Comparison of sequence- and structure-based antibody clustering approaches on simulated repertoire sequencing data

Katharina Waury[1,2] , Stefan Lelieveld[3], Sanne Abeln[1,2*], Henk-Jan van den Ham[3*]

[1]Department of Computer Science, Vrije Universiteit Amsterdam, 1081HV Amsterdam, The Netherlands; [2]AI Technology For Life, Department of Information and Computing Science, and Department of Biology, Utrecht University, 3584CS Utrecht, The Netherlands; [3]ENPICOM B.V., 5211 DA 's-Hertogenbosch, The Netherlands; *s.abeln@uu.nl, h.vandenham@enpicom.com

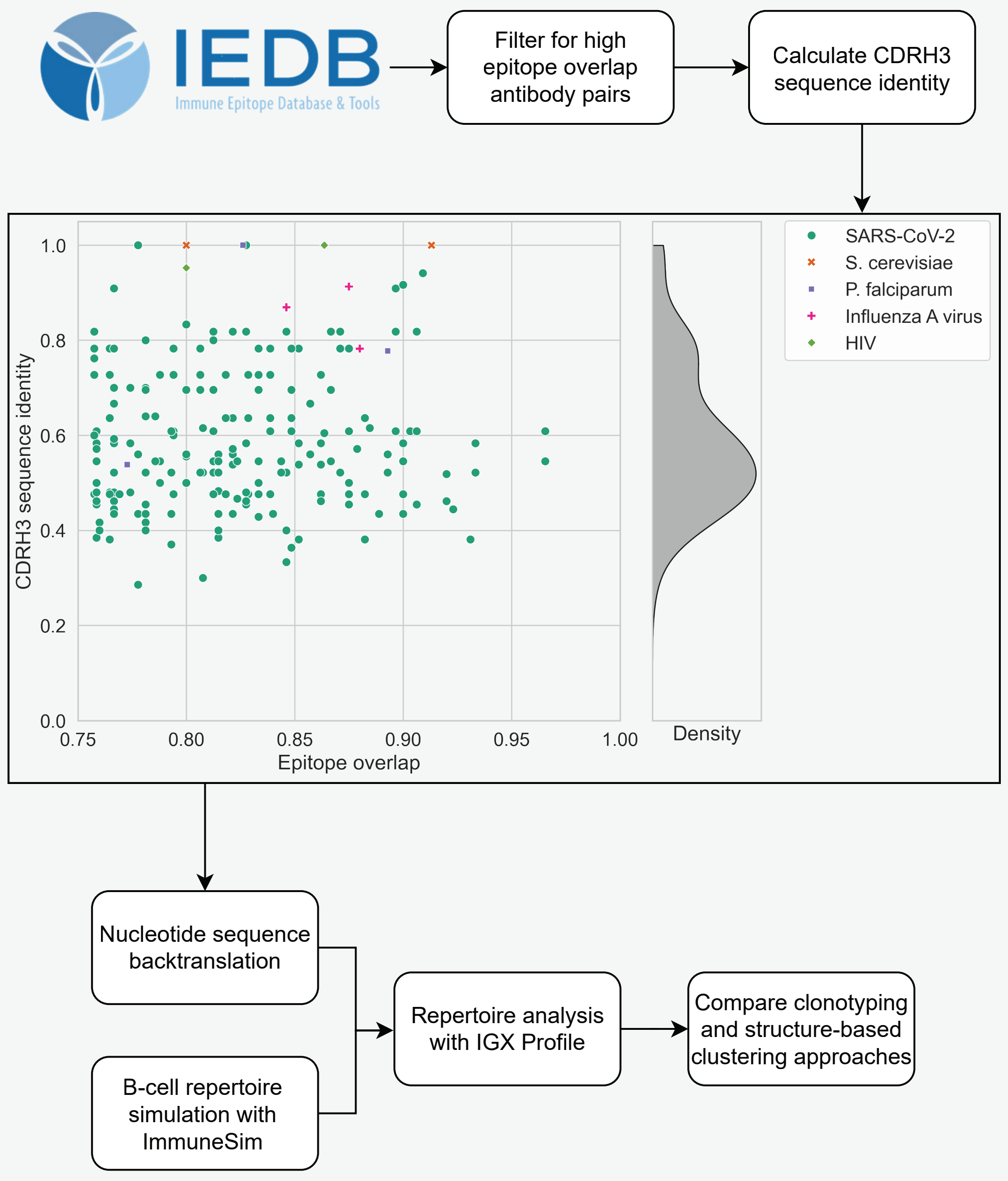
Background

Understanding adaptive immune system responses is crucial for developing vaccines and therapies. High-throughput characterization of antibodies is now an integral part of immune profiling and drug discovery. [1]

One of the main challenges is to identify functionally related antibodies as these enable the formulation of more suitable antibodies for therapies, or aid in characterising immune responses in different individuals. By clustering similar antibody clones, functionally related antibodies can be identified.

