PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Computational detection of antigen-specific B cell receptors following immunization

Maria Francesca Abbate[a,b], Thomas Dupic[c], Emmanuelle Vigne[b], Melody A. Shahsavarian[b], Aleksandra M. Walczak[a,1,2], and Thierry Mora[a,1,2]

Edited by Arup Chakraborty, Massachusetts Institute of Technology, Cambridge, MA; received January 16, 2024; accepted July 10, 2024

B cell receptors (BCRs) play a crucial role in recognizing and fighting foreign antigens. High-throughput sequencing enables in-depth sampling of the BCRs repertoire after immunization. However, only a minor fraction of BCRs actively participate in any given infection. To what extent can we accurately identify antigen-specific sequences directly from BCRs repertoires? We present a computational method grounded on sequence similarity, aimed at identifying statistically significant responsive BCRs. This method leverages well-known characteristics of affinity maturation and expected diversity. We validate its effectiveness using longitudinally sampled human immune repertoire data following influenza vaccination and SARS-CoV-2 infections. We show that different lineages converge to the same responding Complementarity Determining Region 3, demonstrating convergent selection within an individual. The outcomes of this method hold promise for application in vaccine development, personalized medicine, and antibody-derived therapeutics.

adaptive immune system | antigen specificity | repertoire sequencing | influenza vaccine | COVID-19

Significance

The repertoire of antibodies produced by our immune system contains a lot of information about our infection status and protection against current and future pathogens. Extracting this information could help unlock new ways to develop diagnostics and drugs. However, we lack a dictionary to decode which antibodies relate to which target or disease. In this study, we propose a computational method based on graph theory for discovering disease-associated antibodies from repertoire data, using a single blood sample taken from individuals after a flu vaccine or COVID19 infection. We show that the identified antibodies form groups of very similar amino acid sequences that originate from many convergent clones.

B cells in the adaptive immune system express receptors on their surface that bind antigens to initiate an immune response (Fig. 1). Identifying B cell receptors (BCRs) that respond to specific antigens is an important goal for describing the dynamics of B cell immunity following vaccination, with potential applications in vaccine development (1–6), personalized medicine (7–9), and antibody-derived therapeutics (10–14).

The initial diversity of immune receptor repertoires is generated through the random assembly of genomic templates complemented by deletions and insertions at the gene junctions. This initial diversity is further enhanced by affinity maturation. Upon antigenic stimulation, B cells that recognize an antigen migrate to germinal centers, where they acquire somatic hypermutations in their antigen receptor, and undergo selection for antigen binding. This process ultimately results in B cells that better recognize the antigen. Since B cells are released to the periphery throughout, affinity maturation produces lineages of related B cells that can be identified by sequence similarity of their antigen receptor (15–17). As a result, cells carrying different but similar sequences are involved in neutralizing the same antigen with different potencies. In addition, several distinct founder B cells specific to the same antigen can seed distinct but related lineages (18). It is still unclear how affinity maturation further diversifies and focuses the responding repertoire.

Recent advances in high-throughput sequencing technologies make it possible to directly profile the immune repertoire by sequencing the B cell DNA or messenger RNA (mRNA) taken from blood or tissue samples (RepSeq) (19–25). The experiments provide a list of unique sequences with their relative abundances. However, exploiting repertoire information for decoding the immune response is hindered by both our inability to reliably predict the specificity of a given BCR to a given antigen and the fact that only a small fraction of the repertoire is involved in any infection (Fig. 1*A*). Existing methods for identifying responding BCRs combine traditional sorting assays and sequencing (11, 26–29) with computational analyses (30). Approaches based on machine learning (31–36), network analysis (37–39), publicness across multiple individuals (40–42), or using structural information (43–47), have also been proposed to focus on the disease-specific subrepertoire. Longitudinal sampling of repertoires after a strong antigenic stimulation, such as a vaccine or disease (3, 48, 49) is an agnostic way to study the response that does not require prior knowledge about the identity of the triggering antigens, and can identify a polyclonal response against multiple epitopes. Such approaches have also been successfully applied to T cell repertoires (50–53) where the identity of epitopes can sometimes be reverse-mapped using specificity databases (54). These methods directly rely on BCR abundance measured in the repertoire, offering alternatives to traditional

Author affiliations: [a]Laboratoire de physique de l'École normale supérieure, CNRS, Paris Sciences et Lettres University, Sorbonne Université, and Université Paris-Cité, Paris 75005, France; [b]Large Molecule Research, Sanofi, Vitry-sur-Seine 94 400, France; and [c]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138

Author contributions: E.V., M.A.S., A.M.W., and T.M. designed research; M.F.A. performed research; M.F.A. and T.D. contributed new reagents/analytic tools; M.F.A., A.M.W., and T.M. analyzed data; and M.F.A., A.M.W., and T.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]A.M.W. and T.M. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: awalczak@phys.ens.fr or tmora@phys.ens.fr.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2401058121/-/DCSupplemental.

Published August 20, 2024.

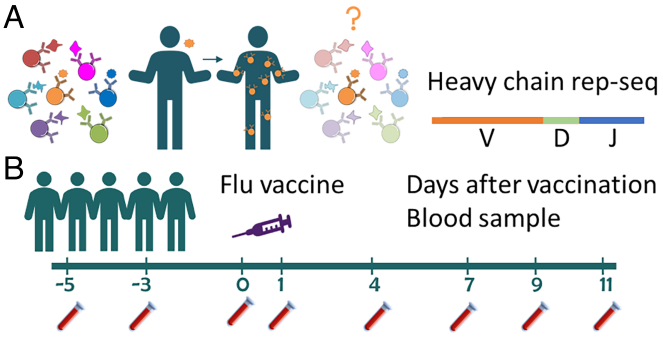

Fig. 1. Identifying responding antibodies from repertoires. (*A*) B cell repertoires exploit a diverse set of antigen receptors (antibodies) with different antigen specificities. Upon an immune challenge, antigen-specific B cells proliferate and mutate. The question addressed here is how to identify these responding clones from repertoire data. (*B*) We exploit bulk BCRs immune repertoire sequencing data (48) that covers the V, D, and J segments of the BCRs heavy chain to detect influenza–responding B cells without knowing the epitope, using the repertoire sampled at a single timepoint. Five healthy humans were vaccinated in late spring of 2012 with the 2011 to 2012 trivalent seasonal flu vaccine. Blood samples were collected before (days −5, −3, and 0) and after (1, 4, 7, 9, and 11) vaccine administration.

screening assays. However, they require to obtain blood samples from individuals prior to the immune challenge, which is not always practical.

We propose a computational approach for identifying clusters of expanded BCRs from the repertoire measured at a single time point, by combining information about sequence similarity and convergent selection. Such convergence has previously been exploited to identify responding receptors in the context of T cell repertoires and is the basis of the ALICE software tool (55). However, affinity maturation and lack of human leukocyte antigen restriction make the problem very different for B cells. We illustrate our method on data from recent studies that track the unbiased B cell responses of 5 healthy individuals after influenza immunization with the 2011 to 2012 trivalent seasonal flu vaccine in late spring of 2012 (48) (Fig. 1*B*), and 18 recent COVID-19 patients at the peak of infection (3). This allows us to study the multiplicity, diversity, and convergent selection of distinct lineages toward viral antigens with great details.

1. Results

1.1. Computational Identification of Responding Clones from a Single Timepoint. When a B cell is involved in an immune response, its BCR undergoes proliferation and mutations. This process yields many copies of the original BCR, while also producing mutated receptors with similar sequence and antigen specificity as the ancestral BCR. At the repertoire level, we expect two main effects: elevated frequencies of responding receptors, and the formation of extensive clusters of similar sequences. These clusters may arise both from the expansion of a single BCR lineage and from the convergent selection of multiple lineages with distinct progenitors (Fig. 2*A*).

