nature biotechnology

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41587-020-0466-7

In the format provided by the authors and unedited.

High-throughput single-cell activity-based screening and sequencing of antibodies using droplet microfluidics

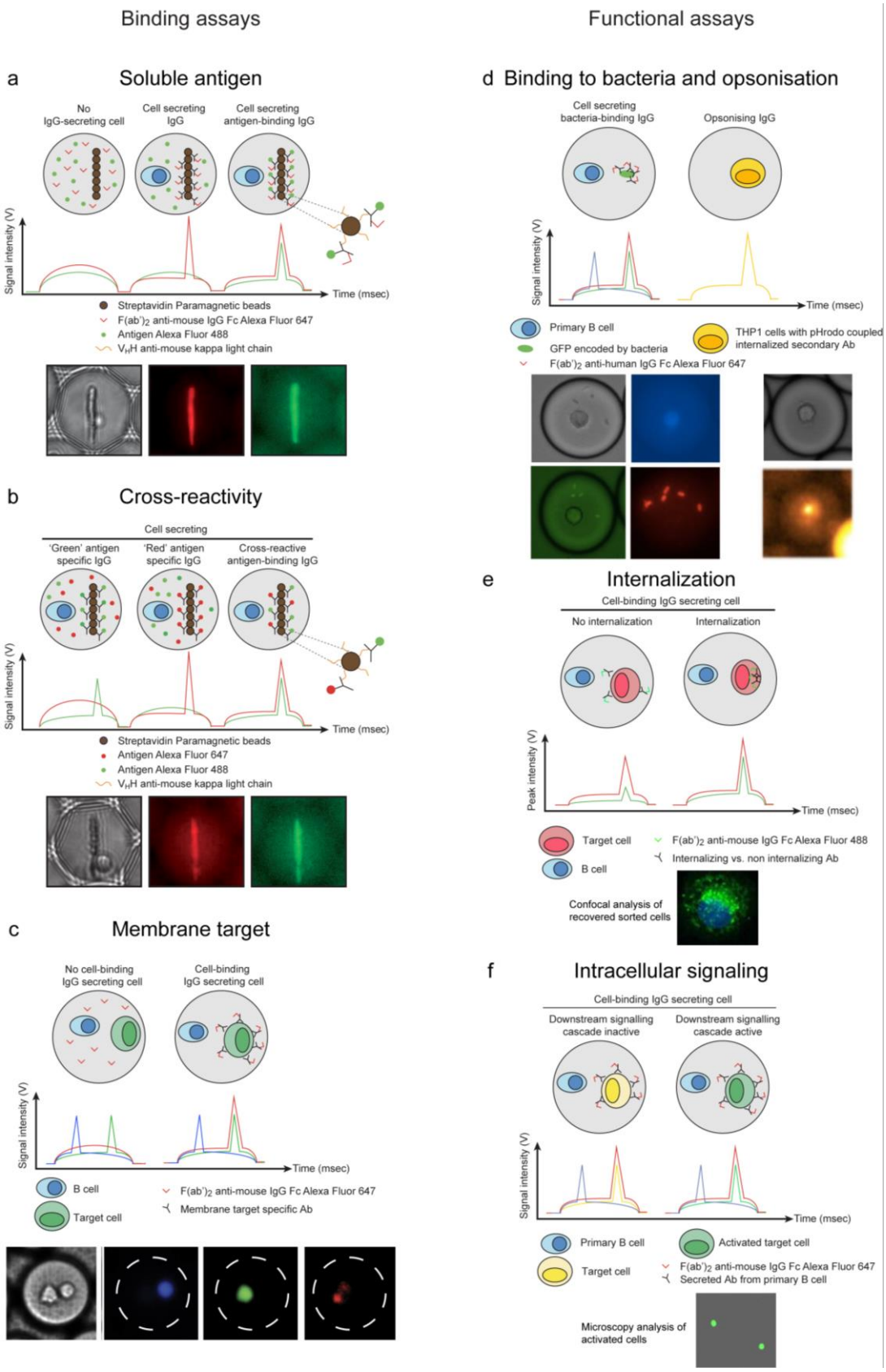
Annabelle Gérard[1,18], Adam Woolfe[1,17], Guillaume Mottet[2,17], Marcel Reichen[1,17], Carlos Castrillon[2,3,4,17], Vera Menrath[1,17], Sami Ellouze[1,17], Adeline Poitou[1,17], Raphaël Doineau[1,3,4,11,17], Luis Briseno-Roa[1,12], Pablo Canales-Herrerias[2,3,5], Pascaline Mary[6], Gregory Rose[6], Charina Ortega[6], Matthieu Delincé[6], Sosthene Essono[6], Bin Jia[7,13], Bruno Iannascoli[2], Odile Richard-Le Goff[2], Roshan Kumar[6], Samantha N. Stewart[6], Yannick Pousse[1], Bingqing Shen[1], Kevin Grosselin[1,4], Baptiste Saudemont[2,4], Antoine Sautel-Caillé[4], Alexei Godina[4], Scott McNamara[1], Klaus Eyer[5,14], Gaël A. Millot[8], Jean Baudry[5], Patrick England[9,10], Clément Nizak[4], Allan Jensen[1,15,18], Andrew D. Griffiths[4,18] ✉, Pierre Bruhns[2,18] ✉ and Colin Brenan[1,6,16,18] ✉

[1]HiFiBiO Therapeutics SAS, Paris, France. [2]Unit of Antibodies in Therapy and Pathology, Institut Pasteur, UMR1222 INSERM, Paris, France. [3]École Doctorale Frontières du Vivant (FdV), Centre de Recherches Interdisciplinaires, Paris, France. [4]Laboratoire de Biochimie (LBC), ESPCI Paris, PSL Research University, CNRS UMR8231 Chimie Biologie Innovation, Paris, France. [5]Laboratoire Colloïdes et Matériaux Divisés (LCMD), ESPCI Paris, PSL Research University, CNRS UMR8231 Chimie Biologie Innovation, Paris, France. [6]HiFiBiO Therapeutics Inc., Cambridge, MA, USA. [7]Pfizer, Cambridge, MA, USA. [8]Institut Pasteur, Hub Bioinformatique et Biostatistique, C3BI, USR 3756 IP CNRS, Paris, France. [9]Plateforme de Biophysique Moléculaire, Institut Pasteur, CNRS UMR3528, Paris, France. [10]Centre d'Innovation et de Technologie, Plateforme de Biophysique des Macromolecules et de leurs Interactions, Institut Pasteur, Paris, France. [11]Present address: Institute for Integrative Nanosciences, Leibniz IFW Dresden e.V., Dresden, Germany. [12]Present address: Medetia Pharma, Institut Imagine, Paris, France. [13]Present address: Abbvie Bioresearch Center, Worcester, MA, USA. [14]Present address: Laboratory for Functional Immune Repertoire Analysis, Institute of Pharmaceutical Sciences, D-CHAB, ETH Zürich, Zurich, Switzerland. [15]Present address: Lundbeck A/S, Valby, Denmark. [16]Present address: 1CellBiO Inc., Watertown, MA, USA. [17]These authors contributed equally: Adam Woolfe, Guillaume Mottet, Marcel Reichen, Carlos Castrillon, Vera Menrath, Sami Ellouze, Adeline Poitou, Raphaël Doineau. [18]These authors jointly supervised this work: Annabelle Gérard, Allan Jensen, Andrew D. Griffiths, Pierre Bruhns, Colin Brenan. ✉e-mail: andrew.griffiths@espci.fr; pierre.bruhns@pasteur.fr; c.brenan@1cell-bio.com

High-throughput single-cell activity-based screening and sequencing of antibodies using droplet microfluidics

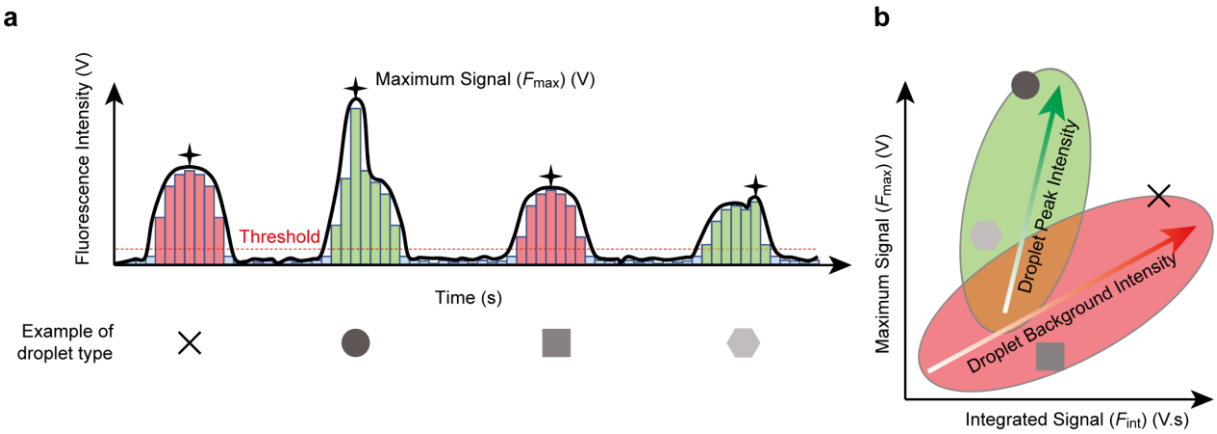
Supplementary Information

Supplementary Figures



Supplementary Figure 1. Example of assays performed in droplets. Droplet microfluidics enables a wide range of assays, ranging from simple antigen binding using soluble, fluorescently labelled, antigen

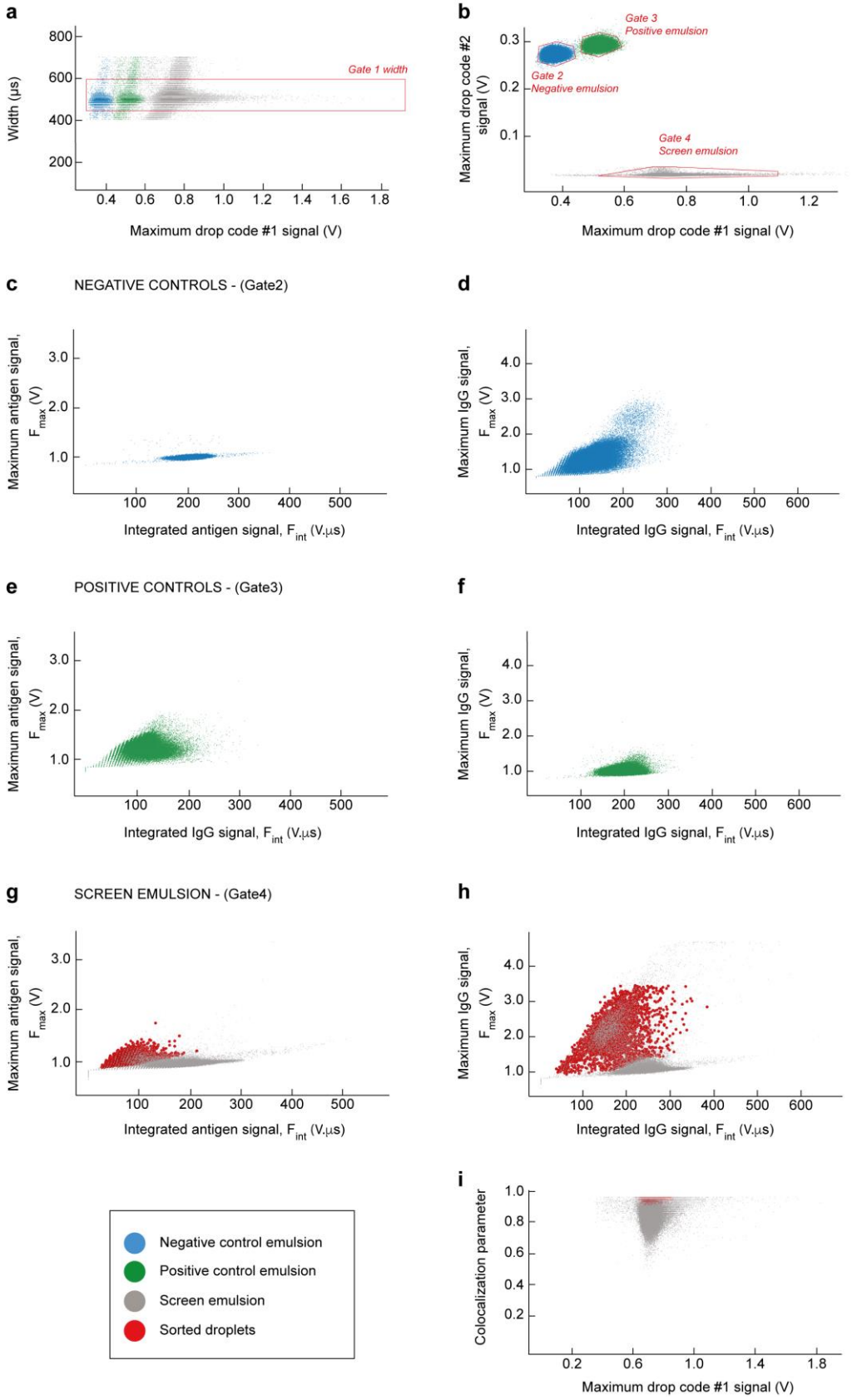
to sophisticated cell-based functional assays. Here we show how we have successfully deployed the platform for both binding assays (left) and functional assays. Each panel shows a schematic of the assay (top), a schematic of the time trace generated in up to 3 colors (blue, green, red) when the droplets are scanned one-by-one as they pass the laser line, and bright field and fluorescence micrographs of droplets (bottom). (**a**) A magnetic beadline assay for soluble antigen (here TT), as detailed in the manuscript for TT and GPI. (**b**) Cross-reactivity assays where soluble antigen (here the extra-cellular domain of a transmembrane glycoprotein) from one species (here human) is labelled with a green fluorophore and the same antigen from a second species (here cynomolgus monkey) is labelled with a red fluorophore. Co-localization of green and red fluorescence to the beadline is indicative of a cross-reactive antibody and has been used as criteria to enrich for cells secreting such antibodies. This assay can also be adapted to identify antibodies that specifically bind one antigen and not another, in which case there is a high ratio of red to green fluorescence (or *vice versa*) on the beadline (**c**) Cell-based assay to detect binding to a membrane-bound antigen (here TSPAN8), as detailed in the manuscript for TSPAN8. (**d**) Antibodies binding gram negative bacterial cells secreted by activated memory B cells (blue fluorescent) from human patients are detected by using multi-peak detection to detect localization of red fluorescent $F(ab')_2$ anti-IgG Fc onto multiple bacterial cells (green) expressing green fluorescent protein (GFP) (left). Direct detection of antibodies inducing opsonization using a functional assay based on detecting acidification of THP1 derived macrophages using pHrodo™ red AM, Thermofisher, (which becomes orange fluorescent at acidic pH) in 100 pL droplets (right) (**e**) Cell-based assay for internalizing antibodies (against a type I transmembrane receptor) based on the higher concentration of internalized antibodies labeled with green fluorescent $F(ab')_2$ anti-IgG Fc in CHO reporter cells, labelled with Deep Red Cell tracker, compared to antibodies on the surface of the cells, which results in a higher green fluorescence peak. (**f**) Detection of a functional cell-based assay for TNFR1 Htr9 agonist antibodies using Jurkat reporter cells stably expressing GFP under control of the NFκB responsive promoter, where GFP expression is monitored and detected in the droplet after 4h in the droplets.



Supplementary Figure 2. Signal analysis method for the fluorescence relocation-based immunoassay for both the magnetic beadline bio-assay for soluble antigen and the cell-based bio-assay for membrane antigen. (**a**) Schematic representation of the raw fluorescence intensities (ordinate) of four distinct droplets flowing across the acquisition point (laser line) as a function of time (abscissa): two droplets showing no fluorescence relocation to the beadline or reporter cell, but different average fluorescence intensity (red bars) and two showing different levels of fluorescence relocation to the beadline or reporter cell (green bars). Droplet sorting is triggered using a two-parameters method. The first parameter is the maximum fluorescence intensity, F_{max} (V). The second parameter is the integrated fluorescence signal of the droplet, F_{int} (V.s), where

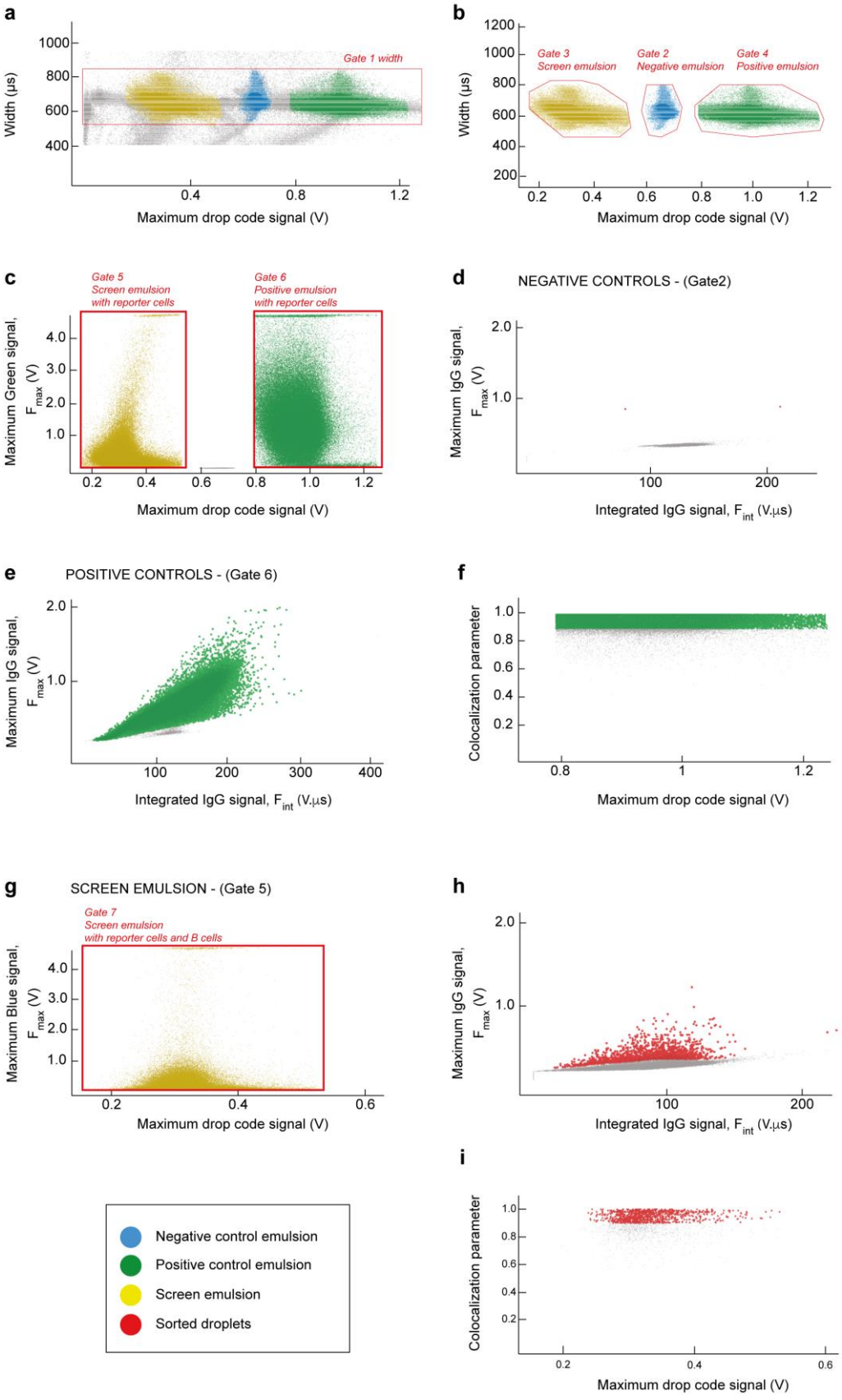
$$F_{int} = \sum_{I>I_{Th}} \frac{I}{f}$$

I is fluorescent intensity at a defined time (V), I_{Th} is a user defined threshold value (V) and f is the acquisition frequency (here 200 kHz). $I_{Th}$is set by the user to be above the inter-droplet fluorescence noise and below the non-relocated fluorescence signal. (**b**) Schematic representation of data distribution as a function of F_{max} and F_{int}. Two distinct classes of droplets are defined: (red) the droplets without fluorescence relocation to the beadline or reporter cell, which have a characteristic approximately Gaussian fluorescence profile and (green) the droplets which display a peak of fluorescence due to relocation of fluorescence to the beadline or reporter cell. The increase in the ratio F_{max}/F_{int} in droplets where there is fluorescence relocation allows them to be reliably sorted even in populations of droplets with differences in average fluorescence intensity.



