The landscape of high-affinity human antibodies against intratumoral antigens

Goran Rakocevic[1], Irina Glotova[1], Ines de Santiago[1], Berke Cagkan Toptas[1], Milena Popovic[1], Milos Popovic[1], Dario Leone[1], Andrew L. Stachyra[1], Deniz Kural[1,*], Daniele Biasci[1,*]

[1] Totient Inc. 1 Alewife Center Suite 120, Cambridge, MA 02140

* Corresponding authors: daniele.biasci@totient.bio, deniz.kural@totient.bio

Abstract

Intratumoral expression of antibody transcripts can be detected across different cancer types and is correlated with positive clinical outcomes. However, the antigens recognized by these antibodies remain largely unknown. Here we inferred the paired sequence of thousands of clonally expanded antibodies using bulk RNA sequencing data from solid tumors in The Cancer Genome Atlas (TCGA). After expressing 283 candidate antibodies in mammalian cells, we individually tested them against a library of twenty thousands full-length human proteins and a library of five thousands membrane-only proteins. Using surface plasmon resonance we confirmed 13 high-affinity antibodies that bind to their targets with K_D in the low nanomolar range. Our work provides insights into the antigens that drive B cell responses in human cancers while also directly identifying fully human, high-affinity antibodies against them.

Main

Expression of antibody transcripts is often detected in solid tumors across different cancer types[1], but their functional relevance remains unclear. Certain characteristics of the intratumoral antibody repertoire (e.g., abundance, clonality and number of detectable somatic mutations) have been associated with positive clinical outcomes, such as longer overall survival and response to immune checkpoint inhibitors[1–3]. Moreover, the presence of intratumoral plasma cells and ectopic germinal centers, which are key components of the antibody selection and production

machinery[4], has been associated with longer overall survival and immunotherapy response[5–8]. Despite these observations, the contribution of intratumoral antibodies to immune responses against cancer remains largely unknown.

The main obstacle to functional characterization of intratumoral antibodies is our ignorance of their target antigens. Previous studies have demonstrated that sequences of individual antibody chains can be reconstructed using bioinformatics methods from bulk RNA sequencing (RNA-Seq) data generated by large-scale cancer genomics efforts[1,3]. Compared to specialized B-cell receptor (BCR) sequencing or single-cell sequencing, bulk RNA-seq has the significant advantage of being readily available for thousands of well characterized tumor samples[9]. However, previous studies did not attempt to pair heavy and light antibody chains, and express the resulting complete antibodies, two key steps required to experimentally identify their target antigens.

In this work we reconstructed and paired thousands of intratumoral antibody chains using legacy bulk RNA-Seq data from TCGA, one of the most comprehensive genomic studies of human cancer to date[9] (Figure 1). We individually performed gene synthesis, mammalian expression, and purification for 283 intratumoral antibodies, obtaining high quality, fully human antibody proteins in most cases. We then individually screened each antibody against two large collections of recombinant proteins, covering the vast majority of the human proteome, in order to obtain the most likely binding targets for each antibody. In selected cases, we confirmed binding using surface plasmon resonance (SPR) in order to characterize their binding kinetics. As a result, we were able to assemble compelling evidence that fully functional intratumoral antibodies can be obtained from legacy bulk RNA-Seq, without the need for specialized BCR or single-cell sequencing, and identify target antigens of high-affinity intratumoral antibodies from a large number of well characterized human tumor samples.

Results

Identification of dominant antibody sequences in the TCGA dataset

It has been observed that in some tumor samples the antibody repertoire is particularly clonal: a small set of antibody transcripts are expressed at high levels and account for the vast majority of

all the reads attributable to antibody genes in that sample[1,3]. We speculated that this phenomenon, which probably results from few B-cell clones winning the competition for the limited access to T follicular helper (Tfh) cells in the germinal center (GC) reaction[4], could be used to successfully pair heavy and light antibody chains from bulk RNA-Seq data, at least in some cases. Accordingly, we computationally assembled antibody chain sequences expressed in each TCGA tumor sample and we calculated the associated Berger–Parker index[10] in order to identify samples with particularly dominant antibody chains. We observed that a small, but not insignificant, proportion of the analyzed samples expressed highly dominant chains (Supplementary Figure 2).

Dominant antibody sequences can be correctly paired from bulk RNA-Seq in-silico

As a next step, we wanted to assess whether the dominance information captured by the Berger–Parker index could be used to correctly pair heavy and light chains from bulk RNA-Seq data, at least in some cases. In order to do so, we used RNA-Seq data from single B cells[11] to simulate bulk RNA-Seq samples for which the expression level, sequence, and correct pairing of all antibody chains was known. Using this approach, we established that for samples with dominance scores larger than 0.382 we could identify and correctly pair the dominant antibody chains with high accuracy (true positive rate 0.46, false positive rate 0.1, Figure 2, AUC=0.757). We used this result to select 191 paired intratumoral antibody sequences from our TCGA analysis for expression in mammalian cells and further characterization (Figure 3). As a control, we also selected for expression and characterization 92 sequences below our selected dominance score cutoff.

