MABS
2021, VOL. 13, NO. 1, e1869406 (14 pages)
https://doi.org/10.1080/19420862.2020.1869406

Taylor & Francis
Taylor & Francis Group

REPORT

OPEN ACCESS Check for updates

A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies

Eve Richardson[a], Jacob D. Galson[b,c], Paul Kellam[d,e], Dominic F. Kelly[f,g], Sarah E. Smith[d], Anne Palser[d], Simon Watson[d], and Charlotte M. Deane[a]

[a]Department of Statistics, University of Oxford, Oxford, UK; [b]Alchemab Therapeutics Ltd, London, UK; [c]Division of Immunology, University Children's Hospital, University of Zurich, Zurich, Switzerland; [d]Kymab Ltd, Cambridge, UK; [e]Department of Infectious Diseases, Faculty of Medicine, Imperial College London, London, UK; [f]Department of Paediatrics, University of Oxford, Oxford, UK; [g]Oxford University Hospitals NHS Foundation Trust, Oxford, UK

ABSTRACT
Due to their shared genetic history, antibodies from the same clonotype often bind to the same epitope. This knowledge is used in immune repertoire mining, where known binders are used to search bulk sequencing repertoires to identify new binders. However, current computational methods cannot identify epitope convergence between antibodies from different clonotypes, limiting the sequence diversity of antigen-specific antibodies that can be identified. We describe how the antibody binding site, the paratope, can be used to cluster antibodies with common antigen reactivity from different clonotypes. Our method, paratyping, uses the predicted paratope to identify these novel cross clonotype matches. We experimentally validated our predictions on a pertussis toxoid dataset. Our results show that even the simplest abstraction of the antibody binding site, using only the length of the loops involved and predicted binding residues, is sufficient to group antigen-specific antibodies and provide additional information to conventional clonotype analysis.

Abbreviations: BCR: B-cell receptor; CDR: complementarity-determining region; PTx: pertussis toxoid

ARTICLE HISTORY
Received 18 September 2020
Revised 22 December 2020
Accepted 22 December 2020

KEYWORDS
Antibody discovery; paratope; pertussis; pertussis toxoid; computational; immune repertoire mining; transgenic mouse; BCR-seq; paired sequencing

Introduction

Next-generation immune repertoire sequencing (BCR-seq or rep-seq[1,2]) can provide comprehensive information about adaptive immune repertoires across individuals[3] and immune states.[4] Progress has been made in the task of interrogating the vast diversity of B-cell receptor (BCR) repertoires, primarily through the analysis of predicted clonal relationships inferred via clonotyping.[5] BCR-seq and associated clonal analysis are finding increasing importance in antibody discovery both as a method of identification of putative antigen-specific antibodies[6–8] and more recently as a method of lead antibody optimization through repertoire mining.[9] The identification of antibodies that are predicted to bind to the same site (epitope) is now a key component of BCR repertoire analysis and antibody discovery.

The starting point for most BCR repertoire analysis is the reduction of thousands or millions of BCRs into orders of magnitude fewer clonotypes.[5] Clonotype definitions vary, primarily through treatment of the complementarity-determining region (CDR) H3, but are intended to capture groups of clonally related antibody sequences derived from common progenitor B cells.[10] Published clonotyping methods use heavy chain information only, which is considered sufficient to capture most clonal relationships.[11] During B cell development, the variable (V), diversity (D) and joining (J) gene segments encoding the variable domain of the antibody heavy chain undergo recombination.[12] A requirement for two sequences to be predicted to share the same clonotype is therefore common V- and J-germline gene assignment.[5] The D gene is not usually included in standard clonotype definitions because its assignment is both difficult and redundant in the clonotype definition, as it is wholly contained within the CDRH3.[13,14] The variable domain of the antibody heavy chain consists of the framework regions and hypervariable CDRs. CDRHs 1 and 2 are encoded by the V gene while the region spanning the recombined V, D and J segments corresponds to the third and most diverse loop on the antibody heavy chain, the CDRH3. The processes of junctional diversification (the insertion of palindromic and random nucleotides at the junction between the V, D and J genes) during recombination act in tandem with somatic hypermutation[15] during affinity maturation to further increase the diversity of the CDRH3. Sequence identity in the CDRH3 is therefore included as a marker of shared origin in most clonotyping tools.[5] The nucleotide or amino acid sequence identity across the CDRH3 required for two sequences to be considered in the same clonotype varies across studies – in studies performing clonotyping with length-normalized amino acid sequence identity thresholds, sequence identity thresholds vary between 80% and 100%.[5]

CONTACT Charlotte M. Deane Email deane@stats.ox.ac.uk Department of Statistics, University of Oxford, 29 St Giles', OX1 Oxford, UK.
Supplemental data for this article can be accessed on the publisher's website.

After recombination, the heavy chain is expressed as a pre-BCR with a surrogate light chain. The light chain is subsequently formed from the recombination of the V and J genes of either of the two light chain loci (lambda or kappa)[15] and is expressed by the immature B cell. While the light chain provides clonal signal, clonotyping has no established precedent using both heavy and light chains. Clonal inference for paired VH/VL sequences from single-cell sequencing has largely used heavy chains only;[16,17] clonal inference within the BraCeR tool defines heavy and light chain clones separately.[18] We therefore refer to "clonotyping" as describing clonotyping using the heavy chain only.

A number of publicly available, well-supported pipelines have made clonotype analysis standard practice.[5,10] This has permitted large advances in the practical utility of BCR-seq data.[19,20] Clinically, it has found use in tracking minimal residual disease in blood cancers,[21] monitoring vaccination responses[22–24] and providing mechanistic insights into immune-mediated diseases.[4,25–27] Clonotyping has also proven useful in antibody discovery as a means of selecting candidate sequences for expression as monoclonal antibodies[6–8] and recently as a method of lead antibody optimization via repertoire mining.[9]

Antibodies within the same clonotype are likely to target a common epitope.[5,10] The majority of antibodies binding to the same epitope in antibody-antigen complex structures in the Structural Antibody Database (SAbDab, a database of experimentally solved antibody and antibody-antigen complex structures) have highly similar CDRH3s.[28] However, it has also been observed that multiple clonotypes may converge on the same epitope. For example, Scheid and colleagues identified clonotypes from distinct immunoglobulin heavy chain variable (IGHV) gene subgroups converging on the CD4 binding site in gp120.[29] Separate clonotypes have also been observed to bind to overlapping epitopes on the hemagglutinin stem[30] or globular head, and on multiple epitopes on the Ebola virus glycoprotein.[31] Wong and colleagues identified 190 pairs of antibodies with sub-80% CDRH3 amino acid identity binding to the same epitope within SAbDab.[28,32] This convergence between clonotypes offers the potential to improve our understanding of the functional landscape of BCR repertoires; large-scale functional convergence between lineages could, for example, explain the apparent scarcity of public clonotypes.[33] This hypothesis is supported by evidence that, while clonotypes are infrequently shared between individuals, the range of antibody structures that these clonotypes generate is more similar between individuals.[34] In the context of antibody discovery, being able to identify binders to the same epitope from different clonotypes would aid in optimization of developability or binding affinity, by allowing hopping between germline scaffolds.

In antibody discovery, clonotyping is used to search for clonal relatives of lead antibodies in bulk BCR-seq data sets in order to identify antibodies that target the same epitope, but which have either an increased affinity or a superior developability profile. This process is referred to as "immune repertoire mining". Hsiao and colleagues performed clonotyping on a set of bulk heavy-chain repertoires and used the resultant clonotypes for hit expansion against two targets.[9] They achieved greater than an order of magnitude improvement in affinity for both targets and between 48% and 100% of tested heavy chain variants retained target-binding.[9] This suggests that sampling within a clonotype can be highly effective as a means of repertoire mining. However, the method does not allow the identification of binders to the same epitope that derive from different clonotypes, which currently limits the sequence distinctness of novel binders that can be recovered from immune repertoires.

Here, we describe a new method to identify functional convergence of antibody sequences that is germline-independent and that considers only the binding site of the antibody sequences, the paratope. We call this approach "paratyping". We show how paratyping allows grouping of antigen-specific sequences from different clonotypes, and is a rapid, structurally intuitive way of grouping functionally related antibodies. Paratyping simplifies the complex phenomenon of antibody–antigen interaction into sets of shared residues. Learning the complexities of antibody–antigen interactions as part of a predictive model of antigen binding has been achieved in the case of the antigen interaction of one therapeutic monoclonal antibody.[35] However, such approaches rely on a large (on the order of 10^4) library of experimentally validated binding and non-binding variants. Paratyping removes the need for large amounts of training data and is generalizable across protein antigens.

We first show the rationale for our paratyping method using the structures of a pair of antibodies from different clonotypes that bind to the same epitope. Paratyping is then applied to a single-cell data set of sequences raised against pertussis toxoid (PTx) in a transgenic mouse platform where it is as accurate as clonotyping but identifies different binders. We then perform a prospective experimental test of the method by expressing as monoclonal antibodies and experimentally testing predicted PTx-binding and non-binding antibodies mined from a set of non-enriched bulk heavy chain sequencing repertoires. Our experimental test demonstrates that paratyping identifies PTx-binding antibodies from different clonotypes to our original hits. This expands the sequence space available through repertoire mining and permits favorable shifts in *in silico* developability metrics. Of particular advantage is paratyping's ability to predict common antigen reactivity of antibodies from different V and J gene backgrounds, which has implications for large-scale repertoire analysis.