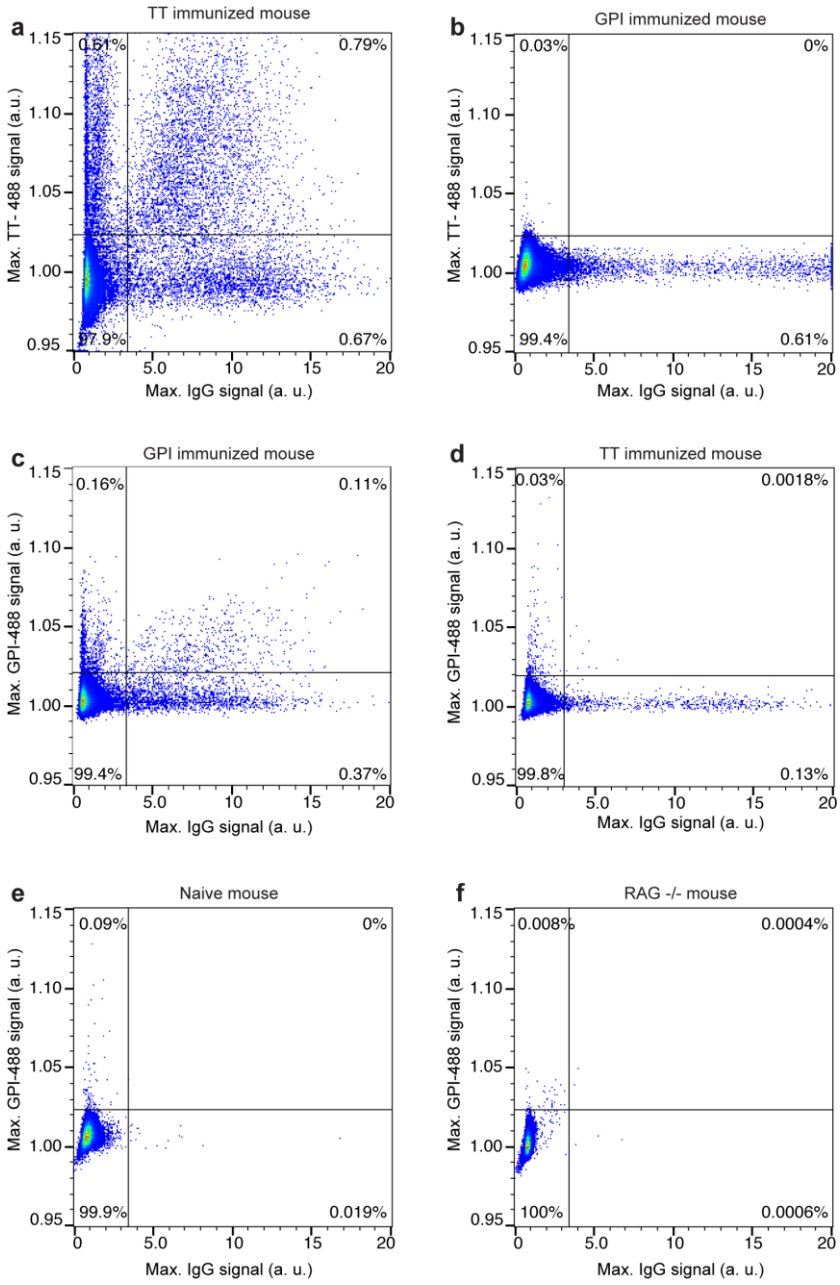
Supplementary Figure 3. Gating strategy for enriching primary B cells producing IgG binding the soluble antigen of interest (here TT). The droplets to be sorted were selected by a sequential gating strategy. (**a**) The droplets were first gated (gate 1) to eliminate coalesced droplets and retain only

droplets of the desired size. (**b**) The droplets in gate 1 were then assigned to one of three gates using the two-color optical droplet barcoding system (200, 400 or 800 nM Sulforhodamine B as Drop code#1 and 0 or 300 nM Dy405 as Drop code#2): gate 2 defines droplets containing 10 nM mouse anti-HRP IgG (Jackson Immuno Research, #223-005-024) (negative control), gate 3 defines droplets containing 10 nM chimeric anti-TT IgG TT10 (positive control, K_d = 0.6 nM, chimeric antibody comprising the V_H and V_L domains of the human antibody SRA:SRR611803.2792688.2 (DeKosky, B.J. et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nature Biotechnology 31, 166-169 (2013).), grafted onto murine IgG2a heavy chain and kappa light chain), and gate 4 defines droplets from the screen emulsion containing enriched B cells. (**c-i**) For the droplets in each of gates 2 to 4, fluorescence relocation to the beadline in droplets was measured by plotting the maximum fluorescence signal (F_{max}) against the integrated fluorescence signal (F_{int}) from the droplets. Relocation of AlexaFluor647 labeled rabbit F(ab')$_2$ anti-mouse IgG Fc (red fluorescent) (left) and AlexaFluor488 labeled TT (green fluorescent) (right) results in an increase in the ratio F_{max}/F_{int} (Supplementary Fig. 8). Droplets containing mouse anti-HRP IgG (negative control) are shown in panel (**c,d**), droplets containing chimeric anti-TT10 IgG (positive control) are shown in panel (**e,f**), and droplets from the screen emulsion containing enriched B cells are shown in panel (**g,h**). Uncoalesced droplets (gate 1) containing splenocytes (gate 4) were sorted (red points) if they satisfied all of the following criteria (see Online Methods): i) relocation of antigen (green fluorescence) to the beadline, ii) relocation of the F(ab')$_2$ anti-mouse IgG Fc (red fluorescence) to the beadline, and iii) colocalisation of the green and red fluorescence peak (**i**). The colocalization parameter, c, was calculated from the time interval in between the peaks in the red and green fluorescence channels, t_p, and the time interval from the beginning to the end of the droplet, t_d,: $c = (1 - t_p)/ t_d$. The colocalization value was thus bounded between 0 and 1, 1 being the perfect colocalization of the two peaks. Droplets were sorted if $c > 0.95$.

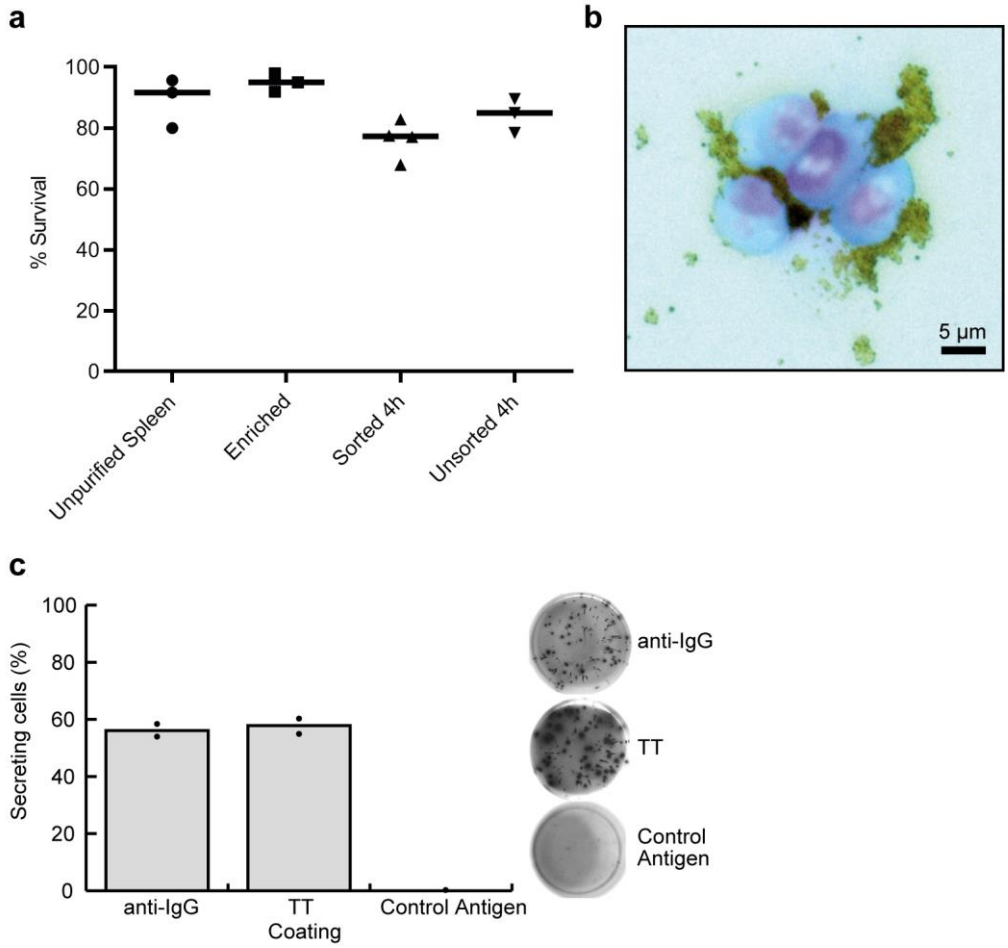


Supplementary Figure 4. Gating strategy for enriching primary B cells producing IgG binding the membrane-expressed antigen of interest (here human TSPAN8). The droplets to be sorted were selected by a sequential gating strategy. (**a**) The droplets were first gated (gate 1) to eliminate coalesced

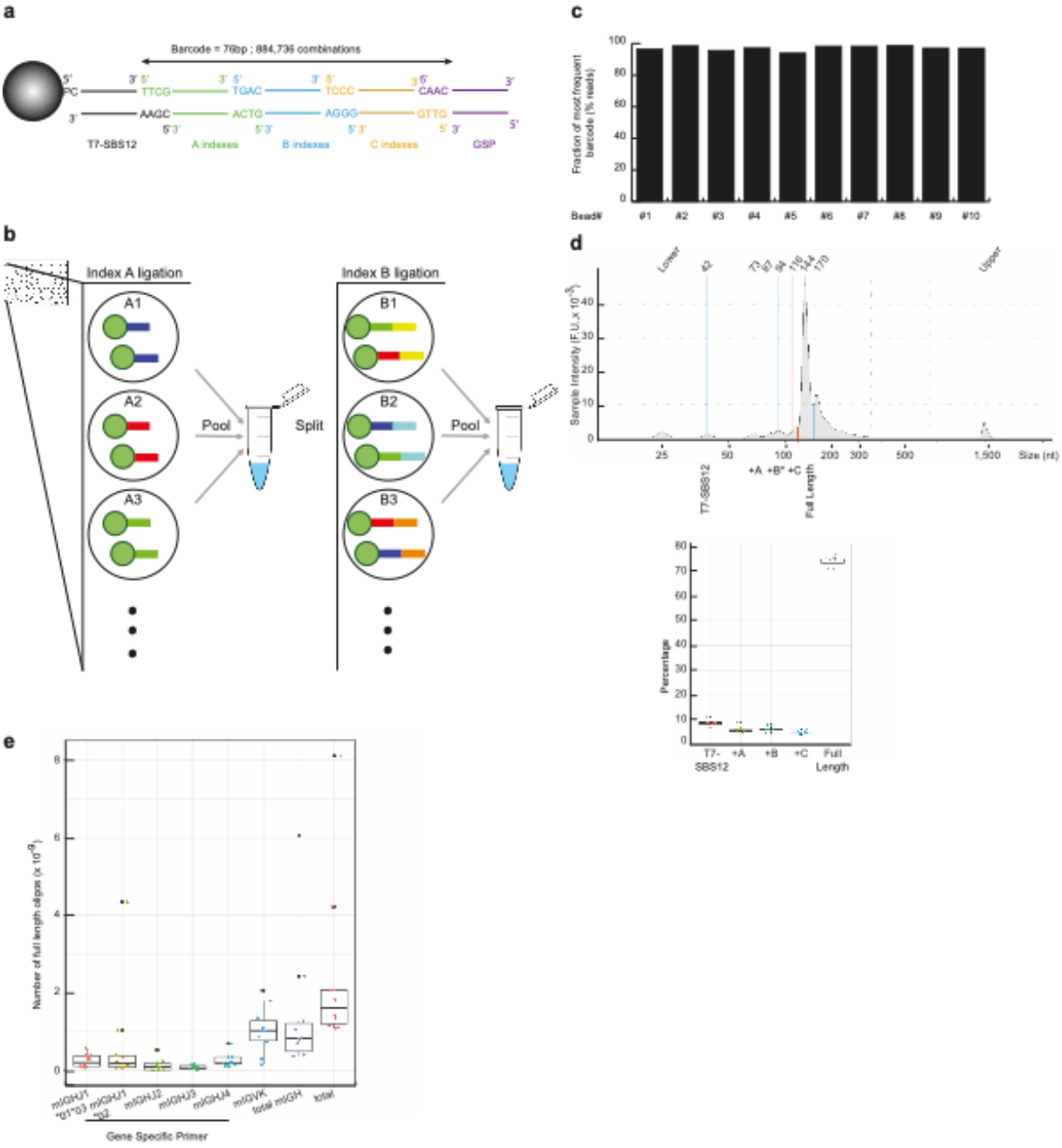
droplets and retain only droplets of the desired size. (**b**) The droplets in gate 1 were then assigned to one of three gates using a one-color optical droplet barcoding system (400, 800 or 1200 nM Sulforhodamine B as drop code); gate 2 defines droplets without reporter cells but containing 10 nM mouse anti-perodixase IgG (Jackson Immuno Research, #223-005-024) (negative control, 800 nM Sulforhodamine B); gate 4 defines droplets containing reporter cells overexpressing human TSPAN8 and 50 nM anti-hTSPAN8 51D3 antibody (Pfizer) (positive control, 1200 nM Sulforhodamine B), and gate 3 defines droplets from the screen emulsion containing reporter cells overexpressing human TSPAN8 and enriched B cells (400 nM Sulforhodamine B). (**c**) Droplets containing a reporter cell overexpressing human TSPAN8 were selected by plotting the maximum fluorescence signal (F_{max}) of CalceinAM Green against the maximum fluorescence signal (F_{max}) of the drop code in the screen and positive control emulsions (gates 5 and 6, respectively). (**g**) Droplets in the screen emulsion containing reporter cells additionally containing enriched B cells were selected by taking the droplets from gate 5 and plotting the maximum fluorescence signal (F_{max}) of Calcein Violet against the maximum fluorescence signal (F_{max}) of the drop code (gate 7). (**d, e, h**) For the droplets in each of gates 2, 6 and 7, fluorescence relocation to the reporter cell in droplets was measured by plotting the maximum fluorescence signal (F_{max}) against the integrated fluorescence signal (F_{int}) from the droplets. Relocation of AlexaFluor647 labeled rabbit F(ab')$_2$ anti-mouse IgG Fc (red fluorescent) results in an increase in the ratio F_{max}/F_{int} (Supplementary Fig. 8). Negative control droplets are shown in panel (**d**), droplets containing 50 nM anti-TSPAN8 51D3 antibody (positive control) are shown in panel (**e**), and droplets from the screen emulsion containing enriched B cells are shown in panel (**h**). (**f, i**) Colocalization of the green (reporter cell) and the red (AlexaFluor647 labelled rabbit F(ab')$_2$ anti-mouse IgG Fc) signals in positive control and screen emulsion. Droplets showing relocation of AlexaFluor647 labelled rabbit F(ab')$_2$ anti-mouse IgG Fc (red fluorescent) on the reporter cells from the positive control and screen emulsion are shown in panel (**f**) and (**i**) respectively. Uncoalesced droplets (gate 1) containing splenocytes (gate 3) were sorted (red points) if they satisfied all of the following criteria (see Online Methods): i) presence of a reporter cell overexpressing TSPAN8 (green fluorescence) in the droplet, ii) presence of an primary B cell (violet fluorescence) in the droplet, iii) relocation of the F(ab')$_2$ anti-mouse IgG Fc (red fluorescence) onto the reporter cell, and iv) colocalization of the green and red fluorescence peaks (**i**). The colocalization parameter, c, was calculated from the time interval in between the peaks in the red and green fluorescence channels, t_p, and the time interval from the beginning to the end of the droplet, t_d,: $c = (1 - t_p)/ t_d$. The colocalization value was thus bounded between 0 and 1, 1 being the perfect colocalization of the two peaks. Droplets were sorted if $c > 0.92$.



Supplementary Figure 5. Specificity of the bioassay for soluble antigen (i.e. using the beadline assay). Normalized maximum antigen fluorescence (AlexaFluor488 labeled TT or GPI) and normalized maximum IgG fluorescence (AlexaFluor647 labeled rabbit $F(ab')_2$ anti-mouse IgG Fc) of droplets loaded with splenocytes isolated from mice immunized with (**a, d**) TT or (**b, c**) GPI, (**e**) from a naive mouse and (**f**) from a B cell deficient ($RAG^{-/-}$) mouse. The percentage of droplets in each quadrant is indicated. As expected, with splenocytes from a $RAG^{-/-}$ mouse, very few droplets were positive for antibody secretion (0.001%). We detected 0.019% of droplets positive for IgG production using splenocytes isolated from the naive mouse: 20-fold higher than a $RAG^{-/-}$ mouse. This low frequency is also expected considering the low antigen exposure of a clean animal facility. For immunized mice the percentage of IgG-producing cells was 24-77 fold higher than for the naive mouse and a large percentage of these were antigen-specific (23% and 54% for GPI-immunized and TT-immunized mice, respectively). (**b**) Splenocytes from a mouse immunized with GPI were also assayed using fluorescent TT, and (**d**) splenocytes from a mouse immunized with TT were assayed using fluorescent GPI. In both cases the percentage of cells secreting antigen-specific IgG was very low: 0.018% and 0% respectively.

