Published online 19 May 2023 *Nucleic Acids Research, 2023, Vol. 51, Web Server issue* ***W17–W24***
https://doi.org/10.1093/nar/gkad400

Abalign: a comprehensive multiple sequence alignment platform for B-cell receptor immune repertoires

Fanjie Zong[1,2,†], Chenyu Long[1,2,†], Wanxin Hu[1,2,†], Shuang Chen[3], Wentao Dai[4], Zhi-Xiong Xiao[1] and Yang Cao[1,2,*]

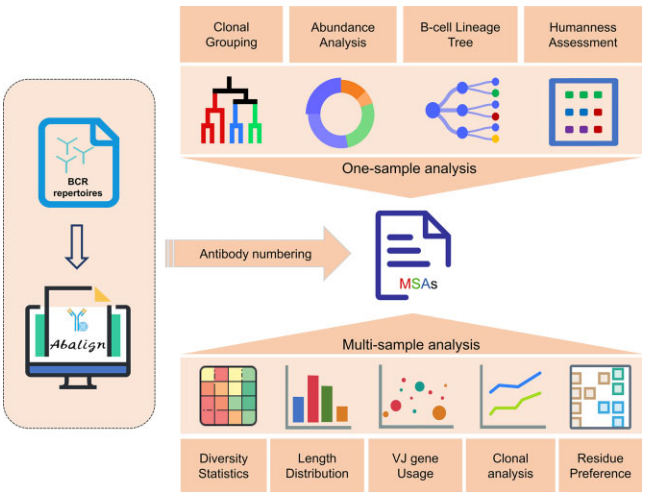
[1]Center of Growth, Metabolism and Aging, Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, China, [2]Animal Disease Prevention and Food Safety Key Laboratory of Sichuan Province, Microbiology and Metabolic Engineering Key Laboratory of Sichuan Province, Chengdu, China, [3]Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, China and [4]NHC Key Laboratory of Reproduction Regulation & Shanghai-MOST Key Laboratory of Health and Disease Genomics, Shanghai Institute for Biomedical and Pharmaceutical Technologies, Shanghai, China

Received February 26, 2023; Revised April 23, 2023; Editorial Decision May 01, 2023; Accepted May 08, 2023

ABSTRACT

The utilization of high-throughput sequencing (HTS) for B-cell receptor (BCR) immune repertoire analysis has become widespread in the fields of adaptive immunity and antibody drug development. However, the sheer volume of sequences generated by these experiments presents a challenge in data processing. Specifically, multiple sequence alignment (MSA), a critical aspect of BCR analysis, remains inadequate for handling massive BCR sequencing data and lacks the ability to provide immunoglobulin-specific information. To address this gap, we introduce Abalign, a standalone program specifically designed for ultra-fast MSA of BCR/antibody sequences. Benchmark tests demonstrate that Abalign achieves comparable or even better accuracy than state-of-the-art MSA tools, and shows remarkable advantages in terms of speed and memory consumption, reducing the time required for high-throughput analysis from weeks to hours. In addition to its alignment capabilities, Abalign offers a broad range of BCR analysis features, including extracting BCRs, constructing lineage trees, assigning VJ genes, analyzing clonotypes, profiling mutations, and comparing BCR immune repertoires. With its user-friendly graphic interface, Abalign can be easily run on personal computers instead of computing clusters. Overall, Abalign is an easy-to-use and effective tool that enables researchers to analyze massive BCR/antibody sequences, leading to new discoveries in the field of immunoinformatics. The software is freely available at http://cao.labshare.cn/abalign/.

GRAPHICAL ABSTRACT



INTRODUCTION

B cells are essential components of the immune system, playing a crucial role in protecting the body against various pathogens. These cells are able to recognize antigens through their B-cell receptors (BCRs), and through somatic hypermutation (SHM) (1,2) and class switch recombination (3), they are able to evolve their BCRs and produce antibodies with increased affinity to the antigens (4). This has led to

*To whom correspondence should be addressed. Tel: +86 02885418843; Email: cao@scu.edu.cn
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

the widespread adoption of BCR sequencing in the study of infectious diseases, allergies, vaccination, autoimmune disorders, tumor immunity, and microbiota(5–9). Sequencing BCR involves the sequencing of the entire BCR gene, or a portion of it, from a large number of individual B cells. This generates a comprehensive dataset of BCR sequences, known as the 'BCR repertoire', demonstrating the picture of BCR landscape. The analysis of the BCR repertoires enables us to understand the diversity and evolution of BCRs over time, and provides insights into the specificity of the B cell population. Furthermore, researchers can identify antigen-specific monoclonal antibodies (mAbs) that can be used in the development of therapeutic antibodies for the treatment of various diseases (10).

BCR repertoire analysis generally involves preprocessing of the data, followed by analysis to infer the B cell population structure and to quantify detailed features (11). Pre-processing assembles high-throughput sequencing (HTS) data into error-corrected BCR sequences, which are then aligned to identify the V(D)J germline genes. It has been addressed by many HTS tools and specialized methods such as TRUST4 (12), CATT (13), MiXCR (14), IgReC (15), pRESTO (16) and IMSEQ (17). The inference of B cell population structure infers the dynamic adaptive B cell response of the BCR repertoire upon clonal expansion and affinity maturation (18,19). The quantitative features characterize complementarity-determining regions, clonal groups, B-cell lineage trees, SHMs, BCR diversity and etc. Numerous BCR tools have been developed independently such as IgBLAST (20), IMGT/HighV-QUEST (21) for V(D)J genes annotation, AbNum (22), ANARCI (23) and AbRSA (24) for antibody numbering, VDJtools (25), IMonitor (26), BRepertoire (27), ImmunediveRsity (28), Vidjil (29), partis (30), IGoR (31), IgAT (32), IMEX (33) for BCR diversity, hypermutation probability and clone tracking. Despite these methods, a crucial method, multiple sequence alignment (MSA), remains inadequate for handling the large number of BCR sequencing data and lacks the ability to provide immunoglobulin-specific information. And it further limits the application of grouping clones, constructing lineage trees, analyzing SHMs, antibody engineering, structure modeling and more.