We first examined the profile of clonotype frequencies in the BCR heavy chain (IgH) repertoires of five healthy individuals who had received the 2011 to 2012 trivalent seasonal flu vaccine (see *Methods* and Dataset S1 for details), both before the vaccine (day 0), and at the peak of the response (day 7) (48). A clonotype is defined by the unique amino acid sequence coding for the Complementarity Determining Region 3 (CDR3) of the heavy chain. While other parts of the sequence also determine specificity, focusing the most variable region gives us more statistical power to detect convergent selection. The comparison of the distributions of clonotype abundances shows very little change between before vaccination and at the peak of the response (Fig. 2*B*), challenging the expectation that the postvaccination repertoire should be dominated by a few large responding clonotypes.

We next consider the *number of neighbors* in amino acid space as a measure of sequence similarity. For each CDR3 amino acid sequence found in the repertoire, we count the number of distinct IgH nucleotide sequences whose CDR3 differ by exactly one amino acid. Accounting for the multiplicity of nucleotide sequences helps capture the diversity of convergent synonymous variants in the response. In contrast to frequencies, the distribution of the number of neighbors does change significantly (Fig. 2*C*) from day 0 to day 7, in agreement with the expectation that mutations and convergent selection can create clusters of related sequences. To explore how much of this observation is due to mutations versus convergent selection, we applied a similar approach to lineages. We inferred lineages at days 0 and 7 using HILARy software (15), and called two lineages neighbors if the average distance between their sequences was below or equal to 2. By contrast to the case of single clonotypes, we did not observe a clear separation between the distributions of the numbers of lineage neighbors at the two time points (*SI Appendix*, Fig. S2). The signal contained in lineage convergence does not have the statistical power to detect the response, meaning that mutations are essential. However, it does not rule out the existence of convergent selection, as we will see later.

We use the observation of Fig. 2*C* to introduce two approaches to identify antigen-specific B cell receptors. The approaches extend previous ideas proposed for T cells (55) to the context of affinity maturation. The first, called fast-STAR (fast Single Timepoint Antibody Recognition) prioritizes computational speed and performs efficient thresholding to output sequences with a high confidence level as responders. The second, less specific method, called full-STAR, assigns a score to each sequence based on a probabilistic approach.

Specifically, fast-STAR calls an IgH CDR3 amino acid sequence a responding clonotype (a "hit") if its number of neighbors, normalized by the total number of unique nucleotide sequences, is higher than a certain threshold. This threshold is set to $9.6 \cdot 10^{-4}$ to obtain a false-discovery rate of ≈5%, as estimated by taking the ratio of the number of hits at days −3 and 0 (only false positives) with the number of hits at day 7 (true positives and false positives), averaged over all 5 subjects. Sequences above the threshold are then grouped by single-linkage clustering, where two amino acid sequences are linked if they have Levenshtein distance 1 or lower. Clusters with fewer than 10 sequences are filtered out to mitigate the effect of potential sequencing errors. As a result, fast-STAR only keeps a small number of hits.

This method results in very high specificity but low sensitivity. In addition, it does not exploit the knowledge that certain sequences are expected to have more neighbors than others due to biases in the generation probability. Full-STAR overcomes these limitations by setting a personalized sequence-dependent threshold for the number of neighbors based on the expectation computed from the previously proposed OLGA method (56), which estimates the probability of observing any given sequence in a random repertoire (*Methods*). We can use two methods to estimate the number of neighbors using OLGA: either exactly, by summing the probabilities of generation of each possible neighbor for each observed sequence, or approximately, by exploiting the fact that neighboring sequences have similar probabilities

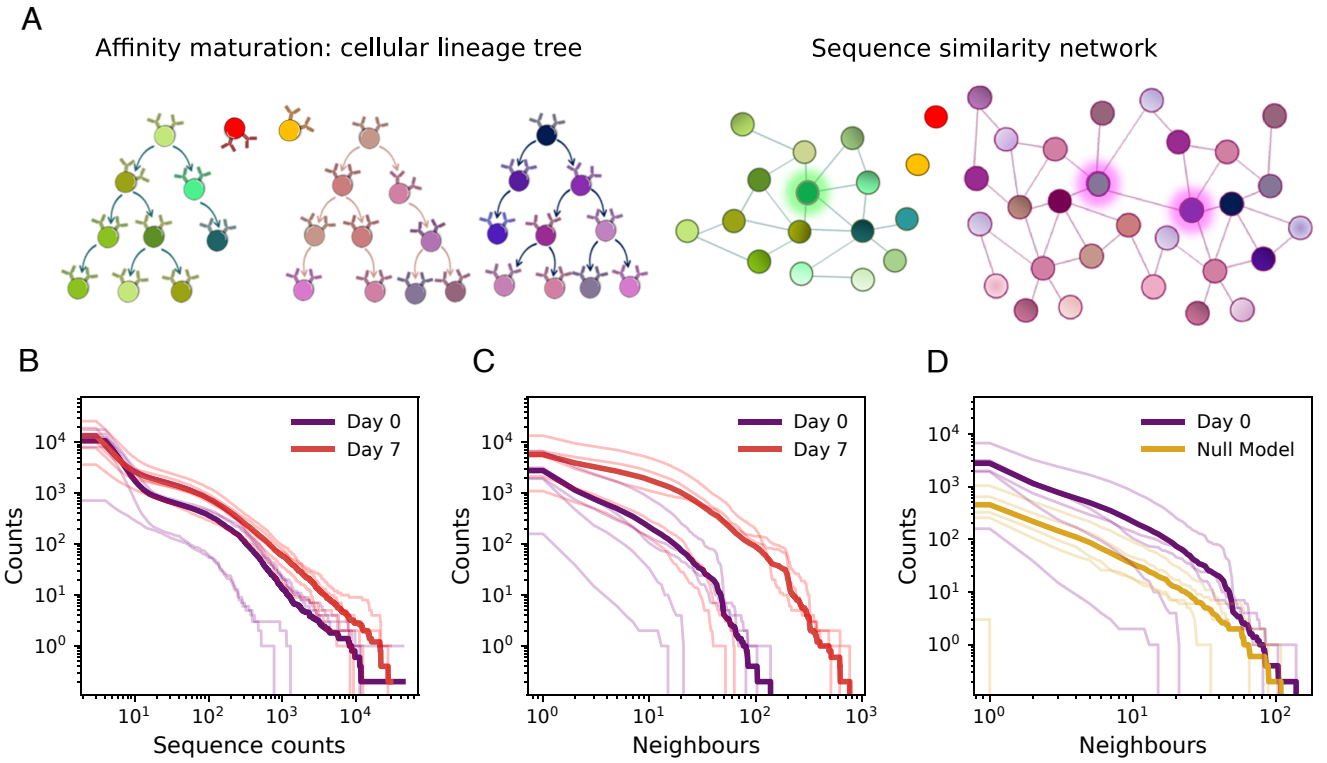

Fig. 2. Similarity network analysis of antibody clonotypes. (*A*) During affinity maturation, distinct naive B cells proliferate and mutate upon recognition of the antigens, giving rise to distinct cell lineages (*Left*). At the sequence level (*Right*), we construct a graph where each node is an IgH nucleotide sequence, and edges connect sequences that differ by at most one amino acid in their CDR3. This may lead to distinct lineages being merged into the same functional cluster (e.g. the pink cluster). The idea of STAR is to identify sequences with high connectivity in the graph (highlighted), which indicates either convergent selection or belonging to a large lineage, or both. (*B*) Distribution of the CDR3 amino acid sequence count for all subjects (background lines) and their average (thick lines), at days 0 and 7. The two distributions are similar. (*C*) The distribution of the number of amino acid neighbors shows a marked difference between days 0 and 7 (same color convention as *B*). (*D*) The distribution of neighbors at day 0 is well described by a computational model of random repertoire generation (56) (see main text).

of generation (*SI Appendix*, Fig. S1) to reduce computational time (*Methods* and *SI Appendix*, Fig. S3). OLGA's prediction for the distribution of neighbors using the second approximate method agrees well with that measured at day 0 (Fig. 2*D*), and so subsequently we will only use that method. Responding sequences are identified as having more neighbors than expected by the model, with significance computed using the Poisson distribution. The threshold on the resulting *P*-value is controlled for multiple testing by the Benjamini–Hochberg procedure by setting a false discovery rate of 0.5%.