In-silico paired antibody sequences express at high levels in mammalian cells

We performed gene synthesis for the 283 selected antibody sequences and attempted expression of the corresponding proteins in mammalian cells (HEK293). For each antibody, we replaced the original heavy constant region with a standard human IgG1 in order to facilitate subsequent detection and screening. We successfully expressed 275 antibodies out of 283 with high yield (mean 2.35, sd 2.03, n=275, Supplementary Table S1). After expression, each antibody was purified using protein A, stored in PBS at a standard concentration of 1ug/uL and used for subsequent target antigen screening.

Identification of target antigens using high-throughput proteomics

Having successfully expressed the intratumoral antibodies we paired in-silico, we decided to use the purified antibodies to attempt identification of their target antigens. To this aim, we used high throughput proteomics to screen individual antibodies against a large collection of wild-type human proteins. We screened 173 antibodies against a collection of about twenty thousand human proteins, each represented in duplicate on the surface of a protein array, covering about 80% of the known human proteome[12]. While protein arrays are a cost effective way to test antibodies against a large number of potential targets, this technology is not designed to display membrane proteins in their correct conformation. Complex membrane proteins are unlikely to fold correctly when immobilized on the surface of the array, outside or their native membrane context. For this reason, we used high-throughput fluorescence-activated cell sorting (HT-FACS) to screen 92 antibodies against a library of five thousand selected membrane proteins expressed on the surface of mammalian cells[13]. As a result of these screenings, we were able to identify high confidence hits for 84 of our antibody candidates screened using protein arrays (48%) and 21 of our antibodies screened using HT-FACS (23%)**.** Reassuringly, targets recognized by our in-silico paired intratumoral antibodies included well known cancer-testis antigens[14] (NY-ESO-1, MAGEA3, GAGE2A) as well as immunomodulatory molecules expressed in the tumor microenvironment (ANXA1, TGFBI, C4BPB; Figure 4 and Supplementary tables 2, 4 and 5).

Intratumoral antibodies bind their target antigens with high-affinity

As a next step, we decided to independently confirm the interaction between selected intratumoral antibodies and their putative target antigens by using surface plasmon resonance (SPR)[15]. For each SPR experiment, we characterized the binding affinity of the tested antibody to its putative target antigen, having sourced the antigen independently of the vendor used for the high-throughput proteomics screening. We confirmed 13 antibody-antigen interactions listed in Supplementary Table S2. We observed that in-silico paired intratumoral antibodies bind to their target antigen with very high-affinity, with K_D in the low nanomolar range. When we compared our in-silico paired fully human antibodies against commercial antibodies obtained from rabbits (a model organism known to generate very high-affinity antibodies[16,17]) immunized with the same antigen, we found no significant differences in their binding affinity (Figure 4).

Intratumoral antibodies can be used to identify their epitope

Having demonstrated that our in-silico paired intratumoral antibodies bind to their target antigens with high-affinity, we decided to assess whether it would be possible to identify their epitope, at least in principle. We selected one antibody and performed epitope mapping using Hydrogen deuterium exchange mass spectrometry (HDX-MS)[18]. Using this technique, we were able to identify the putative binding region of the antibody on its designated target (C4BPB; Figure 5). We observed that the most likely epitope overlaps with the C4BPB binding site for protein S[19], which is crucial for its biological function[20]. Although additional functional studies would be required to clarify whether or not this particular antibody is able to interfere with the binding between C4BPB and protein S, our data shows that in principle it is possible to identify the target antigen and the corresponding epitope for in-silico paired intratumoral antibodies, thus gaining further insights into their possible biological function.

Discussion

Despite their compelling correlation with positive clinical outcomes, the functional role of intratumoral antibodies remains substantially unknown. In this study, we demonstrated for the first time that in-silico pairing can be used to obtain fully functional intratumoral antibodies from legacy tumor RNA sequencing data. Moreover, we showed that it is possible to use high-throughput proteomics techniques to identify the target antigens, characterize their binding kinetics, and discover the corresponding epitopes. These steps are crucial to enable further, much needed, functional characterization studies. Interestingly, we have demonstrated that highly clonal intratumoral antibodies are selected to bind not only cancer-specific antigens (NY-ESO-1, MAGEA3, GAGE2A), but also wild-type proteins expressed in the tumor microenvironment (ANXA1, TGFBI, C4BPB). Despite being directed against non-mutated self-antigens these antibodies bind to their target with very high-affinity, similarly to antibodies obtained by immunizing a different species with the same antigen. The importance of this observation is twofold: on one hand, it highlights the extent to which peripheral tolerance might be compromised during immune responses in cancer-affected tissues; on the other hand, it suggests that high-affinity fully-human antibodies against human proteins can be obtained by sequencing antibody transcripts expressed in tissues affected by chronic inflammation. While we have not

identified the target antigens for many of the antibodies produced in this study, it should be noted that our antigen screening only considered non-mutated proteins. The remaining orphan antibodies could be explained by errors in our in-silico pairing method. Alternatively, it is possible that some of these orphan antibodies bind to antigens for which we could not screen in this study, including neoantigens specific for particular patients, or non-protein antigens such as glycans[20,21]. Advancements in antigen screening methods might in the future allow the deorphaning of additional candidates. Nevertheless, this study generated the largest collection of individually screened fully human intratumoral antibodies generated so far, paves the way to improve our understanding of their functional role during anti-tumor immune responses, and suggests a novel way to extract new immunological insights from legacy RNA sequencing data.