Results

Epitope convergence can be identified at the level of paratope residues

Antibodies from different clonotypes have been observed to converge on the same binding site.[28–31] We hypothesize that these functionally convergent antibodies may use the same paratope for interaction, and examined SAbDab[32] for evidence that antibodies with similar paratopes bind to the same epitope. We defined the epitope and paratope as those residues with any atom within 4.5 Å of any residue in the cognate antibody or antigen, respectively, and consider only the heavy chains of the antibodies because in most BCR sequencing experiments only heavy chain sequences are available.[36]

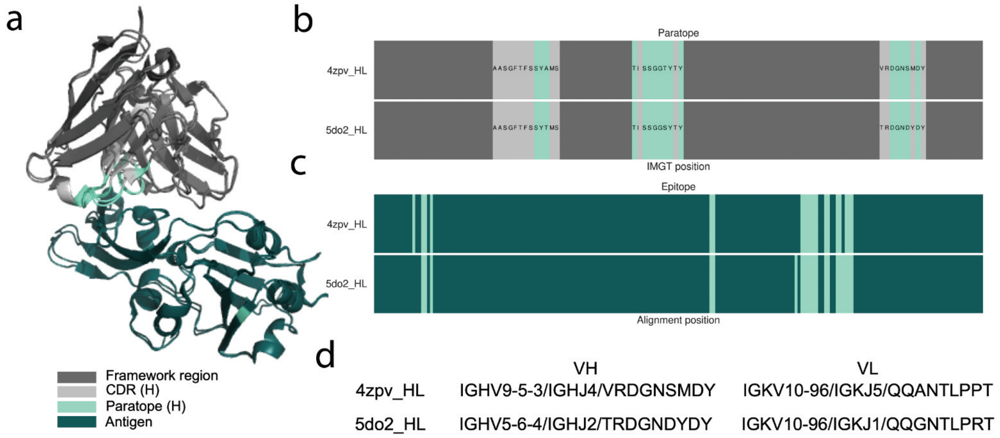

Figure 1. An example of two antibodies that bind to the same epitope but derive from different clonotypes. The murine 4C2 (PDB ID: 5do2) and D12 (PDB ID: 4zpv) anti-MERS-CoV antibodies target the same residues on the receptor-binding domain of the spike protein, which can be seen from alignment of their antibody/antigen complex structures (a). 4C2 and D12 use over 80% of the same heavy chain paratope residues (b) and target 95% of the same epitope residues (c). The antibodies are derived from different IGHV and IGHJ genes and display CDRH3 amino acid identity (66.7%) below standard clonotyping definitions (80–100%)[5] (d), so would not be considered within the same clonotype.

Figure 1 shows an example of a pair of antibody-antigen complex structures where antibodies binding to the same epitope derive from different clonotypes based on their heavy chains, but have very similar heavy chain paratopes. The two monoclonal antibodies 4C2 (PDB ID: 5do2) and D12 (PDB ID: 4zpv) bind to the receptor-binding domain of the MERS-CoV spike protein at the same epitope (94.7% of epitope residues are shared across the pair of structures). In the standard clonotyping definition, an exact VH and JH match is required for grouping of two sequences regardless of treatment of the CDRH3;[5,10] in this instance, the antibodies differ in both IGHV (IGHV5-6-4 vs IGHV5-9-1) and IGHJ gene (IGHJ2/IGHJ4) and have lower CDRH3 identity (66.7%) than common definitions. However, 4C2 and D12 use largely the same paratope residues to achieve this epitope convergence, with 81.3% of heavy chain paratope residues conserved across the structures. Another example of epitope convergence between antibodies with differing IGHJ genes and sub-80% CDRH3 identity can be seen in the anti-lysozyme antibodies HyHel-10 and HyHel-63 (see Supplementary Figure 1; the heavy and light chain CDRs and V and J genes of both examples can be found in Supplementary Table 1).

In this study, our aim was to discover other antibodies from different clonotypes with common antigen reactivity using sequence data alone. In the case of sequence data without accompanying antibody-antigen complex structures, the paratope is not known but can be predicted with high accuracy.[37] We used the state-of-the-art in paratope prediction, Parapred, to annotate sequences with their likely paratope; this predicted paratope is the input to our novel method, "paratyping", as validated in the "Results" section.

Paratyping and clonotyping successfully cluster PTx binders in a single-cell dataset

We defined clonotypes as antibodies with the same heavy chain V and J genes, CDRH3s (North definition[38]) of the same length, and above a threshold level of amino acid identity across the CDRH3.[5] In our novel method, paratyping, antibodies with the same length CDRs (North definition) and above a threshold level of sequence identity across the predicted paratope residues are grouped into the same paratype. To define predicted paratope residues, we used the Parapred model, which is based on convolutional and recurrent neural networks trained on 277 antibody-antigen co-crystal structures, and used the probability threshold of 0.67 as deemed optimal in the original paper[37] (see the "Materials and methods" section for more details).

To test the ability of paratyping and clonotyping to group antibodies that target the same epitope, we performed a test in a single-cell (paired VH/VL) data set of 1290 antibodies isolated from genetically engineered mice that have a full set of human immunoglobulin variable region genes[39] immunized with PTx. Although we had pairing information within the single-cell data set, given that the majority of repertoire sequencing data to date is heavy-chain only,[36] we demonstrated the method using only the heavy chain information. The sequences were annotated with a PTx-binding (364) or non-binding label (926) (using homogeneous time-resolved fluorescence (HTRF) and surface plasmon resonance (SPR) as per the "Single-cell data set" subsection under the ""Materials and methods" section), and we used paratyping (our new method) and clonotyping (the conventional approach) to identify PTx-binding sequences. VH and VK/VL gene frequencies of these antibodies as well as somatic hypermutation counts can be seen in Supplementary Figure 2.

For each of the 364 PTx-binders in turn, we mimicked a repertoire mining experiment by using paratyping or clonotyping to identify binders amongst the remaining 1289 sequences (one-vs-all cross-validation). Each of the PTx-binders is referred to as a "probe" antibody; sequences that are within the same paratype or clonotype as the probe are predicted to bind PTx. Our hypothesis is that antibodies with similar paratopes will bind to similar epitopes, but our method does not require that there are unique paratope/epitope pairings; antibodies with different paratopes may bind to the same epitope, and this would contribute to the false-negative rate of the method. The precision and recall of the two methods

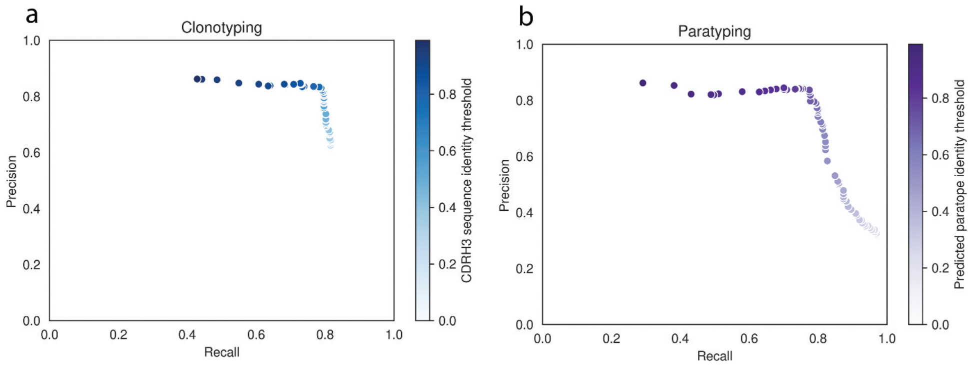

Figure 2. Precision–recall curves for clonotyping (a) and paratyping (b) in the task of predicting PTx binding. Precision and recall values are calculated over a range of CDRH3 sequence identity and predicted paratope identity thresholds respectively between 0.0% and 100.0% identity. The precision–recall profiles are similar.

Table 1. Precision–recall values for prediction of PTx-binding according to paratyping and clonotyping at the optimal thresholds of 75% and 72%, respectively.

Method	Sequence identity threshold	Precision	Recall
Paratyping	75%	84%	76%
Clonotyping	72%	83%	79%

Sequence identity is calculated across the predicted paratope for paratyping and across the CDRH3 for clonotyping. The methods behave comparably over the full precision–recall curve. These thresholds were then used for the prospective repertoire mining experiment.

(calculated over the aggregate of predictions) for repertoire mining are comparable (Figure 2, Table 1). The precision and recall using clonotyping and paratyping with varying CDRH3 sequence identity or paratope sequence identity thresholds, respectively, are shown in Figure 2a and 2b. The methods require different sequence identity thresholds for optimum performance but have similar precision–recall profiles.

For clonotyping, the optimal heavy-chain only threshold is 72% CDRH3 amino acid identity. For paratyping, the optimum occurs at 75% paratope identity. These thresholds are optimal as they maximize both precision and recall, corresponding to the “shoulder” of the precision–recall curve. At these optimal thresholds, clonotyping recovers binders with 83% precision and 79% recall (meaning that 21% of the binders in this data set are not related by clonotype to any other). Paratyping recovers binders with a precision of 84% and a recall of 76% (meaning that 24% of binders in this data set have distinct paratopes). We expect that this would be an overestimation of performance in the bulk data set due to the enrichment of PTx-binders created through antigen-specific sorting.

For each probe, a prediction can be made by both paratyping and clonotyping (hereafter labeled a “Both” prediction), paratyping alone (labeled “Paratype-only”) or clonotyping alone (labeled “Clonotype-only”). Out of all of these predictions, 76.6% are made by both methods with 89 “paratype-only” predictions (precision: 76%) and 345 “clonotype-only” predictions (precision: 84%). Paratyping and clonotyping make a number of method-exclusive predictions, as shown in Figure 3, where the probe antibodies with the largest number of “paratype-only” and “clonotype-only” predictions are shown in dendrograms with these predictions. Figure 3 also shows the variability of the method’s performance across probes, with the probe antibodies yielding the lowest and highest precision shown. One probe identified 20 other PTx-binding antibodies with 100% precision, while another had just 33% precision. The heterogeneity in predictions across probe emphasizes the utility of either method when only a small number of binders are known, despite the similar behavior of the methods over the aggregate of probe antibodies.