Conventional antibody clustering approaches rely on sequence-based information, particularly CDRH3 sequence identity and V/J gene usage, to group antibodies. However, it is known that sequences of different clonal origin may lead to antibodies with similar binding properties, because they are similar in structure but not sequence. [2] Recent advances have made structure-based clustering methods feasible on a high-throughput repertoire scale. However, so far, performance of these methods has only been evaluated on single-antigen sets of antibodies. Here, we benchmark sequence- and structure-based clustering methods and highlight strengths and weaknesses, and suggest avenues for future development. Two structure-based clustering methods, SPACE2 [3] and SAAB+ [4], are evaluated against IGX-cluster, a conventional sequence-based clustering method.

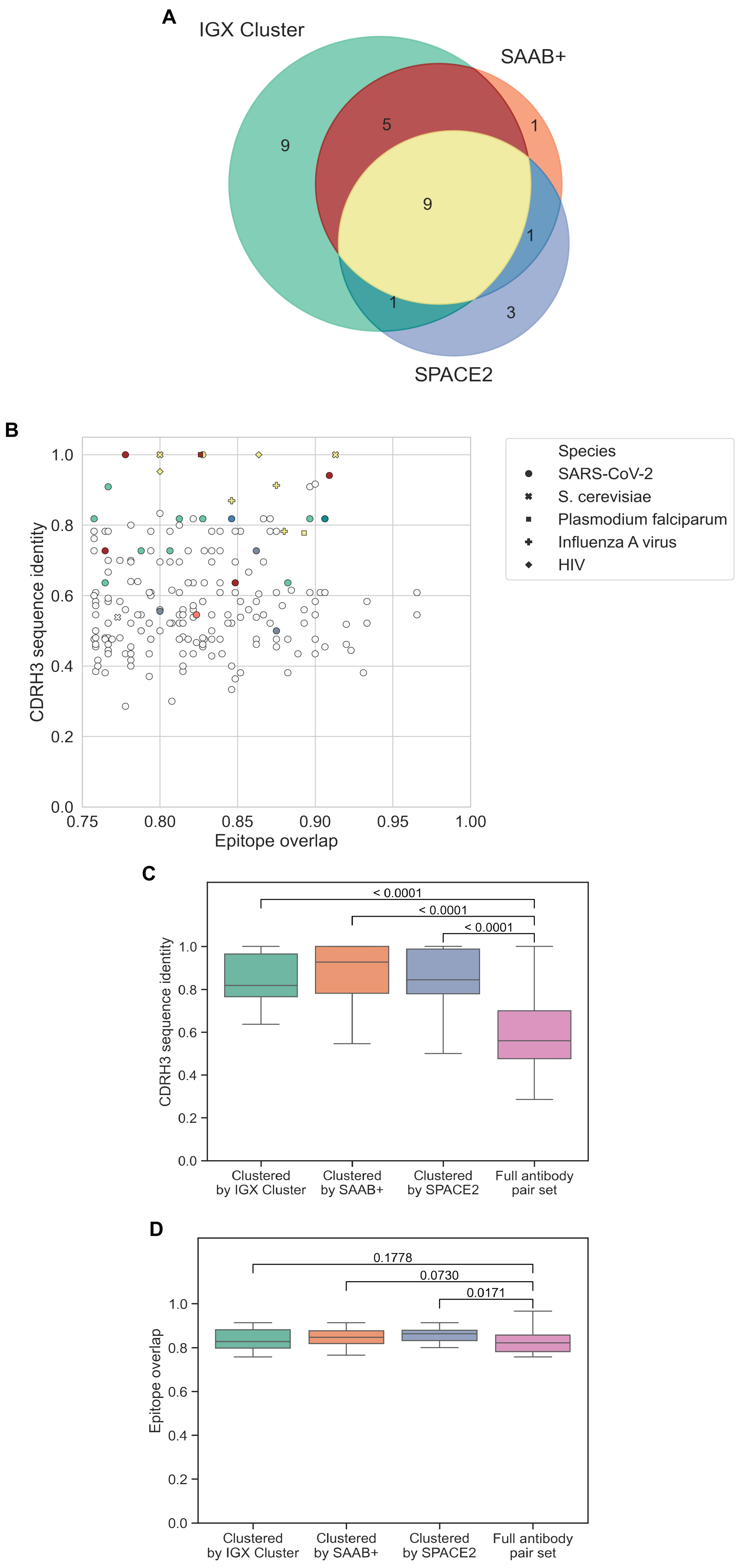
Data & Analysis


IEDB
Immune Epitope Database & Tools
Filter for high epitope overlap antibody pairs
Calculate CDRH3 sequence identity
CDRH3 sequence identity
1.0
0.8
0.6
0.4
0.2
0.0
0.75
0.80
0.85
0.90
0.95
1.00
Epitope overlap
Density
SARS-CoV-2
S. cerevisiae
P. falciparum
Influenza A virus
HIV
Nucleotide sequence backtranslation
B-cell repertoire simulation with ImmuneSim
Repertoire analysis with IGX Profile
Compare clonotyping and structure-based clustering approaches