Methods

Antibody reconstruction. For the purpose of reconstruction of antibody chain sequences from RNA sequencing data we implemented a computational workflow following the same basic idea underlying previously described approaches [11,22]. We reduced the number of sequencing reads not originating from immunoglobulin (Ig) transcripts through three filtering steps. First, we selected only the reads that either map to the immunoglobulin genes, or fail to map using Kallisto (version 0.44) [23] and Gencode [24] (version 22) protein coding sequences as the reference. Second, we mapped the remaining reads to the full human genome (version GRCh38, [24,25]), and retained only the reads that mapped to the immunoglobulin loci, or failed to map at all. In order to extract the selected reads and revert to FASTQ format we used Sambamba [26] view (version 0.5.9) and Samtools [27] Fastq (version 1.8) commands. Third, we depleted the reads originating from viral and bacterial sources by running kraken2[28] and retained the unclassified reads. We assembled the retained reads using Trinity RNA Seq assembler (version 2.8.4)[29] with custom parameters "--no_normalize_reads", "--max_chrysalis_cluster_size 100", and "--max_reads_per_graph 5000000". We then mapped the assembled sequences germline V, D, and J gene regions using IgBLAST[28,30] (version 1.13). Sequences with high V gene match scores (cutoff = 100) were kept as putative immunoglobulin chains.

Antibody reconstruction performance evaluation. To evaluate the performance of this workflow and evaluate the possibility of identifying which heavy and light chain form a pair in cases where there is a strongly expressed clone we generated a synthetic benchmark using single cell sequencing data for sample PW2 from Lindman et.al.[11]. We estimated the distribution of Ig reads in TCGA data and found it to closely follow a lognormal model (Supplementary Figure 1). We then generated a set of 1000 synthetic samples. We generated each sample by sampling 25 numbers from this distribution and assigning them as the total number of Ig reads to 25 random B cells from the PW2 data. We then randomly selected the corresponding number of RNA sequencing reads from those mapped to Ig chains in each of the 25 B cells. To this mix we added 10 million reads randomly selected from a bulk RNA sequencing sample (TCGA-04-1348-01A), to act as background. In evaluation, we considered the reconstruction to be a success if all of the CDR regions (CDR1, CDR2, and CDR3) for both the heavy and light chain of the top clone were correct and the pair was correctly selected as the most abundant antibody.

Antibody quantification, pairing, and somatic hypermutation analysis. For each processed TCGA sample we used Kallisto to quantify the amount of RNA sequencing reads originating from each of the putative immunoglobulin chain transcripts. We separately calculated the *Berger-Parker dominance index*[10] for both the heavy and light Ig chains, as the proportion of reads attributed to the most common chain of the corresponding type. We then calculated the dominance score for each sample as the geometric mean of the heavy and light chain Berger-Parker indices. Using the results of the synthetic benchmark we determined the threshold on this score (>0.382) above which we expect the sequence reconstruction to be accurate, and the top (most abundant) heavy and light chain to form a paired antibody. For each such antibody we assigned the antibody isotypes by mapping the constant portions of the heavy chains (sections of the assembled sequences post the annotated J regions) to the Gencode v22 reference immunoglobulin C sequences. We used the IgBlast mappings to the germline segment to identify the mutated positions in each chain. For every position as defined by the Martin numbering scheme[31] we calculated the mutation frequency across all of the reconstructed immunoglobulin chains.

Processing of TCGA RNA sequencing data. TCGA BAM files were retrieved and processed in Cancer Genomics Cloud[32] (CGC). We applied sambamba-sort to sort the BAM files and

196 Samtools fastq to convert them to FASTQ files, followed by Salmon quant in mapping-based
197 mode (GENOME v27 (GRCh38.p10) transcript assembly was used) with GC bias correction. We
198 obtained TPM values for 10,533 tumor samples and 730 normal samples across 33 TCGA cancer
199 types (Supplementary Table S3).

200 **Statistical analysis of antigen arrays.** Raw signal intensities (mean F635 values) were
201 correlated for background signal (mean B635 values) using the function background from R
202 package limma[33], with the correction method set to 'normexpr' with 'mle' parameter estimation
203 strategy. The replicate spots for the same target clone were summarized by calculating the
204 geometric mean followed by logarithmic transformation. We noticed that a number of spots
205 reported high signal intensities across multiple arrays likely due to non-specific binding. To
206 counter this issue we subtracted the mean signal value across all of the 172 arrays for each target
207 clone. We then centered and scaled the signal intensities and performed multiple testing
208 corrections by calculating the false discovery rate according to the Benjamini-Hochberg
209 method[33,34]. We used a stringent q-value cutoff of 0.01 in order to control the number of false
210 positive hits.