Paratyping-only predictions are PTx-binding heavy chain sequences from different clonotypes to the probe antibody. There are 89 predictions, which can be split into three groups: 1) those with different V genes (4); 2) those with the same V genes but different J genes (38); and 3) those with the same V and J genes but with CDRH3 identity below 72.0% (47). For example, a pair of binders with just 40% H3 sequence identity but 80% predicted paratope identity were clustered together. Homology modeling of the sequences using ABodyBuilder[42] suggests that the pair of sequences would also be predicted to be highly structurally similar (see Supplementary Table 2 for model information). Antibodies within the same paratype are predicted to bind to the same epitope, as it is highly improbable that sequence and structurally similar antibodies would bind to completely separate epitopes on the same antigen. This hypothesis could be experimentally validated by solving structures of antibodies in complex with antigen, or with epitope competition data.

Paratyping enables the identification of novel anti-PTx antibodies from different clonotypes to known binders in prospective experiment

Paratyping and clonotyping identify different binders in a bulk repertoire experiment

In order to test how the methods would scale in a larger and non-enriched data set, we performed a prospective repertoire mining experiment in which the heavy chain sequences from PTx-binding antibodies were used to identify novel PTx-binding heavy chains from a set of bulk heavy chain repertoires.

We first looked at how the methods scale in terms of number of predictions, and then experimentally validated the PTx-reactivity of a number of the predictions made by both clonotyping and paratyping.

We used the thresholds determined in the single-cell section. Sequences were considered to be in the same paratype if they have the same length CDRs (North definition) and above 75% amino acid sequence identity in the predicted paratope. Sequences were considered to be in the same clonotype if they

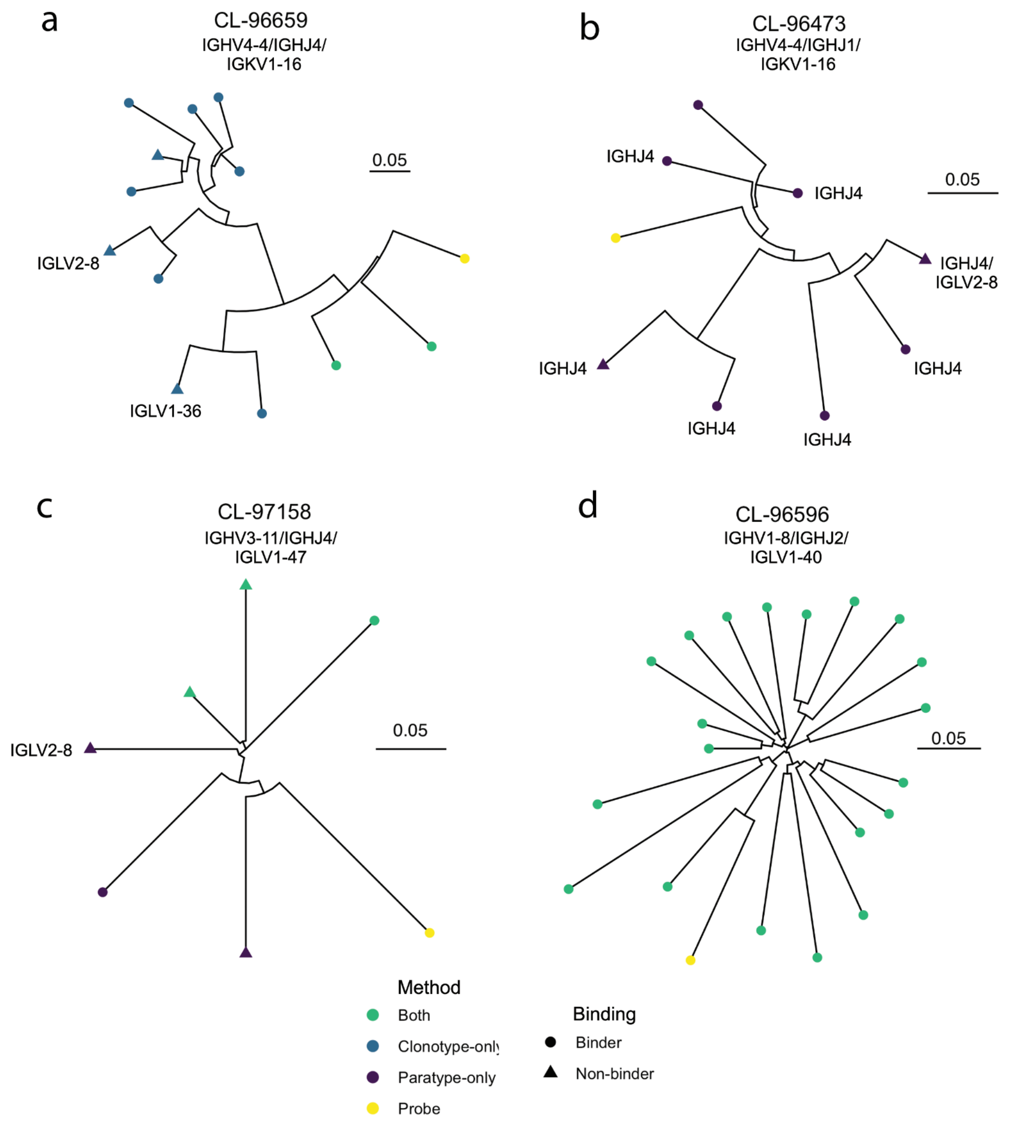

Figure 3. Representative dendrograms from the single-cell repertoire mining experiment. Probe sequences (yellow) are known PTx binding sequences. Other sequences which are in the same paratype or clonotype are predicted to also bind PTx. These predicted PTx binding sequences are colored according to whether they are identified by both paratyping and clonotyping ("Both"), paratyping but not clonotyping ("Paratype-only paratyping but not clonotyping") or clonotyping but not paratyping ("Clonotype-only"). Circular leaves represent true PTx binding antibodies (i.e., true positives) while triangular leaves represent sequences that do not bind PTx (false positives). Dendrogram A shows the probe which had the most "clonotype-only" predictions, of which 70% are true positives; dendrogram B shows the probe with the most "paratype-only" predictions, of which 75% are true positives; dendrogram C is the probe antibody which is associated with the most false positives (33% true positives) while dendrogram D is the probe antibody associated with the most true positives (20 true positives). Performance is heterogeneous across probes ranging between 33% and 100% precision; precision and recall values reported elsewhere in this manuscript consider the performance in aggregate of all 364 probes. IGHV, IGHJ or IGK/LV genes are annotated where this changes within a dendrogram. It can be seen that PTx-binding is abrogated by all instances of a change in IGK/LV gene. Dendrograms are constructed using the full VH sequence with the neighbor-joining algorithm of the R package ape,[40] plotted using the R package ggtree.[41] The units of the scale bar are amino acid substitutions per residue.

have the same V and J genes, the same length CDRH3 (North definition) and above 72% amino acid sequence identity across the CDRH3. The 72% value is below the most common thresholds used for clonotyping (80–100%), but it was chosen for fair comparison with paratyping as the threshold providing the best performance in the single-cell section of "Results".

Using the 364 known PTx binders as probes, 59,107 heavy chain sequences from bulk sequencing repertoires were searched for paratype- or clonotype-related sequences. A total of 4269 sequences were identified by both clonotyping and paratyping, 1113 by paratype-only and 1077 by clonotype-only.

The 1113 paratype-only predictions can be categorized as those with a different V gene with respect to the probe (179); of the 934 predictions with the same V gene, 396 have a different J gene; the remaining 538 predictions have sub-72% CDRH3 identity to the probe.

Prospective experimental validation of novel PTx-binding sequences

For the experimental validation of predicted PTx binders and non-binders, we created a category of prediction more stringent than the "paratype-only" or "clonotype-only" categories to show the utility of paratyping in a real antibody discovery experiment context where multiple probe sequences are available. As shown in Figure 3, a particular probe antibody may make predictions via clonotype or paratype alone. As reflected in the similarity in the precision–recall values calculated over the aggregate of probe antibodies in the single-cell one-versus-all cross-validation, these "method-unique" predictions become rarer when considering a larger number of probes. Such predictions, which could not be found by another method even when using the full complement of known binders, are referred to as "paratype-unique" or "clonotype-unique" predictions.

Using the 97 probes of highest SPR affinity from the single-cell data set, 2193 heavy chains were predicted to bind PTx. Of the potential 2193 predicted PTx binders, 139 were selected for expression (see the "Materials and methods" section) and PTx-binding assay. An additional 48 antibodies predicted to not bind PTx (due to paratope identity or shared clonotype with confirmed non-binders) were also tested. The predictions were split into thirds according to whether they were predicted by both methods (labeled "both") or were unique across all binders for paratyping ("paratype-unique") or clonotyping ("clonotype-unique") (see the "Materials and methods" section).

Of the 43 novel heavy chains predicted to bind PTx by both paratyping and clonotyping, 39 (90%) were experimentally confirmed PTx binders. Paratope identity between known and predicted binders ranged between 83% and 100%, with CDRH3 identity ranging between 77% and 100%. Thirty-one of the 48 (65%) of the clonotype-unique predictions were true PTx binders, with minimally 57% predicted paratope identity to a known binder, 72% CDRH3 identity and 74% CDRH identity (amino acid identity calculated over the heavy chain CDRs). Fourteen of the 48 (30%) of the paratype-unique PTx binders bound PTx. The minimal CDRH3 identity of a PTx binder to any known binder was 56% with 76% paratope identity. The distribution of CDRH3, total CDR and total VH amino acid identity of novel PTx-binding heavy chains to known PTx-binding antibodies is shown in Figure 4 (see Supplementary Table 3 for a comparison of VH/JH/CDRH3 combinations in probe antibodies and the novel PTx-binding antibodies discovered via paratyping, clonotyping or both methods). None of the 48 predicted PTx non-binders bound PTx.

As in the single-cell data set, the success rate in the predictions that were made by both clonotyping and paratyping is higher than either method alone. The success rate of paratype-unique predictions is significantly lower than that of clonotype-unique predictions. However, it may not be appropriate to compare performance across different probe antibodies, some of which may be liable to activity cliffs (a concept from small-molecule chemistry where a compound exhibits a large change in activity given only a small change in structure[43]). A direct comparison can be made where both clonotype-unique and paratype-unique predictions were made using the same known PTx binder. This occurred for six PTx probes, and across these, an average precision of 75% for paratype-unique and 92% for clonotype-unique was observed.