MSA is the alignment of multiple sequences to illustrate their functional, structural, or evolutionary relationships. It is one of the most widely used modeling methods in biology today (34). However, generating accurate MSA can be challenging, as it is known to be an NP-complete problem. As a result, most MSA methods rely on heuristics that use an estimated guide tree to perform progressive alignment (35). Even though the state-of-the-art large-scale MSA methods such as Clustal Omega (ClustalO) (36), MAFFT (37), MUSCLE(38), PASTA (39), SATé-II (40), are ineffective in handling massive BCR sequencing data. Processing a typical 1GB BCR sequencing data set can be a time-consuming task, which often needs several weeks as well as huge memory demands using computing clusters. What is more, the high diversity of BCRs, generated by the processes of V(D)J gene segments recombination and SHMs, results in an over-insertion problem (41), which has become a bottleneck in the analysis of BCR data. To address the issue, it would be beneficial for MSA to introduce the human knowledge about BCR sequence patterns. Unfortunately, to the best of our knowledge, there are no such MSA methods currently available.

BCR/antibody sequences exhibit region-specific features, with complementarity-determining regions (CDRs) displaying a high degree of diversity and framework regions (FRs) being more conserved. Certain residue sites tend to remain invariant, while others are prone to insertion and deletion. These features can be summarized by antibody numbering, which is an immunoinformatic technique that enables standardization of the residue indexes of an antibody variable domain. We previously developed AbRSA (24), a highly robust method for antibody numbering. Building on AbRSA, we have now proposed Abalign, a new MSA method specifically designed for BCRs. Unlike existing MSA tools that rely on guide trees, Abalign performs the alignment using well-characterized antibody numbering information. Independent benchmark tests show that Abalign outperforms existing MSA tools in terms of speed, alignment quality, and memory consumption, reducing the time required for high-throughput analysis from weeks to hours. Abalign has been implemented in user-friendly graphic interfaces and integrated with a range of approaches for BCR repertoire analysis based on MSA, including constructing B-cell lineage trees, assigning VJ genes, analyzing clonotypes, humanizing antibody, and comparing BCR immune repertoires. Details on these features will be provided in the following sections.

MATERIALS AND METHODS

MSA algorithm

Abalign performs the alignment guided by well-characterized antibody numbering schemes such as IMGT (42), Kabat (43), Chothia (44) and Martin (22). This process is divided into two steps (Supplementary Figure S1). Firstly, each of the query sequences is aligned to the pre-designed consensus sequences of the antibody as described in AbRSA. This alignment marks the residues of the query sequences with standardized index numbers, which are then stored in a matrix. Each row of the matrix stores a sequence, while each column contains roughly aligned residues. Secondly, the matrix is expanded, since one numbering position may contain multiple residues according to the scheme. To extend these positions, gaps must be inserted into the positions which have fewer residues in the same column. Abalign takes into account the different levels of conservativeness in FRs and CDRs and employs different strategies accordingly. In FRs, Abalign searches for the maximum number of residues at each position in a column and inserts the gaps at position $\lfloor \frac{L_{i,j}}{2} \rfloor + 1$ in row i and column j, where $L_{i,j}$ is the number of residues in (i, j) position. The number of gaps to be inserted at this position is given by:

$$\mathrm{num}_{i,j} = \max\left(\mathrm{L}_j\right) - L_{i,j} \quad (1)$$

In CDRs, Abalign inserts gaps at pre-annotated positions (Supplementary Table S1), rather than inserting at any positions which contain multiple residues in FRs. Abalign counts the CDRs length $CL_{i,k}$ ($k = 1, 2, 3$ for CDR1/2/3

respectively) for all sequences, and records the largest length mL_k. The positions where gaps are inserted at the pre-annotated insertion positions defined by the antibody numbering schemes. The number of gaps to be inserted is given by:

$$\mathrm{num}_{i,k} = \mathrm{mL}_k - \mathrm{CL}_{i,k} \quad (2)$$

It is worth noting that the results of the MSA may exhibit minor discrepancies depending on the selected numbering scheme, owing to differences in the definition of CDRs and insertion positions. In the following work, we have used the Chothia scheme as the default.

Germline genes

The germline gene is identified by comparing a query sequence to the germline sequence set taken from IMGT (45) database following the method of IgBLAST (20). To accelerate this process, Abalign performed antibody numbering for the germline sequences in advance. Therefore, after MSA, the query sequences are mapped to these germline sequences automatically.

MSA benchmark sets

Structure-based alignments have been widely used as the reference for MSAs (46–48). To evaluate our method, we constructed a structure-based benchmark set containing 1800 non-redundant antibody crystal structures selected from the antibody structural database SAbDab (49) using the criteria of resolution <3.0 Å. Those structures were split into heavy and light chains and divided into 36 groups (18 groups each for heavy and light chains) randomly, each of which contains 100 structures. For each group, the chains followed by structural alignment using a rapid structure alignment tool Caretta (50) to generate the structure-based reference MSAs (Supplementary Table S2). In some cases that Caretta cannot correctly align the structures, we employed TM-align (51) to correct these alignments.