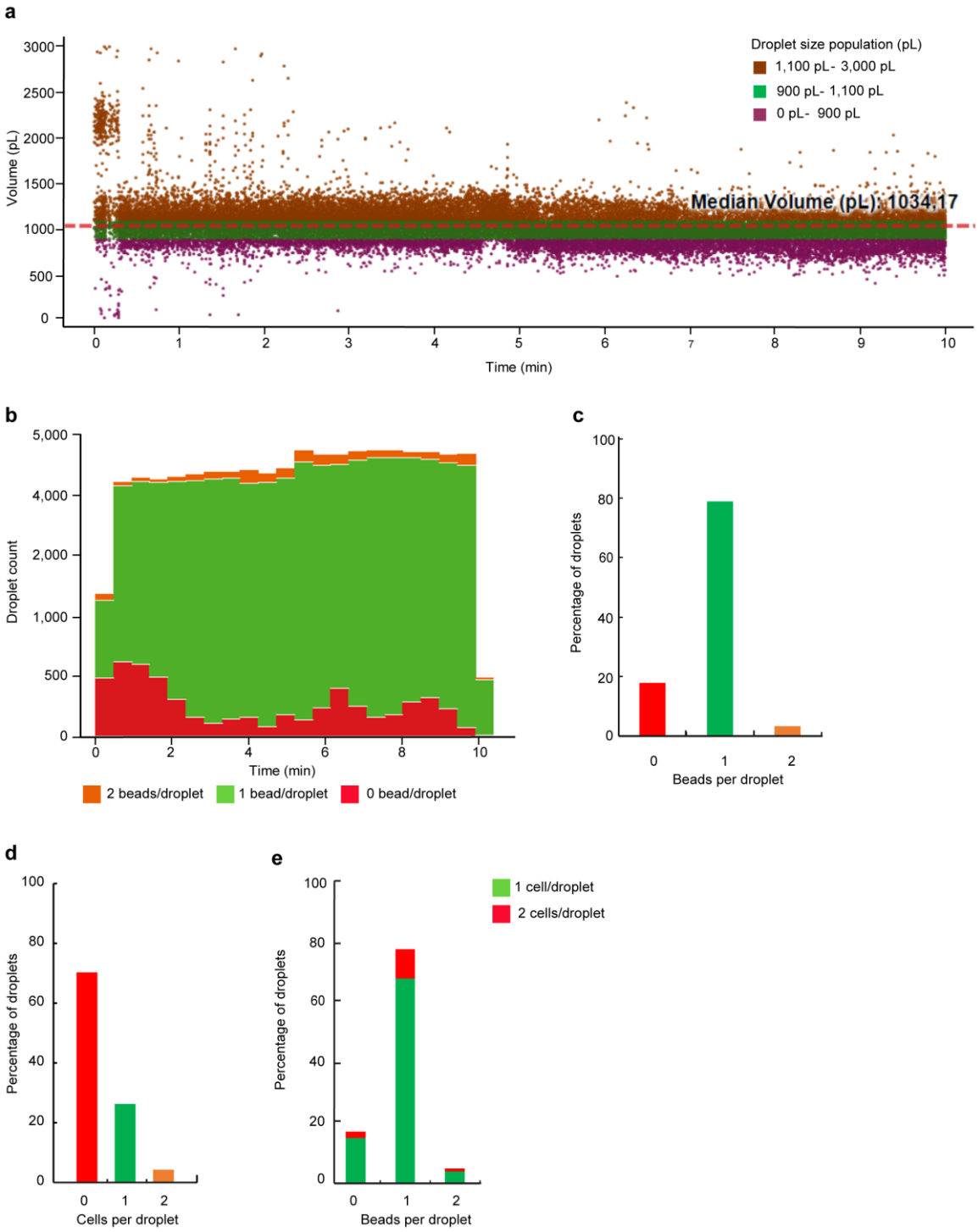


Supplementary Figure 6. Cell analysis. (**a**) Cell viability was followed throughout the process by flow cytometry using a live/dead assay (LIVE/DEAD Fixable Aqua stain, Thermo Fisher Scientific). Horizontal bars represent the median (n=3, 3, 4, 3). (**b**) Cytospin analysis of sorted cells from a GPI-immunized mouse showing expanded cytoplasm (shades of light blue), a characteristic of plasmablast and plasma cells (n=3). (**c**) Left, percentage of mean of IgG-secreting cells producing antigen-specific IgGs from ELISpot analysis of sorted cells from a TT-immunized mouse (TT#1) using wells coated with IgG (n=2), TT (n=2) or a negative control antigen (n=1). Right, pictures of representative ELISpot wells.

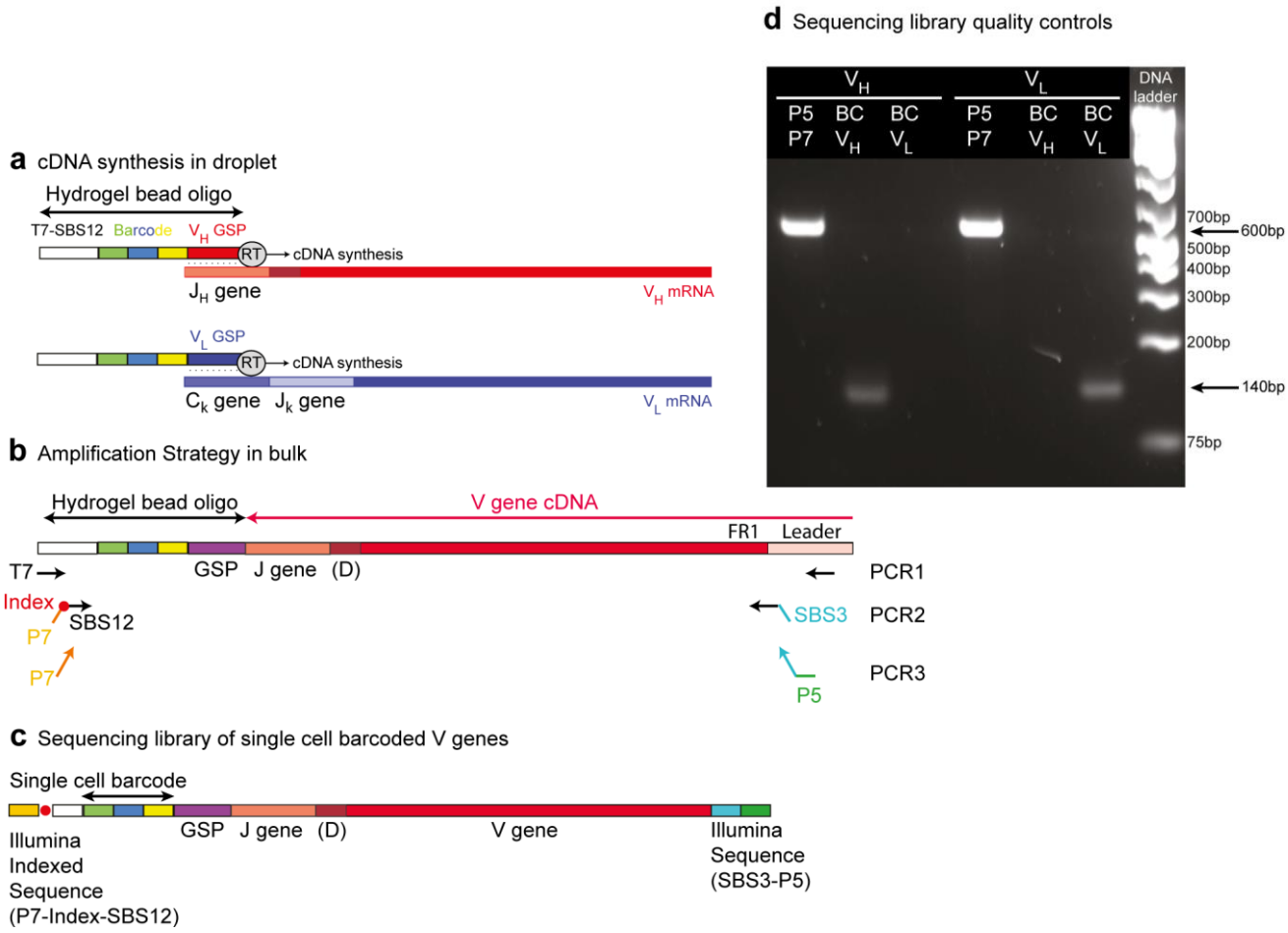


Supplementary Figure 7. Quality controls and characterization of single-cell barcoding. (**a**) Schematic of barcoded oligonucleotides on hydrogel beads. Each barcode comprises three 20 residues long indexes, Index A, Index B and Index C (see Supplementary File 1). The barcoded are produced by ligation of double-stranded indexes containing 5' overhangs of 4 residues by three cycles of split-and-pool synthesis using 96 Index As, 96 Index Bs, and 96 Index Cs (see Online Methods). The barcode is flanked at one end by a universal sequence (containing the T7 promoter and the Illumina SBS12 sequence) which is attached to the hydrogel beads via a photocleavable (PC) oligonucleotide linker. At the other end of the barcodes there is a mixture of sequences in which the top strand corresponds to Gene Specific Primers (GSPs) complementary to the four mouse J_H-genes and the mouse Cκ region that serve as primers for reverse transcription. After adding the GSP region the bottom strand is removed by alkaline hydrolysis. Each bead carries ~10^9 primers. (**b**) Split-and-pool synthesis principle. (**c**) Barcoded beads were created as above carrying 10% Cy5 labeled oligonucleotides which allowed single-beads to be identified by epifluorescence microscopy. Approximately 3,000 single beads were isolated and

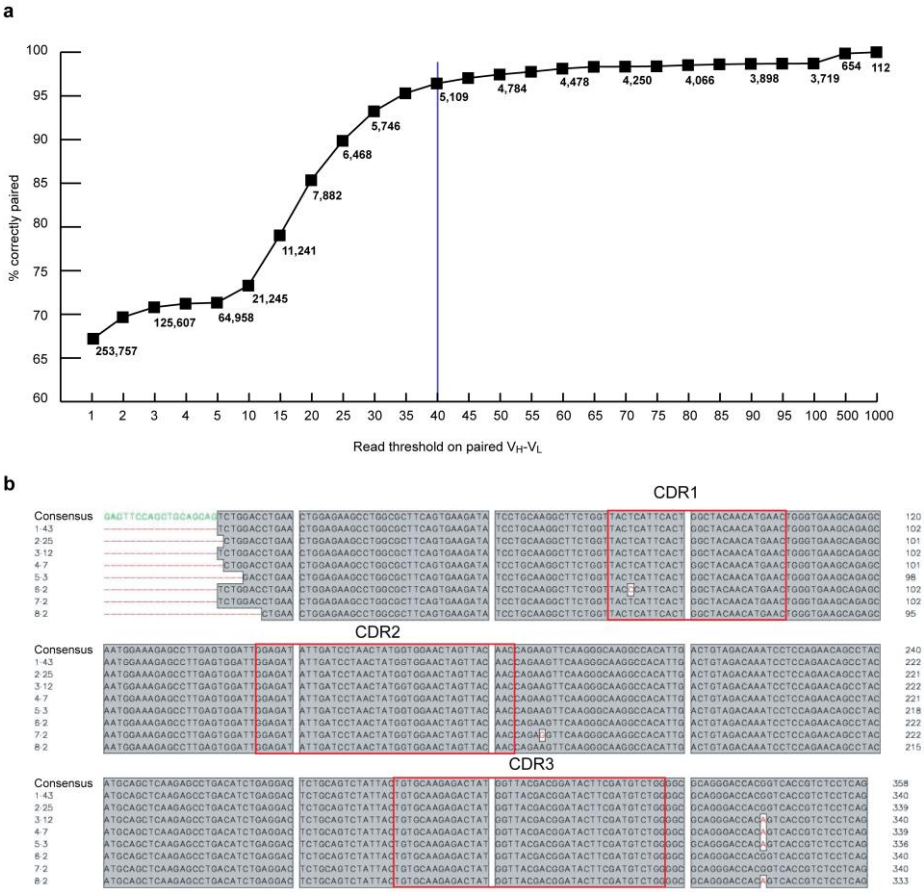
diluted down to <1 bead per well in a 384-wells plate (#240074, Thermofisher) in 20 ul RNAse/DNAse free water. The barcoded primers were then photocleaved and amplified by PCR for sequencing. The amplified products were analyzed on a 2% agarose gel, size selected, purified with a gel purification kit (28606, Qiagen). The sequencing libraries from each bead, each carrying specific Illumina indexes, were pooled and sequenced on a MiSeq Nano v2 PE75 flow cell. The percentage of reads corresponding to the most abundant barcode present on an individual bead is indicated for each bead. The mean is >97%. (**d**) To quantify full-length barcoded primers the oligonucleotides were released from 3,000 beads by photo-cleavage and analysed on a 2200 TapeStation (Agilent). (Top) Plot of sample intensity versus size (base pairs) (example representative of >10 library preparations). (Bottom) Box-plot of quantification based on TapeStation analysis representing the percentage of incomplete and full-length oligos bound to beads. Boxes extend from the 25^{th} to 75^{th} percentile and the line in the middle of the box is the median of 10 library preparations (n=10). (**e**) To quantify the number of full-length barcoded primers per bead, qPCR using forward primers with sequence identical to the 5' end of the bead oligo (SBS12) and each individual backward primer complementary to the 3'-end of the bead oligo (GSP region, specific for the 4 mouse J_H genes [and alleles] and the mouse C_κ region) was performed on the same samples as used for the TapeStation analysis, diluted to the equivalent of 10, 1 and 0.1 bead volumes per qPCR. Brilliant III ultra-fast qPCR master Mix (600880, Agilent Technologies) kit was used together with standard templates as a standard curve reference for both quantitative analysis and melting curve analysis. The sum of heavy chain specific primers (mIGH) and the kappa light chain primers (mIGK) shows balanced V_H and V_L specific primers on the beads. Data are the mean of 10 libraries. Boxes extend from the 25^{th} to 75^{th} percentile and the line in the middle of the box is the median.



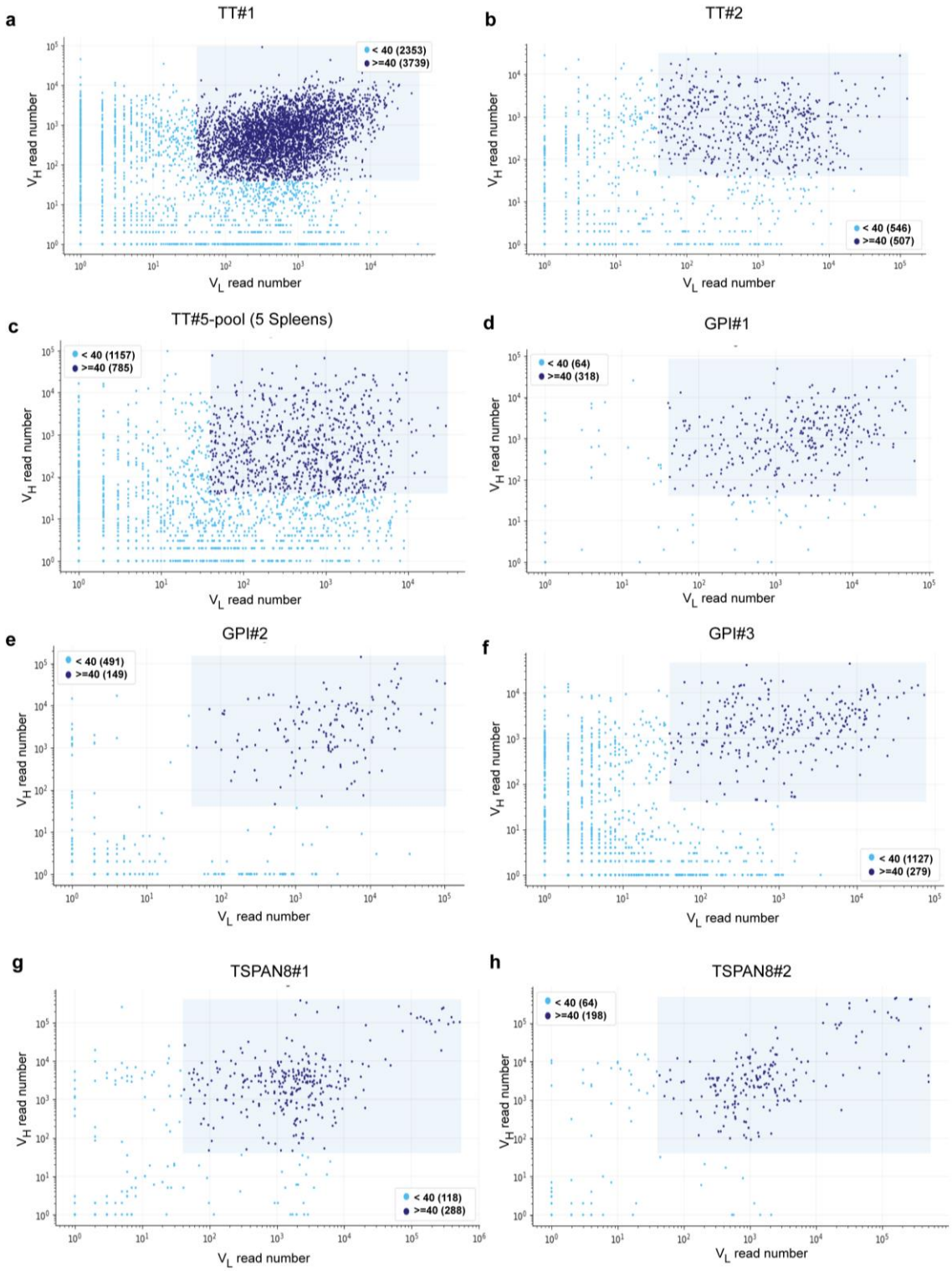
Supplementary Figure 8. Co-compartmentalization of cells and barcoded hydrogel beads in droplets. A representative example (TT mouse #1) is shown. (**a**) Droplet size distribution over the entire run (93,572 droplets). The median volume was 1034 pL with a coefficient of variation (CV) of 4.9%. (**b**) Number of beads per droplet monitored during the entire run and represented as a function of droplet count measured every 30 sec. Hydrogel beads were labeled with Cy5 to monitor the number of beads loaded into each droplet. (**c**) Statistical analysis of hydrogel bead loading in droplets. (**d**) Statistical analysis of cell loading in droplets (sorted cells plus Chinese hamster ovary (CHO) carrier cells). (**e**) Statistical analysis of hydrogel bead and cell co-compartmentalization.