Discovery of novel anti-PTx antibodies from different clonotypes

Paratyping identified PTx-binding antibodies that could not be found using clonotyping ("paratype-unique"), for example, those using a different V gene to any of the known PTx binders. An example is shown in Figure 5), where the original antibody used the inherently autoreactive V gene V4-34,[44] which may be problematic in development. However, paratyping recovers seven PTx-reactive antibodies that use the V4-59 gene segment instead.

Paratyping also recovered novel PTx-binding heavy chains that derive from different J genes and examples with CDRH3 identities well below most clonotyping thresholds (commonly 80–100%). The minimal CDRH3 identity of a validated PTx-binding antibody to any known binder was 56%, suggesting that paratyping can identify antibodies that bind to the same epitope that could not be found by any clonotyping method.

Repertoire mining can improve in silico developability metrics

One of the limitations of clonotyping as a method for immune repertoire mining is the relatively narrow sequence space within which it is capable of making predictions, meaning that the discovered antibodies may have conserved developability problems. We have already seen one example where paratyping's ability to jump between germlines allows us to avoid an autoreactive V gene-derived antibody; other developability problems such as aggregation propensity may also be improved by using paratyping.

Of the original 97 antibodies used for repertoire mining, 38 antibodies were flagged by the Therapeutic Antibody Profiler (TAP) tool as having possible developability issues due to CDR length, high density of charge or hydrophobicity, or charge asymmetry between the heavy and light chains. As paratyping only groups of antibodies with the same length CDRs, we considered only the latter four developability metrics. Twenty-six of the original probe antibodies were flagged with extreme (2) or unobserved (24) values. Values are considered extreme (flagged as amber by TAP) if they fall within either the top or bottom 5% of the distribution of values observed in a set of 377 clinical-stage therapeutics (CSTs).[45] Values are unobserved (flagged as red by TAP) if they are outside of the range of values observed in existing CSTs. Seventeen of these probes were used in the identification of novel antibodies assayed for PTx binding, of which 13 successfully identified new PTx binders. For four of these probes, one of the new PTx-binding antibodies identified showed a sufficiently large change in the flagged developability metric that the flag was removed. There were a further five new PTx-binding antibodies showing a shift in the flagged developability metric toward the mean value among CSTs.

Figure 6 shows the improvement in patch surface hydrophobicity achieved by immune repertoire mining using CL-95375 as a probe antibody. CL-95375 had an amber flag for this metric. Repertoire mining was used to identify a number of predicted PTx-binding antibodies. It can be seen that the more sequence-distinct paratype-only predictions are able to achieve greater changes in patch surface hydrophobicity (PSH). These predictions were not assayed as they were within the clonotype of another known binder, and therefore not "paratype-unique".

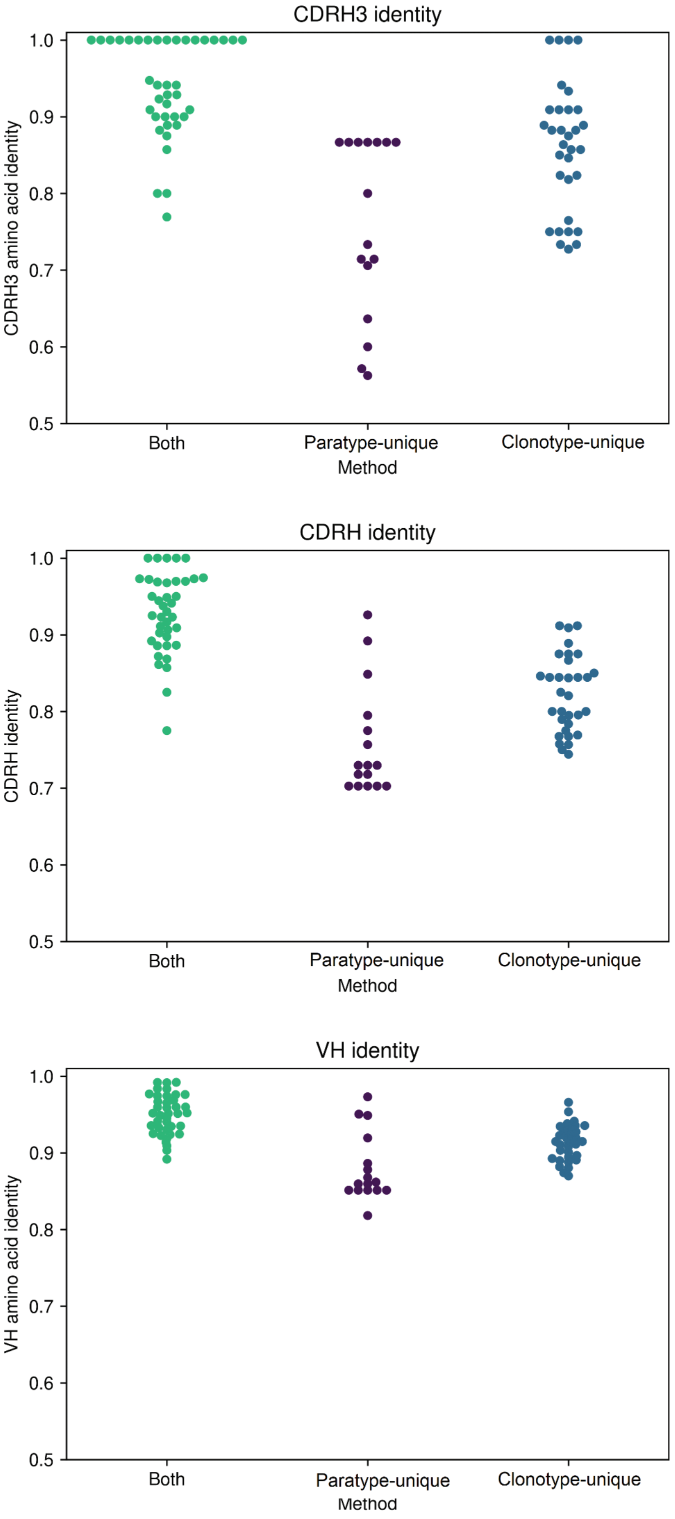

Figure 4. CDRH3, total CDR and total VH amino acid identity of novel PTx-binding antibodies to the known PTx-binding antibody by which they were identified, according to method by which they were identified. Paratyping enables the discovery of PTx-binding sequences with lower sequence identity across each of these regions with minimally 56% CDRH3 identity, 70% CDRH identity and 80% total VH identity.

Discussion

Characterizing the functional relationship between sequence-distinct antibodies is an important step in our understanding of the adaptive immune landscape. Mapping antigen preference to antibody repertoires will allow us to identify epitope convergence of antibodies at a large scale. In a test system of transgenic mice immunized with PTx, we show for the first time that prediction and comparison of paratopes can be used to group antigen-specific antibodies in both an enriched, single-cell data set and non-enriched bulk heavy chain repertoires. We demonstrate the utility of the method in the context of an antibody discovery experiment alongside the conventional approach of clonotyping and discover new anti-PTx antibodies from different clonotypes to any known binders.

We first developed the method in a single-cell data set, where paratyping and clonotyping were able to group PTx-specific antibody heavy chains with high precision (84% and 82%, respectively). These results may not map to the bulk sequencing data set given that the sequences derived from both plasma-, memory- and antigen-sorted cells (leading to ca. 30% of sequences being PTx-reactive). To validate the method in a less enriched data set, we performed a prospective experimental test of paratyping in a non-antigen sorted set of bulk heavy chain sequencing repertoires.

In the prospective experimental test, paratyping allowed us to discover new PTx-binding antibodies that we could not have found using clonotyping. These include antibodies that use different germline genes as well as antibodies with lower CDRH3 identity than in common definitions (80–100%). In terms of antibody discovery, paratyping allows us to identify sequence-distinct antibodies that bind to the same epitope and that can differ significantly in developability or affinity. In the terms of repertoire analysis, paratyping expands our ability to functionally group antibodies beyond clonotypes, and therefore allows us to detect specific cases of epitope convergence between clonotypes. For example, we found epitope convergence between IGHV4-34/IGHJ6 and IGHV4-59/IGHJ6 clonotypes, IGHV4-39/IGHJ6 and IGHV4-39/IGHJ4 clonotypes, IGHV4-34/IGHJ5 and IGHV4-34/IGHJ4 clonotypes and IGHV5-51/IGHJ4 and IGHV5-51/IGHJ6 clonotypes in PTx binders. We did not observe pairs of antibodies in the same paratype using different V gene subgroups, but antigen-reactive antibodies with identical CDRH3s deriving from different V gene subgroups have been observed, suggesting that this convergence does occur.[46] Paratope identity across the germline-encoded CDRHs 1 and 2 is equal to or in excess of 75% across members of IGHV1 and IGHV7, and IGHV3 and IGHV4 (see Supplementary Figure 4), so we predict that it should be possible to use paratyping to find binders from different V gene subgroups, should a large enough sequencing data set be mined. The implications of this for immune repertoire clustering are as yet unexplored. If such convergence is widespread, it is possible that clonotyping overestimates functional diversity, and this may account for low proportions of clonotypes shared across individuals.[3,47]

Success rates in the prospective experimental validation were variable across the categories of prediction. We found that 90% of sequences predicted by both clonotyping and paratyping to be binders bound PTx. The success rate was considerably lower in clonotype-unique predictions (65%) and even lower in paratype-unique predictions (30%). It should be noted that these method-unique predictions