1.2. Clones Identified by STAR Recapitulate the Immune Response Dynamics. We applied both fast-STAR and full-STAR to the IgH repertoires of all 5 subjects at each timepoint (Dataset S1). Although neither pipeline used any information about the time course of clonal abundances, they could both detect a marked increase of the number of hits following vaccination, with a peak on day 7 and rapid decay after that (Fig. 3 *A* and *B*). This response peak is consistent with previous observations based on longitudinal analysis (48), and is characteristic of a memory recall response following vaccination. The fast-STAR method, which is more conservative and robust, even finds no hits at all for all the subjects prior to day 7. This suggests that the pipeline specifically identifies responding clonotypes. The most common isotype found in the repertoires is IgM, except at day 7 where it is still substantially represented (*SI Appendix*, Fig. S4*A*). However, almost all clonotypes identified by fast-STAR in subject 1 were IgG (*SI Appendix*, Fig. S4*B*), while full-STAR hits were IgA or IgG (*SI Appendix*, Fig. S4*C*), with most IgA found before the response peak. This serves as further validation that the clonotypes identified by STAR are involved in the memory recall response, and suggests that those identified by full-STAR also include previously expanded clones associated to distinct immune challenges.

Clonotypes identified by STAR on day 7 can be used to retrospectively study the dynamics of the response. Examining the frequency–time courses of single clonotypes identified as responding at day 7 shows a consistent pattern across all subjects, with no detected presence before day 7, and a sharp peak at day 7 followed by rapid decay (Fig. 3 *C* and *D*). This again validates the approach, as neither pipeline used frequency information. To go beyond single clonotypes, we aggregated the frequencies of all clonotypes identified as responsive, and plotted their cumulative frequency as a function of time for both pipelines (Fig. 3 *E* and *F*). These time traces show again a marked peak on day 7 for both pipelines. The less conservative full-STAR captures a larger fraction of the responding repertoire, identifying as much as 75% of the repertoire being involved in the response at its peak.

In addition, we directly validated the ability of some of our hits to bind the virus. In ref. 57, 21 antibodies belonging to 5 distinct lineages found to be expanded in subject 1, and separately single-cell sequenced, were tested for affinity against the vaccine as well as various viral strains using enzyme-linked immunosorbent assays. Among those 5, 2 lineages contained vaccine-binding antibodies, only one of which, called L1 and containing 5 antibodies, was also present in the bulk longitudinal data analyzed in the current paper. These 5 antibodies use 4 distinct CDR3s. All of them belonged to the cluster of hits identified by STAR in

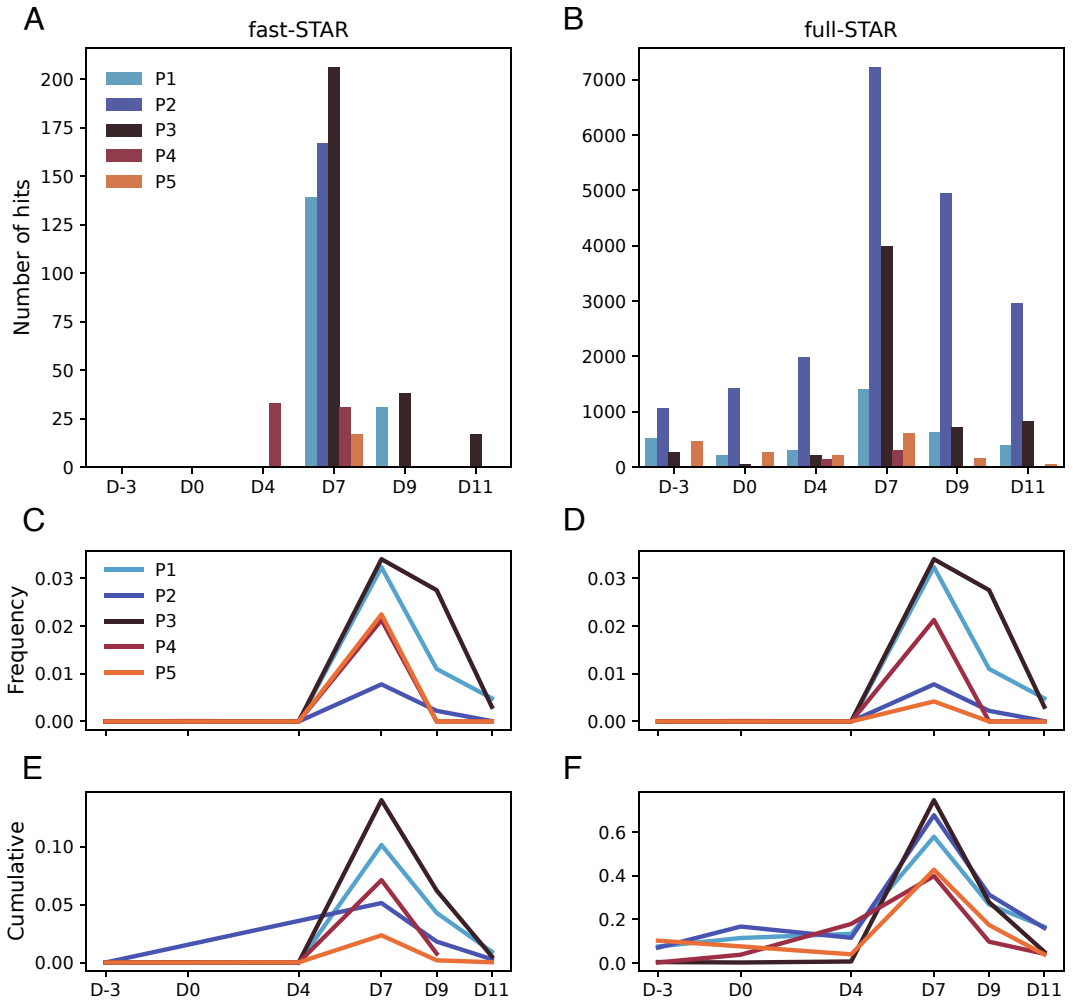

Fig. 3. Results and validation. (*A* and *B*) Number of putative responding sequences identified by (*A*) fast-STAR and (*B*) full-STAR for each day and each subject. We observe very few hits on the days before vaccination and a large peak on day 7. (*C* and *D*) Frequency–time traces of the top-scoring sequences identified on day 7 for each subject found by (*C*) fast-STAR (with largest number of neighbors) and (*D*) full-STAR (with lowest *P*-value). Note that the best-scoring clonotype is the same for the two methods for all subjects except subject 5. (*E* and *F*) Sum of the frequencies of all responding clonotypes according to (*E*) fast-STAR and (*F*) full-STAR.

subject 1 (Fig. 4*A*), providing direct evidence that this group of antibodies specifically target the viral proteins.