We curated a dataset of well-annotated pairs of antibodies that show high overlap in epitope residues and thus bind the same region within their respective antigen. This set of antibodies was introduced into a simulated synthetic repertoire dataset. After clonotype annotation with IGX-Profile [5], the performance of clustering approaches is evaluated on this generated repertoire dataset.

Sequence-based clustering is limited but highly accurate

Performance comparison of different clustering strategies. The three clustering approaches, namely sequence-based clustering, SAAB+ and SPACE2, were applied to the repertoire. Their performance on the annotated set of 213 functionally similar antibody pairs was evaluated.



A: Euler diagram showing the overlap of correctly clustered antibody pairs between methods.

B: A scatter plot shows the CDRH3 sequence identity and epitope overlap of each antibody pair. The majority of antibody pairs have not been clustered together by any methods (gray, 184 antibody pairs).

C: All three clustering strategies group antibody pairs with a significantly higher CDRH3 sequence identity compared to the full antibody pair set. CDRH3 sequence similarity within groups is similar between all methods.

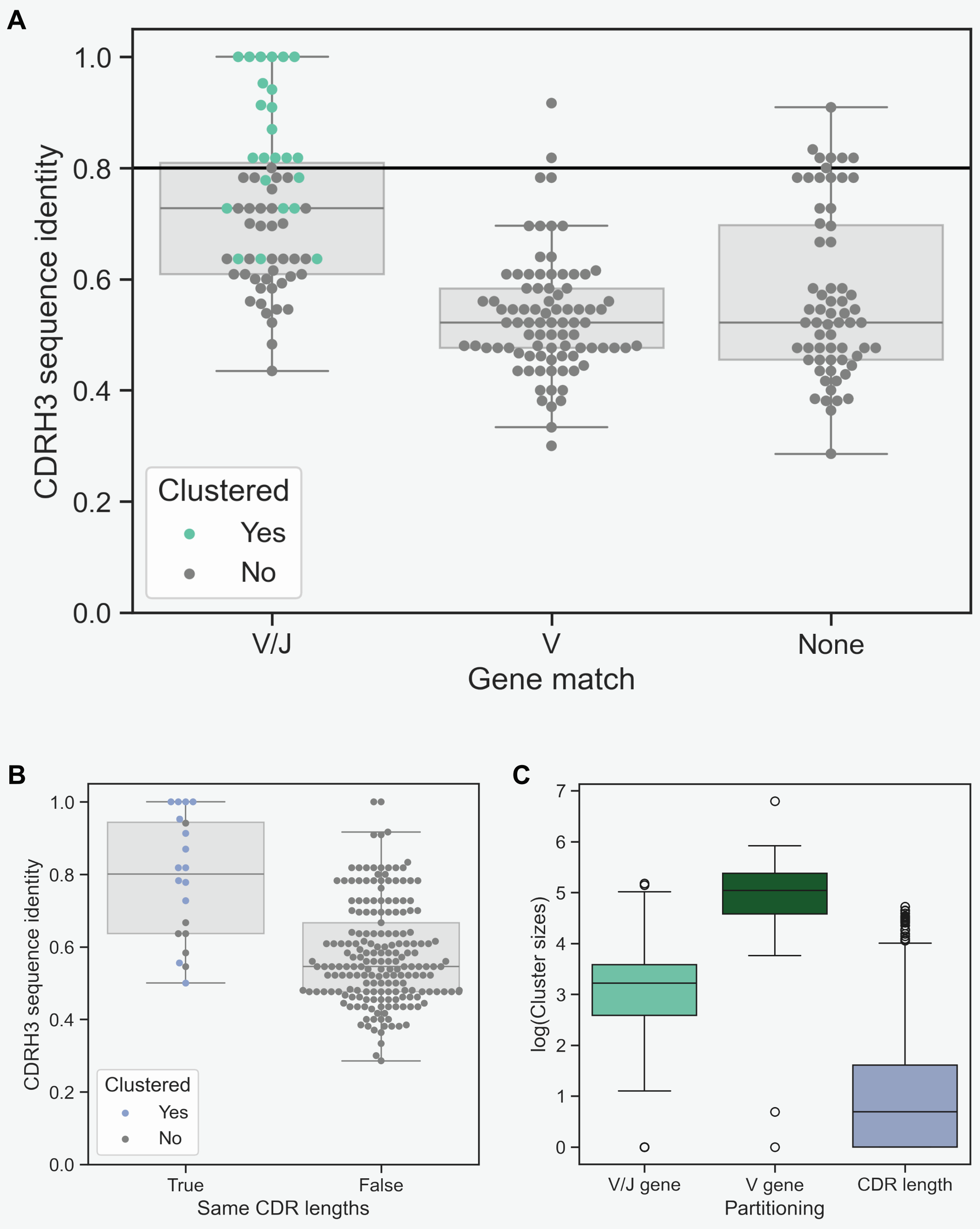
D: The epitope overlap of clustered antibody pairs is similar between the methods, albeit SPACE2 identified antibody pairs with a slightly higher epitope overlap compared to the other methods.