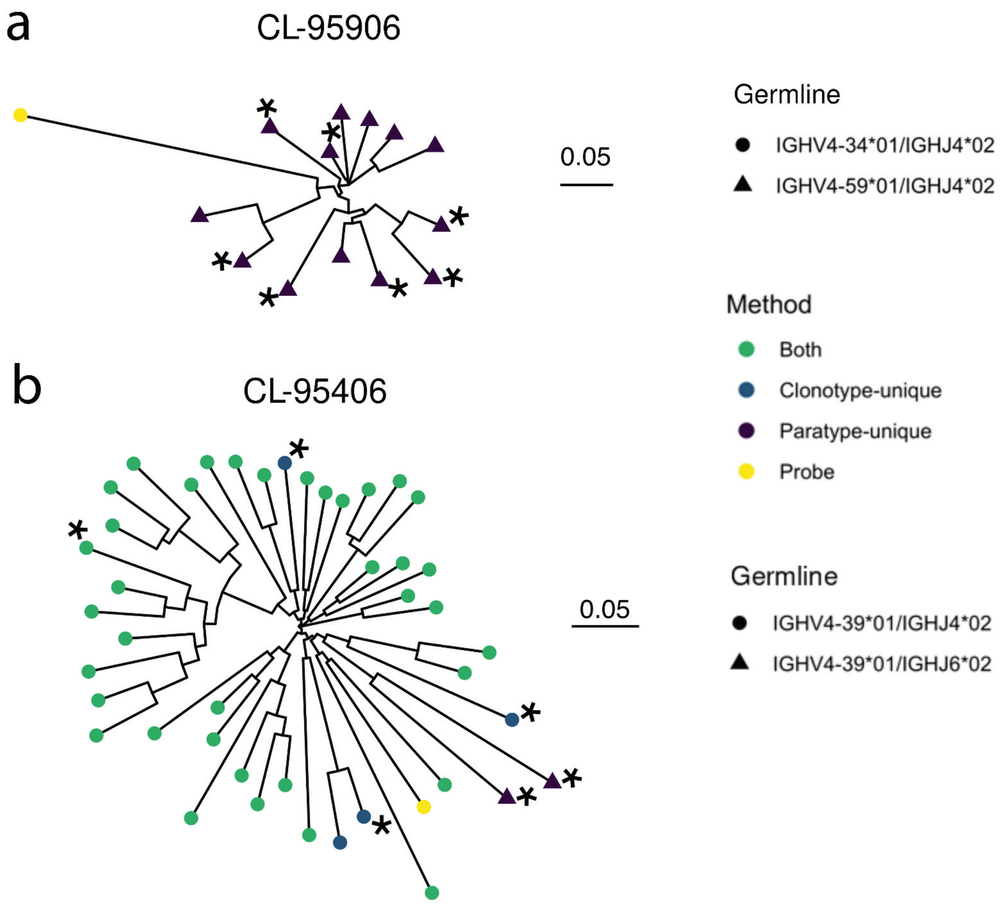

Figure 5. Dendrograms showing two examples of immune repertoire mining, using known PTx binders CL-95906 and CL-95940 (leaves colored yellow). Heavy chains predicted to bind PTx are colored as green, blue or purple depending on whether they are predicted to bind via both methods or were clonotype- or paratype-unique. Asterisks indicate heavy chains selected for testing, all of which were validated as PTx binding. A shows sequences predicted to bind PTx that use a different V gene (V459) than the known PTx binder used for prediction (V4-34). Sequences using a different J gene to the known PTx binder are shown in B.

form the minority of predictions for either method (15–22% of predictions). The lower success rate of paratyping versus clonotyping may be attributable to the particularly low CDRH3 identity of these predictions to known binders – paratyping does not give special weight to the CDRH3 in the paratope identity calculation despite the particular role it plays in antigen complementarity. Predictions with as little as 33% CDRH3 identity to a known binder were assayed, but no antibody with a CDRH3 amino acid identity below 56% bound to PTx, suggesting that paratyping could be further improved by the use of CDRH3 weighting.

The method relies upon the accuracy of the paratope prediction step. We have used a previously published sequence-based paratope prediction model, Parapred,[37] which achieved the highest F-score in our benchmark (see Supplementary Table 4) and is fully open-source. We tested other paratope prediction methods as inputs to paratyping using the data from the single-cell experiment, and found that the results were comparable to those with Parapred (see Supplementary Table 5). Parapred's convolutional and recurrent neural networks were trained and validated on 277 antibody-antigen complexes with protein antigens. While it is reasonable to expect paratyping to generalize across protein antigens, it may not be reasonable to expect paratyping to generalize across classes of antigens that did not feature in Parapred's training set. Among a number of metrics shown to discern carbohydrate- and protein-binding antibodies, paratope size is a key discriminant.[48–52] The application of paratope prediction models trained on protein antigens to carbohydrate or peptide antigens could lead to lower success rates in paratyping's predictions.

Paratope prediction only takes around 0.1 s per sequence (of which 0.02 s correspond to CDR extraction) as opposed to 0.05 s per sequence for VDJ annotation using IgBlast[53]), which means it is tractable for large datasets unlike homology modeling (on average 30 seconds per sequence, 600 times slower than germline gene annotation[34]). It has the advantage that it does not rely on the upkeep of consistent and complete germline databases (a leading cause of disagreement between germline annotation tools[13]), but rather on the distribution of a pretrained paratope prediction model.[37] The paratope prediction step is purely sequence-based and structural modeling is not required, meaning that immune repertoires can be annotated without accurate germline alignment and without access to large computing power.

Paratyping was shown to identify functional relationships between PTx-binding antibodies that are not related by clonotype, as per the hypothesis that antibodies with similar paratopes will bind to the same epitope; we would expect this to generalize across protein antigens. As an example, we looked at

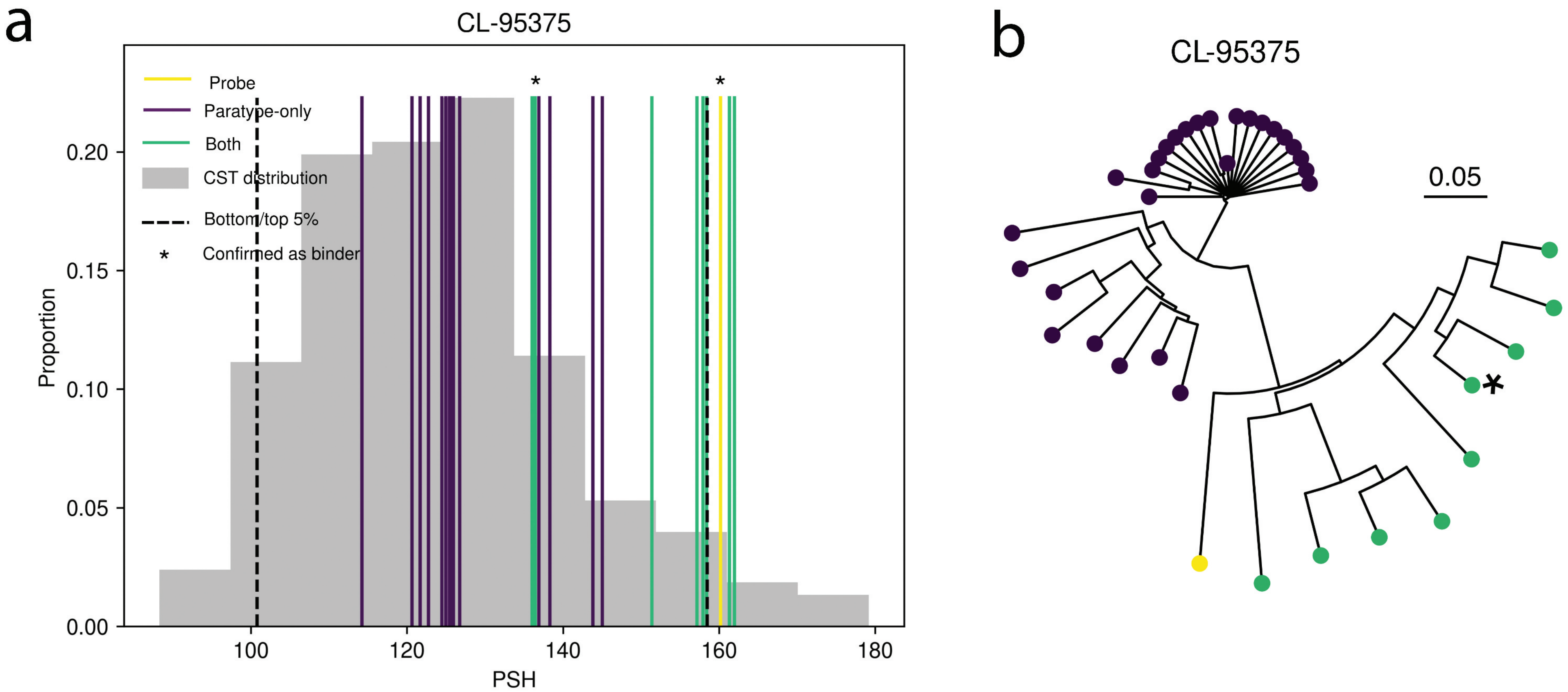

Figure 6. An example of a probe antibody with a flagged developability issue (in this instance, patch surface hydrophobicity (PSH)) which could be significantly improved by immune repertoire mining. Clonotyping and paratyping were used to discover a novel antibody targeting the same epitope but without a developability flag. This shift is shown in A, where asterisks indicate the probe antibody, CL-95375 (yellow), and the validated novel PTx-binding antibody which has a shift in PSH toward the mean value among 377 clinical-stage therapeutics.[45] The extremes of the distribution (upper and lower 5% for PSH) are shown as the black dashed line. The distribution of the metric among antibodies within the same paratype and/or clonotype (colored according to the legend) is shown; the only antibody assayed, which was validated as PTx-binding, is marked with an asterisk. Other examples can be seen in Supplementary Figure 3. The predicted PTx binders in A are shown in the dendrogram in B, with the confirmed novel PTx binder marked by an asterisk.

Cov-AbDab,[54] a database of antibodies and nanobodies known to bind to betacoronavirus proteins. Paratyping identified a number of pairs of antibodies from different clonotypes binding to the same epitope, for example, MERS-4 and MERS-4-V2, D12 and MERS-27, and 27D and 6A. We further analyzed five separate antigen examples (hemagglutinin, gp120, SARS CoV-2 spike protein, gp160 and lysozyme) to evaluate the hypothesis that antibodies within the same paratype bind to the same epitope. All pairs of antibodies within the same paratype shared some epitope residues, and would therefore be likely to complete for the same epitope (see Supplementary Table 6).