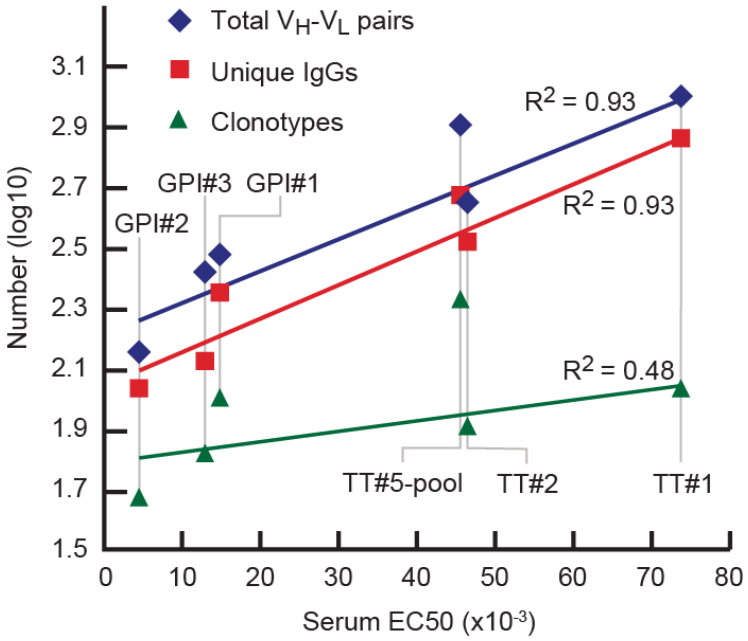
Supplementary Figure 9. Sequencing library preparation and quality control. (**a**) The reverse transcription of V_H and V_L mRNAs from individual cells in droplets was primed using barcoded primers carrying the T7-SBS12 sequences followed by the 3 index barcode and the gene-specific primer (GSP) sequences complementary to heavy chain J genes and light chain constant region sequences. The cDNAs generated from each cell carry a different barcode, allowing cognate V_H and V_L pairs to be identified after NGS. (**b**) Strategy for amplification of barcoded cDNA. After cDNA purification from the broken emulsion, the barcoded V_H and V_L genes were amplified by nested PCR, with forward primers priming sequentially on the T7 and on the Illumina SBS12 sequence of the barcode and reverse primers priming sequentially on the leader and framework 1 sequences of the V gene. In the second PCR, the forward primer appends the first part of the Illumina P7 and Illumina index sequences by priming on the SBS12 sequence and the reverse primer appends the Illumina SBS3 sequence. A third PCR adds the P5 sequence by priming on the SBS3 sequences. (**c**) Schematic of the final product, ready for Illumina sequencing, obtained after the 3 step PCR. (**d**) Presence of Illumina sequences for bridge PCR (P5 and P7) as well as specific single-cell barcodes for V_H (BC V_H) and V_L (BC V_L) in the sequencing libraries and purity of sequencing libraries were controlled by a semi-quantitative 12 cycle PCR, starting from 0.01 ng of material. The amplified products (n=5) were analyzed on a 2% agarose gel alongside 1kb+ DNA ladder (ThermoFisher Scientific). The arrows indicate the expected size of the V_H and V_L PCR products (600 bp) and the size of the amplified barcode QC product (140 bp).



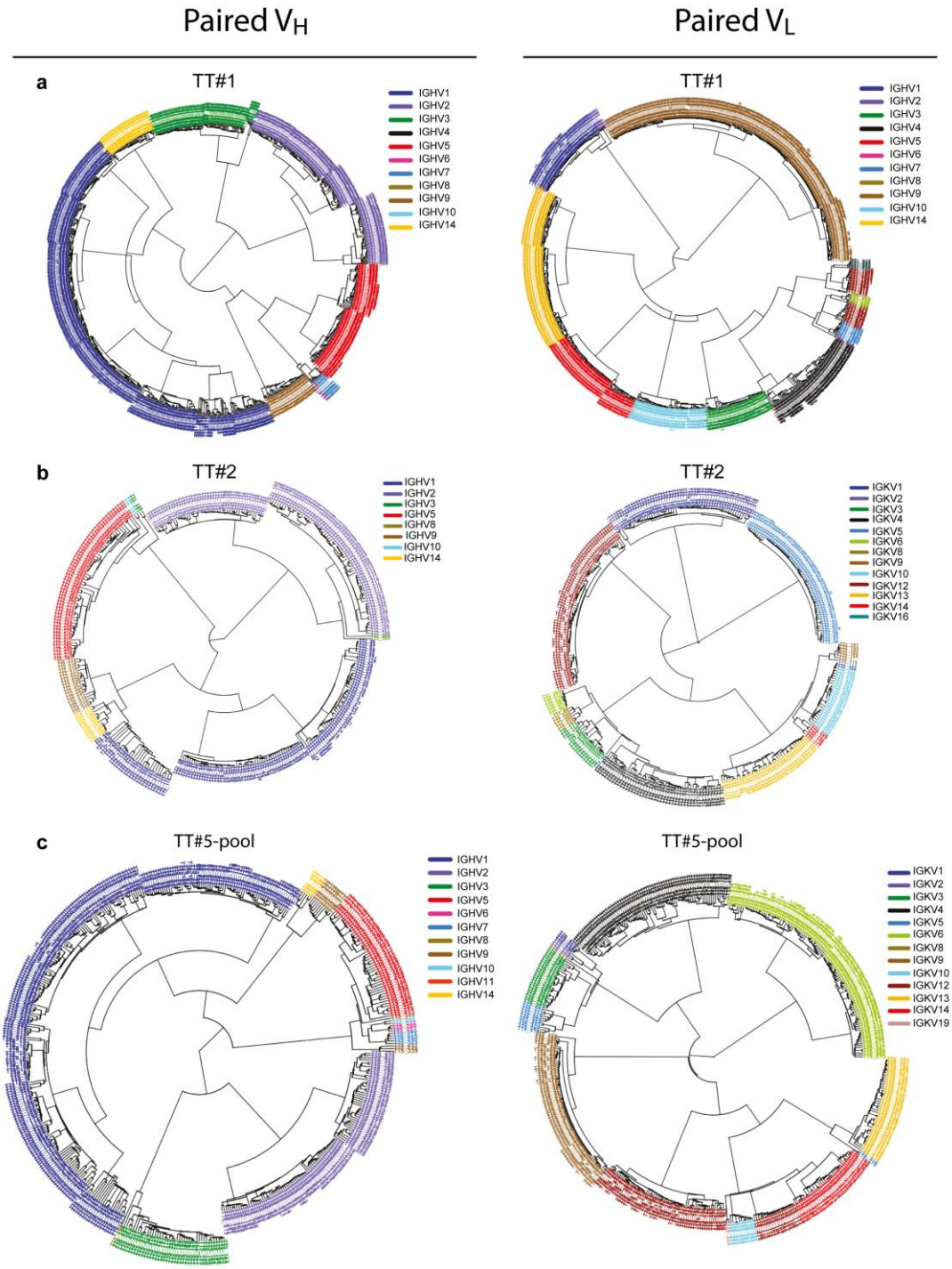
Supplementary Figure 10. Assigning the read threshold for correct V_H-V_L pairing and creation of consensus sequences within a barcode cluster. (a) Assigning the read threshold for correct V_H-V_L pairing using a mixture of hybridomas. Single-cell barcoded V_H-V_L sequencing was performed on a mixture of three mouse hybridomas (1:3 each) whose V_H and V_L sequences are known. Increasing read thresholds were placed on V_H-V_L pairs with the same barcode, requiring that both the V_H and V_L consensus sequences of the pair were made up from the number of reads at or above the threshold. At each threshold, we measured both the number of pairs (indicated next to each point) and the percentage of pairs that matched the known hybridoma pairing. The percentage of correctly paired sequences remains low (67-73%) up to a threshold of 10 but increases rapidly thereafter suggesting these lower read pairs represent noise. The percentage of correct pairs begins to plateau at a read threshold of 40 (blue line), which gives a high rate of pairing (>96%) while still providing a high number of pairs. We therefore set the read thresholding for high confidence V_H-V_L pairs at 40 **(b)** Creation of consensus sequences within a barcode cluster to filter out errors introduced by PCR or sequencing. To exemplify the concept, the figure shows the top 8 sequence contributors (sharing the same barcode) to a heavy chain consensus (with the number of reads represented by each sequence indicated on the left, e.g. 1.43 is the top contributor (1) with 43 reads). The top sequence is the consensus created from the alignment to show how the sequences that made up the consensus relate to it. Sequences showing significant sequence similarity within a barcode cluster were first identified through a DNA clustering algorithm. The primer region (first 18 bp) was removed prior to multiple alignment and a consensus was created using the most represented base in each column. In this way, errors, shown in white boxes with red text, are removed. The rest of framework 1 was "filled-in" using germline sequence from the closest germline gene (sequence in green), removing any bases from the start of the consensus sequence that come from an incomplete codon, to avoid introducing a frameshift. The position of CDRs in the sequence are indicated by red boxes.



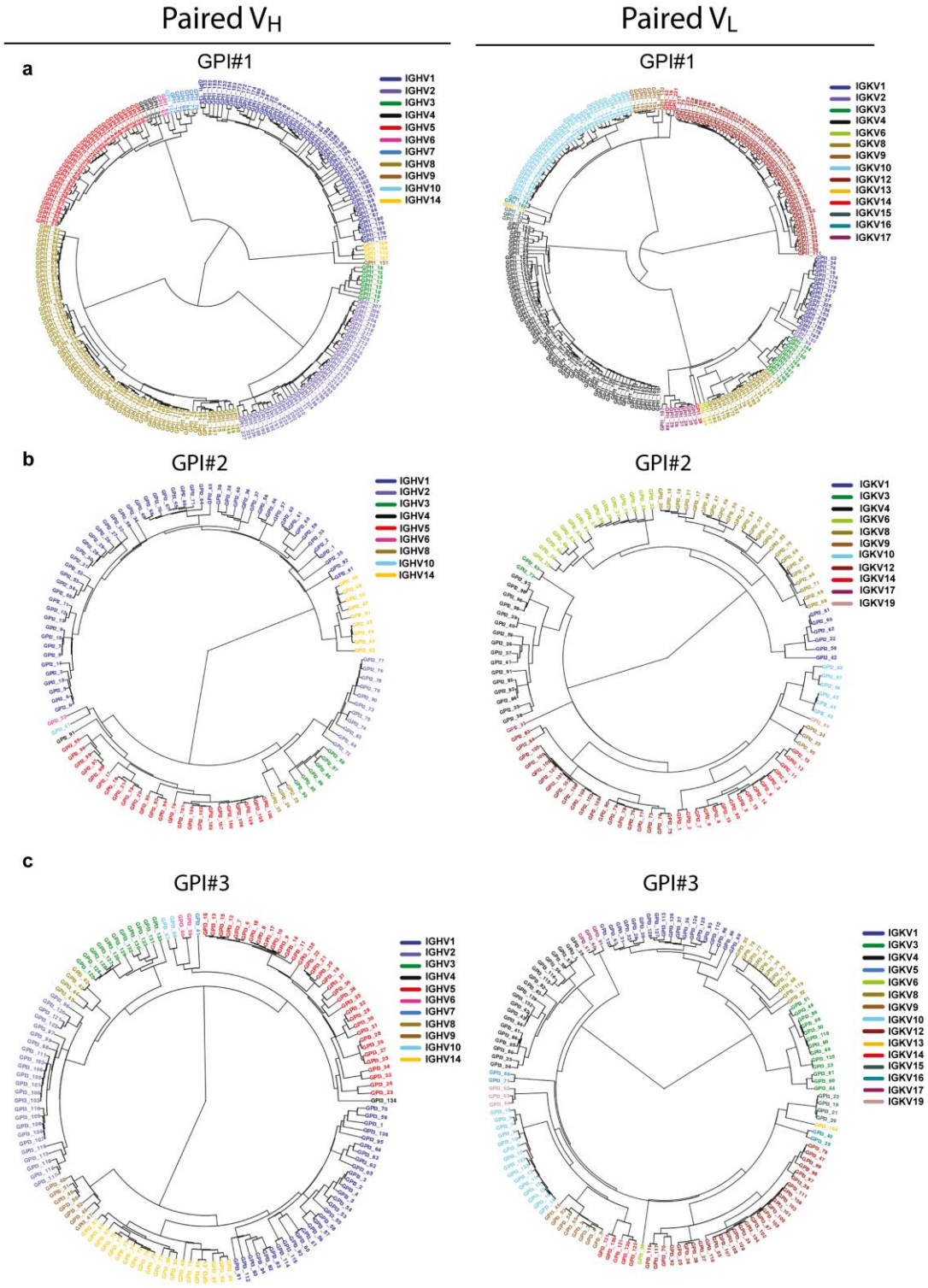
Supplementary Figure 11. Read abundances for all V_H-V_L pairs identified from two TT, three GPI and two TSPAN8-cell immunized mice. Read abundances for V_H and V_L in V_H-V_L pairs with the same barcode from screens of TT-immunized mice, (**a**) TT #1, (**b**) TT#2, (**c**) the pool of spleens from five TT-immunized mice (TT#5-pool), GPI-immunized mice, (**d**) GPI#1, (**e**) GPI#2, (**f**) GPI#3 and TSPAN8-cell immunized mice, (**g**) TSPAN8#1 and (**h**) TSPAN8#2 are shown. We set a stringent read threshold of 40 reads (Supplementary Figure 16) for both V_H and V_L. Pairs that cross this threshold are located in the blue shaded box and plotted in dark blue. Pairs that fall below this threshold are in light blue. The number of pairs above and below the threshold is indicated.



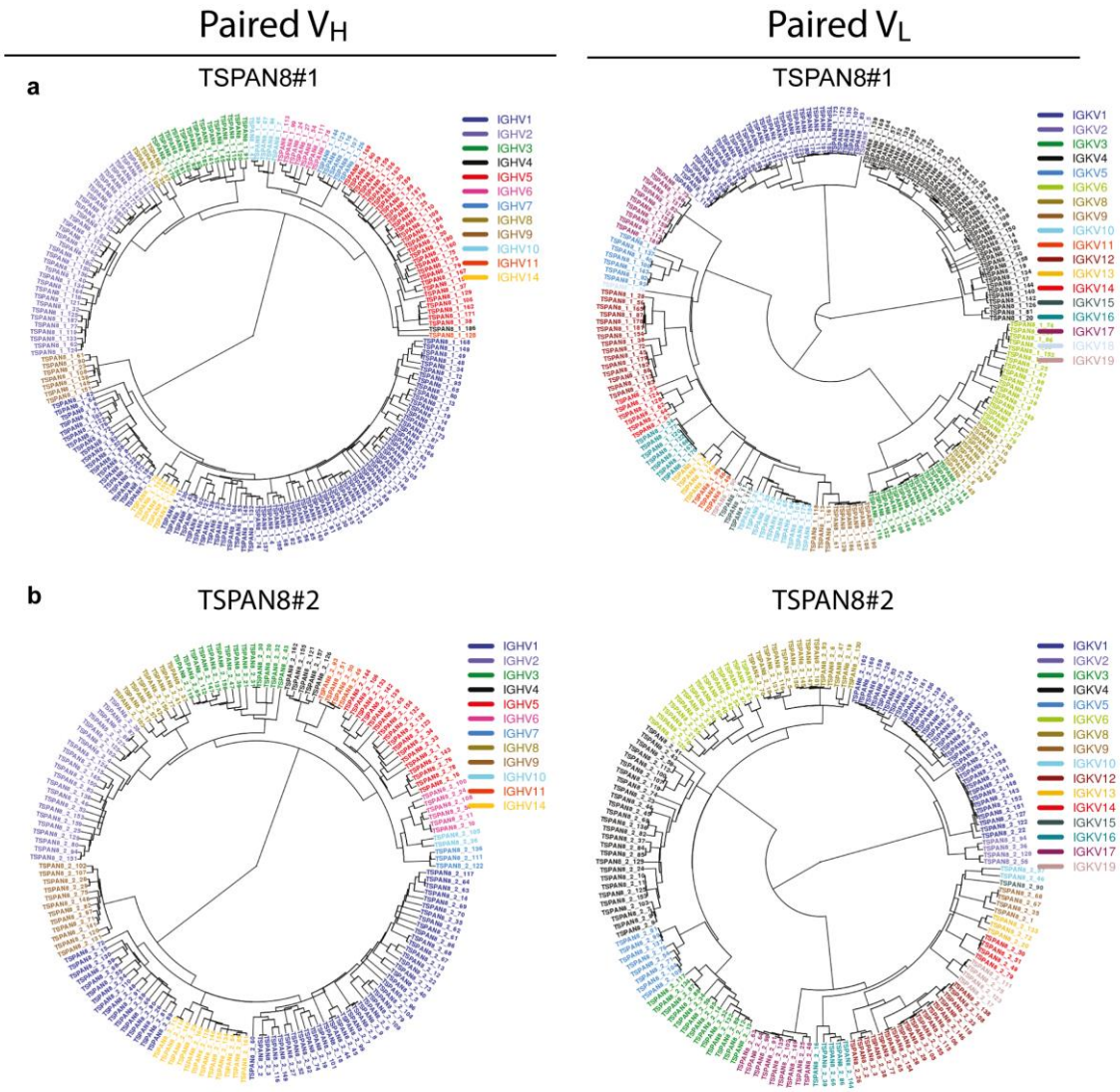
Supplementary Figure 12. V_H-V_L pair numbers and unique IgG numbers correlate positively with serum IgG titer. Graph of number of V_H-V_L pairs associated with different barcodes, the total number of non-redundant IgGs, and the total number of clusters of non-redundant IgGs derived from the same primary germline antibody via somatic hypermutation (clonotypes) versus serum IgG titer (EC50) in individual GPI-immunized mice (n=3) or TT-immunized mice (n=2) mice and TT#5-pool (n=1 pool of 5 spleens), for which the mean serum EC50 is used (refer to Supplementary Table 1 for the number of pairs and clonotypes analyzed). The exponential fit (the best fit) to the data from individual mice (i.e. not for TT#5-pool) and R^2 values are shown.