211 **Antibody synthesis.** The variable region of the antibodies were recombined with a constant
212 region of human IgG class I using AbAb's recombinant platform (Absolute Antibody Ltd,
213 Oxford, UK). The antibodies were expressed into HEK293 cells exploiting the Absolute
214 Antibody transient expression system and purified by one-step affinity chromatography. We
215 performed Quality Control (QC) analysis to assess concentration, total amount, level of
216 endotoxin, and sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)
217 (Supplementary Table S1).

218 **HuProt™ Protein arrays.** We tested each candidate antibody using the human proteome
219 microarray HuProt™ 4.0, a comprehensive library of GST-tagged recombinant human protein
220 expressed in yeast. Briefly, the HuProt arrays were blocked with 5% BSA/1×TBS-T at room
221 temperature for 1 h. Then antibodies were incubated at 4 °C overnight, washed and labelled with
222 secondary antibodies prior detection.

Membrane Proteome Array (MPA)[13]. The recombinant antibodies were further probed using MPA, a library that covers 94% of the plasma membrane proteome including G protein-coupled receptors (GPCRs), channels protein and transporters. MPA is a high-throughput cell based assay composed of 4,500 unique human membrane proteins expressed at the plasma membrane of a mammalian cell line. The recombinant antibodies were used to stain the unfixed cell-based-array and analyzed by flow-cytometry to determine their target and off-target binding.

Surface Plasmon Resonance (SPR) assay. A Biacore 8K instrument, exploiting the SPR principle, was used to calculate the dissociation constant (K_D). Selected antibodies (ligand) were individually immobilized to a sensor chip coated with Protein A to ensure the correct orientation of the antibody while the antigens (analyte) were injected within the flow stream at 30 µl/min. We used 90s association time for each experiment with 600s dissociation time. Data analysis was performed using the R package *pbm*[35].

Epitope mapping. Hydrogen deuterium exchange mass spectrometry (HDX-MS) was used to determine the epitope recognized by the antibodies. Linear peptides spanning the target antigen length were incubated in a deuterium containing buffer in presence or absence of the relevant antibody to observe differential exchange with hydrogen at the putative binding sites.

Data availability

The raw RNA sequencing reads used in this study are generated by the TCGA Research Network: https://www.cancer.gov/tcga, and available through the TCGA data portal. All relevant processed data are included as Source Data linked to this paper.

Code availability

The code for generation of the synthetic data used in the benchmarks, statistical analyses of the arrays, and plots is from the authors available upon a reasonable request.

Authors contributions

G.R., I.S., and B.C.T. developed the sequence reconstruction workflow and generated the Ig chain sequences. G.R. performed the performance evaluation of the workflow. D.B. and G.R. analysed the protein array data. A.L.S. prepared supplementary figures depicting sensorgram

plots for SPR assay experiments. Milena P. provided statistical analysis comparing K_D values of human vs. commercially available rabbit-derived antibodies. Milos P. and I.G. processed the C4BPB epitope mapping and generated the visualization. D.B. and D.K. designed research and D.B. oversaw the implementation. G.R., D.B., and D.A.L. prepared the manuscript.

Competing interests

G.R., I.G., B.C.T., Milena P., Milos P., A.L.S., D.K., and D.B. are participants in Totient, Inc.'s Equity Incentive Plan.

References

1. Hu, X. *et al.* Landscape of B cell immunity and related immune evasion in human cancers. *Nat. Genet.* **51**, 560–567 (2019).
2. Zhao, J. *et al.* Author Correction: Immune and genomic correlates of response to anti-PD-1 immunotherapy in glioblastoma. *Nat. Med.* **25**, 1022 (2019).
3. Bolotin, D. A. *et al.* Antigen receptor repertoire profiling from RNA-seq data. *Nat. Biotechnol.* **35**, 908–911 (2017).
4. De Silva, N. S. & Klein, U. Dynamics of B cells in germinal centres. *Nat. Rev. Immunol.* **15**, 137–148 (2015).
5. Helmink, B. A. *et al.* B cells and tertiary lymphoid structures promote immunotherapy response. *Nature* **577**, 549–555 (2020).
6. Cabrita, R. *et al.* Author Correction: Tertiary lymphoid structures improve immunotherapy and survival in melanoma. *Nature* **580**, E1 (2020).
7. Petitprez, F. *et al.* B cells are associated with survival and immunotherapy response in sarcoma. *Nature* **577**, 556–560 (2020).
8. Siliņa, K. *et al.* Germinal Centers Determine the Prognostic Relevance of Tertiary Lymphoid Structures and Are Impaired by Corticosteroids in Lung Squamous Cell

Carcinoma. *Cancer Res.* **78**, 1308–1320 (2018).

9. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* vol. 45 1113–1120 (2013).

10. Berger, W. H. & Parker, F. L. Diversity of Planktonic Foraminifera in Deep-Sea Sediments. *Science* vol. 168 1345–1347 (1970).

11. Lindeman, I. *et al.* BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nat. Methods* **15**, 563–565 (2018).

12. Jeong, J. S. *et al.* Rapid identification of monospecific monoclonal antibodies using a human proteome microarray. *Mol. Cell. Proteomics* **11**, O111.016253 (2012).

13. Tucker, D. F. *et al.* Isolation of state-dependent monoclonal antibodies against the 12-transmembrane domain glucose transporter 4 using virus-like particles. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4990–E4999 (2018).

14. Chen, Y.-T. *et al.* Genomic cloning and localization of CTAG, a gene encoding an autoimmunogenic cancer-testis antigen NY-ESO-1 to human chromosome Xq28. *Cytogenetic and Genome Research* vol. 79 237–240 (1997).

15. Oliveira, L. C., Lima, A. M. N., Thirstrup, C. & Neff, H. F. *Surface Plasmon Resonance Sensors: A Materials Guide to Design, Characterization, Optimization, and Usage*. (Springer, 2019).

16. Weber, J., Peng, H. & Rader, C. From rabbit antibody repertoires to rabbit monoclonal antibodies. *Exp. Mol. Med.* **49**, e305 (2017).

17. Peng, H. *et al.* Mining Naïve Rabbit Antibody Repertoires by Phage Display for Monoclonal Antibodies of Therapeutic Utility. *J. Mol. Biol.* **429**, 2954–2973 (2017).

18. Masson, G. R. *et al.* Recommendations for performing, interpreting and reporting hydrogen

297 deuterium exchange mass spectrometry (HDX-MS) experiments. *Nat. Methods* **16**, 595–602
298 (2019).
299 19. Fernández, J. A. & Griffin, J. H. A protein S binding site on C4b-binding protein involves
300 beta chain residues 31-45. *J. Biol. Chem.* **269**, 2535–2540 (1994).
301 20. Rezende, S. M., Simmonds, R. E. & Lane, D. A. Coagulation, inflammation, and apoptosis:
302 different roles for protein S and the protein S–C4b binding protein complex. *Blood* vol. 103
303 1192–1201 (2004).
304 21. Dube, D. H. & Bertozzi, C. R. Glycans in cancer and inflammation — potential for
305 therapeutics and diagnostics. *Nature Reviews Drug Discovery* vol. 4 477–488 (2005).
306 22. Upadhyay, A. A. *et al.* BALDR: a computational pipeline for paired heavy and light chain
307 immunoglobulin reconstruction in single-cell RNA-seq data. *Genome Med.* **10**, 20 (2018).
308 23. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
309 quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
310 24. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE
311 Project. *Genome Res.* **22**, 1760–1774 (2012).
312 25. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies
313 demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864
314 (2017).
315 26. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing
316 of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
317 27. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
318 2078–2079 (2009).
319 28. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2.

Genome Biol. **20**, 257 (2019).

29. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* vol. 29 644–652 (2011).

30. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34–40 (2013).

31. Abhinandan, K. R. & Martin, A. C. R. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Mol. Immunol.* **45**, 3832–3839 (2008).

32. Lau, J. W. *et al.* The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research. *Cancer Res.* **77**, e3–e6 (2017).

33. Silver, J. D., Ritchie, M. E. & Smyth, G. K. Microarray background correction: maximum likelihood estimation for the normal-exponential convolution. *Biostatistics* **10**, 352–363 (2009).

34. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* vol. 57 289–300 (1995).

35. Davies, J. Protein Binding Models [R package pbm version 1.1.0]. (2018).

Figures

Fig. 1 | Antibody reconstruction and characterization protocol.
a. Computational workflow using RNA sequencing data from tumor samples as the input. Initially, the reads not originating from immunoglobulin transcripts are depleted; in a second step, the immunoglobulin chain sequences are computationally assembled and used to identify the pairs of chains that form clonally dominant antibodies. **b**. Experimental workflow using computationally paired antibodies as the input. First, the reconstructed antibodies are expressed; the putative targets are then identified via protein arrays or membrane protein screen. Finally, the detected antibody-antigen interactions are confirmed and characterised using surface plasmon resonance.

Fig. 2 | Evaluation of reconstruction performance on synthetic data.
a. Distribution of correctly and incorrectly reconstructed samples depending on the dominance score (geometric mean of the Berger-Parker indices for heavy and light Ig chains) estimated by the workflow. Above the threshold of 0.382 the top antibodies for 90% of the synthetic samples were correctly reconstructed. **b**. ROC curve for the evaluation at different values of the dominance score. The red cross marks the dominance score of 0.382 selected as the point with the highest true positive rate (0.46) where the false positive rate is below 0.1.