Deciphering the functional landscape of immune repertoires will greatly improve our understanding of the adaptive immune system. Improving our ability to group antibodies binding to the same epitope is a step toward this. Our results show here that the simple and computationally rapid abstraction of the antibody binding site used by paratyping is sufficient to group antigen-specific antibodies in a way that provides us with additional information beyond clonotyping. This additional information is particularly significant in the context of antibody discovery, where it allows us to recover different novel antigen-specific antibodies from immune repertoires.

Materials and methods

Data sets

Single-cell data set

Five genetically engineered mice that have a full set of human immunoglobulin variable region genes (Intelliselect transgenic mice)[39] were immunized with PTx. This study was carried out under the Project Licenses 70/8718 issued by the UK Government Home Office under Animal (Scientific Procedures) Act (A(SP)A), 1986, incorporating Directive 2010/63/EU of the European Parliament, and with the approval of the Sanger Institute Animal Welfare and Ethical Review Body. The institute complied with the Code of Practice issued by the UK Government which aids compliance with the A(SP) A. The institute has a PHS assurance F16-00128 (WTSI).

A total of 1290 paired (VH/VL) sequences were recovered from antigen-sorted, plasma and memory cells via a previously published method.[55] These 1290 sequences were expressed in HEK293 cells. Antibody supernatant was collected on d 8 after transfection and screened for binding to wildtype PTx by HTRF. Positive control anti-PTx antibody (ab37574, abcam) was diluted in Expi293^{TM} Expression Medium (Gibco) over an 11-point titration using one in three dilutions to generate a standard curve. Titrations of 5 µl of antibody ab37574 were added to a 384-well white-walled assay plate (Greiner Bio-One) . Negative control wells received 5 µl of Expi293^{TM} Expression Medium only. Five icroliters of HEK293 antibody supernatants (undetermined concentration) were added to one well of the 384-well plate. Five microliters of PTx conjugated to Alexa 647 (Lightning-Link, Innova Bioscience) (3.75 nM final concentration) was added to all wells of the assay plate except negative control wells, which instead received 5 µl Expi293^{TM} Expression Medium. Finally, 10 µl of anti-mouse IgG donor antibody (Southern Biotech; to bind to murine constant region in control and chimeric Intelliselect transgenic mouse antibodies) labeled with europium cryptate (Cis Bio), (1:4000 final concentration) was added to each well and the assay was left in the dark at room temperature to incubate for 2 h. After incubation, the assay was read on an Envision plate reader (Perkin Elmer) using a standard HTRF protocol. Then, 620 and 665 nm channel values were exported to Microsoft Excel (Microsoft) and F calculations performed ((665/620 nm ratio − signal negative control)/signal negative control) × 100). Percent effect values were calculated for each antibody by comparing its F value against a positive control antibody (ab37574) at 6.66 nM. The number of PTx-positive antibodies as a function of percentage effect values are shown in Figure 7a.

Antibody sequences with greater than 10% effect value relative to the positive control were labeled as binders. This resulted in 364 PTx-binders and 926 non-binders. SPR was performed on each of the 364 binders (Figure 7b).

Bulk data set

Heavy chains from sorted splenic B cells from the same individual five mice as the single-cell data set were sequenced using standard protocols[25] and processed using the pRESTO/Change-O pipeline.[56,57] This resulted in 259,151 heavy chain sequences. For quality control, only sequences with read count

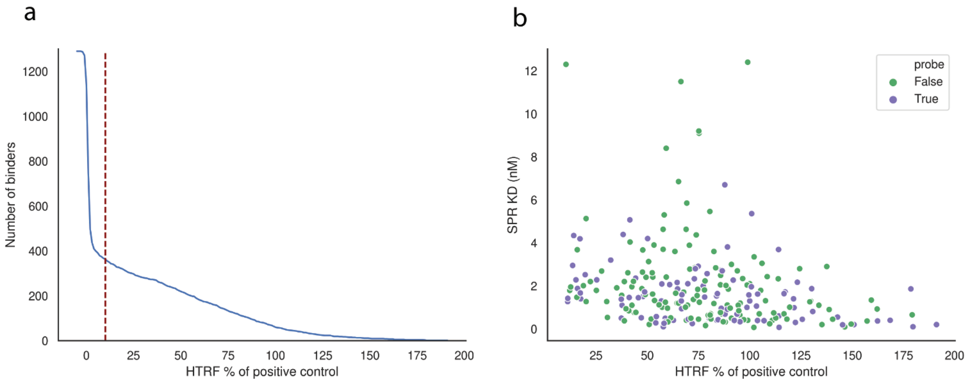

Figure 7. (a) Number of binders among the 1290 antibodies from the single-cell experiment as a function of percentage effect value relative to the positive control antibody, ab37574. At the chosen percentage effect of 10%, 364 of the 1290 binders are deemed PTx reactive. (b) Affinity measurements across the 364 PTx-reactive antibodies; antibodies in purple were used as probe antibodies for repertoire mining as per the "Repertoire mining experiment" subsection under the "Materials and methods" section.

(reads with a particular unique molecular identifier (UMI)) above or equal to two or a consensus count (reads with different UMIs but the same nucleotide sequence) above or equal to 10 were considered, reducing the size of the data set to 59,107 sequences.

Clonotyping

Clonotypes were defined as groups of heavy chain sequences sharing the same V and J genes, with identical CDRH3 lengths and a number of amino acid mismatches equal to or below a threshold sequence identity. VJ annotation was performed with IgBLAST[53] within Change-O.[57] CDRH3s were extracted according to the North definition[38] with IMGT numbering[58] performed with ANARCI.[59]

Paratyping

Paratypes were defined as heavy chain sequences sharing the same CDR lengths and greater than a threshold sequence identity across the predicted paratope regions. CDRs were extracted according to North definitions[38] with IMGT numbering[58] performed via ANARCI.[59]

Parapred[37] was used for paratope prediction using the model as distributed by E. Liberis at https://github.com/eliberis/parapred. To convert the output of Parapred, binding probabilities, into a binary label, we selected a threshold of 0.67 as deemed optimal by the authors of the original paper,[37] i.e., residues with a predicted probability of being in the paratope of above 0.67 were annotated as paratope residues. Paratope identity was defined as the number of identical paratope residues (residues that are predicted to be in the paratope in both cases) divided by the smallest number of paratope residues of either sequence being compared (Figure 8). The performance of Parapred on a set of 552 antibody-antigen structures (SAbDab October 2020) is evaluated in Supplementary Table 4.

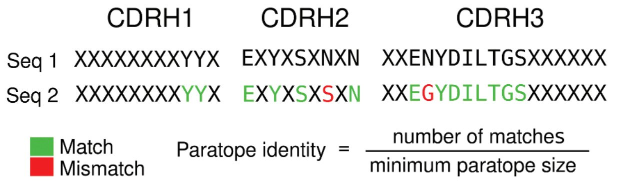
CDRH1
CDRH2
CDRH3
Seq 1 XXXXXXXXYYX EXYXSXNXN XXENYDILTGSXXXXXX
Seq 2 XXXXXXXXYYX EXYXSXSXN XXEGYDILTGSXXXXXX
Match
Mismatch
Paratope identity = number of matches / minimum paratope size

Figure 8. Method of calculating predicted paratope identity. X indicates CDR residues not predicted to constitute the paratope. In the example shown, the paratope identity is 87.5%.

Repertoire mining experiment

In the repertoire mining experiment, we selected a number of hit antibodies from the single-cell data set to use as probes. SPR was carried out on the HTRF-positive antibodies and the highest affinity representatives were selected from each clonotype containing only sequences labeled as binders equating to 97 antibodies, in order to align with previous repertoire mining experiments.[9] We also selected a number of non-PTx binding antibodies to be used as a negative control. The non-PTx binding antibodies were selected as representatives of clonotypes containing only non-PTx binding sequences (551 antibodies).

The heavy chains from these probe antibodies were used as probes to mine the bulk repertoires via both paratyping and clonotyping, using the optimal sequence identity thresholds from the single-cell data set (75% predicted paratope identity for paratyping and 72% CDRH3 sequence identity for clonotyping). The paratyping process is illustrated in Figure 9.

Predictions were labeled as "paratype-only", "clonotype-only" or "both" as detailed in the "Results" section. If an antibody is within the same paratype as a particular probe but not within its clonotype, it is a "paratype-only" prediction and vice-versa. A prediction is labeled as "both" if it is within both the paratype and clonotype of the same probe. In the following subsection, the more stringent "paratype-unique" and "clonotype-unique" definitions were used.

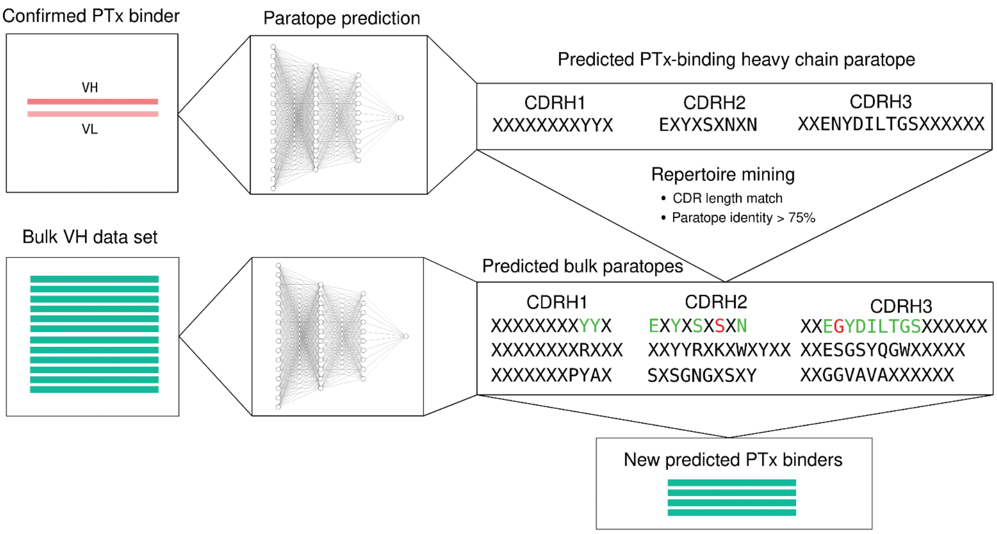

Figure 9. Graphical illustration of the process of repertoire mining via paratyping. A probe antibody is selected. Paratope prediction is performed both on the probe antibody and the bulk data set. Heavy chain predicted paratope identity is used to mine the bulk repertoire for new predicted binders.