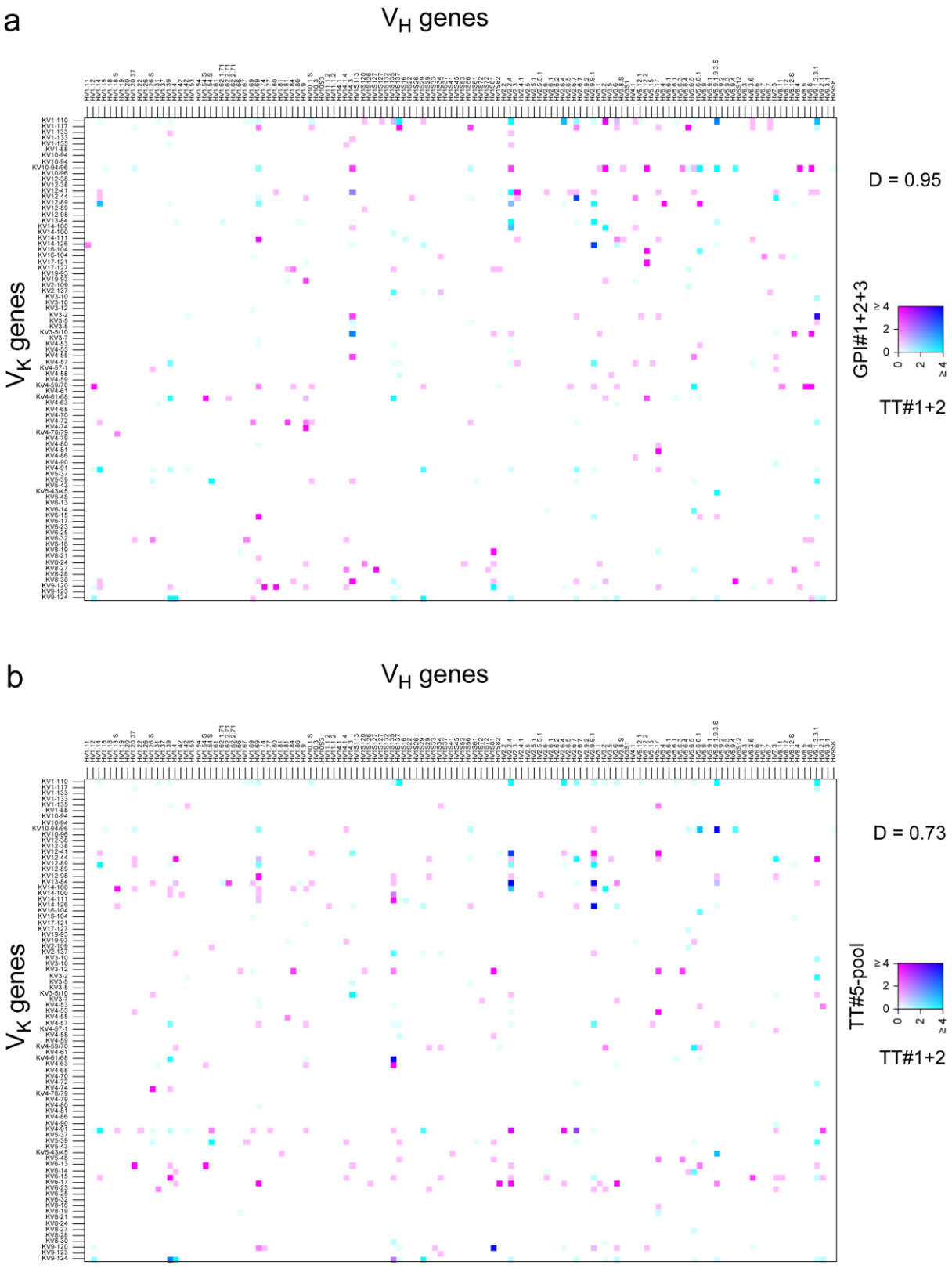
Supplementary Figure 13. Phylogeny of captured V_H-V_L pairs identified in screens of splenocytes from two TT immunized mice. Phylogenetic trees showing V-gene usage from screens of TT-immunized mice, (**a**) TT#1, (**b**) TT#2 and (**c**) TT#5-pool. Paired V_H genes are shown on the left and paired V_κ genes on the right. Each antibody sequence was assigned to its closest germline V-gene. Full length amino acid sequences of all V_H or V_L from each screen were aligned and are displayed as a phylogenetic tree. The number of sequences used for the alignment used in each tree can be found in column 12 of Supplementary Table 1. Labels are colored according to the gene family of the germline gene.



Supplementary Figure 14. Phylogeny of captured V_H-V_L pairs identified in screens of splenocytes from three GPI immunized mice. Phylogenetic trees showing V gene usage in (**a**) GPI mouse #1, (**b**) GPI mouse #2 and (c) GPI mouse #3. Paired V_H genes are shown on the left and paired V_κ genes on the right. Each antibody sequence was assigned to its closest germline V-gene. Full length amino acid sequences of all V_H or V_L from each screen were aligned and are displayed as a phylogenetic tree. The number of sequences used for the alignment used in each tree can be found in column 12 of Supplementary Table 1. Labels are colored according to the gene family of the germline gene.

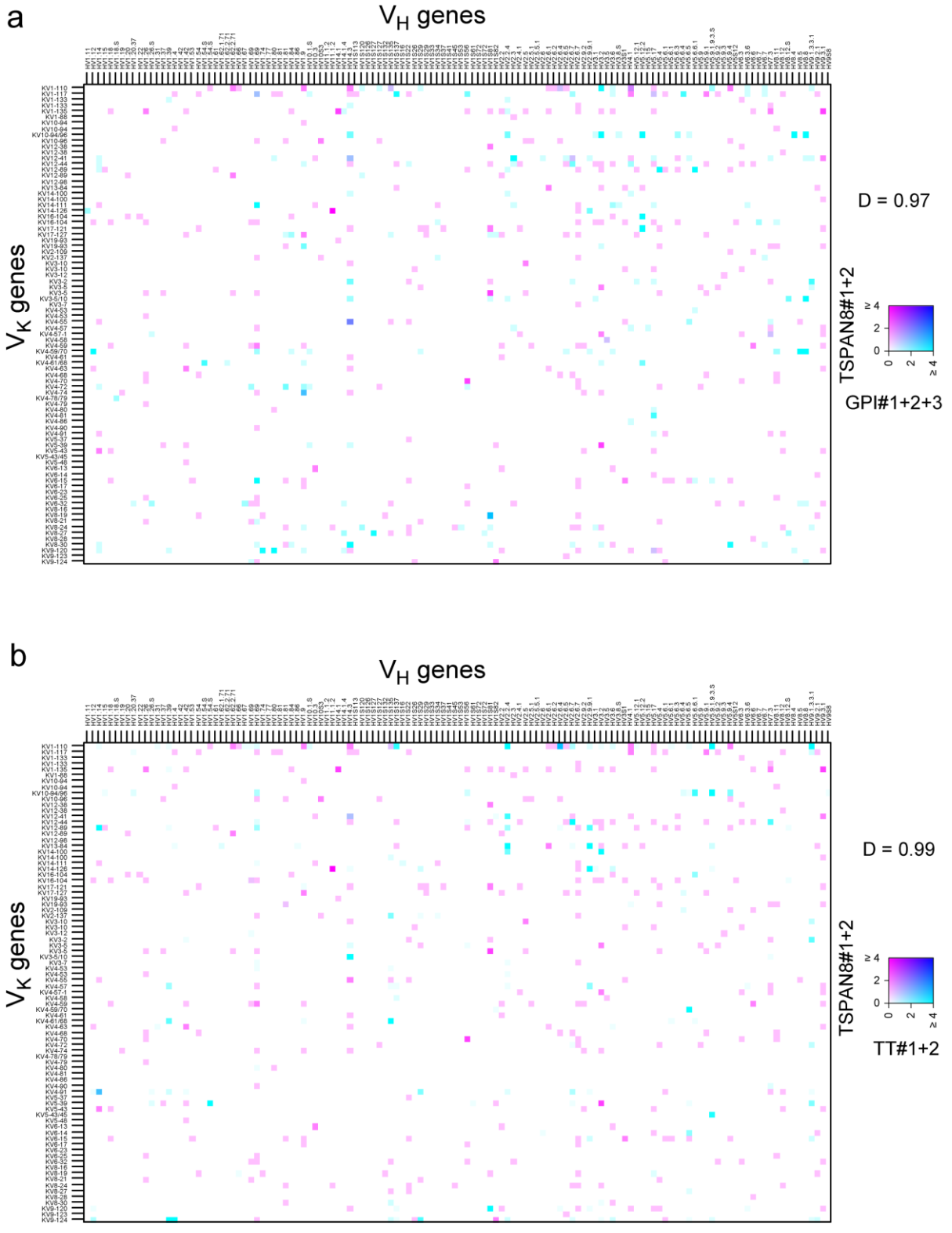


Supplementary Figure 15. Phylogeny of captured V_H-V_L pairs identified in screens of splenocytes from two TSPAN8-overexpressing cell immunized mice. Phylogenetic trees showing V gene usage in (**a**) TSPAN8 mouse #1, (**b**) TSPAN8 mouse #2. Paired V_H genes are shown on the left and paired V_κ genes on the right. Each antibody sequence was assigned to its closest germline V-gene. Full length amino acid sequences of all V_H or V_L from each screen were aligned and are displayed as a phylogenetic tree. The number of sequences used for the alignment used in each tree can be found in column 12 of Supplementary Table 1. Labels are colored according to the gene family of the germline gene.



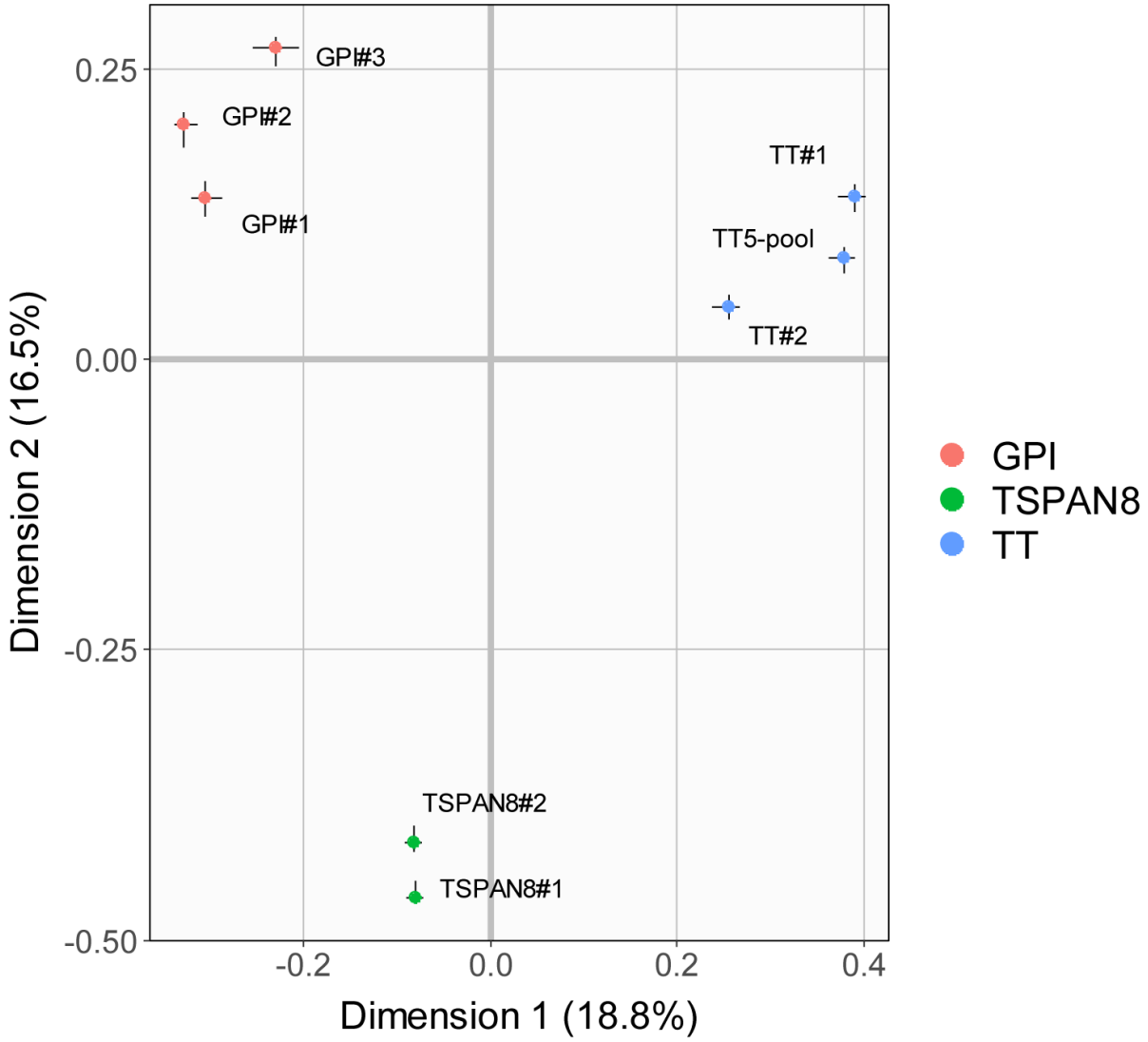
Supplementary Figure 16. Heatmaps of pairwise V_H-V_L gene usage after TT or GPI immunization. The number of sequences used for the analysis can be found in column 12 of Supplementary Table 1. **(a)** Comparison of the combined data from two individual TT-immunized mice (TT#1+2) and three individual GPI-immunized mice (GPI#1+2+3). (**b**) Comparison of the combined data from two individual TT-immunized mice (TT#1+2) and from the pooled splenocytes of five TT-immunized mice

(TT#5-pool). (**a,b**) V_H genes are represented in columns and V_L genes (V_κ genes) in rows. The legends beside each heatmap depict relative gene abundances between the two compared groups of mice, as well as overlays. For instance, in panel a, the white-to-cyan x-axis indicates the relative gene abundance in TT#1+2, between 0 and 4 cells (i.e. unique barcodes), the white-to-magenta y-axis indicates the relative gene abundance of GPI mice #1+2+3, between 0 and 4 cells, and the white-to-dark blue diagonal represents a full gene abundance overlay. Dynamic ranges of the two panels have been set between 0 and 4 to facilitate visual comparisons, but the real normalized range is 0-57 for TT#1+2 and 0-48 for GPI#1+2+3 (a), 0-58 for TT#1+2 and 0-61 for TT#5-pool (b). Distances, *D*, between the compared V_H-V_L gene usage matrices are also shown (see Online Methods). *D* is higher between TT#1+2 and GPI#1+2+3 in panel a, than between TT#1+2+3 and TT#5-pool in panel b, and consistent with this result, the frequency of overlapping V_H-V_L gene usage (dark blue cells) is higher in panel b than in panel a.

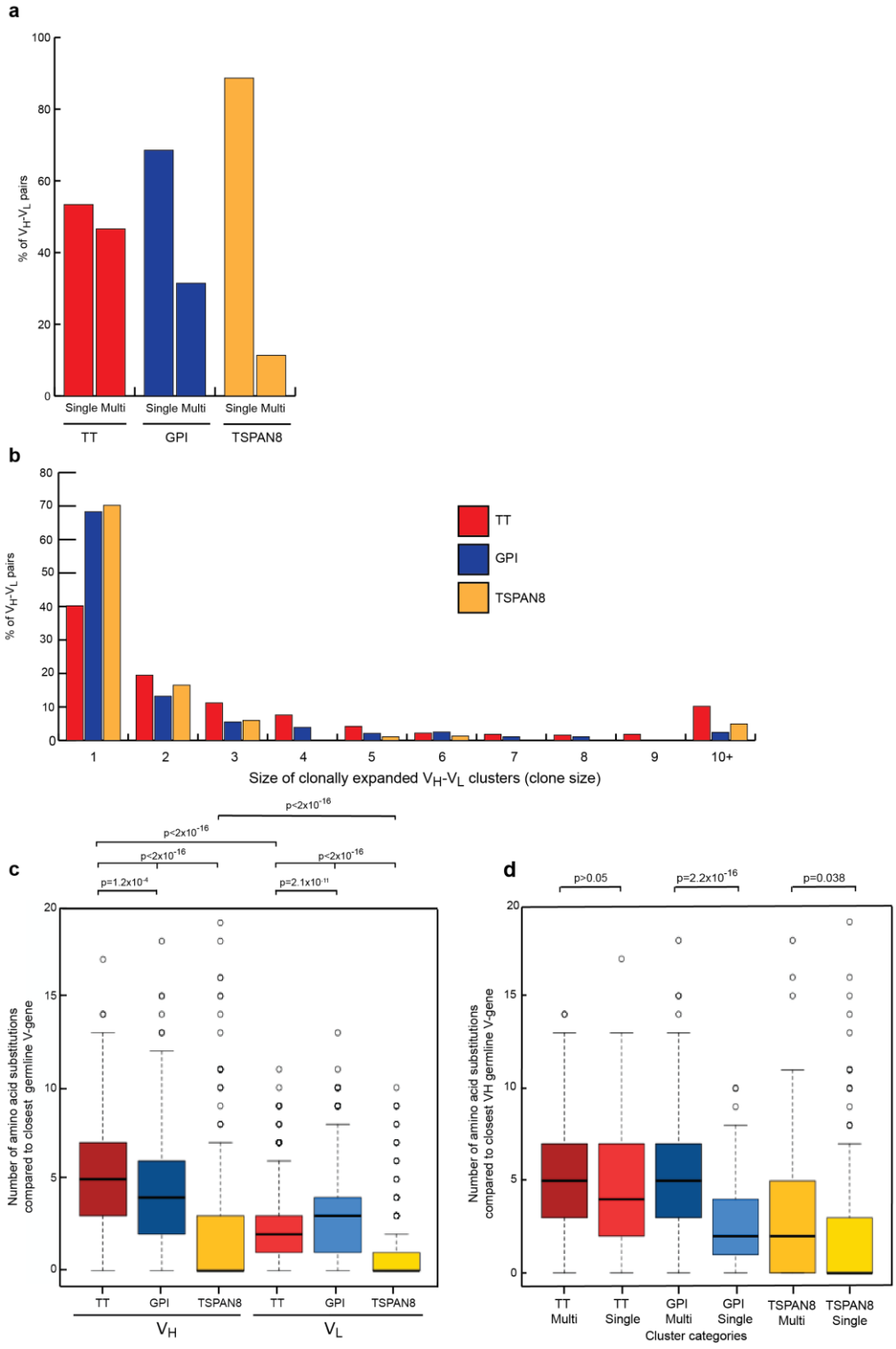


Supplementary Figure 17. Heatmaps of pairwise V_H-V_L gene usage after TT, GPI and TSPAN8 immunization. The number of sequences used for the analysis can be found in column 12 of Supplementary Table 1. **(a)** Comparison of the combined data from two individual TSPAN8-immunized mice (TSPAN8#1+2) and three individual GPI-immunized mice (GPI#1+2+3). (**b**) Comparison of the combined data from two individual TSPAN8-immunized mice (TSPAN8#1+2) and two individual TT-immunized mice (TT#1+2). (**a,b**) As in Supplementary Figure 16, V_H genes are represented in columns

and V_L genes (V_κ genes) in rows. The legends beside each heatmap depict relative gene abundances between the two compared group of mice, as well as overlays. Dynamic ranges of the two panels have been set between 0 and 4 to facilitate visual comparisons, but the real normalized range is 0-56 for GPI#1+2+3 and 0-4 for TSPAN8#1+2 (a), 0-43 for TT#1+2 and 0-4 for TSPAN8#1+2 (b). Distances, *D*, between the compared V_H-V_L gene usage matrices are also shown (see Online Methods). *D* is very high and similar between TSPAN8#1+2 and GPI#1+2+3 (panel a), and between TSPAN8#1+2 and TT#1+2 (panel b), and consistent with this result, the frequency of overlapping V_H-V_L gene usage (blue cells) is almost null in panels a and b.

