frequent inconsistency in output formats and user experience, an objective comparison of AIRR data aligners is still missing. Thus, here we aimed to benchmark the most widely used AIRR-Seq data analysis tools — HighV-Quest, IgBlast, and MiXCR — and evaluate their alignment accuracy [1, 2]. **Methods.** We simulated T-cell receptor alpha chain (TRA) sequences, from which we constructed TRA clonotypes and reads, which served as an input dataset for all tools we ran. The dataset mimicked the diversity of the AIRR. **Results.** Our analysis demonstrated a notable difference in the alignment accuracy of the tools. Based on V-region overlap, IgBlast achieved an accuracy of 67–74%, HighV-Quest ranged from 59–66%, and MiXCR reached up to 9%, indicating that IgBlast and HighV-Quest excel in V-region alignment, while MiXCR approaches homogenous alignment across all regions, sacrificing higher specificity for V-region alignment. We also observed that MiXCR experienced fewer tool failures — instances where the tool couldn't process reads — (from 0.5% to 0.1%) and achieved higher overall alignment success rates — the percentage of reads correctly aligned — (from 2% to 14.5%) as we increased the read lengths in our simulated dataset from 50 bp to 100 bp, demonstrating that longer reads provide more sequence information and improve alignment accuracy. Additionally, we calculated the percentage of mapped reads out of the expected reads to be mapped for each clonotype reconstructed by MiXCR to test the null hypothesis that variation in V-region overlap within a read is not correlated with read mapping failure. We examined the distribution of V and J region overlap in reads that mapped to the two clonotypes with the highest percentage of reads mapped (80% and 75%, respectively) and two clonotypes with the lowest percentage of reads mapped (0%). Notably, we found no statistical significance ($p > 0.05$) in the differences of distribution of V-region genes among the reads that mapped to these four clonotypes, thereby supporting our null hypothesis. **Conclusions.** Our study shows that based on the V-region overlap call, IgBlast achieves the best performance, with HighV-Quest and MiXCR following respectively. We also demonstrate that the distribution of V-region genes in reads has no statistically significant effect on the mapping outcome of each read.

Keywords: Adaptive immune receptor repertoire, T cells, Junctional Diversity.

REFERENCES

1. *Dmitriy A et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nature methods.* 2015; **12**(5): 380–1.
2. *Ye J et al.* IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* 2013; **41**(Web Server issue):W34–W40.