To evaluate the high-throughput performance of Abalign, we constructed another benchmark data set using the BCR sequencing data from the immune repertoire database OAS (52). We obtained BCR immune repertoire sequencing data from SARS-CoV-2 patients (Supplementary Table S3) to measure the running speed and memory consumption. We randomly selected 1000(1K), 2000(2K), 4000(4K), 8000(8K), 16 000(16K), 32 000(32K), 64 000(64K), 128 000(128K), 256 000(256K), and 512 000(512K) sequences to construct 10 subsets. Additionally, we utilized the STRIKE(53) score to assess the quality of MSAs with a huge number of BCR sequences, which requires additional antibody structures as templates. 100 antibodies were selected randomly from the structure-based benchmark sets (Supplementary Table S2) for repeating STRIKE tests 100 times.

MSA assessment criteria

SP (Sum of Pairs) and TC (Total Column) scores have been widely used in assessing MSA results (54–56). SP score measures the fraction of aligned residue pairs in the reference alignment that agree to the tested alignment. TC score measures the fractions of columns that are totally aligned.

STRIKE (53) score is also employed to assess MSA results. Unlike SP and TC scores which need reference alignment and annotation information, STRIKE score only requires one protein structure for a whole MSA. Hence, STRIKE score can be applied to assess MSA with a huge number of BCR sequences.

Software implementation

Abalign is a versatile visualization software that can be used across multiple platforms. The program utilizes standard C++ to implement MSA and germline gene identification, which allows for efficient parallel computing. Additionally, Abalign's user-friendly GUI and other advanced features are implemented with Python 3.8. The plotting is done by using matplotlib and seaborn. The B-cell lineage tree is constructed with FastTree2 (57) and ETE3 (58). Other modules for plotting include WebLogo (59), seqLogo (60) and UpSet (61). To ensure software reliability, Abalign has been tested on Linux (Ubuntu versions 18.04, 20.04, 22.04), Windows (versions 7, 10, 11), and macOS (versions 10.14, 10.15, 11.6, 12.4).

RESULTS

MSA benchmarking

Most MSA benchmarking methods use 3D-structure-based reference MSAs as the gold standard and evaluate the differences between the tested and referenced MSAs using statistical scores such as SP and TC (54–56). The SP score measures the fraction of aligned residue pairs in the alignment that agree to the tested alignment, while the TC score counts the fraction of columns that are exactly aligned. To evaluate Abalign's performance, we constructed 3D-structure-based reference MSAs using 1800 antibody structures from the SAbDab database (Materials and Methods). We also compared Abalign with three state-of-the-art MSA tools (MUSCLE, Clustal Omega (ClustalO), and MAFFT). Our results demonstrate that Abalign achieves the SP (Figure 1A) and TC (Figure 1B) scores of 0.955 and 0.653 for heavy chains, and 0.976 and 0.889 for light chains (Supplementary Figure S2), respectively. Abalign shows comparable or even better performance than MUSCLE (0.949, 0.645 for heavy chains and 0.973, 0.833 for light chains) and ClustalO (0.950, 0.638 for heavy chains and 0.971, 0.845 for light chains), and surpasses MAFFT (0.770, 0.568 for heavy chains and 0.946, 0.778 for light chains) significantly in this test. Further analysis demonstrates that Abalign excels in accurately aligning highly conserved motifs, while preserving the clear demarcation between FRs and CDRs (Supplementary Figure S3). This sets it apart from other methods which often introduce blurring in the boundaries between these regions. These findings highlight the effectiveness of Abalign in conducting BCR-specific MSA tasks.

In order to evaluate Abalign's high-throughput performance, we tested it on another dataset containing 1 023 000 BCR sequences (Materials and Methods). We measured the accuracy using the STRIKE score, which allows for 3D-structure-based evaluation without reference alignment. To