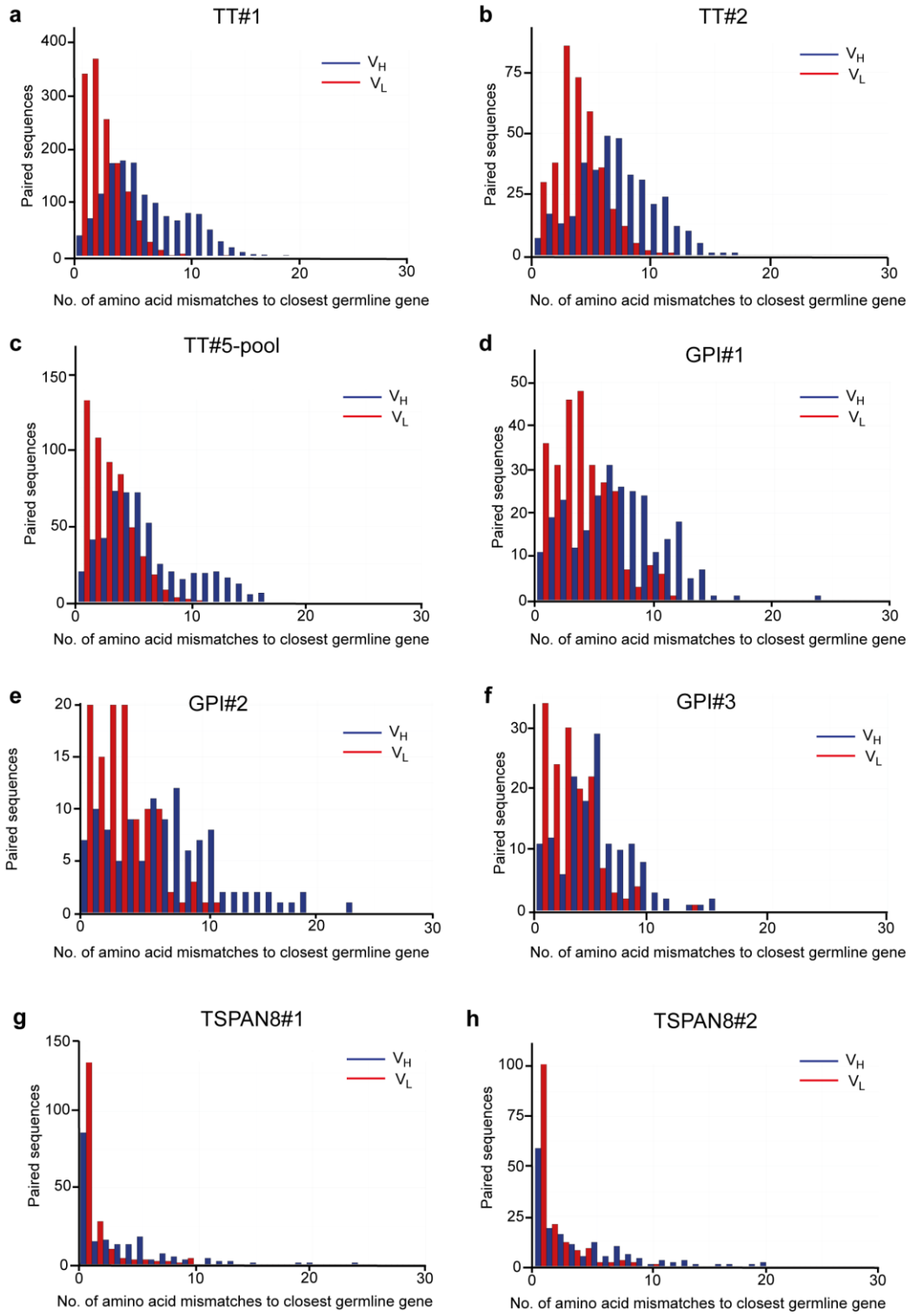


Supplementary Figure 18. Multidimensional scaling (MDS) of the distance matrix of V_H-V_L gene usage. The metric MDS is similar to Principal Component Analysis (PCA) but starting from a distance matrix instead of a correlation matrix, and without initial coordinates of individuals because no individuals are available in MDS, only variables relationship. The graph depicts the projection of the initial TT#1, TT#2, GPI#1, GPI#2, GPI#3, TSPAN8#1 and TSPAN8#2 variables on the 2 first components, with the associated percentage of the total variance between brackets. Vertical and horizontal bars behind dots represent the 95% confidence interval of the dot positions. Non-overlapping intervals indicate a significant distance between the dots on the projection plan. The number of sequences n of each TT#1, TT#2, GPI#1, GPI#2, GPI#3, TSPAN8#1 and TSPAN8#2 variables can be found in column 12 of Supplementary Table 1, and correspond to n=7 different V_H-V_L gene usage repertoires.

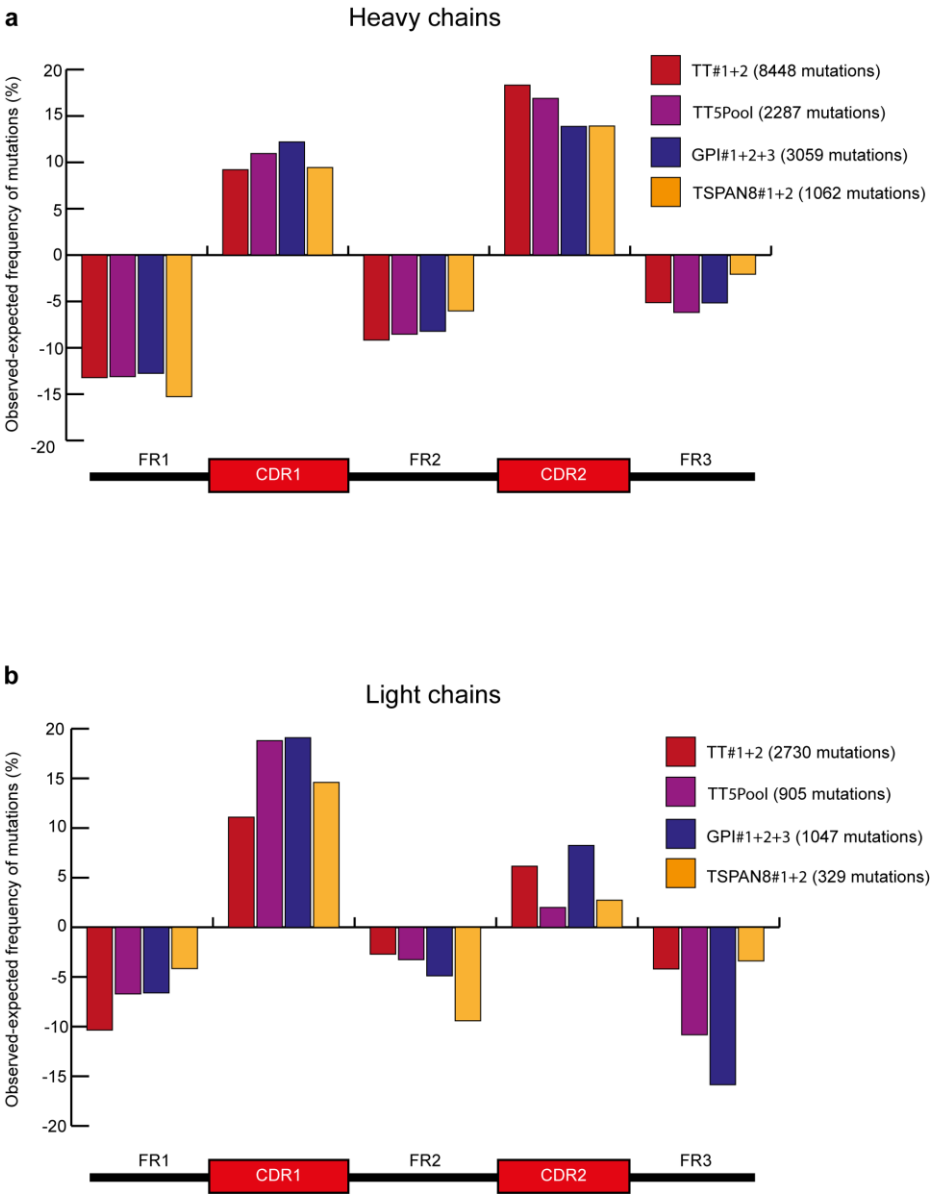


Supplementary Figure 19. Somatic hypermutation and clonal expansion. Data from the two TT-immunized (TT#1+2), three GPI-immunized mice (GPI#1+2+3) and two TSPAN8-immunized mice (TSPAN8#1+2) were pooled to create a single dataset for each target. (**a**) Proportion of non-redundant V_H-V_L pairs that fall within families containing a single member ('Single') or multiple members ('Multi') derived from the same germline gene rearrangement. The proportion of V_H-V_L pairs in families containing multiple members was higher in TT-immunized mice than in GPI-immunized, but the GPI-immunized mice contained more families with multiple members than TSPAN8-immunized mice. Even after correcting for sample size (which would bias the probability of V_H-V_L pairs being within single or

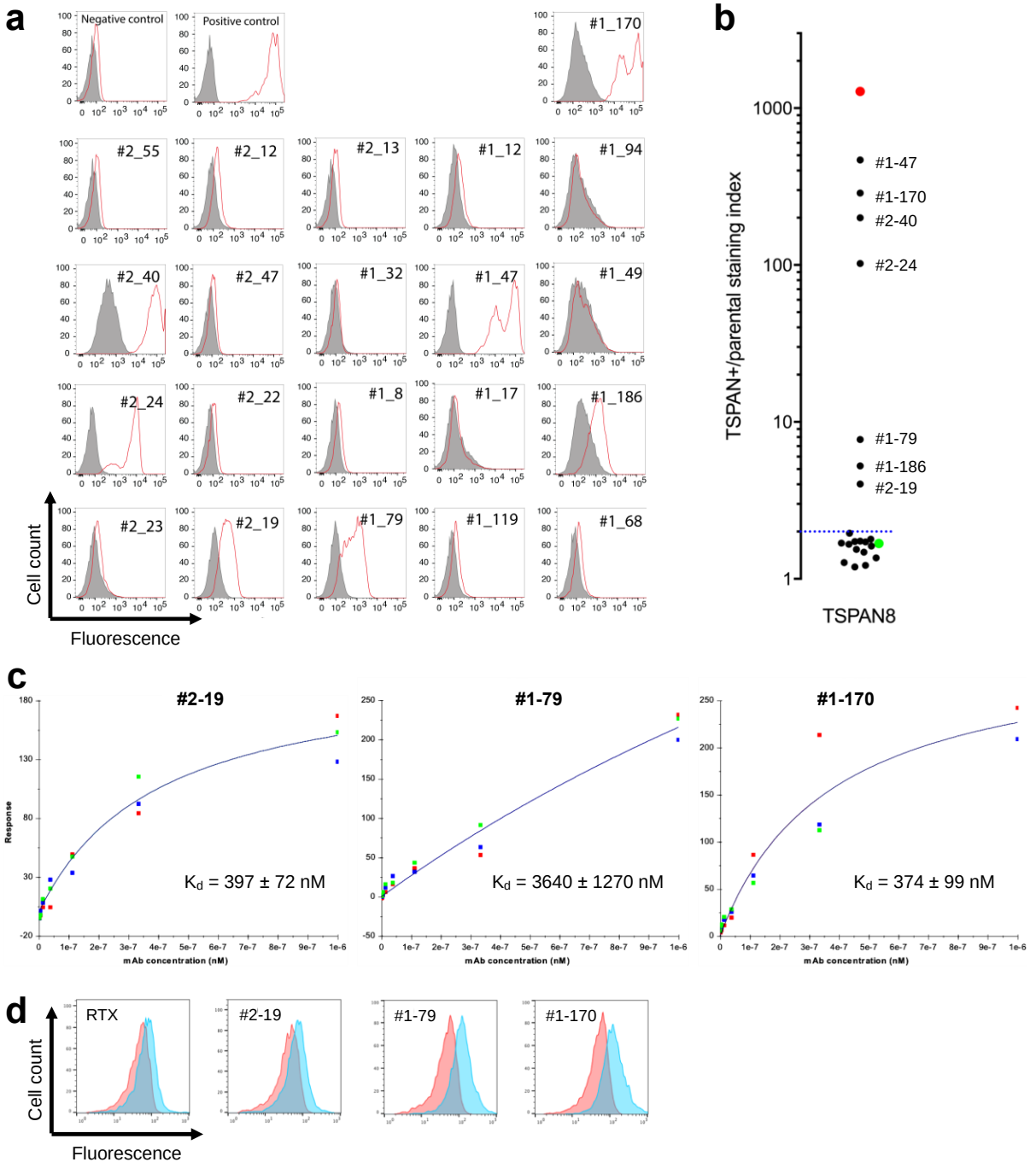
multi-member clusters), by random subsampling 100,000 groups of sequences of the same number as that for GPI dataset, we found that the probability of the TT dataset having the same proportion of single or multi-member clusters as the GPI dataset was $p = 4 \times 10^{-5}$ (Bootstrap method, see Online Methods). V_H-V_L pairs from the TSPAN8 dataset were largely singletons. (**b**) Distribution of clone size (the number of V_H-V_L pairs with identical amino acid sequence but associated with different barcodes), corresponding to the same IgG expressed by different cells due to recent clonal expansion, from TT, GPI and TSPAN8-immunized mice. (**c**) Box plots of distributions of amino acid substitutions compared to the closest germline V-gene (corresponding to non-synonymous somatic hypermutations) for V_H and V_L genes in V_H-V_L pairs identified in TT samples (n=1390 IgGs, VH: min=0, max=17, mean=5.18; VL: min=0, max=11, mean=2.08), GPI (n=473 IgGs, V_H: min=0, max=18, mean=4.55; V_L: min=0, max=13, mean=2.92) and TSPAN8-cell immunized mice (n=353 IgGs, V_H: min=0, max=19, mean=2.01; V_L: min=0, max=10, mean=-0.93) (see Supplementary File 3). The frequency of substitutions was similar in V_H or V_L in TT and GPI immunized mice ($p > 0.05$; two-sided Welch's test), but with a significantly higher frequency in V_H than in V_L ($p < 2x10^{-16}$ for all V_H-V_L comparisons in each target, two-sided Welch's test). In contrast, TSPAN8 showed significantly lower numbers of mutations than TT or GPI ($p < 2x10^{-16}$, two-sided Welch's test) indicative of a weaker immune response. (**d**) Box plots of distributions of amino acid substitutions in heavy chains compared to the closest germline V-gene from V_H-V_L pairs from the same dataset as (c) that derive from either a singleton ('Single'), or multi-member families ('Multi') of non-redundant IgGs (TT-Multi n=1280 IgGs, min=0, max=14, mean=5.22; TT-Single n=110 IgGs, min=0, max=17, mean=4.67; GPI-Multi n=327 IgGs, min=0, max=18, mean=5.4; GPI-Single n=146 IgGs, min=0, max=10, mean=2.66; TSPAN8-Multi n=40 IgGs, min=0, max=18, mean=3.38; TSPAN8-Single n=313 IgGs, min=0, max=19, mean=1.84). While there was no statistically significant difference in the mean number of substitutions between these two groups in V_H-V_L pairs from TT-immunized mice ($p > 0.05$; two-sided Welch's test), in V_H-V_L pairs from GPI-immunized mice and to a lesser extent TSPAN8-immunized mice heavy chains from singleton families were significantly lower than those coming from multi-member families ($p = 1x10^{-8}$ (GPI) and p=0.038 (TSPAN8) two-sided Welch's test). In the box plots the middle line is the median, the height of the box is the interquartile range, and the whiskers represent 1.5 times the interquartile range from the box in each direction.



Supplementary Figure 20. Distribution of non-synonymous somatic hypermutations. The distribution of amino acid changes compared to the closest germline V gene (corresponding to non-synonymous somatic hypermutations) in V_H and V_L genes in V_H-V_L pairs from screens of TT-immunized mice, (**a**) TT #1, (**b**) TT#2, (**c**) TT#5-pool, and of GPI-immunized mice, (**d**) GPI#1, (**e**) GPI#2 and (**f**) GPI#3, and of TSPAN8-immunized mice (**g**) TSPAN8#2 and (**h**) TSPAN8#2. There are a higher number of non-synonymous mutations in V_H than in V_L in all mice, although the difference is greater for soluble antigens (mean 6.0.3±3.65 vs. mean 2.86±2.25) than for TSPAN8 (mean 3.1±4.21 vs. mean 0.93±1.89) and for both target types this difference is highly significant ($p < 2\times10^{-16}$, two-sided Welch's test, n=2325 IgGs for soluble, n=353 IgGs for TSPAN8).



Supplementary Figure 21. Distribution of non-synonymous somatic hypermutations in framework (FR) and complementarity-determining regions (CDRs) of V-genes. The frequency of amino acid changes compared to the closest germline V-gene (corresponding to non-synonymous somatic hypermutations) in complementarity-determining regions 1 and 2 (CDR1 and CDR2) and frameworks 1, 2 and 3 (FR1, FR2 and FR3) of (**a**) V_H and (**b**) V_L from all sequences of the two TT-immunized (TT#1+2), the pooled splenocytes from five TT-immunized mice (TT5Pool), all three GPI-immunized (GPI#1+2+3) mice or both TSPAN8-immunized mice (TSPAN8#1+2). The expected frequency for each functional region was calculated by multiplying the total number of mutations by the proportion of bases represented by that functional region. The positions of the CDR1, CDR2, FR1, FR2 and FR3 were identified and the per residue frequency of non-synonymous somatic hypermutations in each region was determined across all sequences in each group. The graphs show a clear enrichment of non-synonymous somatic hypermutations within the CDR1 and CDR2 regions of both V_H and V_L across all samples, although the degree of enrichment of each CDR differs between samples.