Fig. 3 | Immunoglobulin isotypes, SHM, and chain pairing in the selected set of antibodies.
a. Distribution of the antibodies across TCGA cancer types. **b**. Distribution of amino acid lengths of IgH CDR3 regions. **c**. Number of selected antibodies by isotype. **d**. Different germline V segments for the heavy and light chains are used at varying frequencies. Thickness of the line linking a heavy and a light V segment represents the relative frequency with which the two are paired. **e.** and **f**. Mean SHM rate for each position in the Martin numbering scheme across the heavy and light chains in the selected set of antibodies. For each chain, per amino-acid mutation rates were estimated by mapping the chains to germline sequence constructs by IgBlast and numbered following the Martin numbering scheme. For chains where multiple amino acids map

363 to the same Martin number we used the mean value across those amino acids.

364 **Fig. 4 | Distribution of K_D values for Totient antibodies and commercially available**
365 **antibodies against the same antigens.**
366 **a**. Distribution of empirically determined K_D values for human-derived antibodies (Abs) and
367 their rabbit-derived counterparts for the same antigens, showing no statistically significant
368 difference using a paired t-test. As Abs have been tested in multiple experiments, to calculate the
369 p-value, we first averaged the -log10(K_D) across the experiments with a specific analyte
370 (different source of antigen), then paired each of these averages from human-derived Abs with
371 averages from rabbit-derived Abs for the same analyte and applied paired t-test. **b**. K_D values for
372 human and rabbit-derived antibodies presented in a. When multiple experiments have been
373 performed for a human or rabbit Ab, we calculated the average K_D from these experiments. For
374 detailed information about each experiment consult Supplementary Table S2.

375 **Fig. 5 | Epitope mapping results for C4BPB show overlap with protein S binding site.**
376 HDX-MS was used to measure the level of deuterium (D) uptake by C4BPB alone or in presence
377 of CYC214 antibody. **a.** Relative D uptake difference per residue (blue shading) across the entire
378 protein surface. **b**. Details of the protein region containing the known binding site[20] for protein S.
379 High D uptake difference (dark blue) is detected in the region containing the binding site for
380 protein S, thus suggesting that CYC214 might be disrupting the interaction between C4BPB and
381 protein S. Each C4BPB fragment is measured up to three times and each residue can be covered
382 by one or more overlapping fragments: the uptake difference per residue was calculated using the
383 mean of the uptake differences of the fragments covering the residue.

Supplementary figures

Figure S1 | Distribution of Ig reads mapping to different chains in TCGA data follows a lognormal model.

A lognormal distribution is used to fit the counts of reads mapping to different Ig chains in the TCGA data (estimated mean=5.78, standard deviation=1.52).

Figure S2 | Distribution of Berger-Parker index across different cancer types in the TCGA data

Berger-Parker index from the heavy and the light chains in the TCGA data. Some samples exhibit top clone dominance for both chains.

Figure S3 | Surface plasmon resonance assay sensorgram plots providing antibody-antigen dissociation constants

Sensorgrams plots depict the SPR response of the Biacore 8K instrument in relative units (RU) vs. time, as commercially sourced antigens are flowed across surface-captured antibodies during an initial 90 s measurement interval, and then permitted to dissociate without antigen flowing, during a subsequent 600 s measurement interval. Antigens are prepared across a range of two-fold serially diluted concentrations, and within a given set of concentration curves comprising a single measurement attempt, data from all concentrations are fitted simultaneously using non-linear least squares regression to arrive at a single unifying K_D estimate (or a pair of K_D estimates, when a biphasic binding model provides a better fit) valid for that measurement attempt. Solid colored lines indicate raw data observed by the apparatus, and overlaid solid black lines indicate the fit result estimated using the model; a closer match between the two indicates a better fit, with correspondingly superior explanatory power. In some cases, indicated by the absence of an overlaid solid black regression curve, data corresponding to one or more

concentration curves may be discarded from the input to the regression model, if the result leads to a significantly better fit across the remaining concentration curves. Parameter estimates provided by the fit model are listed in the table at the top of each page, while lettered panels enumerate cases in which multiple replicate measurements were attempted for a given antibody-antigen pairing.

Figure S4 | Binding scores for top 5 hits from protein array screening.
Plots depict the scaled binding scores from top 5 hits from each of the protein arrays. Binding scores were calculated by fitting random effects models to the signal intestines, weighting each target by its mean rank across the arrays (to eliminate common promiscuous binders) and scaled by taking the z scores (Methods section). Asterisks mark hits with Benjamini and Hochberg false discovery rate < 0.01.

Figure S5 | Figure representing the results of the Integral molecular screening.
Each plot depicts the binding score detected by the Integral molecular MPA screening for one candidate antibody. Each dot represents a membrane protein screened during the assay. Blue dots positive binding sites above assay background according to the assay vendor. The Y axis represents the MPA binding score.

Supplementary tables

Table S1 | Antibody expression and antigen screening summary.