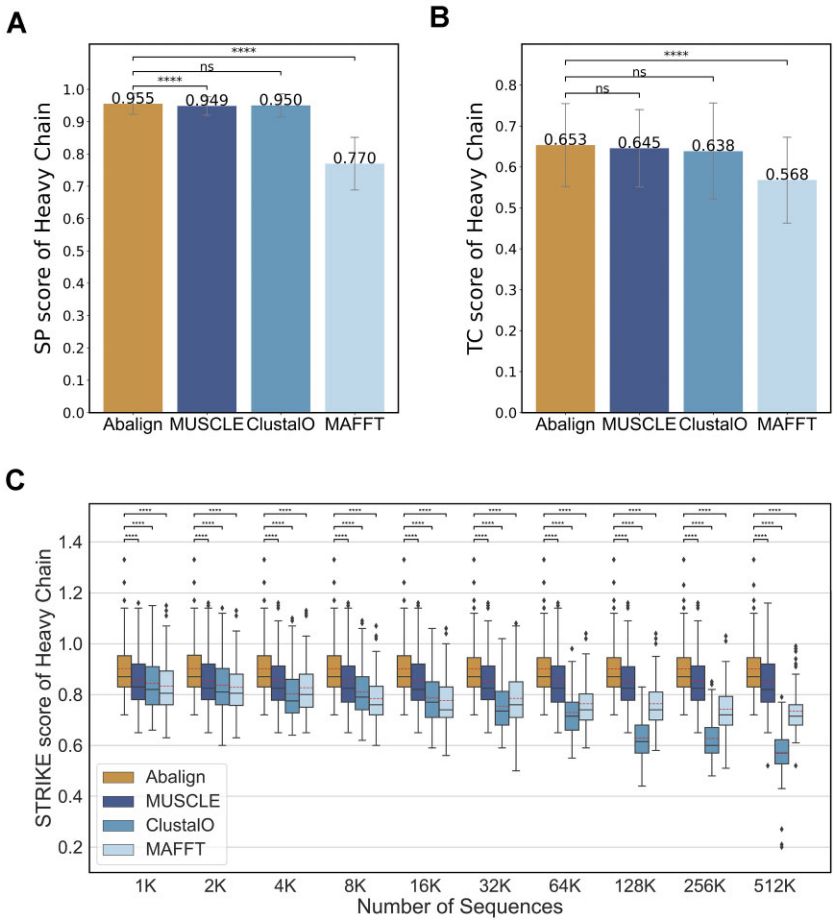


Figure 1. The performance of Abalign, MUSCLE, ClustalO, and MAFFT on the heavy chain variable domains using three benchmark scores: SP score, TC score, and STRIKE score. (**A**) SP scores of testing the 1800 heavy chains from a 3D-structure-based benchmark set. We split the sequences into 10 datasets, and each dataset produced a multiple sequence alignment and a SP score. (**B**) TC scores of testing the same benchmark set. We split the sequences into 10 datasets, and each dataset produced a multiple sequence alignment and a TC score. (**C**) STRIKE scores of testing the 10 subsets of BCR sequencing data from COVID-19 patients, ranging from 1000 to 512 000 sequences. Each subset produced a MSA and 100 STRIKE scores. We used Wilcoxon signed-rank tests to assess the significance of the results, with P values denoted as $^{****}P < 0.0001$, $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$, and 'ns' indicating not significant.

assess the performance as the data size increased, we divided the dataset into 10 subsets, each containing 1000, 2000, 4000, …, 512 000 sequences, respectively. Our results show that Abalign significantly outperforms the other three state-of-the-art methods in terms of STRIKE score, and maintains its high stability in all the subsets (Figure 1C). This result can be attributed to the fact that Abalign utilizes antibody numbering to assign the positions of each residue. As a result, the positions are not influenced by other sequences in the MSA, and therefore, the accuracy of the results is maintained even when the data size increases. Conversely, ClustalO and MAFFT demonstrate a significant decrease in performance, primarily because of their tendency to blur the boundaries between FRs and CDRs as the data size increases (Supplementary Figure S3A). We also observed the performance of MUSCLE, exhibiting relative stability in heavy chains, which may attribute to its divide-and-conquer algorithm (38). However, similar to Clustal Omega and MAFFT, its performance decreases in light chains due to its inability to accurately align FR4 (Supplementary Figure S3B).

Efficiency is a crucial factor when it comes to processing large-scale BCR sequencing data. Using a Ryzen 2990WX CPU with 16 threads, Abalign is able to complete the alignment of the above 512000 sequences in just 6 minutes. This is notably faster than MUSCLE, Clustal Omega, and MAFFT, which takes 53, 13 and 7 times longer, respectively (Figure 2A). In terms of memory consumption, Abalign also outperforms these other methods. It uses only 2.3 GB of memory, which is 20%, 34% and 67% size of the others, respectively (Figure 2B). These differences become even more pronounced as the size of the input data increases.

Overall, the performance of Abalign in terms of accuracy, speed and memory usage makes it an ideal choice for the MSA of large-scale BCR sequencing data.

Software features

Leveraging the advantages of its unique MSA method, Abalign has integrated a range of useful functions and presented them in a user-friendly manner through visual interfaces (Figure 3A and Supplementary Figure S4). Users

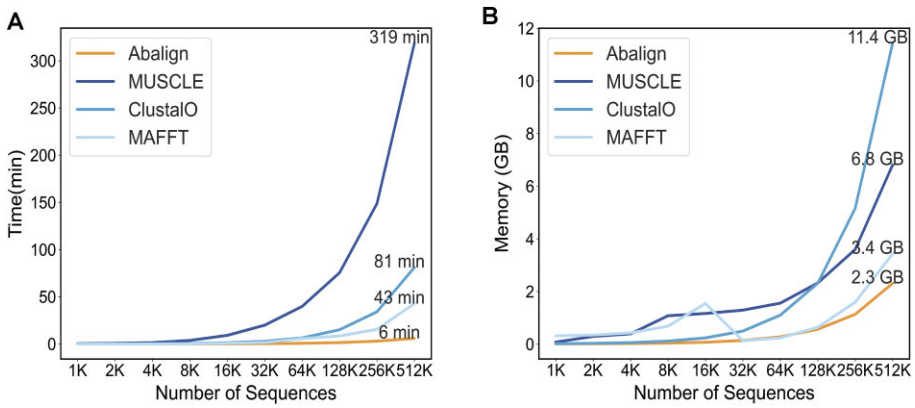


Figure 2. The run time and memory consumption of Abalign, MUSCLE, ClustalO and MAFFT in testing 10 subsets of BCR sequencing data from COVID-19 patients, ranging from 1000 to 512 000 sequences. (**A**) Line graph of the relationship between time and sample size. (**B**) Line graph of the relationship between memory consumption and sample size.