As a final validation test, we compared our putative influenza-responding clonotypes to those reported in ref. 58, independently obtained from subjects vaccinated with the inactivated influenza vaccine. In ref. 58, the authors computationally identified responsive sequences as those that significantly expanded between a pre- and postvaccination timepoint. We evaluated the overlap between the IgH amino acid CDR3 sequence of their 1,513 vaccine-responding candidate antibodies, and our STAR hits. In general, overlap of responding clonotypes between distinct datasets is known to be low because of their rarity (42), and we do not expect our list of responding clonotypes to be able to classify specific from nonspecific sequences in other individuals. While we found a small overlap of 2 sequences with our fast-STAR hits, and 9 with our full-STAR hits, these numbers are much larger than expected by chance: we evaluated the overlap between STAR hits and a control dataset of healthy individuals (24), and we found $2 \cdot 10^{-4}$ for fast-STAR and 0.03 for full-STAR, after normalizing for dataset size. Applying a Poisson test to these overlaps gave a *P*-value of respectively $P = 3 \cdot 10^{-8}$ and $P = 10^{-19}$ (*Methods*).

1.3. Convergent Selection. The sequences identified by fast-STAR can be organized as clusters of closely related sequences (*SI Appendix*, Fig. S5). Three our of five subjects (1, 2, and 5) have a single cluster, suggesting an immunodominant response against a single epitope, while the other two (3 and 4) have four and three clusters respectively, suggesting a polyclonal response. To better understand the structure of the repertoire response in sequence space, we analyzed in more detail the sequence structure of fast-STAR hits found in subject 1. These hits formed a single cluster of densely connected and highly conserved sequences (Fig. 4*A* and weblogo therein).

We wondered whether this diversity of similar CDR3 amino acid sequences arose from a single lineage, or from distinct selection events originating from different naive sequences. To explore this question, we applied HILARy software (15) to separate the main cluster of subject into 23 different lineages, and RAxML (59) to reconstruct the corresponding phylogenetic trees (Fig. 4*B*). We observe one predominant lineage, a few smaller lineages (6 of which are represented), and some isolated sequences. Each lineage was associated with a unique V gene (given by MiXCR (60), ignoring allelic variants), and each V gene represented by a single lineage. The same CDR3

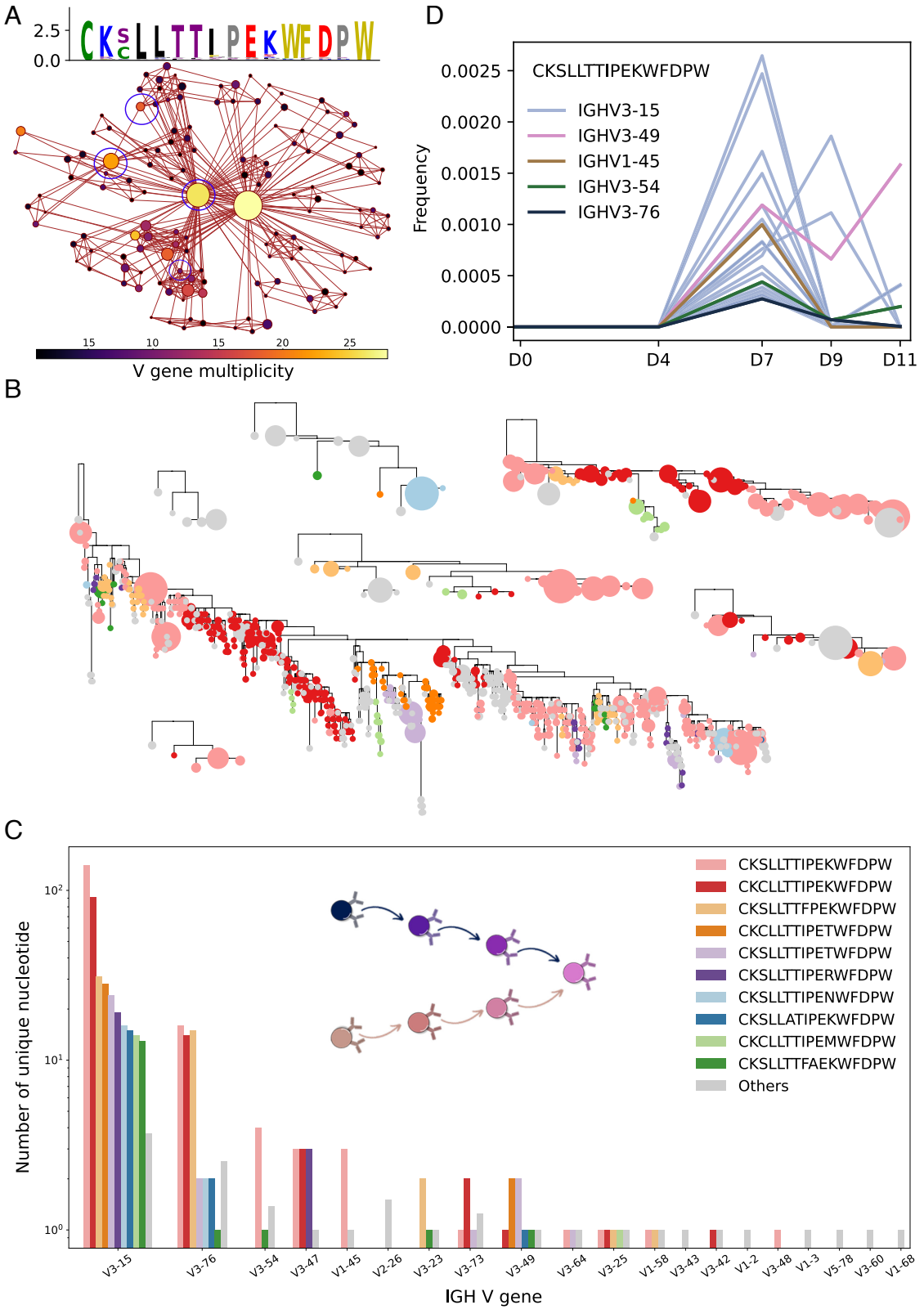

Fig. 4. Convergent antibody response toward a conserved IgH CDR3 motif. (*A*) Graph structure of fast-STAR hits found in subject 1. Each node is an amino acid IgH CDR3 sequence. Node size is proportional to the number of distinct IgH nucleotide sequences with that CDR3, and color indicates the number of distinct V genes used. An edge is drawn between two CDR3 if they differ by one amino acid. The blue circle indicates the four amino acid CDR3s identified in the experimental testing as responding. *Top*: sequence logo of the sequences of the cluster. Height corresponds to the entropy of the amino acid choice at each site, and relative letter size to amino acid frequencies. (*B*) Reconstructed lineage trees of the main cluster of fast-STAR hits in subject 1. Branch length represents numbers of mutations, and node sizes frequencies of individual nucleotide sequences. The root is the unmutated naive sequence reconstructed from genomic templates. Each color represents a distinct amino acid CDR3 sequence (legend in *C*). (*C*) Number of distinct nucleotide sequences with the same amino acid CDR3, grouped by V gene usage. Each CDR3 sequence can be formed using distinct V genes, and within each V gene group, up to hundreds of nucleotide variants. (*D*) Frequency–time course of the most abundant nucleotide sequences associated with the amino acid CDR3 sequence CKSLLTTIPEKWFDPW. Sequences with different V genes are shown in different colors. Each sequence corresponds to a distinct B cell clone that expanded independently at day 7.

amino acid sequences appear multiple times in distinct lineages, with different V genes, suggesting independent expansion and evolution of distinct B cell clones. The most frequent CDR3 amino acid sequence could be formed by all 23 different V genes (Fig. 4*C*). This diversity of V gene assignment was not a spurious consequence of misassignment due to hypermutations and trimming of the 3′ end of the V gene, since the observed hypermutation rate (3 base pairs on average over the sequenced V region) was small compared to the pairwise distance between nearby V genes over the same region after trimming (*SI Appendix*, Fig. S6). It could also not be explained by hybridization of distinct gene rearrangements during PCR, based on the analysis of unique molecular identifiers appended to both ends of the gene products during repertoire profiling (*Methods*).