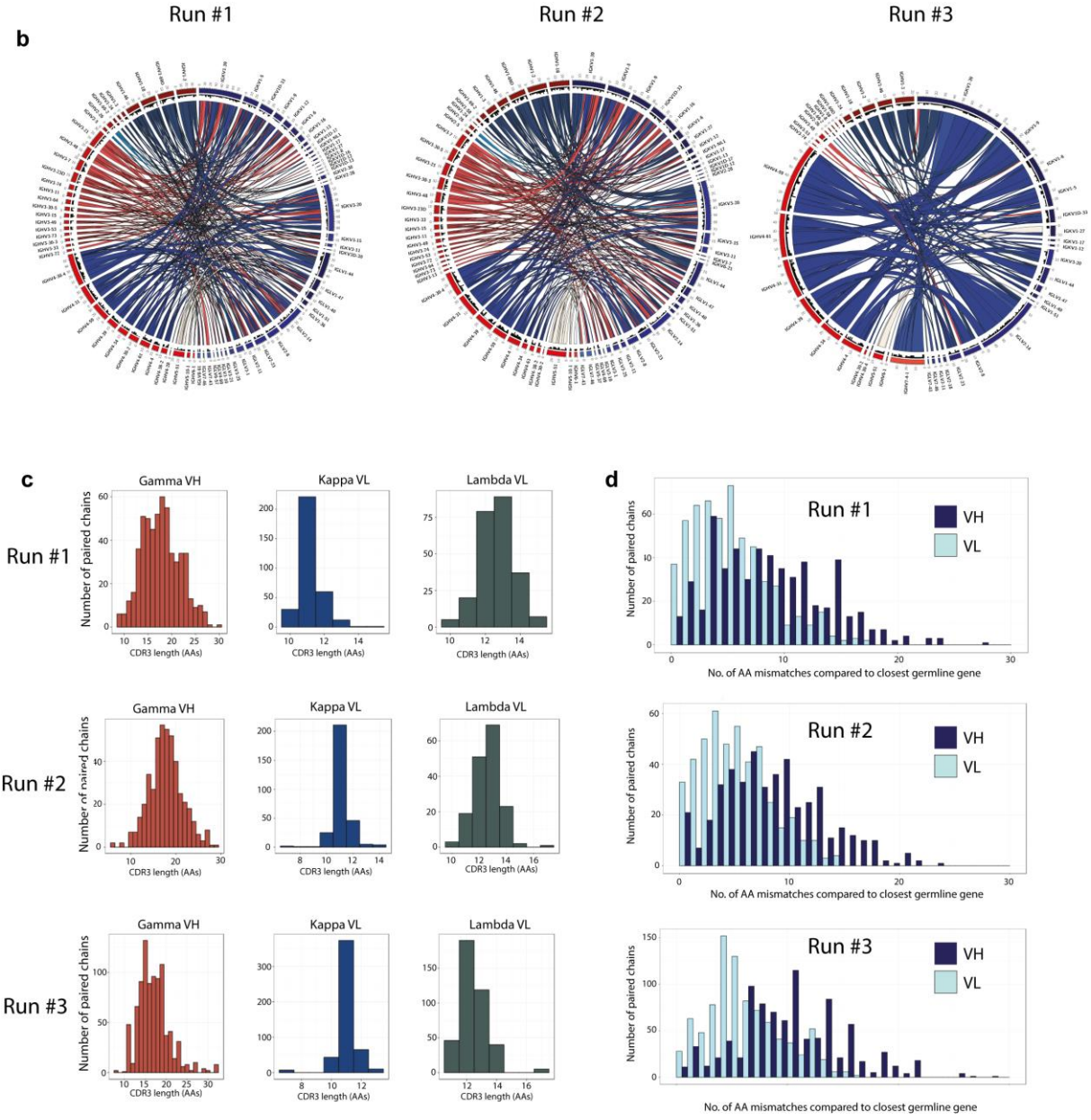


Supplementary Figure 22. Binding to TSPAN8 of recombinant IgGs from VH-VL pairs from the TSPAN8 sorts. (a) Flow cytometry analysis of binding of 21 purified recombinant mouse/human chimeric IgGs expressed from cloned genes identified from TSPAN8-immunized mice, or anti-TSPAN8 mAb 43A6 (positive control) or no primary antibody (negative control). Binding to parental M300.19 cells (filled grey histogram) or TSPAN8-expressing M300.19 cells (red histogram) was revealed by a fluorescently labelled anti-human IgG secondary antibody. The IgG ID number is indicated (refer to Supplemental File 3). **(b)** Median fluorescence index ($Median^{TSPAN8+}/Median^{parental}$) of the data from (a) with the IgG ID number next to its value (black dot). The red dot corresponds to the positive control and the green dot to no primary antibody. The threshold (dotted horizontal line) of binding to $TSPAN8^+$ cells was set at ≥ 2 times the maximum mean fluorescence of the negative control. **(c-d)** *Flow cytometry-based affinity measurement of anti-TSPAN8 IgGs.* Anti-TSPAN8 IgGs binding $TSPAN8^+$ cells over the threshold set in (b) and as a negative control anti-human CD20 rituximab (RTX) were directly labelled

with Alexa488, and incubated with TSPAN8^{+} M300.19 cells at increasing concentrations. (**c**) K_d values (±SD) were calculated from the antibody-concentration dependence of the response, *R*, the mean fluorescent signal of the indicated antibodies minus the mean fluorescent signal of the negative control (RTX) for IgGs where the maximum mean fluorescence was ≥ 2 times the maximum mean fluorescence of the negative control. The K_d values (±SD), calculated by fitting the pooled data from triplicate experiments (each triplicate is represented by dots of a given color) to Equation 1 (see Online Methods), are indicated, together with the standard deviation of the fit of the pooled data (n=3). (**d**) Histograms of the data analysed in (c) and of the negative control anti-human CD20 rituximab (RTX) at 0 (red histogram) or 1 µM (blue histogram).

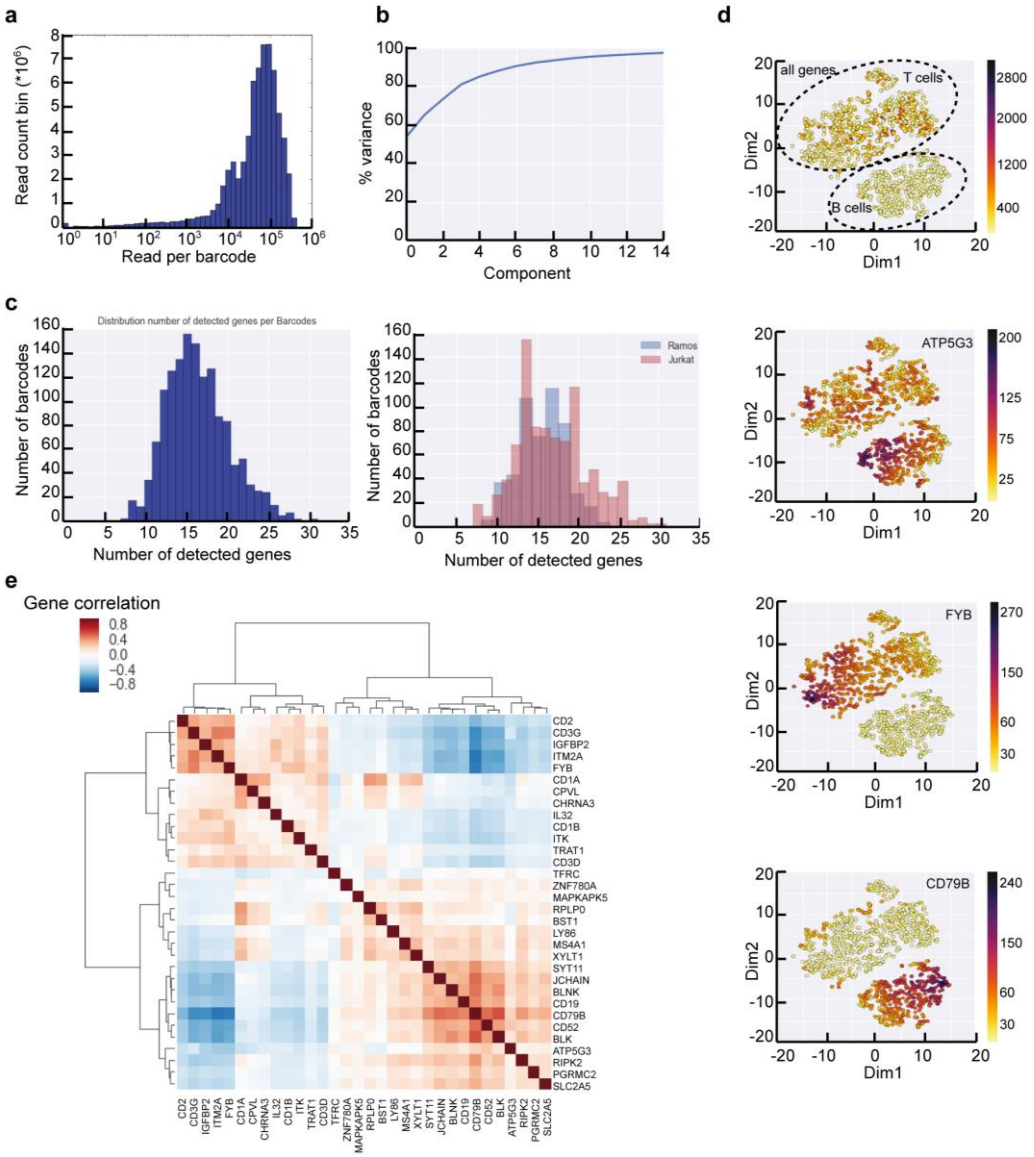
a

Human Activated Memory B cells	Starting Cell Number	% IgG Secreting Cells	IgG Producing Cells	Number of Unique CDR3		Number of IgG Pairs
				Heavy	Light	
#1	4,600	50	2,300	501	813	463
#2	4,600	50	2,300	612	984	562
#3	3,760	57	2,143	529	326	899

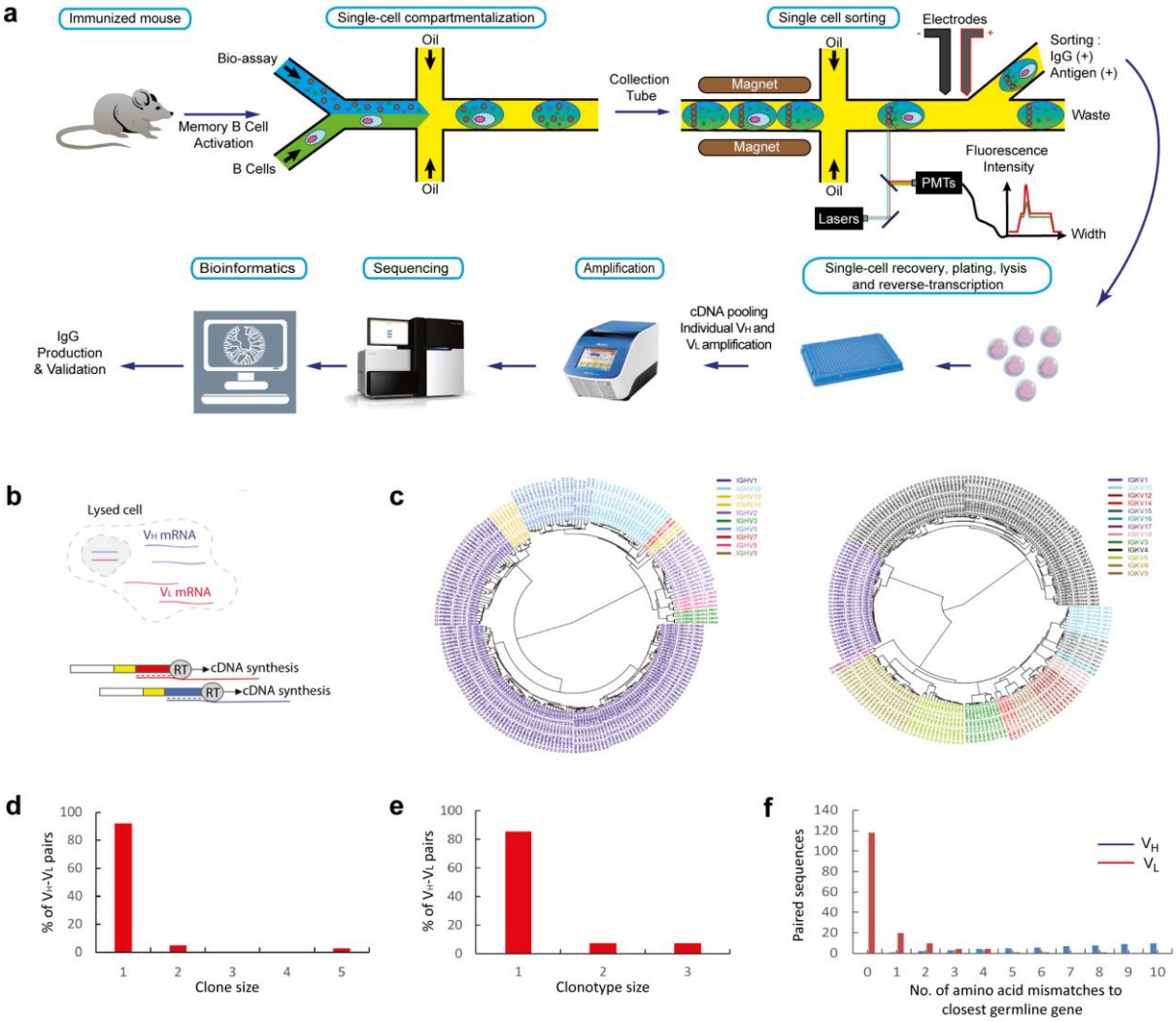


Supplementary Figure 23. Paired V_H-V_L repertoires of human peripheral B cells using CelliGo. Human peripheral blood mononuclear cells (PBMCs) collected by cytapheresis from two donors were enriched in memory B cells using the Memory B Cell Isolation Kit (Miltenyi), incubated with anti-CD40 antibodies and CpG oligodeoxynucleotides (CpG) for 4 days, and the indicated cell numbers ("starting cell number") encapsulated with barcoded hydrogel beads for paired V_H-V_L sequencing. Two runs (#1 and #2) were performed on PBMCs from Donor A and one run (#3) on polymorphonuclear cells (PMNCs) from donor B. The percentage and number of IgG-secreting cells encapsulated in drops was determined using an aliquot of the cell suspension immediately prior to encapsulation for an IgG-

ELISpot assay. (**a**) Statistics on the three runs including starting cell number, number of IgG producing cells, number of unique heavy and light chain CDR3s (both unpaired and paired at a read threshold of 40) and the number of V_H-V_L pairs at a 40-read threshold. (**b**) Circus plots of V_H-V_L detailing the associations of V-genes between pairs. Heavy chain V-genes are spanned to the left of the circle (red bars), while light chain V-genes are spanned to the right of the circle (blue bars). The width of the ribbon connecting heavy and light chain V-genes is proportional to the number of pairs using those genes. Ribbons from particular heavy chain V-gene families (e.g. IGHV1, IGHV4, etc.) are colored differently. A histogram of the number of somatic hypermutations on the amino acid level are plotted in the inner circle of the circus plot for each V_H-V_L pair. (**c**) Distribution of CDR3 lengths (in amino acids) for IgG heavy chains and kappa and lambda light chains in the three runs. (d). Distribution of non-synonymous somatic hypermutations (measured as the difference in the number of amino acids across the V-gene up to the CDR3 compared to the closest germline V-gene) in paired V_H-V_L.

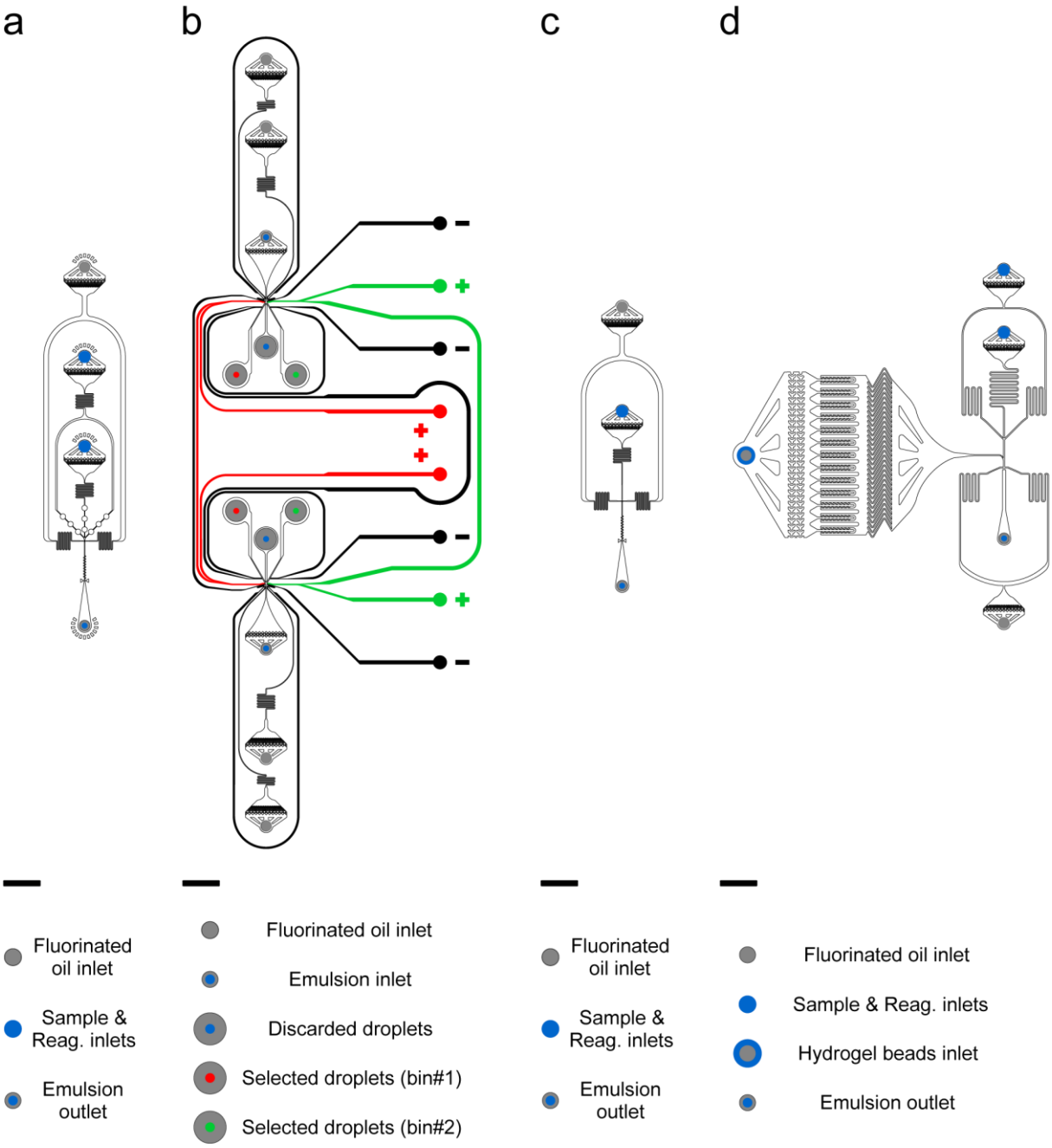


Supplementary Figure 24. Single-cell profiling using targeted RNA-seq. A 50:50 ratio of a human B cell line (Ramos) and T cell line (Jurkat) were mixed and single cells encapsulated in droplets together with single hydrogel beads carrying barcoded, gene-specific, cDNA primers together with reagents for cell lysis and reverse transcription. The protocol was similar to for paired antibody V_H-V_L sequencing, except that on each bead the 3'-end of the barcoded cDNA primers comprised a mixture of primer regions specific for 32 transcripts, selected to cover B and/or T specific genes based on public datasets, among which 12 where B cell specific, 14 were T cell specific and 6 were control genes expressed in both cell types. (**a**) Histograms of the distribution of targeted scRNA-seq raw sequencing reads per cell. (**b**) Percentage of variance as a function of principal components. (**c**) Histogram plot of mean of number of detected genes per cell (left) and per cell type (right), after their classification based on PCA dimensionality reduction. (**d**) tSNE dimensionality reduction of the targeted scRNA-seq dataset, highlighting B and T cells (n=1,428) based on their respective gene expression. The top panel shows total UMI count per cell. The three panels beneath show points colored according to UMI count on each cell for three genes, including a house-keeping gene (ATP5G3), top; one gene expressed in T cells (FYB), middle panel; and one gene expressed in B cells (CD79B), lower panel. Scale indicated on the right of each panel. (**e**) Heatmap of Pearson's correlation score between genes, grouped by hierarchical clustering (dark blue: not correlated, dark red: correlated) (n=1,428 cells analyzed). ***Note**: This Figure demonstrates that the cell phenotype following CelliGO can be distinguished based on a limited set of primers with high precision, where >80% of variance is explained by the first two principal components. Hence, the protocol can easily be adapted to the analysis of any transcript panel (or total RNA-seq).*



Supplementary Figure 25. Paired V_H-V_L repertoires following single-cell sorting into 384-well plates. (**a**) Schematic of selection of activated memory B cells and single-cell paired V_H-V_L sequence recovery using microplates: 30 Million frozen splenocytes from an antigen-immunized mouse were thawed and incubated *in vitro* for 3 days with mouse IL-2 and R-848. Cells were harvested, pre-enriched by pan B purification (Miltenyi), encapsulated in droplets and sorted using the CelliGO beadline assay, detecting binding of IgG to the soluble antigen used during immunization. Following emulsion breaking, cells were distributed into wells of a 384 well-plate using an acoustic liquid dispenser (Echo 555, Labcyte) at an average of one IgG-secreting cell per well. Each well of the plate contained lysis buffer, barcoded cDNA primers (with a unique barcode per well), reverse transcription buffer, dNTPs and reverse-transcriptase enzyme. After reverse transcription, the barcoded cDNAs were pooled, purified, and the V_H and V_L genes amplified separately using three rounds of PCR, and resulting in amplicons possessing the Illumina adaptors P7-SBS12 followed by the Illumina index sequence on one end of the fragment and P5-SBS3 on the other end of the fragment. Sequencing was performed using Illumina MiSeq 2x300bp paired-end sequencing. (**b**) Single-cell barcoding of V_H and V_L cDNA in a well of a 384 well-plate. Cells from sorted droplets were recovered and single cells distributed into wells in the presence of reverse transcription reagents and primers for reverse transcription of V_H and V_L mRNA tagged with a well-specific DNA barcode. (**c-f**) Example from one mouse immunized with a TNFR superfamily member. (**c**) Phylogeny of captured V_H-V_L pairs identified in screens of activated memory B cells member. Phylogenies were created using the full-length amino acid sequences of the V regions. Sequences are color-coded by the closest germline V gene family. A total of 177 V_H-V_L pairs were

identified (**d**) Distribution of the size of clones (families comprising V_H-V_L pairs with identical amino acid sequence but associated with different barcodes), corresponding to the same IgG expressed by different cells due to clonal expansion. (**e**) Relative frequency of the paired V_H-V_L sequences versus clonotype size. The clonotype size is equal to the number of non-redundant V_H-V_L sequences derived from the same ancestral V_H-V_L germline recombination event (i.e. derived from the same germline V-J genes, with CDR3s of the same length and differing by at least one amino acid across the variable region). (**f**) Distribution of non-synonymous somatic hypermutations. The distribution of amino acid changes compared to the closest germline V gene (corresponding to non-synonymous somatic hypermutations) in V_H and V_L genes in V_H-V_L pairs.