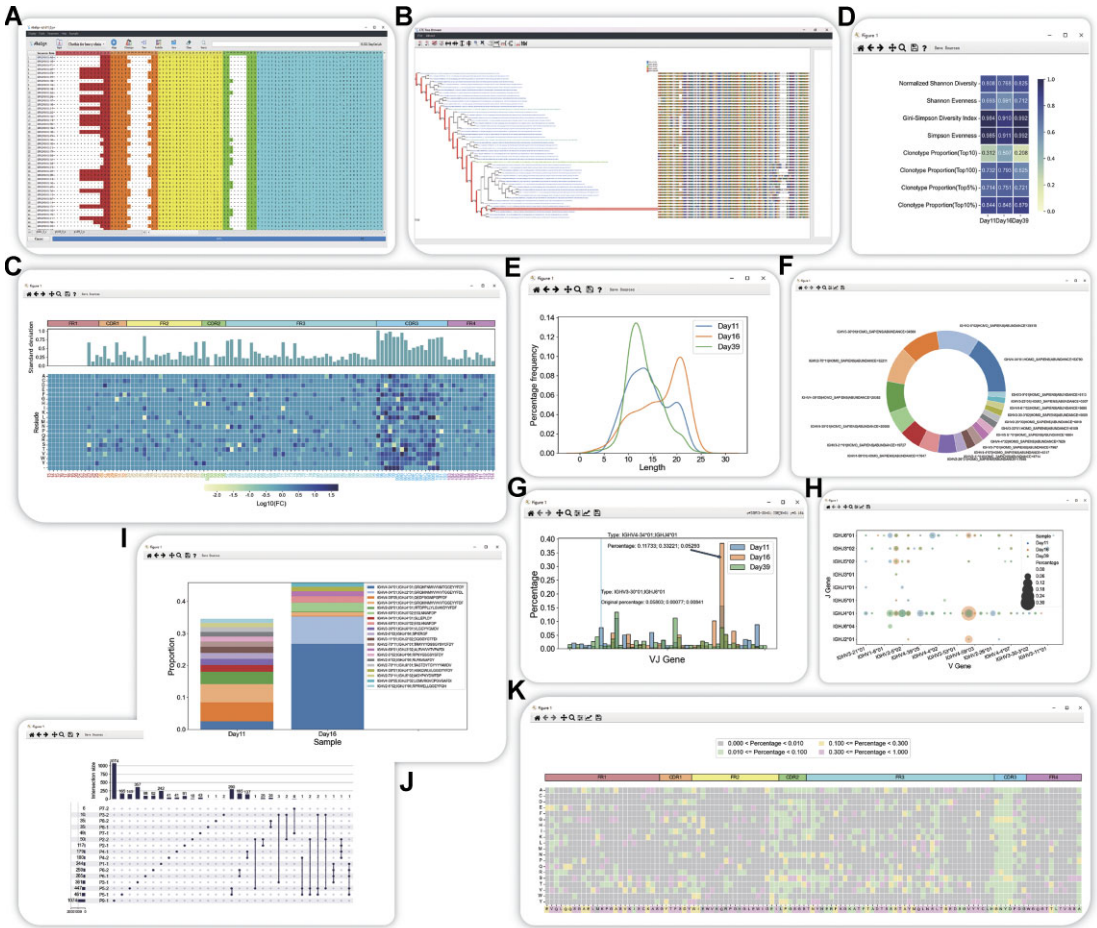


Figure 3. Software features of Abalign. (**A**) The graphic interface displaying the aligned sequences after MSA. The FR1/2/3/4 and CDR1/2/3 are highlighted in colors. (**B**) B-cell lineage tree, annotated with the aligned sequences as well as their proportions. The most abundant sequence is highlighted in red. (**C**) Residue preferences obtained from BCR repertoire analysis. The column chart on the upper panel displays the frequencies of residue mutation at each position. The heatmap in the lower panel shows the frequencies of the 20 residues at each position. The dark colors indicate positive selection, while the light colors indicate negative selection. (**D**) Diversity statistics of BCR repertoire, including Shannon–Weiner index, Gini-Simpson index, Shannon evenness and Simpson evenness. (**E**) CDR3 length distribution of multiple BCR repertoires. (**F**) V gene usage distribution. (G-H) Dynamic VJ gene usage of multiple BCR repertoires. The annotation will be displayed when users hover the mouse over the target. (**I**) The variation of clonotypes between two BCR repertoires. The top 20 most variable clonotypes are displayed in colors. (**J**) The numbers of public clonotypes in multiple BCR repertoires. (**K**) Heatmap of observed residues in human BCRs obtained from the OAS immune repertoire database. The query sequence is displayed at the bottom of the heatmap for comparison. Residue frequencies for the human antibody are indicated by colors. Unusual residues are marked in grey.

can interact with Abalign in real-time with just a few clicks of buttons. The results can be easily presented and exported as image or text files, with various customization options available to users.

Input format

Abalign accepts nucleotide or protein sequence files in the FASTA format. To ensure accuracy, we recommend performing an error-correction process for BCR repertoire sequencing data with unique molecular identifiers (UMIs) using tools such as pRESTO (16), or MIGEC (62). Once the sequences are loaded into Abalign, they will be immediately displayed in the visualization window (Supplementary Figure S5).

MSA

Abalign is capable of automatically identifying whether the input sequence is a BCR heavy or light chain sequence. Upon performing MSA, Abalign will obtain the antibody numbering indexes, delimited FRs and CDRs, as well as the VJ germline genes simultaneously. The resulting MSA will be displayed in Abalign's visual interface (Figure 3A), enabling users to easily investigate and explore the details of their results.

Clonal grouping

To comprehensively assess the composition of BCR repertoires, sequences from the same primitive B cell that differ only by their SHM processes are typically grouped together (63). One common approach is with high CDR3 homology, as well as identical CDR3 length and VJ gene usage (64,65). Abalign offers tunable clonotype definition schemes, including those based on identical VJ genes, identical CDR3 length, and CDR3 sequence identity to cluster sequences.

B-cell lineage tree

B-cell lineage trees are branching diagrams that display the ancestral relationships and developmental stages among individual B cells, providing valuable information on the temporal ordering of mutations, affinity maturation, and selection processes (66). Abalign employs an evolutionary tree module called FastTree2 (57) that allows for the construction of B-cell lineage trees for aligned sequences and visualizes it with ETE3 (58). Additionally, Abalign enables specific annotations on the tree, including abundance of branches, maturation tracks, and aligned sequences. Figure 3B provides an example of a B-cell lineage tree generated and displayed by Abalign using the BCR repertoires from COVID-19 patients.