To further test the hypothesis of independent expansion of distinct B cell clones, we examined the frequency dynamics of individual IgH nucleotide sequence variants coding for the top-scoring amino acid CDR3 of subject 1 (CKSLLTTIPEKWFDPW) (Fig. 4*D*). For illustration purposes, we took the most frequent nucleotide sequences (frequency $\geq 2.5 \cdot 10^{-4}$), and color coded time traces by the germline V gene the sequences use. All traces show a clear and independent expansion between days 4 and 7. By contrast, randomly picked sequences with the same frequency at day 7 do not show the same stereotyped behavior (*SI Appendix*, Fig. S7). This observation supports the notion that multiple lineages independently converged toward a shared functional outcome. The fact that multiple distinct V genes are represented excludes the possibility of an artifact due to sequencing errors or hypermutations. This observation provides direct evidence for antigen-specific convergent selection of multiple lineages and introduces potentially a novel feature for investigating and exploiting antigen-specific receptors.

1.4. Application to COVID-19 Repertoires. To assess the generalizability of our computational pipeline, we applied the same methodology to a distinct dataset from subjects naturally infected with COVID19. We used IgH repertoire data collected from Galson et al. (40), comprising samples from 18 subjects at the infection peak. Note that no longitudinal data were available, providing a test case for our method. We applied the STAR pipelines to identify candidate CDR3 amino acid heavy chain BCR sequences responding to COVID-19 (full list in Dataset S2). Applying fast-STAR we obtained an average of 131 hits (range 0 to 661), while with full-STAR we obtained an average of 10,169 (range 918 to 27,024). The clusters obtained with fast-STAR for each patient are shown in *SI Appendix*, Fig. S8. In certain patients (10 out of 18), a polyclonal response was not detected, indicated by either zero or only one cluster. This observation may be attributed, in part, to variations in sequencing depth, as detailed in Dataset S2. Conversely, some patients exhibited an exceptionally diverse response, with up to 11 clusters observed in the case of patient 10. Notably, the highly conservative fast-STAR pipeline did not detect any shared hits between different patients.

The source study did not test antibodies for specificity, making a direct validation of our candidate COVID-specific sequences difficult. Nonetheless, to assess the specificity of our method, for each subject, we evaluated the overlap between our hits and a comprehensive COVID19 database containing known antibodies associated to various variants of the SARS-CoV-2 virus (61). For each subject, we compared the fraction of full-STAR hits within the full repertoire to the fraction of full-STAR hits among sequences from the repertoire that were also found in the COVID19 database. We found a significant enrichment of hits in the part of the repertoire that overlapped with the database in 11 out of the 18 subjects that we analyzed (Fig. 5). Dataset S1 shows the number of sequences from the COVID-19 database (40) found in each patient and how many of them were isolated by the full-STAR method. This application demonstrates the versatility of our pipeline in identifying responsive BCR sequences, including in the context of a natural infection.

2. Discussion

The main interest of STAR is its ability to detect responding clonotypes from a single repertoire snapshot, without the need of longitudinal data. This method outperforms the naive approach of selecting clones with the highest frequencies as the likely responders. We validated our results using a variety of tests, including binding assays and an analysis of overlap with previously published influenza datasets. We could not validate all of our predicted hits for influenza responding BCR since the data we used (48) only had BCR heavy chain sequences. Without paired light and heavy chain, we cannot directly test for binding. We nevertheless verified that paired BCR sequences validated in the original study were among our predicted hits. We also demonstrated the effectiveness of our method by applying it to COVID-19 data. Although we did not conduct experimental tests on isolated sequences, we found that some of the BCRs we identified as responding using our method were present in the COVID-19 database (61) and were identified as responders to COVID-19 epitopes (Fig. 5). In particular, in 11 out of 18 patients, the overlap with the database is statistically significant. The method can easily be applied to paired chain data which will allow for full experimental validation.

Our approach is inspired by the ALICE method developed for T cells (55). Both methods share the idea of examining amino acid sequence neighbors as a measure of similarity, and, in the case of full-STAR, to compare the number of neighbors to an expectation computed using a generative model of receptor sequences. The key difference arises from the nature of T cell responses, which do not form lineages. In T cells the emergence of clusters of neighboring sequences can only occur through the convergent selection for a common function of distinct lineages, while for B cells this effect is confounded by hypermutations, which create neighbors in sequence space belonging to the same lineage. Distinguishing the two effects is in general difficult, in particular, because it is often hard to separate distinct lineages that have the same or similar CDR3 (15, 17, 62–64).

The STAR pipelines can in principle be extended to light-chain or paired-chain repertoire data. However, the low diversity of the light chain, both in terms of sequence length and variability, makes it less informative than the heavy chain. High-throughput paired-chain data (with at least 100,000 unique sequences) would ideally be the most informative; however, most single-cell sequencing data are restricted to relatively small dataset (a few thousands), where the neighbors are only sparsely sampled, making the method inapplicable. When both massive bulk single-chain repertoire (>100,000) and a smaller number (1,000 to 10,000) of single-cell data are available, hits identified by our method from the bulk data can be matched with paired chains from the single-cell data to infer the full antibody sequence, which can be subsequently tested for binding or neutralization.

Our findings highlight the importance of focusing on days close to the peak of infection, particularly day 7 in the case of influenza, for robust identification of responsive BCR

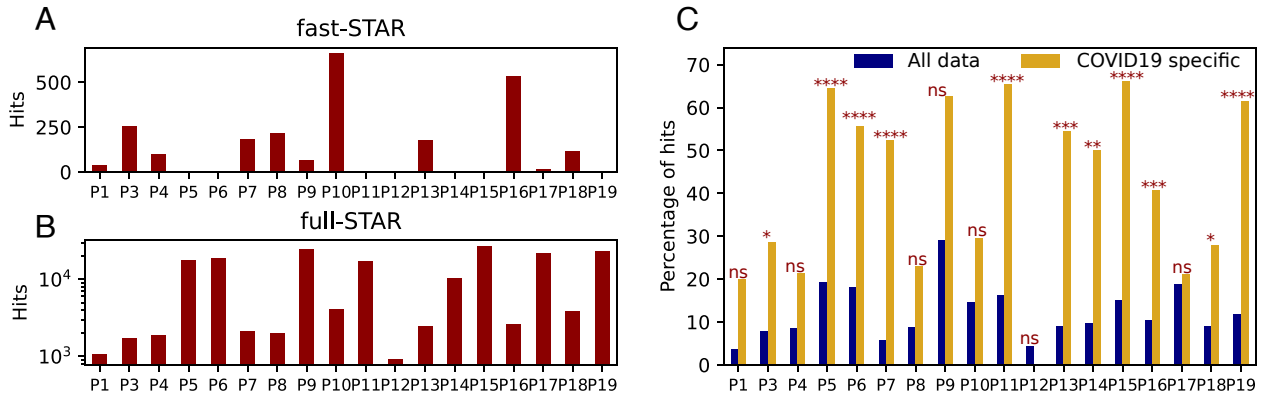

Fig. 5. COVID19 specific sequences. (*A*) Number of hits obtained with fast-STAR per patient. (*B*) Number of hits obtained with full-STAR per patient. (*C*) Percentage of hits obtained with full-STAR in the entire dataset for each patient versus the percentage of hits obtained with full-STAR in the subsample of the dataset with SARS-CoV-2 specific sequences taken from the COVID19 antibody database, with significance. The nonsignificant patients are labeled with ns ($P > 0.05$), one star corresponds to a P value $P \leq 0.05$, two stars $P \leq 0.01$, three stars $P \leq 0.001$ and four stars $P \leq 0.0001$.

sequences. The algorithm's efficacy decays rapidly after the peak, underscoring the sensitivity of our method to the precise day when the sample is taken. Future investigations could explore the decay dynamics following the peak, and its implications for long-term immune responses.