Experimental validation of predicted binders

We selected 192 heavy chain sequences from the bulk data set for expression. A total of 144 of the sequences were predicted to be PTx binders using the 97 PTx-binding probe sequences. We created a new way of categorizing predictions in order to fully test the method in the context of an antibody discovery experiment. We define "paratype-unique" predictions as predictions which are not within the paratype of not only the probe in question (paratype-only) but of any of the full complement of 97 probes, and similarly for "clonotype-unique" predictions. Paratype-unique predictions are a more stringent subset of paratype-only predictions and present a greater challenge of the method (as such predictions tend to be corroborated by fewer probes). Forty-eight predicted PTx-binding heavy chain sequences were selected from each of the three categories of prediction ("paratype-unique", "clonotype-unique" and "both"). Five of the "both" predictions were identical to their probes. The remaining 48 sequences assayed were predicted non-PTx binding heavy chain sequences with 16 sequences selected from "clonotype-unique", "paratype-unique" and "both" categories of prediction.

The heavy chain sequences selected for expression from the bulk repertoire were paired with the cognate light chain of the probe sequence by which they were identified. Within-clonotype VH/VL pairing has been validated[9,60] as a method of reconstituting binding where the cognate light chain is not sequenced. Pairings were only made where there was greater than 82% sequence identity across the residues considered to constitute the VH/VL interface, based on solvent accessibility,[45] in order to maximize the probability of expression. These positions lie outside of the CDRs and therefore do not enforce any constraint on binding site sequence identity.

The predicted binding and non-binding antibodies were expressed in HEK293 cells. The assay is as described in the "Single-cell data set" "subsection under the "Materials and methods" section with the exception that the anti-PTx antibody 1B7 was used as positive control. Antibody sequences with an F value exceeding 100 were labeled as PTx-reactive antibodies.

Performance evaluation

To evaluate the performance of either method in grouping PTx-binding sequences, we calculated precision and recall of the method in the single-cell data set according to the following definitions of true positives, false positives and false negatives.

The task is to group antibodies that bind PTx. However, the method is hypothesized to work by grouping by epitope. There will be multiple epitopes on PTx so not all PTx-binding antibodies will be grouped into a single paratype or clonotype. As a result, we do not expect perfect recall. Further, we do not classify antibodies as non-binders if they do not group with a particular binder – we only predict that they do not bind at the same epitope as the binder in question. As a result, we do not calculate a "true negative" rate. Precision and recall are calculated as per standard definitions (TP/FP+TP, TP/TP+FN, respectively).

True positive (TP): A PTx-binding sequence that was identified by another PTx-binding sequence.

False positive (FP): A non-PTx-binding sequence that was identified by a PTx-binding sequence.

False negative (FN): A PTx-binding sequence that was not identified by any PTx-binding sequence.

In silico developability assessment

We calculated the in-silico developability metrics of total CDR length, PSH, patches of positive charge, patches of negative charge and structural Fv charge symmetry parameter using the Therapeutic Antibody Profiler tool (TAP).[45] The tool calculates these metrics from homology models built using ABodyBuilder[42] and compares them to a database of 277 CSTs.[61] According to the metric in question, values are flagged as "amber" (extreme) if they lie outside the upper or lower, or upper and lower, 5% of observed values among CSTs. An antibody is flagged as "red" if the value of a particular metric is outside of the observed range in CSTs.[45]

Disclosure statement

P.K., S.S., A.P. and S.W. are employees of Kymab Ltd. at the time of writing.

Funding

The research was supported by the Medical Research Council [grant number: MR/R015708/1], the Gates Foundation, and the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

ORCID

Charlotte M. Deane http://orcid.org/0000-0003-1388-2252

Data and code availability

The code and data associated with this study are available at http://opig.stats.ox.ac.uk/resources.

References

1. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol. 2014;32(2):158–68. doi:10.1038/nbt.2782. PMID: 24441474.
2. Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. Genome Med. 2015;7(1):121. doi:10.1186/s13073-015-0243-2. PMID: 26589402.
3. Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature. 2019;566(7744):393–97. doi:10.1038/s41586-019-0879-y. PMID: 30664748.
4. Bashford-Rogers RJM, Bergamaschi L, McKinney EF, Pombal DC, Mescia F, Lee JC, Thomas DC, Flint SM, Kellam P, Jayne DRW, et al. Analysis of the B cell receptor repertoire in six immune-mediated diseases. Nature. 2019;574(7776):122–26. doi:10.1038/s41586-019-1595-3. PMID: 31554970.
5. Greiff V, Miho E, Menzel U, Reddy ST. Bioinformatic and statistical analysis of adaptive immune repertoires. Trends Immunol. 2015;36(11):738–49. doi:10.1016/j.it.2015.09.006. PMID: 26508293.

6. Reddy ST, Ge X, Miklos AE, Hughes RA, Kang SH, Hoi KH, Chrysostomou C, Hunicke-Smith SP, Iverson BL, Tucker PW, et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. Nat Biotechnol. 2010;28(9):965–69. doi:10.1038/nbt.1673. PMID: 20802495.
7. Zhu J, Ofek G, Yang Y, Zhang B, Louder MK, Lu G, McKee K, Pancera M, Skinner J, Zhang Z, et al. Mining the antibodyome for HIV-1–neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. Proc Natl Acad Sci U S A. 2013;110(16):6470–75. doi:10.1073/pnas.1219320110. PMID: 23536288.
8. Tsioris T, Gupta NT, Ogunniyi AO, Zimnisky RM, Qian F, Yao Y, Wang X, Stern JNH, Chari R, Briggs AW, et al. Neutralizing antibodies against West Nile virus identified directly from human B cells by single-cell analysis and next generation sequencing. Integr Biol (Camb). 2015;7(12):1587–97. doi:10.1039/c5ib00169b. PMID: 26481611.
9. Hsiao Y, Shang Y, DiCara DM, Yee A, Lai J, Kim SH, Ellerman D, Corpuz R, Chen Y, Rajan S, et al. Immune repertoire mining for rapid affinity optimization of mouse monoclonal antibodies. mAbs. 2019;11(4):735–46. doi:10.1080/19420862.2019.1584517. PMID: 30900945.
10. Hershberg U, Luning Prak ET. The analysis of clonal expansions in normal and autoimmune B cell repertoires. Philos Trans R Soc Lond B Biol Sci. 2015;370(1676):20140239. doi:10.1098/rstb.2014.0239. PMID: 26194753.
11. Zhou JQ, Kleinstein SH. Cutting edge: ig H chains are sufficient to determine most B cell clonal relationships. J Immunol. 2019;203(7):1687–92. doi:10.4049/jimmunol.1900666. PMID: 31484734.
12. Tonegawa S. Somatic generation of antibody diversity. Nature. 1983;302(5909):575–81. doi:10.1038/302575a0. PMID: 6300689.
13. Smakaj E, Babrak L, Ohlin M, Shugay M, Briney B, Tosoni D, Galli C, Grobelsek V, D'Angelo I, Olson B, et al. Benchmarking immunoinformatic tools for the analysis of antibody repertoire sequences. Bioinformatics. 2020;36(6):1731–39. doi:10.1093/bioinformatics/btz845. PMID: 31873728.
14. Chen Z, Collins AM, Wang Y, Gaëta BA. Clustering-based identification of clonally-related immunoglobulin gene sequence sets. Immunome Res. 2010;6(1):S4. doi:10.1186/1745-7580-6-S1-S4. PMID: 20875155.
15. Lefranc MP, Lefranc G. The immunoglobulin factsbook. 2001. London (UK): Academic Press.
16. Roy B, Neumann RS, Snir O, Iversen R, Sandve GK, Lundin KEA, Sollid LM. High-throughput single-cell analysis of B cell receptor usage among autoantigen-specific plasma cells in celiac disease. J Immunol. 2017;199(2):782–91. doi:10.4049/jimmunol.1700169. PMID: 28600290.
17. Horns F, Dekker CL, Quake SR. Memory B cell activation, broad antiinfluenza antibodies, and bystander activation revealed by single-cell transcriptomics. Cell Rep. 2020;30(3):905–913.e6. doi:10.1016/j.celrep.2019.12.063. PMID: 31968262.
18. Lindeman I, Emerton G, Mamanova L, Snir O, Polanski K, Qiao SW, Sollid LM, Teichmann SA, Stubbington MJT. BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. Nat Methods. 2018;15(8):563–65. doi:10.1038/s41592-018-0082-3. PMID: 30065371.
19. Robins H. Immunosequencing: applications of immune repertoire deep sequencing. Curr Opin Immunol. 2013;25(5):646–52. doi:10.1016/j.coi.2013.09.017. PMID: 24140071.
20. Robinson WH. Sequencing the functional antibody repertoire—diagnostic and therapeutic discovery. Nat Rev Rheumatol. 2015;11(3):171–82. doi:10.1038/nrrheum.2014.220. PMID: 25536486.
21. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. Sci Transl Med. 2009;1(12):12ra23–12ra23. doi:10.1126/scitranslmed.3000540. PMID: 20161664.
22. Jiang N, He J, Weinstein JA, Penland L, Sasaki S, He XS, Dekker CL, Zheng NY, Huang M, Sullivan M, et al. Lineage structure of the human antibody repertoire in response to influenza vaccination. Sci Transl Med. 2013;5(171):171ra19–171ra19. doi:10.1126/scitranslmed.3004794. PMID: 23390249.
23. Jackson KJL, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, Marshall EL, Gurley TC, Moody MA, Haynes BF, et al. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. Cell Host Microbe. 2014;16(1):105–14. doi:10.1016/j.chom.2014.05.013. PMID: 24981332.
24. Galson JD, Pollard AJ, Trück J, Kelly DF. Studying the antibody repertoire after vaccination: practical applications. Trends Immunol. 2014;35(7):319–31. doi:10.1016/j.it.2014.04.005. PMID: 24856924.
25. Stern JNH, Yaari G, Vander Heiden JA, Church G, Donahue WF, Hintzen RQ, Huttner AJ, Laman JD, Nagra RM, Nylander A, et al. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. Sci Transl Med. 2014;6(248):248ra107–248ra107. doi:10.1126/scitranslmed.3008879. PMID: 25100741.
26. Bashford-Rogers RJM, Smith KGC, Thomas DC. Antibody repertoire analysis in polygenic autoimmune diseases. Immunology. 2018;155(1):3–17. doi:10.1111/imm.12927. PMID: 29574826.
27. Roskin KM, Simchoni N, Liu Y, Lee JY, Seo K, Hoh RA, Pham T, Park JH, Furman F, Dekker CL, et al. IgH sequences in common variable immune deficiency reveal altered B cell development and selection. Sci Transl Med. 2015;7(302):302ra135. doi:10.1126/scitranslmed.aab1216. PMID: 26311730.
28. Wong WK, Robinson SA, Bujotzek A, Georges G, Lewis AP, Shi J, Snowden J, Taddese B, Deane CM. Ab-Ligity: identifying sequence-dissimilar antibodies that bind to the same epitope. BioRxiv. 2020;2020:24.004051.
29. Scheid JF, Mouquet H, Ueberheide B, Diskin R, Klein F, Oliveira TYK, Pietzsch J, Fenyo D, Abadir A, Velinzon K, et al. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. Science. 2011;333(6049):1633–37. doi:10.1126/science.1207227. PMID: 21764753.
30. Joyce MG, Wheatley AK, Thomas PV, Chuang GY, Soto C, Bailer RT, Druz A, Georgiev IS, Gillespie RA, Kanekiyo M, et al. Vaccine-induced antibodies that neutralize group 1 and group 2 influenza A viruses. Cell. 2016;166(3):609–23. doi:10.1016/j.cell.2016.06.043. PMID: 27453470.
31. Rijal P, Elias SC, Rosendo Machado S, Xiao J, Schimanski L, O'Dowd V, Baker T, Barry E, Mendelsohn SC, Cherry CJ, et al. Therapeutic monoclonal antibodies for ebola virus infection derived from vaccinated humans. Cell Rep. 2019;27(1):172–186.e7. doi:10.1016/j.celrep.2019.03.020. PMID: 30943399.
32. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM. SAbDab: the structural antibody database. Nucleic Acids Res. 2014;42(D1):D1140–D1146. doi:10.1093/nar/gkt1043. PMID: 24214988.
33. Briney B, Burton DR. Massively scalable genetic analysis of antibody repertoires. BioRxiv. 2018;447813.
34. Raybould MIJ, Marks C, Kovaltsuk A, Lewis AP, Shi J, Deane CM. Evidence of antibody repertoire functional convergence through public baseline and shared response structures. BioRxiv. 2020;2020:17.993444.
35. Mason DM, Friedensohn S, Weber CR, Jordi C, Wagner B, Meng S, Gainza P, Correia BE, Reddy ST. Deep learning enables therapeutic antibody optimization in mammalian cells by deciphering high-dimensional protein sequence space. BioRxiv. 2019;617860.
36. Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K. Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. J Immunol. 2018;201(8):2502–09. doi:10.4049/jimmunol.1800708. PMID: 30217829.
37. Liberis E, Veličković P, Sormanni P, Vendruscolo M, Liò P. Parapred: antibody paratope prediction using convolutional and recurrent neural networks. Bioinformatics. 2018;34(17):2944–50. doi:10.1093/bioinformatics/bty305. PMID: 29672675.
38. North B, Lehmann A, Dunbrack RL. A new clustering of antibody CDR loop conformations. J Mol Bio. 2011;406(2):228–56. doi:10.1016/j.jmb.2010.10.030. PMID: 21035459.