Mutation profiling

During adaptive immune responses, B cells undergo SHM. A critical step in B cell sequence analysis is identifying these somatic mutations by comparing them to their germline state. Abalign accounts for biases in observed mutation patterns due to positive and negative selection pressures using the MSA results. Abalign demonstrates the mutation profiles through colored blocks, indicating the increased or decreased frequency of replacement mutations, as evidence of antigen-driven positive or negative selection, respectively. Users can gain insights into the contribution of residue mutations to affinity maturation. Figure 3C provides an example of the mutation profile of COVID-19 patients.

Diversity statistics

The quantification of BCR repertoire diversity is informative of an individual's immune status (67). Abalign provides commonly used diversity assessments of clonotypes, such as the Shannon-Weiner index, Gini-Simpson index, Shannon evenness and Simpson evenness (68). These assessments can be used to infer potential changes in clonotypes during the generation of adaptive immune responses (as depicted in Figure 3D). Abalign also allows for counting dynamic changes in the length of the variable domain and differences in VJ gene usage (as illustrated in Figure 3E–H), which provides a comprehensive view of the dynamics of immune status.

Repertoire comparison

Abalign enables users to track longitudinal changes in clonotypes, which is useful for studying antibody maturation and evolution during disease progression. Furthermore, quantifying the overlap of clonotypes across multiple samples can provide essential insights into developing monoclonal antibodies and screening disease markers. Abalign can screen for expanded clonotypes by longitudinally profiling BCR repertoires and generate public sets from multiple individuals. Additionally, Abalign can draw different types of plots that highlight specific features of these data (Figure 3I, J, and Supplementary Figure S6). Detailed information on each clonotype, including germline genes, CDR3, member sequences, and degree of variation, among others, is available and can be displayed in tabular form.

Humanness assessment

Antibody humanization involves the engineering of antibodies to make them more effective and less immunogenic in humans. Abalign includes the MSA result of human BCR/antibody sequences from the OAS immune repertoire database, as well as residue probabilities at each position. Abalign can assess the humanness of the query antibody by detecting the popularity of residues and plotting a comparison figure to aid in engineering unusual residues (Figure 3K).

CONCLUSIONS AND FUTURE PERSPECTIVES

High-throughput sequencing of B-cell immune repertoires is revolutionizing our ability to profile the adaptive immune response against a wide range of diseases. However, this produces hundreds of millions of sequences that require specialized bioinformatics tools for analysis. In this work, we propose an integrated analyzing platform capable of handling massive BCR sequencing data through

ultrafast MSA. The platform enables intensive computation on personal computers, eliminating the need for high-performance computing clusters. Abalign's unique feature of ultrafast MSA produces results consistent with well-established antibody numbering schemes, enabling it to seamlessly connect with a broad range of BCR data analysis. Users can perform various analyses, including clonal grouping, lineage tree construction, mutation profiling, diversity statistics, VJ gene assignment, repertoire comparison and more, without the need for programming scripts. Overall, Abalign represents a powerful and accessible tool for analyzing BCR sequencing data, providing a much-needed solution for researchers in the field of immunology.

Although Abalign offers advanced features, there are still some shortcomings that we aim to address in our subsequent work. One of these limitations is that Abalign ignores annotation information such as isotype (e.g. IgM, IgG) correlated with the sequence. Another limitation is that Abalign cannot currently handle T-cell receptor sequences, which are another key component in the analysis of immune repertoires. We are working to address these limitations in future versions of Abalign. By incorporating user feedback and our efforts, we hope to make Abalign an increasingly useful tool for the analysis of immune repertoires.

DATA AVAILABILITY

Abalign is publicly available at http://cao.labshare.cn/abalign/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank Prof. X. Shirley Liu and Dr Li Song of Harvard University for invaluable discussion. And the authors also wish to thank Dr Xiaoqiong Wei, Dr Qingyang Ding, Dr Jing Xiao, Dr He Li and Dr Zhangyi Ouyang for their assistance in testing Abalign.

FUNDING

National Natural Science Foundation of China [81973243, 32000472]; Biological Resources Progaramme, Chinese Academy of Sciences [KFJ-BRP-008], Innovation Promotion Program of NHC and Shanghai Key Labs SIBPT [RC2023-01]; Shanghai Academy of Science&Technology [SKY2022003].

Conflict of interest statement. None declared.