While our pipeline exhibits promising results, certain limitations should be acknowledged. Fast-STAR is very specific, but likely misses a large fraction of responding clones. Full-STAR is more comprehensive, but may contain a substantial number of false positives, even if those contribute moderately to the cumulative frequency. The sensitivity of both methods strongly depends on sequencing depth, which must be sufficient to sample enough neighbors of responding sequences. Further experimental tests should be applied to the candidate sequences proposed by STAR to properly assess their function. The algorithm relies on a computational null model that assumes independence between B cells and ignores their possible lineage relationship. Because memory B cells have undergone hypermutations, we expect them to violate this assumption even in the absence of an immune challenge. In practice, this means that the null model should systematically underestimate the number of neighbors. Luckily, this effect is compensated by another inaccuracy. The approximate expected number of neighbors we use in full-STAR is actually an overestimation compared to the exact computation, because many amino acid neighbors are not proper antibody sequences and thus have extremely low probabilities, which is not accounted for in the approximation. As a result of these two errors canceling each other, the approximate estimate is in fact more accurate than the exact one. A multiplicative factor may be manually added to the exact estimate to correct for the error (*SI Appendix*, Fig. S3). However, future progress should rely on refining the null model to incorporate the generation of correlated lineages and thus provide a more accurate model of the human B cell repertoire before immunization. A more detailed treatment of the expected dynamics of B cell clones in absence of stimulation could also help improve the null model predictions.

We developed and validated our method in the context of immunization, and showed that it could also be applied to the case of a natural infection. Future research could test its applicability to other immune contexts, such autoimmune diseases (65), allergies (66), or chronic infections such as HIV (67). Compared to an acute challenge, the signal may be too weak to detect a response, unless repertoires are collected during a burst. Chronic disorders are expected to produce a higher number of highly mutated lineages (68), which may be more easily detected by the method.

The notion of convergent selection demonstrated here is related to that of public repertoires, defined as the set of receptors shared between individuals afflicted with the same condition (40–42). While many receptors are shared by chance due to their high generation probability, even in healthy individuals (42, 69), individuals with a common condition tend to share more receptors, owing to the shared selective pressures that they are subjected to. This can lead to the sharing of the same CDR3 with different V genes across individuals infected with COVID (42). Here, we show that convergence can happen even in the same individual. Because of this shared convergent selection, one could expect the clusters of reactive BCRs identified by STAR to overlap between patients. However, we found no such sharing among COVID19 or influenza-vaccine patients. This implies that, while there exists a public repertoire responding to the same disease or vaccine, the primary immune response remains predominantly private. It is however unclear what individual biological factors could explain this observation. We were careful to check that the data were free from alignment errors (*SI Appendix*, Fig. S6) and hybridization artifacts during library preparation (*Methods*), both of which could have confounded our result. However, we cannot rule out other artifacts in the data, and further investigation is needed to fully confirm that conclusion. Understanding how to combine information from both private and public contributions to the response could help design better predictors of immune status.

3. Methods

We develop our method using data from mRNA-based heavy chain sequencing of the BCRs of 5 healthy human subjects vaccinated against influenza (48). The sequences were tagged with unique molecular identifiers (UMI) to correct for the PCR amplification bias. Preprocessed data were aligned to V, D, and J templates with MiXCR (60), and then filtered to remove singletons. Each repertoire contained on average 3.2×10^5 (range 15,035 to 814,033) unique UMI sequences, 6.0×10^4 (range 3,767 to 121,608) unique nucleotide sequences, and 5.2×10^4 (range 3,367 to 101,254) unique amino acid CDR3 sequences, refer to *SI Appendix*, Table S1 for more details.

Since our analysis reports convergence of the CDR3 with different V genes, we checked that this effect was not due to hybridization during PCR amplification. The UMI is composed of a 8-nt UMI on the 3′ end (UMI3), and another 8-nt UMI on the 5′ end (UMI5). Since the CDR3 and J genes are on the 3′ side of the read, in the case of hybridization we expect that the CDR3 sequence to be strongly linked to UMI3. Hybridization during PCR would result in the same CDR3 nucleotide sequence and UMI3 to be

associated with many distinct UMI5. A given CDR3 nucleotide sequence is often represented by multiple RNA molecules and thus distinct UMI3s and UMI5s. However, hybridization during PCR amplification would create an even larger diversity of UMI5 for each UMI3, resulting in an overall larger UMI5 diversity. We checked the UMI diversity of the most abundant CDR3 nt sequence, TGCAAGTCTCTGTTGACTACTATTCCGGAAAAGTGGTTCGACCCCTGG. We found 7,835 distinct UMI3 and 7,111 distinct UMI5. We thus conclude that the diversity of UMI5 is consistent with that of UMI3, and thus that hybridization is negligible.

In each repertoire, for each amino acid heavy-chain CDR3 sequence we counted the number of its neighbors using ATrieGC software (15), which uses indexing trees to efficiently span neighbors. Each neighbor is weighted by its multiplicity in terms of unique nucleotide sequences that have the corresponding amino acid CDR3.

The expected number of neighbors for each sequence s is computed using the software OLGA (56), which gives the probability of occurrence of an amino acid sequence (called Pgen), i.e. the sum of probabilities of all nucleotide sequences translating into that amino acid sequence. We sum that probability (Pgen) over all amino acid sequences s' that belong to the ensemble of neighbors of the query sequence s to obtain the probability that a random nucleotide sequence translates into a neighbor s' of s. We multiply this cumulative probability by the total number N of unique nucleotide sequences in the dataset, which gives the expected number λ of nucleotide sequences translating into neighbors of s, in a dataset of size N. Mathematically:

$$\lambda(s) = N \sum_{s' \in V(s)} P_{\text{gen}}(s'), \qquad [1]$$

where $\lambda(s)$ is the expected number of neighbors of the CDR3 amino acid sequence s, $V(s)$ is the set of neighbors of s (one amino acid difference), N the total number of unique nucleotide sequences of the repertoire, and $P_{\text{gen}}(s')$ the generation probability of a CDR3 amino acid sequence s' as given by the OLGA model. Because of the high number of neighbors for each sequence, applying the formula above directly is computationally too expensive. It may be estimated using a synthetic dataset generated by OLGA, with a Monte-Carlo sample of 10^8 sequences. We can then count the number of neighbors of each s in this synthetic dataset, normalized by 10^8, and multiplied by N, the size of the real dataset. In practice, we used an approximation where we assume $P_{\text{gen}}(s) \approx P_{\text{gen}}(s')$, as justified by *SI Appendix*, Fig. S1, which yields $\lambda(s) \approx \hat{\lambda}(s) = 19L(s)NP_{\text{gen}}(s)$, where $L(s)$ is the CDR3 length of s. The results of the Monte-Carlo estimate of the exact formula, and of the approximate formula that we used in the paper, are compared with the prevaccination data in *SI Appendix*, Fig. S3.