39. Lee EC, Liang Q, Ali H, Bayliss L, Beasley A, Bloomfield-Gerdes T, Bonoli L, Brown R, Campbell J, Carpenter A, et al. Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. Nat Biotechnol. 2014;32 (4):356–63. doi:10.1038/nbt.2825. PMID: 24633243.
40. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35 (3):526–28. doi:10.1093/bioinformatics/bty633. PMID: 30016406.
41. Yu G, Smith DK, Zhu H, Guan Y, Lam TY. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8 (1):28–36. doi:10.1111/2041-210X.12628.
42. Leem J, Dunbar J, Georges G, Shi J, Deane CM. ABodyBuilder: automated antibody structure prediction with data–driven accuracy estimation. mAbs. 2016;8(7):1259–68. doi:10.1080/19420862.2016.1205773. PMID: 27392298.
43. Hu Y, Stumpfe D, Bajorath J. Advancing the activity cliff concept. F1000Res. 2013;2:199. doi:10.12688/f1000research.2-199.v1. PMID: 24555097.
44. Pugh-Bernard AE, Silverman GJ, Cappione AJ, Villano ME, Ryan DH, Insel RA, Sanz I. Regulation of inherently autoreactive VH4-34 B cells in the maintenance of human B cell tolerance. J Clin Invest. 2001;108(7):1061–70. doi:10.1172/JCI12462. PMID: 11581307.
45. Raybould MIJ, Marks C, Krawczyk K, Taddese B, Nowak J, Lewis AP, Bujotzek A, Shi J, Deane CM. Five computational developability guidelines for therapeutic antibody profiling. Proc Natl Acad Sci U S A. 2019;116(10):4025–30. doi:10.1073/pnas.1810576116. PMID: 30765520.
46. D'Angelo S, Ferrara F, Naranjo L, Erasmus MF, Hraber P, Bradbury ARM. Many routes to an antibody heavy-chain CDR3: necessary, yet insufficient, for specific binding. Front Immunol. 2018;9. doi:10.3389/fimmu.2018.00395. PMID: 29568296.
47. Soto C, Bombardi RG, Branchizio A, Kose N, Matta P, Sevy AM, Sinkovits RS, Gilchuk P, Finn JA, Crowe JE. High frequency of shared clonotypes in human B cell receptor repertoires. Nature. 2019;566(7744):398–402. doi:10.1038/s41586-019-0934-8. PMID: 30760926.
48. Marillet S, Lefranc MP, Boudinot P, Cazals F. Novel structural parameters of Ig–Ag complexes yield a quantitative description of interaction specificity and binding affinity. Front Immunol. 2017;8. doi:10.3389/fimmu.2017.00034. PMID: 28232828.
49. Di Rienzo L, Milanetti E, Lepore R, Olimpieri PP, Tramontano A. Superposition-free comparison and clustering of antibody binding sites: implications for the prediction of the nature of their antigen. Sci Rep. 2017;7(1):45053. doi:10.1038/srep45053. PMID: 28338016.
50. Raghunathan G, Smart J, Williams J, Almagro JC. Antigen-binding site anatomy and somatic mutations in antibodies that recognize different types of antigens. J Mol Recognit. 2012;25(3):103–13. doi:10.1002/jmr.2158. PMID: 22407974.
51. Collis AVJ, Brouwer AP, Martin ACR. Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. J Mol Biol. 2003;325(2):337–54. doi:10.1016/S0022-2836(02)01222-6. PMID: 12488099.
52. Lee M, Lloyd P, Zhang X, Schallhorn JM, Sugimoto K, Leach AG, Sapiro G, Houk KN. Shapes of antibody binding sites: qualitative and quantitative analyses based on a geomorphic classification scheme. J Org Chem. 2006;71(14):5082–92. doi:10.1021/jo052659z. PMID: 16808494.
53. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 2013;41 (W1):W34–W40. doi:10.1093/nar/gkt382. PMID: 23671333.
54. Raybould MIJ, Kovaltsuk A, Marks C, Deane CM. CoV-AbDab: the coronavirus antibody database. Bioinformatics. 2020. doi:10.1093/bioinformatics/btaa739. PMID: 32805021.
55. Bradley A, Lee EC, Liang Q, Liu H, Wood A, Wong V Expression vector production and high-throughput cell screening, July 2019.
56. Vander Heiden JA, Yaari G, Uduman M, Stern JNH, O'Connor KC, Hafler DA, Vigneault F, Kleinstein SH. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics. 2014;30(13):1930–32. doi:10.1093/bioinformatics/btu138. PMID: 24618469.
57. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics. 2015;31(20):3356–58. doi:10.1093/bioinformatics/btv359. PMID: 26069265.
58. Lefranc MP, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol. 2003;27 (1):55–77. doi:10.1016/S0145-305X(02)00039-3. PMID: 12477501.
59. Dunbar J, Deane CM. ANARCI: antigen receptor numbering and receptor classification. Bioinformatics. 2016;32(2):298–300. doi:10.1093/bioinformatics/btv552. PMID: 26424857.
60. Zhu J, Wu X, Zhang B, McKee K, O'Dell S, Soto C, Zhou T, Casazza JP, Mullikin JC, Kwong PD, et al. De novo identification of VRC01 class HIV-1-neutralizing antibodies by next-generation sequencing of B-cell transcripts. Proc Natl Acad Sci U S A. 2013;110(43):E4088–E4097. doi:10.1073/pnas.1306262110. PMID: 24106303.
61. Raybould MIJ, Marks C, Lewis AP, Shi J, Bujotzek A, Taddese B, Deane CM. Thera-SAbDab: the therapeutic structural antibody database. Nucleic Acids Res. 2020;48(D1):D383–D388. doi:10.1093/nar/gkz827. PMID: 31555805.