Table S2 | Biacore results.

Table S3 | Gene expression table TCGA (Salmon, TPM, no other normalization).

Table S4 | Protein array results.

Table S5 | Integral Molecular results

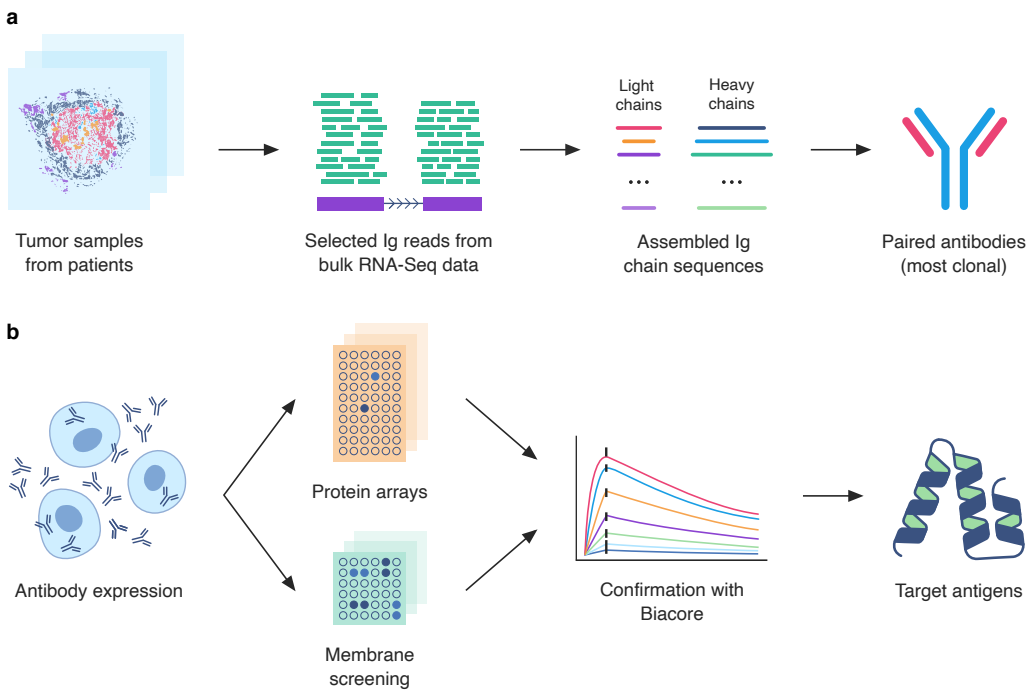


Fig. 1 | Antibody reconstruction and characterization protocol.

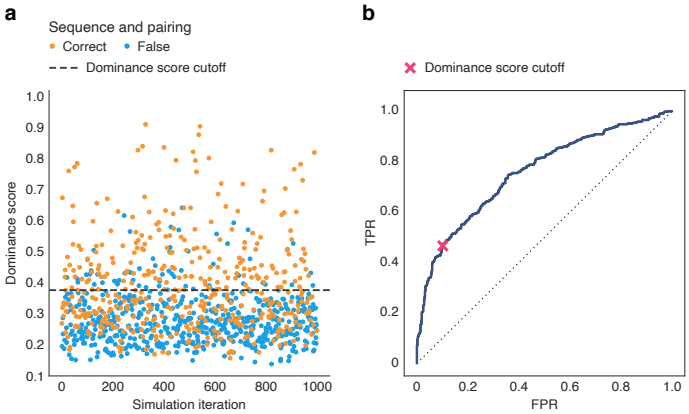


Fig. 2 | Evaluation of reconstruction performance on synthetic data.