The full-STAR pipeline compares the expected number of neighbors $\lambda(s)$ with the observed one, denoted by $n(s)$. The P-value is computed as the probability to find at least as many neighbors as observed, where that number is assumed in the null model to follow a Poisson distribution of mean $\hat{\lambda}(s)$, $p = \sum_{n=n(s)}^{\infty} \frac{e^{-\hat{\lambda}(s)}}{n!} \hat{\lambda}(s)^n$. From the operational point of view, the full-STAR method is identical to the ALICE software introduced for TCR (55), with the difference that ALICE uses a Monte-Carlo estimate of $\lambda(s)$ instead of the $\hat{\lambda}(s)$ approximation. Another difference with ALICE is the motivation for why responding clonotypes should have more neighbors than expected (*Discussion*). The fast-STAR method differs from both ALICE and full-STAR in that it does not compare the number of neighbors to a sequence-specific expected value, but to a repertoire-wide baseline.

Poisson tests performed on the overlap with independent influenza (58) and COVID19 (61) datasets were performed in the following way. The expected overlap of our list of STAR hits with a control (non-disease-specific) dataset was computed, and turned into a coincidence probability $x_{\text{null}} = n_{\text{null}}/N_{\text{null}}$, where N_{null} the size of the control dataset. Then the P-value on the observed overlap of the list with the disease-specific dataset, n_{specific}, is then computed as $p = \sum_{n=n_{\text{specific}}}^{\infty} \frac{e^{-x_{\text{null}} N_{\text{specific}}}}{n!} (x_{\text{null}} N_{\text{specific}})^n$, where N_{specific} is the size of the disease-specific dataset.

The lineages are inferred using HILARy software (15), and the corresponding trees are reconstructed with RAxML (59) and represented with iTOL (70).

Data, Materials, and Software Availability. The trivalent vaccine influenza bulk data (48) and the single-cell data (57) are available on the European Nucleotide Archive with accession number, Sequence Read Archive: PRJNA512111, BioProject–PRJNA512111 (71). The vaccine influenza single-cell RNA sequencing and V(D)J data of ref. 58 have been deposited in NCBI's Gene Expression Omnibus and are available at the GEO Series accession number GSE175524 (72). The raw COVID-19 BCRs sequence data (40) are available on the European Nucleotide Archive under BioProject Accession PRJNA638224 (73). The sequences of the antibodies tested for being COVID-19 specific are available at https://opig.stats.ox.ac.uk/webapps/covabdab/ (74). The code to reproduce fast and full STAR is freely available at https://github.com/statbiophys/STAR (75).

ACKNOWLEDGMENTS. This work was supported by Sanofi, the European Research Council consolidator grant no 724208 (A.M.W., T.M., and M.F.A.), and the Agence Nationale de la Recherche grant no ANR-19-CE45-0018 "RESP-REP" (A.M.W., T.M., and M.F.A.). E.V. and M.A.S. are Sanofi employees and may hold shares and/or stock options in the company. We declare that this study received funding from Sanofi. The funder collaborated directly in the study and was involved in the study design, analysis, and interpretation of data, the writing of this article, and the decision to submit it for publication. We are grateful for the discussions and suggestions from Natanael Spisak, Emanuele Loffredo, Antoine Aragon, and Maria Ruiz Ortega.

1. K. Fink, Can we improve vaccine efficacy by targeting T and B cell repertoire convergence? *Front. Immunol.* **10**, 110 (2019).
2. C. Wang *et al.*, B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 500–505 (2015).
3. J. D. Galson *et al.*, B-cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B-cell activation. *Genome Med.* **8**, 1–13 (2016).
4. K. Wang *et al.*, Memory B cell repertoire from triple vaccinees against diverse SARS-CoV-2 variants. *Nature* **603**, 919–925 (2022).
5. C. Kreer, H. Gruell, T. Mora, A. M. Walczak, F. Klein, Exploiting B cell receptor analyses to inform on HIV-1 vaccination strategies. *Vaccines* **8**, 13 (2020).
6. C. Kreer *et al.*, Longitudinal isolation of potent near-germline SARS-CoV-2-neutralizing antibodies from COVID-19 patients. *Cell* **182**, 843–854 (2020).
7. K. B. Hoehn *et al.*, Cutting edge: Distinct B cell repertoires characterize patients with mild and severe COVID-19. *J. Immunol.* **206**, 2785–2790 (2021).
8. M. Wu *et al.*, Systemic lupus erythematosus patients contain B-cell receptor repertoires sensitive to immunosuppressive drugs. *Euro. J. Immunol.* **52**, 669–680 (2022).
9. M. Ota *et al.*, Multimodal repertoire analysis unveils B cell biology in immune-mediated diseases. *Ann. Rheum. Dis.* **82**, 1455–1463 (2023).
10. C. Kreer *et al.*, Probabilities of developing HIV-1 bNAb sequence features in uninfected and chronically infected individuals. *Nat. Commun.* **14**, 7137 (2023).
11. L. Gieselmann *et al.*, Effective high-throughput isolation of fully human antibodies targeting infectious pathogens. *Nat. Protoc.* **16**, 3639–3671 (2021).
12. A. Agrafiotis *et al.*, Generation of a single-cell B cell atlas of antibody repertoires and transcriptomes to identify signatures associated with antigen specificity. *iScience* **26**, 106055 (2023).
13. A. Pedrioli, A. Oxenius, Single B cell technologies for monoclonal antibody discovery. *Trend. Immunol.* **42**, 1143–1158 (2021).
14. W. H. Robinson, Sequencing the functional antibody repertoire-diagnostic and therapeutic discovery. *Nat. Rev. Rheum.* **11**, 171–182 (2015).
15. N. Spisak, T. Dupic, T. Mora, A. M. Walczak, Combining mutation and recombination statistics to infer clonal families in antibody repertoires. arXiv [Preprint] (2022). http://arxiv.org/abs/2212.11997 (Accessed 22 December 2022).
16. D. K. Ralph, F. A. Matsen IV, Likelihood-based inference of B cell clonal families. *PLoS Comput. Biol.* **12**, e1005086 (2016).
17. N. Nouri, S. H. Kleinstein, Somatic hypermutation analysis for improved identification of B cell clonal families from next-generation sequencing data. *PLoS Comput. Biol.* **16**, e1007977 (2020).
18. L. Mesin, J. Ersching, G. D. Victora, Germinal center B cell dynamics. *Immunity* **45**, 471–482 (2016).
19. J. A. Weinstein, N. Jiang, R. A. White III, D. S. Fisher, S. R. Quake, High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807–810 (2009).
20. P. D. Baum, V. Venturi, D. A. Price, Wrestling with the repertoire: The promise and perils of next generation sequencing for antigen receptors. *Euro. J. Immunol.* **42**, 2834–2839 (2012).
21. H. Robins, Immunosequencing: Applications of immune repertoire deep sequencing. *Curr. Opin. Immunol.* **25**, 646–652 (2013).
22. C. Vollmers, R. V. Sit, J. A. Weinstein, C. L. Dekker, S. R. Quake, Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 13463–13468 (2013).
23. G. Georgiou *et al.*, The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32**, 158–168 (2014).