Supplementary Figure 26. Design of microfluidic chips. (**a**) Device to produce 40 pL droplets. (**b**) Device to sort 40 pL droplets by dielectrophoresis. Potential and ground electrodes are indicated with the "+" and "-" signs, respectively. Due to its symmetry, the same device can also be used to perform two-way sorting, sorting droplets into either bin#1 (left outlet), or bin#2 (right outlet), by activation of the red and green potential electrodes on left or right of the channel, respectively. However, only one-way sorting was used in this study. (**c**) Device to produce 125 pL droplets to make hydrogel beads, as described in (Klein, A.M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187-1201 (2015)). (**d**) Device for co-compartmentalization of single cells and single hydrogel bead in ~1 nL droplets, as described in (Klein, A.M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187-1201 (2015); Abate, A.R., Chen, C.H., Agresti, J.J. & Weitz, D.A. Beating Poisson encapsulation statistics using close-packed ordering. Lab Chip 9, 2628-2631 (2009)). Scale bars are all 2 mm.

Supplementary Tables

Supplementary Table 1. Summary of results throughout the Celli*GO* process.

	ELISA serum		#Cells (x 10^{-6})		Droplet data			Phenotype	Genotype[17]				
	ELISA EC50[1]	Avg Kd (nM)[2]	Total raw[3]	Post PanB*[4]	Encapsulated[5]	% IgG[6]	% specific[7]	%IgG / specific[8]	Hits[9]	Recovered ASCs[10]	V_H-V_L pairs[11]	Unique IgGs[12]	Families[13]	Germline[14]	Seq Multi[15]	Seq Single[16]
TT#1	73,698	0.9	524	23.8	2.5	2.2	1.1	50	22,113	9,650	1,727 (1,036)	1,053 (731)	123 (107)	5	987	65
TT#2	46,421	1.3	525	39.1	1.6	3.15	1.39	44	5,200	2,220	448	337	83	1	294	43
					Mean[18]	2.67	1.24	47		Mean[18]	742	534	95			
TT#5-pool	45,302	n.d.	n.d.	112	1.5	4.57	2.65	58	12,000	n.d.	806	480	219	18	307	232
GPI#1	14,559	0.82	137	7.8	1.5	2.03	0.91	44	4,400	n.d.	305	228	99	7	156	72
GPI#2	4,182	1.28	115	3.6	1.5	1.28	0.35	27	6,530	n.d.	146	109	48	6	78	31
GPI#3	12,880	1.17	318	2.7	1.5	4.36	0.62	14	3,500	n.d.	261	136	68	9	93	43
					Mean[18]	2.82	0.49	28		Mean[18]	237	158	72			

	FC EC50[1] serum	
	TSPAN8	parental
TSPAN8#1	6,976	8,786
TSPAN8#2	14,548	15,557

1. Serum antigen-specific EC50 values (by ELISA or FC, Flow Cytometry)
2. Average of serum IgG affinity (nM) against the target (measured by Surface Plasmon Resonance)
3. Number of cells harvested from spleen
4. Number of cells after enrichment for B cell lineages (post Pan B* or Post Pan B)
5. Number of cells encapsulated in 40 pL droplets
6. Percentage of droplets containing IgG secreting cells detected on the station
7. Percentage of droplets containing IgG secreting cells where antigen binding is detected on the station
8. Percentage of IgG secreting cells secreting antigen binding IgG compared to total IgG
9. Number of sorted droplets (hits) corresponding to cells secreting antigen binding IgG
10. Number of recovered IgG secreting cells (by EliSpot)
11. The number of total V_H-V_L pairs with ≥40 reads for V_H and ≥40 reads for V_L. This represents the number of cells captured (includes identical antibody sequences).
12. The number of different IgGs (V_H-V_L paired sequences with ≥40 reads for V_H and ≥40 reads for V_L differing by at least one amino acid)
13. Number of clusters of IgGs from column 11 where each cluster is a group of IgGs thought to derive from the same V(D)J rearrangement event and differing only by somatic hypermutation
14. The number of germline IgGs identified (non-redundant V_H-V_L pairs with ≥40 reads for V_H and ≥40 reads for V_L) and no non-synonymous changes compared to the closest germline V_H-gene and V_L-gene.
15. Number of IgG falling into multi-members clusters (clusters from column 14 that contain > 1 member)
16. Number of IgG falling into single-member clusters
17. Data in parentheses are normalized to 1.5 million encapsulated cells by random sampling. To normalize for the larger number of cells used, we randomly subsampled 60% (1.5/2.5) of VH/VL pairs and created 1000 sets. We then calculated the mean number of unique IgGs and families from that 1000 sampled sets.
18. Means and values used to calculate means are highlighted in yellow.

Supplementary Table 2. P values of V_H and V_L usage between different immunized mice

		VH usage										
		TT#1	TT#2	GPI#1	GPI#2	GPI#3	TSPAN8#1	TSPAN8#2	TT#5-pool	TT#1+2	GPI#1+2+3	TSPAN8#1+2
VL usage	TT#1		<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001			
	TT#2	<0.00001		<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001			
	GPI#1	<0.00001	<0.00001		0.0039	0.00063	<0.00001	<0.00001	<0.00001			
	GPI#2	<0.00001	<0.00001	0.013		0.20	0.0020	0.00015	0.00012			
	GPI#3	<0.00001	<0.00001	0.17	0.034		<0.00001	<0.00001	<0.00001			
	TSPAN8#1	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001		0.64	<0.00001			
	TSPAN8#2	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	0.14		<0.00001			
	TT#5-pool	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001		<0.00001	<0.00001	<0.00001
	TT#1+2								<0.00001		<0.00001	<0.00001
	GPI#1+2+3								<0.00001	<0.00001		<0.00001
	TSPAN8#1+2								<0.00001	<0.00001	<0.00001	

P values were adjusted according to Benjamini & Hochberg. See column 12 of Supplementary Table 1 for the number of sequences n of each TT#1, TT#2, GPI#1, GPI#2, GPI#3, TSPAN8#1 and TSPAN8#2 mouse. See the statistical section in the online methods for the description of the two-sided approximate permutation test used. Cells with blue background: p>0.05.

Supplementary Table 3. Expressed recombinant IgGs tested for binding to TT, GPI or TSPAN8-overexpressing cells

From the set of high-confidence V_H-V_L pairs (Supplementary File 2), 42 V_H-V_L pairs from a TT-immunized mouse (TT#1), 14 V_H-V_L pairs from a GPI-immunized mouse (GPI#2) and 20 V_H-V_L pairs from both TSPAN8-overexpressing cell immunized mice (TSPAN#1 + #2) were selected, cloned and expressed as mouse/human chimeric IgG1,κ antibodies as described in the Online Methods. For soluble antigens (TT and GPI) each candidate pair was tested for binding to their cognate antigen using ELISA and/or Octet (see Online Methods), column D. For TSPAN8, candidate pairs were tested for specific binding to the TSPAN8-overexpressing m300.19 cell line over the parental m300.19 cell line. In the "Antibodies tested" tab K_d values in column D (Octet for GPI affinity measurements, Flow cytometry for TSPAN8 measurements), the EC50 values are shown in column E (ELISA). For ELISA, non-binders are shown in red (-), weak binders are shown in light green (+) and strong binders in dark green (++) in Column C. For TSPAN8, two binding indexes are shown: (i) a $TSPAN8^+$/parental binding index is show in column C, corresponding to values presented in Supplementary Fig. 22b that were calculated from histogram plots represented in Supplementary Fig. 22a; the "binding index" is calculated from the ratio of median fluorescence from staining of TSPAN8-expressing M300.19 cells compared to staining of parental M300.19 cells ($Median^{TSPAN8+}$/$Median^{parental}$). Non-binders are indicated in red (index < 2). (ii) a mAb/Isotype control binding index is show in column D that was calculated from histogram plots represented in Supplementary Fig. 22d; the "binding index" is calculated from the ratio of median fluorescence from staining of TSPAN8-expressing M300.19 cells with the indicated mAb compared to staining with an isotype control (Rituximab, RTX) ($Median^{mAb}$/$Median^{RTX}$). Non-binders are indicated in red (index < 2). The K_d values of anti-TSPAN8 IgGs were calculated by fitting the data on antibody-concentration dependence of the fluorescent signal (see methods). Summary sequence information about each antibody is given in subsequent columns. Column J designates whether the V_H-V_L pair falls within a multi-member family of IgG that have undergone clonal expansion and affinity maturation. Columns M an N show the number of amino acid changes compared to the closest germline V_H-gene and V_L-gene, respectively (corresponding to non-synonymous somatic hypermutations), colored according to whether they are germline (bright red) or have low (1-

3, pink), medium (4-9, yellow) or high (10+, green) levels of mutation. The “Summary Table” tab gives a tabulated summary of the number of antibodies that were binders, classed based on whether they fell within multiple member families that derived from the same germline gene rearrangement (“Member of multi-member family”), or not.

Supplementary Table 4. List of unique VH-VL pairs identified across sorts of splenocytes from TT, GPI and TSPAN8-overexpressing cell immunized mice.
The table provides detailed information on each non-redundant VH-VL pair (containing at least one amino acid difference in either the VH or VL sequence between any two pairs) with ≥40 reads for VH and ≥40 reads for VL from screens of TT immunized mice (TT#1, TT#2, and TT#5-pool), GPI immunized mice (GPI#1, GPI#2 and GPI#3) and TSPAN8-overexpressing cell immunized mice (TSPAN8#1 and TSPAN8#2). VH-VL pairs are clustered into families if they derive from the same V-gene and J-gene and the same CDR3 lengths in both VH and VL. The cluster number is indicated in column C. All other attributes of the antibodies are included in subsequent columns including germline V-gene and J-gene, number of nucleotide and amino-acid substitutions compared to the germline, read abundances associated with the consensus sequence, CDR1-3 amino acid sequences and full length amino and nucleotide sequences of the VH and VL.

Supplementary Table 5. Sequences of single cell barcode indexes.
The first column indicates the position of the top-strand and bottom-strand index oligonucleotides in the 96-well plates, the second column the sequence identity (ID) and the third column the sequence of the oligonucleotide. The double stranded DNA was generated by annealing at equimolar ratio the top-strand oligonucleotide with the bottom-strand oligonucleotide.

Supplementary Movie Legend

Supplementary Movie. Movie of cell and hydrogel bead co-compartmentalization into ~1 nL droplets. The aqueous phase containing cells and hydrogel beads, as well as lysis and reverse transcription reagents (right) were co-compartmentalized at the T-junction (middle) to form ~1 nL droplets (left) containing reagents necessary to lyse cells and initiate reverse transcription of V_H-V_L mRNA to form barcoded cDNA.

Supplementary Note

Supplementary Note. Methods pertaining to Bioinformatics Analysis

We developed a dedicated bioinformatics pipeline (Absolution^{TM}) for analysis of single-cell antibody sequences. Paired-end reads were trimmed at the 3' end to remove bad quality bases that were below a Phred quality score of Q20 using the program trim_galore [https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/]. If the read was trimmed below 155 bp, both paired-end reads were removed. Remaining paired-end reads were merged using the program FLASH (Magoc, T. & Salzberg, S.L. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957-2963 (2011)) requiring at least a 10 bp overlap. Barcodes were extracted from merged reads by first searching for the constant 4 bp linkers found between the 20-mer indices of the barcode (Supplementary Fig. 10a and Supplementary File 1), allowing up to 1 mismatch in each linker. If the correct linkers were identified, the entire barcode sequence was separated from the rest of the IgG-containing sequence, and the three interspersed 20-mer indices were extracted and concatenated together to form a 60 bp non-redundant barcode sequence. A library of all 884,736 combinations of the 3 sets of 96 indices (96^3) was used to map barcode sequences using the sensitive read mapper Cushaw3 (Liu, Y., Popp, B. & Schmidt, B. CUSHAW3: sensitive and accurate base-space and color-space short-read alignment with hybrid seeding. PLoS One 9, e86869 (2014)). Each set of indices is error-correcting because it takes more than an edit-distance of 3 to convert one index into another. We therefore set a total mismatch threshold of 4 across the entire barcode, with a maximum of 2 per index to avoid mis-assigning sequences to the wrong barcode identity. In a second, slower step, sequences that could not be mapped to the Cushaw3 index-library were split into their individual indices and each index compared against the set of 96 possible indices, allowing up to 2 mismatches in each individual index. Any sequences not assigned to a particular barcode ID by these two steps was discarded. DNA barcoding allows us to derive those sequences that came from a single cell and therefore to use that information to help us filter out variants introduced by the process of reverse transcription, PCR amplification and sequencing errors by creating a consensus that should remove mutations introduced at low level. To identify all the distinct sequence species in each barcode cluster, we used a

DNA clustering algorithm, dnaclust (Ghodsi, M., Liu, B. & Pop, M. DNACLUST: accurate and efficient clustering of phylogenetic marker genes. BMC Bioinformatics 12, 271 (2011)), requiring that sequences share at least 93% to group sequences likely to represent the same IgG chain. To avoid sequence from diverse PCR amplification primers and facilitate the clustering process the first 18 bp of each sequence was removed. After clustering, all groups with at least 40 sequences have a consensus sequence created from them, by aligning up to 200 sequences using ClustalO (Sievers, F. & Higgins, D.G. Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol Biol 1079, 105-116 (2014)) and then creating a consensus by taking the most abundant base for each position that has nucleotide sequence in at least 50% of positions in the alignment (i.e. not a gap). The resulting consensus sequences were characterized for immunoglobulin content (identifying the closest V(D)J germline genes and the identity of the CDR1, 2 and 3) using VDJFasta [https://sourceforge.net/projects/vdjfasta/] against an updated and modified database of V,D and J gene sequences derived from IMGT (downloaded 1st April 2019, http://www.imgt.org/vquest/refseqh.html#VQUEST). To create this database, we removed all pseudogenes (sequences with premature stop codons) to avoid any spurious mis-assignment to productive antibodies. The mice strain we carried out our immunizations in was BALB/c and while IMGT contains sequences from several mice strains including BALB/c (22% of the sequences), it is clear from the literature that additional strain-specific BALB/c variants exist that might affect the analysis (particularly of SHMs) (Collins, A.M., Wang, Y., Roskin, K.M., Marquis, C.P. & Jackson, K.J. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. Philos Trans R Soc Lond B Biol Sci 370 (2015)). To add additional BALB/c variant V-gene sequences we used a dataset of BALB/c variants (Corcoran, M.M. et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. Nat Commun 7, 13642 (2016)) using the IgDiscover algorithm. We compared this dataset of 223 nucleotide sequences with our pre-existing database of V-genes from IMGT using BLASTn. We identified 77 non-redundant, productive sequences that contained at least a single mismatch to a closely related pre-existing germline gene (>=90% identity) but no novel genes (<90% identical to a pre-existing gene) in our IMGT database. We integrated these new sequences into our IMGT V-gene database and named the new variant genes with the same name as their closest pre-existing gene (e.g. IGHV2-3*05, IGHV2-3*06). After immune characterization of consensus reads by VDJFasta, reads containing frameshifts, stop-codons or lacking identifiable CDRs were filtered out. V_H-V_L pairing was carried out by identifying the most abundant V_H and V_L consensus sequence (by number of reads that contributed to that consensus) in each barcode cluster. The paired V_H and V_L sequences must be larger than any other V_H or V_L present in the cluster by at least 1 read. To minimize V_H-V_L mispairing, antibody sequences were only considered for further analysis if both the paired VH/VL consensus sequences comprised at least 40 reads (Supplementary Fig. 10). Low-level mispairing (wrong assignment of light chain with heavy chain) was removed by clustering all heavy chains with the same V-J gene combination and a CDR3 amino acid sequence within

a hamming distance of 2 and using the paired light chain associated with the largest number of independent barcodes.

Single-cell analysis has a distinct advantage over bulk repertoire analysis in that we can measure evidence of affinity maturation and clonal expansion of progenitor B cells. Clonal expansion was analysed by clustering V_H-V_L pairs in two ways (see Fig. 3f inset). First, V_H-V_L pairs with identical sequence but associated with different barcodes, corresponding to identical IgGs expressed by different cells due to clonal expansion were clustered (termed 'clones'; Fig. 3f inset, dark blue regions). The number of members of each such cluster was used to quantify clonal expansion of cells expressing specific IgGs. The common IgG sequence of each such cluster was classed as a non-redundant IgG sequence. Second, V_H-V_L pairs from different cells, but are all likely derived from the same primary germline gene V(D)J rearrangement following clonal expansion and somatic hypermutation were clustered if they derived from the same V and J gene in both heavy and light chains and their CDR3 amino acid lengths were the same (termed 'clonotypes'; Fig. 3f inset, light blue region). Unlike bulk analyses which are based on unpaired heavy and light chains, we were able to make these assumptions because we now have natively paired light chain information. Therefore, V_H-V_L pairs which derive from the same V and J gene and have the same CDR3 length in the heavy and light chain respectively are highly unlikely to derive from two independent recombination events, at least not within the same repertoire. The number of members of each such cluster was used to quantify both clonal expansion of cells expressing primary germline antibodies and their progeny expressing antibodies with somatic hypermutations.