REFERENCES

1. Kim,S., Davis,M., Sinn,E., Patten,P. and Hood,L. (1981) Antibody diversity: somatic hypermutation of rearranged VH genes. *Cell*, **27**, 573–581.
2. Tonegawa,S. (1983) Somatic generation of antibody diversity. *Nature*, **302**, 575–581.
3. Hwang,J.K., Alt,F.W. and Yeap,L.-S. (2015) Related mechanisms of antibody somatic hypermutation and class switch recombination. *Microbiol. Spectr.*, **3**, MDNA3-0037–2014.
4. Melchers,F. (2015) Checkpoints that control B cell development. *J. Clin. Invest.*, **125**, 2203–2210.
5. Mikocziova,I., Greiff,V. and Sollid,L.M. (2021) Immunoglobulin germline gene variation and its impact on human disease. *Genes Immun.*, **22**, 205–217.
6. Hou,X. L., Wang,L., Ding,Y. L., Xie,Q. and Diao,H. Y. (2016) Current status and recent advances of next generation sequencing techniques in immunological repertoire. *Genes and immunity*, **17**, 153–164.
7. Lindau,P. and Robins,H.S. (2017) Advances and applications of immune receptor sequencing in systems immunology. *Current Opinion in Systems Biology*, **1**, 62–68.
8. Song,L., Ouyang,Z., Cohen,D., Cao,Y., Altreuter,J., Bai,G., Hu,X., Livak,K.J., Li,H., Tang,M., Li,B. and Liu,X.S. (2022) Comprehensive Characterizations of Immune Receptor Repertoire in Tumors and Cancer Immunotherapy Studies. *Cancer Immunol. Res.*, **10**, 788–799.
9. Kiyotani,K., Mai,T. H., Yamaguchi,R., Yew,P.Y., Kulis,M., Orgel,K., Imoto,S., Miyano,S., Burks,A.W. and Nakamura,Y (2018) Characterization of the B-cell receptor repertoires in peanut allergic subjects undergoing oral immunotherapy. *J. Hum. Genet.*, **63**, 239–248.
10. Marks,C. and Deane,C.M. (2020) How repertoire data are changing antibody science. *J. Biol. Chem.*, **295**, 9823–9837.
11. Yaari,G. and Kleinstein,S.H. (2015) Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.*, **7**, 121.
12. Song,L., Cohen,D., Ouyang,Z., Cao,Y., Hu,X. and Liu,X.S. (2021) TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nat. Methods*, **18**, 627–630.
13. Chen,S.-Y., Liu,C.-J., Zhang,Q. and Guo,A.-Y. (2020) An ultra-sensitive T-cell receptor detection method for TCR-seq and RNA-seq data. *Bioinformatics*, **36**, 4255–4262.
14. Bolotin,D.A., Poslavsky,S., Mitrophanov,I., Shugay,M., Mamedov,I.Z., Putintseva,E.V. and Chudakov,D.M. (2015) MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods*, **12**, 380–381.
15. Shlemov,A., Bankevich,S., Bzikadze,A., Turchaninova,M.A., Safonova,Y. and Pevzner,P.A. (2017) Reconstructing antibody repertoires from error-prone immunosequencing reads. *J. Immunol.*, **199**, 3369–3380.
16. Vander Heiden,J.A., Yaari,G., Uduman,M., Stern,J.N.H., O'Connor,K.C., Hafler,D.A., Vigneault,F. and Kleinstein,S.H. (2014) pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, **30**, 1930–1932.
17. Kuchenbecker,L., Nienen,M., Hecht,J., Neumann,A.U., Babel,N., Reinert,K. and Robinson,P.N. (2015) IMSEQ—a fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics*, **31**, 2963–2971.
18. Gupta,N.T., Vander Heiden,J.A., Uduman,M., Gadala-Maria,D., Yaari,G. and Kleinstein,S.H. (2015) Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics*, **31**, 3356–3358.
19. Barak,M., Zuckerman,N.S., Edelman,H., Unger,R. and Mehr,R. (2008) IgTree: creating immunoglobulin variable region gene lineage trees. *J. Immunol. Methods*, **338**, 67–74.
20. Ye,J., Ma,N., Madden,T.L. and Ostell,J.M. (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.*, **41**, W34–W40.
21. Alamyar,E., Duroux,P., Lefranc,M.-P. and Giudicelli,V. (2012) IMGT(®) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol. Biol.*, **882**, 569–604.
22. Abhinandan,K.R. and Martin,A.C.R. (2008) Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Mol. Immunol.*, **45**, 3832–3839.
23. Dunbar,J. and Deane,C.M. (2016) ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, **32**, 298–300.
24. Li,L., Chen,S., Miao,Z., Liu,Y., Liu,X., Xiao,Z.-X. and Cao,Y. (2019) AbRSA: a robust tool for antibody numbering. *Protein Sci.*, **28**, 1524–1531.
25. Shugay,M., Bagaev,D.V., Turchaninova,M.A., Bolotin,D.A., Britanova,O.V., Putintseva,E.V., Pogorelyy,M.V., Nazarov,V.I., Zvyagin,I.V., Kirgizova,V.I. *et al.* (2015) VDJtools: unifying