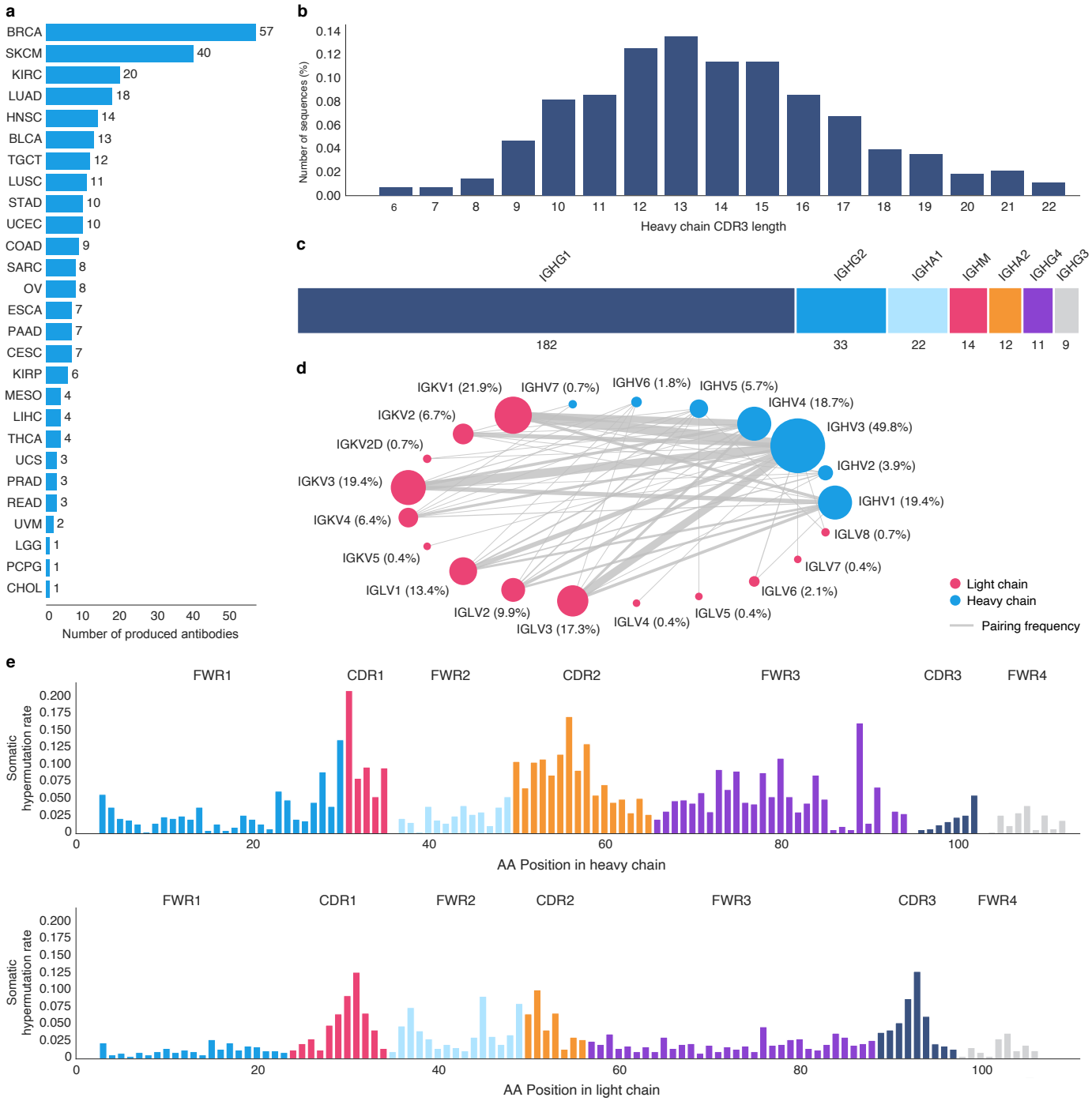
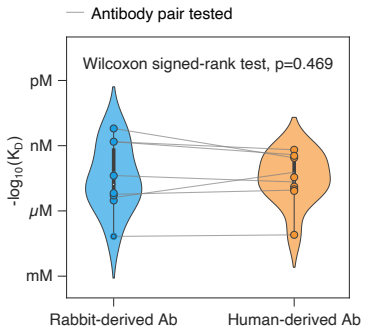


Fig. 3 | Immunoglobulin isotypes, SHM, and chain pairing in the selected set of antibodies.

a



b

Antigen	Analyte	Human-derived Ab	K_D (Human-derived Ab), nM	Rabbit-derived Ab	K_D (Rabbit-derived Ab), nM
ANXA1	APrEST71575	CYC188	7.333	HPA011271	7.629
ANXA1	APrEST71576	CYC188	4.900	HPA011271	4.828
GAGE2A	AR51161PU-S	CYC014	8.812	HPA043232	9.183
GAGE2A	AR51161PU-S	CYC025	8.583	HPA043232	9.183
MAGEA3	ab99958	CYC066	8.442	ab223162	9.792
TGFBI	TP300411	CYC168	7.777	HPA008612	6.615
TGFBI	TP300411	CYC168	7.777	HPA017019	6.748
TXLNA	TP310301	CYC086	6.958	HPA045383	6.766

Fig. 4 | Distribution of KD values for Totient antibodies and commercially available antibodies against the same antigens.

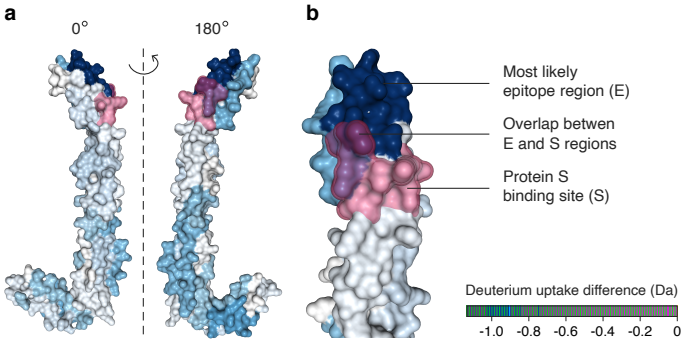


Fig. 5 | Epitope mapping results for C4BPB show overlap with protein S binding site.