24. B. Briney, A. Inderbitzin, C. Joyce, D. R. Burton, Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393–397 (2019).
25. C. Kreer *et al*., Probabilities of HIV-1 bNAb development in healthy and chronically infected individuals. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.07.11.499584 (Accessed 28 December 2022).
26. C. Kreer, H. Gruell, T. Mora, A. M. Walczak, F. Klein, Exploiting B cell receptor analyses to inform on HIV-1 vaccination strategies. *Vaccines* **8**, 13 (2020).
27. J. S. Babcook, K. B. Leslie, O. A. Olsen, R. A. Salmon, J. W. Schrader, A novel strategy for generating monoclonal antibodies from single, isolated lymphocytes producing antibodies of defined specificities. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 7843–7848 (1996).
28. F. A. Harding, M. M. Stickler, J. Razo, R. DuBridge, The immunogenicity of humanized and fully human antibodies: Residual immunogenicity resides in the CDR regions. *MAbs* **2**, 256–265 (2010).
29. M. Vakhitova *et al*., A rapid method for detection of antigen-specific B cells. *Cells* **12**, 774 (2023).
30. J. Trück *et al*., Identification of antigen-specific B cell receptor sequences using public repertoire analysis. *J. Immunol.* **194**, 252–261 (2015).
31. V. Greiff, G. Yaari, L. G. Cowell, Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Curr. Opin. Syst. Biol.* **24**, 109–119 (2020).
32. R. Akbar *et al*., A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Rep.* **34**, 108856 (2021).
33. V. Kunik, S. Ashkenazi, Y. Ofran, Paratome: An online tool for systematic identification of antigen-binding regions in antibodies based on sequence or structure. *Nucleic Acids Res.* **40**, W521–W524 (2012).
34. M. C. Jespersen, S. Mahajan, B. Peters, M. Nielsen, P. Marcatili, Antibody specific B-cell epitope predictions: Leveraging information from antibody-antigen protein complexes. *Front. Immunol.* **10**, 298 (2019).
35. E. Liberis, P. Veličković, P. Sormanni, M. Vendruscolo, P. Liò, Parapred: Antibody paratope prediction using convolutional and recurrent neural networks. *Bioinformatics* **34**, 2944–2950 (2018).
36. D. M. Mason *et al*., Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* **5**, 600–612 (2021).
37. E. Miho, R. Roškar, V. Greiff, S. T. Reddy, Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nat. Commun.* **10**, 1321 (2019).
38. K. B. Hoehn *et al*., Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 22664–22672 (2019).
39. Y. H. Chang *et al*., Network signatures of IGG immune repertoires in hepatitis B associated chronic infection and vaccination responses. *Sci. Rep.* **6**, 26556 (2016).
40. J. D. Galson *et al*., Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. *Front. Immunol.* **11**, 605170 (2020).
41. Z. Montague *et al*., Dynamics of B cell repertoires and emergence of cross-reactive responses in patients with different severities of COVID-19. *Cell Rep.* **35**, 109173 (2021).
42. M. Ruiz Ortega, N. Spisak, T. Mora, A. M. Walczak, Modeling and predicting the overlap of B-and T-cell receptor repertoires in healthy and SARS-CoV-2 infected individuals. *PLoS Genet.* **19**, e1010652 (2023).
43. W. K. Wong *et al*., Ab-ligity: Identifying sequence-dissimilar antibodies that bind to the same epitope. *MAbs* **13**, 1873478 (2021).
44. A. Kovaltsuk *et al*., How B-cell receptor repertoire sequencing can be enriched with structural antibody data. *Front. Immunol.* **8**, 1753 (2017).
45. K. Krawczyk *et al*., Structurally mapping antibody repertoires. *Front. Immunol.* **9**, 1698 (2018).
46. M. I. Raybould, W. K. Wong, C. M. Deane, Antibody-antigen complex modelling in the era of immunoglobulin repertoire sequencing. *Mol. Syst. Des. Eng.* **4**, 679–688 (2019).
47. S. A. Robinson *et al*., Epitope profiling using computational structural modelling demonstrated on coronavirus-binding antibodies. *PLoS Comput. Biol.* **17**, e1009675 (2021).
48. F. Horns, C. Vollmers, C. L. Dekker, S. R. Quake, Signatures of selection in the human antibody repertoire: Selective sweeps, competing subclones, and neutral drift. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 1261–1266 (2019).
49. U. Laserson *et al*., High-resolution antibody dynamics of vaccine-induced immune responses. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 4928–4933 (2014).
50. M. V. Pogorelyy *et al*., Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 12704–12709 (2018).
51. A. A. Minervina *et al*., Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection. *eLife* **10**, e63502 (2021).
52. M. Puelma Touzel, A. M. Walczak, T. Mora, Inferring the immune response from repertoire sequencing. *PLoS Comput. Biol.* **16**, e1007873 (2020).
53. A. A. Minervina *et al*., Primary and secondary anti-viral response captured by the dynamics and phenotype of individual T cell clones. *eLife* **9**, e53704 (2020).
54. M. V. Pogorelyy *et al*., Resolving SARS-CoV-2 CD4$^+$ T cell specificity via reverse epitope discovery. *Cell Rep. Med.* **3**, 100697 (2022).
55. M. V. Pogorelyy *et al*., Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLoS Biol.* **17**, 1–13 (2019).
56. Z. Sethna, Y. Elhanati, C. G. Callan Jr., A. M. Walczak, T. Mora, Olga: Fast computation of generation probabilities of B-and T-cell receptor amino acid sequences and motifs. *Bioinformatics* **35**, 2974–2981 (2019).
57. F. Horns, C. L. Dekker, S. R. Quake, Memory B cell activation, broad anti-influenza antibodies, and bystander activation revealed by single-cell transcriptomics. *Cell Rep.* **30**, 905–913.e6 (2020).
58. M. Wang *et al*., High-throughput single-cell profiling of B cell responses following inactivated influenza vaccination in young and older adults. *Aging* **15**, 9250 (2023).
59. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
60. D. A. Bolotin *et al*., MiXCR: Software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).
61. M. I. Raybould, A. Kovaltsuk, C. Marks, C. M. Deane, CoV-AbDab: The coronavirus antibody database. *Bioinformatics* **37**, 734–735 (2021).
62. D. K. Ralph, F. A. Matsen IV, Inference of B cell clonal families using heavy/light chain pairing information. *PLoS Comput. Biol.* **18**, e1010723 (2022).
63. K. B. Hoehn, O. G. Pybus, S. H. Kleinstein, Phylogenetic analysis of migration, differentiation, and class switching in B cells. *PLoS Comput. Biol.* **18**, e1009885 (2022).
64. N. Nouri, S. H. Kleinstein, A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data. *Bioinformatics* **34**, i341–i349 (2018).
65. R. Jiang *et al*., Thymus-derived B cell clones persist in the circulation after thymectomy in myasthenia gravis. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30649–30660 (2020).
66. M. Levin *et al*., Persistence and evolution of allergen-specific IGE repertoires during subcutaneous specific immunotherapy. *J. Allergy Clin. Immunol.* **137**, 1535–1544 (2016).
67. E. L. Johnson *et al*., Sequencing HIV-neutralizing antibody exons and introns reveals detailed aspects of lineage maturation. *Nat. Commun.* **9**, 4136 (2018).
68. K. B. Hoehn *et al*., Human B cell lineages associated with germinal centers following influenza vaccination are measurably evolving. *eLife* **10**, e70873 (2021).
69. Y. Elhanati, Z. Sethna, C. G. Callan Jr., T. Mora, A. M. Walczak, Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunol. Rev.* **284**, 167–179 (2018).
70. I. Letunic, P. Bork, Interactive tree of life (iTOL) v6: Recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).
71. F. Horns, C. Vollmers, C. L. Dekker, S. R. Quake, Antibody repertoire sequencing after influenza vaccination. BioProject. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA512111. Deposited 28 December 2018.
72. M. Wang *et al*., High-throughput single-cell profiling of B cell responses following inactivated influenza vaccination in young and older adults. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE175524. Deposited 20 August 2023.
73. J. D. Galson *et al*., BCR repertoire sequencing from COVID-19 patients. BioProject. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA638224. Deposited 9 June 2020.
74. M. I. J. Raybould, A. Kovaltsuk, C. Marks, C. M. Deane, Coronavirus-binding antibody sequences & structures. Cov-AbDab. https://opig.stats.ox.ac.uk/webapps/covabdab/. Deposited 17 August 2020.
75. M. F. Abbate, STAR. Github. https://github.com/statbiophys/STAR. Deposited 18 December 2023.