post-analysis of T cell receptor repertoires. *PLoS Comput. Biol.*, **11**, e1004503.
26. Zhang,W., Du,Y., Su,Z., Wang,C., Zeng,X., Zhang,R., Hong,X., Nie,C., Wu,J., Cao,H. *et al.* (2015) IMonitor: a robust pipeline for TCR and BCR repertoire analysis. *Genetics*, **201**, 459–472.
27. Margreitter,C., Lu,H.-C., Townsend,C., Stewart,A., Dunn-Walters,D.K. and Fraternali,F. (2018) BRepertoire: a user-friendly web server for analysing antibody repertoire data. *Nucleic Acids Res.*, **46**, W264–W270.
28. Cortina-Ceballos,B., Godoy-Lozano,E.E., Sámano-Sánchez,H., Aguilar-Salgado,A., Velasco-Herrera,M.D.C., Vargas-Chávez,C., Velázquez-Ramírez,D., Romero,G., Moreno,J., Téllez-Sosa,J. *et al.* (2015) Reconstructing and mining the B cell repertoire with ImmunediveRsity. *MAbs*, **7**, 516–524.
29. Duez,M., Giraud,M., Herbert,R., Rocher,T., Salson,M. and Thonier,F. (2016) Vidjil: a web platform for analysis of high-throughput repertoire sequencing. *PLoS One*, **11**, e0166126.
30. Ralph,D.K. and Matsen,F.A. (2016) Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. *PLoS Comput. Biol.*, **12**, e1004409.
31. Marcou,Q., Mora,T. and Walczak,A.M. (2018) High-throughput immune repertoire analysis with IGoR. *Nat. Commun.*, **9**, 561.
32. Rogosch,T., Kerzel,S., Hoi,K.H., Zhang,Z., Maier,R.F., Ippolito,G.C. and Zemlin,M. (2012) Immunoglobulin analysis tool: a novel tool for the analysis of human and mouse heavy and light chain transcripts. *Front. Immunol.*, **3**, 176.
33. Schaller,S., Weinberger,J., Jimenez-Heredia,R., Danzer,M., Oberbauer,R., Gabriel,C. and Winkler,S.M. (2015) ImmunExplorer (IMEX): a software framework for diversity and clonality analyses of immunoglobulins and T cell receptors on the basis of IMGT/HighV-QUEST preprocessed NGS data. *BMC Bioinf.*, **16**, 252.
34. Van Noorden,R., Maher,B. and Nuzzo,R. (2014) The top 100 papers. *Nature*, **514**, 550–553.
35. Bawono,P., Dijkstra,M., Pirovano,W., Feenstra,A., Abeln,S. and Heringa,J. (2017) Multiple sequence alignment. *Methods Mol. Biol.*, **1525**, 167–189.
36. Sievers,F. and Higgins,D.G. (2014) Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.*, **1079**, 105–116.
37. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
38. Edgar,R.C. (2022) Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.*, **13**, 6968.
39. Collins,K. and Warnow,T. (2018) PASTA for proteins. *Bioinformatics*, **34**, 3939–3941.
40. Liu,K., Warnow,T.J., Holder,M.T., Nelesen,S.M., Yu,J., Stamatakis,A.P. and Linder,C.R. (2012) SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst. Biol.*, **61**, 90–106.
41. Talavera,G. and Castresana,J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.*, **56**, 564–577.
42. Lefranc,M.-P., Pommié,C., Ruiz,M., Giudicelli,V., Foulquier,E., Truong,L., Thouvenin-Contet,V. and Lefranc,G. (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and ig superfamily V-like domains. *Dev. Comp. Immunol.*, **27**, 55–77.
43. Wu,T.T. and Kabat,E.A. (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.*, **132**, 211–250.
44. Chothia,C., Lesk,A.M., Tramontano,A., Levitt,M., Smith-Gill,S.J., Air,G., Sheriff,S., Padlan,E.A., Davies,D. and Tulip,W.R. (1989) Conformations of immunoglobulin hypervariable regions. *Nature*, **342**, 877–883.
45. Giudicelli,V., Chaume,D. and Lefranc,M.-P. (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.*, **33**, D256–D261.
46. Edgar,R.C. and Batzoglou,S. (2006) Multiple sequence alignment. *Curr. Opin. Struct. Biol.*, **16**, 368–373.
47. Pei,J., Kim,B.-H. and Grishin,N.V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.*, **36**, 2295–2300.
48. Zhou,C.L.E. (2015) CombAlign: a code for generating a one-to-many sequence alignment from a set of pairwise structure-based sequence alignments. *Source Code Biol. Med.*, **10**, 9.
49. Dunbar,J., Krawczyk,K., Leem,J., Baker,T., Fuchs,A., Georges,G., Shi,J. and Deane,C.M. (2014) SAbDab: the structural antibody database. *Nucleic Acids Res.*, **42**, D1140–D1146.
50. Akdel,M., Durairaj,J., de Ridder,D. and van Dijk,A.D.J. (2020) Caretta - A multiple protein structure alignment and feature extraction suite. *Comput. Struct. Biotechnol. J.*, **18**, 981–992.
51. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
52. Olsen,T.H., Boyles,F. and Deane,C.M. (2022) Observed antibody space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci.*, **31**, 141–146.
53. Kemena,C., Taly,J.-F., Kleinjung,J. and Notredame,C. (2011) STRIKE: evaluation of protein msas using a single 3D structure. *Bioinformatics*, **27**, 3385–3391.
54. Edgar,R.C. (2010) Quality measures for protein alignment benchmarks. *Nucleic Acids Res.*, **38**, 2145–2153.
55. Thompson,J.D., Koehl,P., Ripp,R. and Poch,O. (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.
56. Chatzou,M., Magis,C., Chang,J.-M., Kemena,C., Bussotti,G., Erb,I. and Notredame,C. (2016) Multiple sequence alignment modeling: methods and applications. *Brief Bioinform*, **17**, 1009–1023.
57. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
58. Huerta-Cepas,J., Serra,F. and Bork,P. (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.*, **33**, 1635–1638.
59. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
60. Bembom,O. (2007) Sequence logos for DNA sequence alignments. *R Package Version*, 1.
61. Lex,A., Gehlenborg,N., Strobelt,H., Vuillemot,R. and Pfister,H. (2014) UpSet: visualization of intersecting sets. *IEEE Trans... Vis. Comput. Graph.*, **20**, 1983–1992.
62. Shugay,M., Britanova,O.V., Merzlyak,E.M., Turchaninova,M.A., Mamedov,I.Z., Tuganbaev,T.R., Bolotin,D.A., Staroverov,D.B., Putintseva,E.V., Plevova,K. *et al.* (2014) Towards error-free profiling of immune repertoires. *Nat. Methods*, **11**, 653–655.
63. Hershberg,U. and Luning Prak,E.T. (2015) The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **370**, 20140239.
64. Greiff,V., Miho,E., Menzel,U. and Reddy,S.T. (2015) Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol.*, **36**, 738–749.
65. López-Santibáñez-Jácome,L., Avendaño-Vázquez,S.E. and Flores-Jasso,C.F. (2019) The pipeline repertoire for ig-seq analysis. *Front. Immunol.*, **10**, 899.
66. Kepler,T.B., Munshaw,S., Wiehe,K., Zhang,R., Yu,J.-S., Woods,C.W., Denny,T.N., Tomaras,G.D., Alam,S.M., Moody,M.A. *et al.* (2014) Reconstructing a B-cell clonal lineage. II. Mutation, selection, and affinity maturation. *Front. Immunol.*, **5**, 170.
67. Greiff,V., Bhat,P., Cook,S.C., Menzel,U., Kang,W. and Reddy,S.T. (2015) A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med*, **7**, 49.
68. Patil,G.P. and Taillie,C. (1982) Diversity as a concept and its measurement. *J. Am. Statist. Assoc.*, **77**, 548–561.