Conclusions

- ✓ All methods capture antibodies with distinct sequences that bind the same epitope; all methods have a relatively high specificity.
- ✓ Most antibodies that bind the same epitope are missed by all methods; all methods have low sensitivity.
- ✓ A priori stratification of datasets by V-gene or CDR length leads to fewer pairs being identified. Improvement in current performance is is difficult if this point is not addressed.

Stratification limits clustering

Both sequence-based clustering and SPACE2 partition the antibodies based on CDRH3 sequence identity or CDR RMSD, respectively. This a priori division of repertoire repertoire data into separate sections is a major impediment for improvement as it segregates similar antibodies into separate groups before clustering.



A: Clonotyping partitions clones based on matching V and J genes. Antibodies with identical V and J gene usage have a higher sequence identity than antibodies with identical gene usage in only the V or none of the genes. Partitioning based on only the V gene can improve the coverage of clustering slightly, but is limited by the low sequence identity between these antibodies. Colored dots indicate the correctly clustered antibody pairs for each method.

B: SPACE2 partitions the repertoire data based on same length in all six CDR regions. The majority of antibody pairs (193, 90.61%) do not meet this requirement. Of the antibody pairs with same CDR lengths, 70% (14 out of 20) are correctly grouped together.

C: The natural logarithm of cluster sizes across the full repertoire dataset indicates how stringent different partitioning strategies are. The criterion of same CDR region lengths is the most stringent, while requiring solely the same V gene is the least stringent and leads to the largest cluster sizes.

References

1. Arnaout RA, Prak ETL, Schwab N, Rubelt F. The Future of Blood Testing Is the Immunome. Frontiers in Immunology. 2021;12. doi:10.3389/fimmu.2021.626793.

2. Raybould MIJ, Rees AR, Deane CM. Current strategies for detecting functional convergence across B-cell receptor repertoires. mAbs. 2021;13(1). doi:10.1080/19420862.2021.1996732.

3. Spoendlin FC, Abanades B, Raybould MIJ, Wong WK, Georges G, Deane CM. Improved computational epitope profiling using structural models identifies a broader diversity of antibodies that bind to the same epitope. Frontiers in Molecular Biosciences. 2023;10. doi:10.3389/fmolb.2023.1237621.

4. Kovaltsuk A, Raybould MIJ, Wong WK, Marks C, Kelm S, Snowden J, et al. Structural diversity of B-cell receptor repertoires along the B-cell differentiation axis in humans and mice. PLOS Computational Biology. 2020;16(2):e1007636. doi:10.1371/journal.pcbi.1007636.

5. ENPICOM. IGX Platform: Unlock the full potential of your repertoire data; 2024. Available from: https://enpicom.com/igx-platform/.