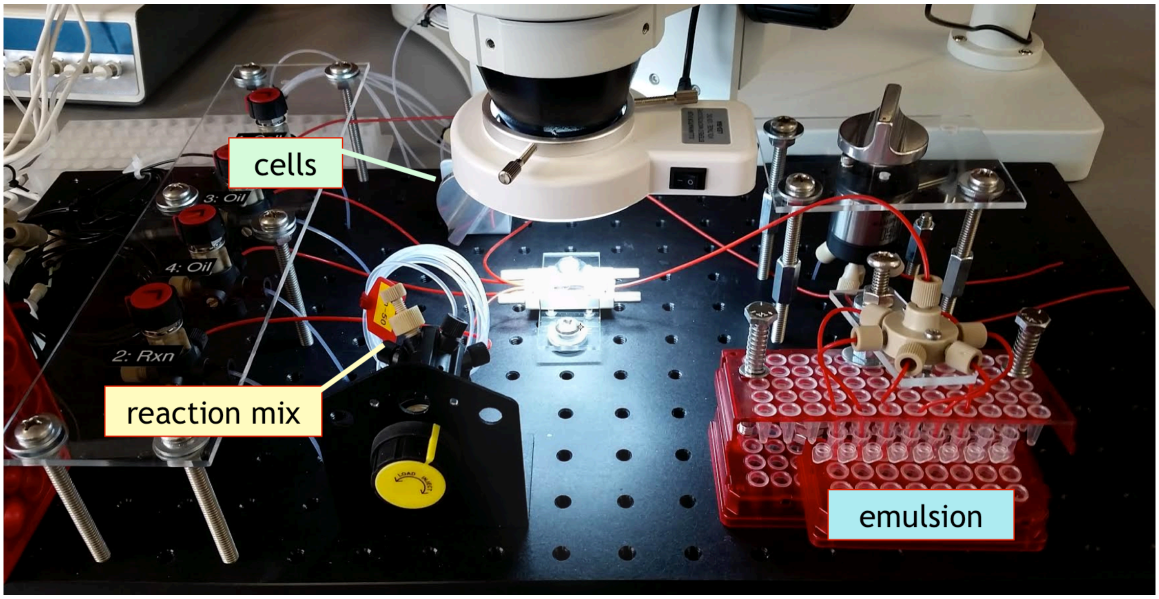
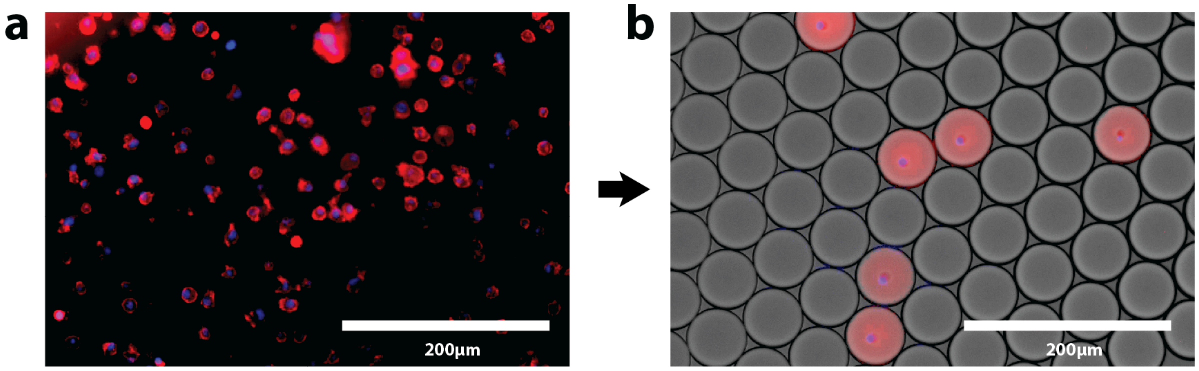
Supplementary Information

This file contains 3 Supplementary Figures, 1 Supplementary Table, and Supplementary Methods.

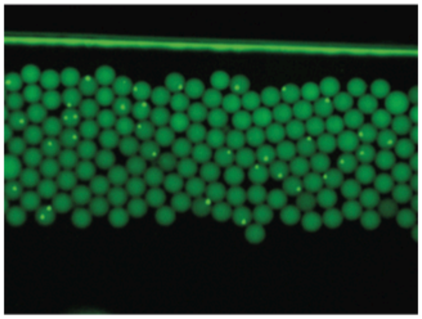


Supplementary Figure 1. Microfluidic emulsion generation platform. Injection loops for cell and and reaction aqueous phases, respectively, are shown.

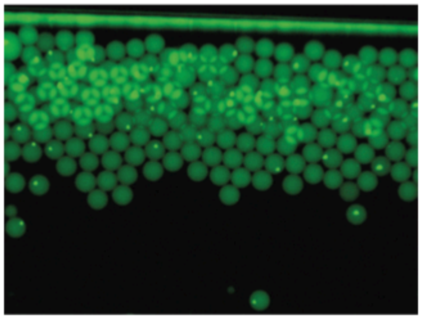


Supplementary Figure 2. Controlled cell lysis in emulsion using 2-reagent input chip (Main Text Figure 1a, chip from Dolomite Microfluidics). Immortalized B-cells were stained with NucBlue nuclear stain (Life Technologies, blue) and CellMask Deep Red Plasma membrane Stain (Life Technologies, red) according to manufacturer's instructions before entry into emulsion. Note the presence of lysed plasma membrane only within the droplets containing nuclei, which themselves have not completely degraded.

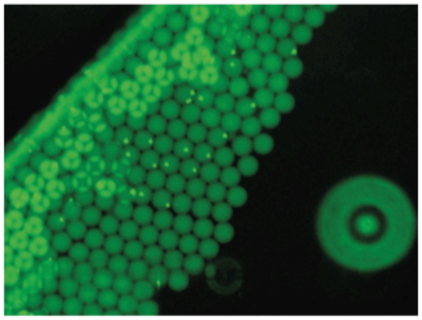
Non-overlapped droplets:	**148**
1-cell droplets:	**30**
2-cell droplets:	**2**
3-cell droplets:	**0**
average cells per droplet:	**0.23**



Non-overlapped droplets:	**65**
1-cell droplets:	**11**
2-cell droplets:	**0**
3-cell droplets:	**0**
average cells per droplet:	**0.16**



Non-overlapped droplets:	**94**
1-cell droplets:	**14**
2-cell droplets:	**0**
3-cell droplets:	**1**
average cells per droplet:	**0.15**

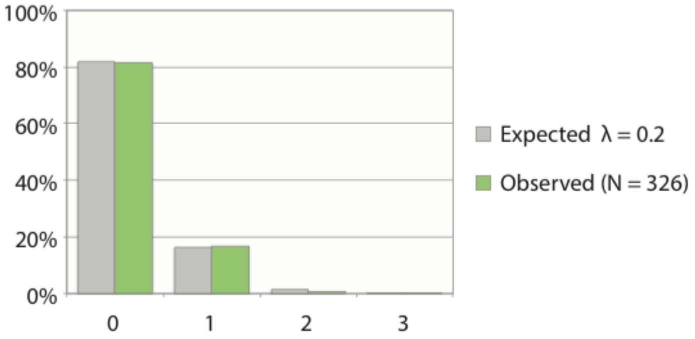


TOTAL

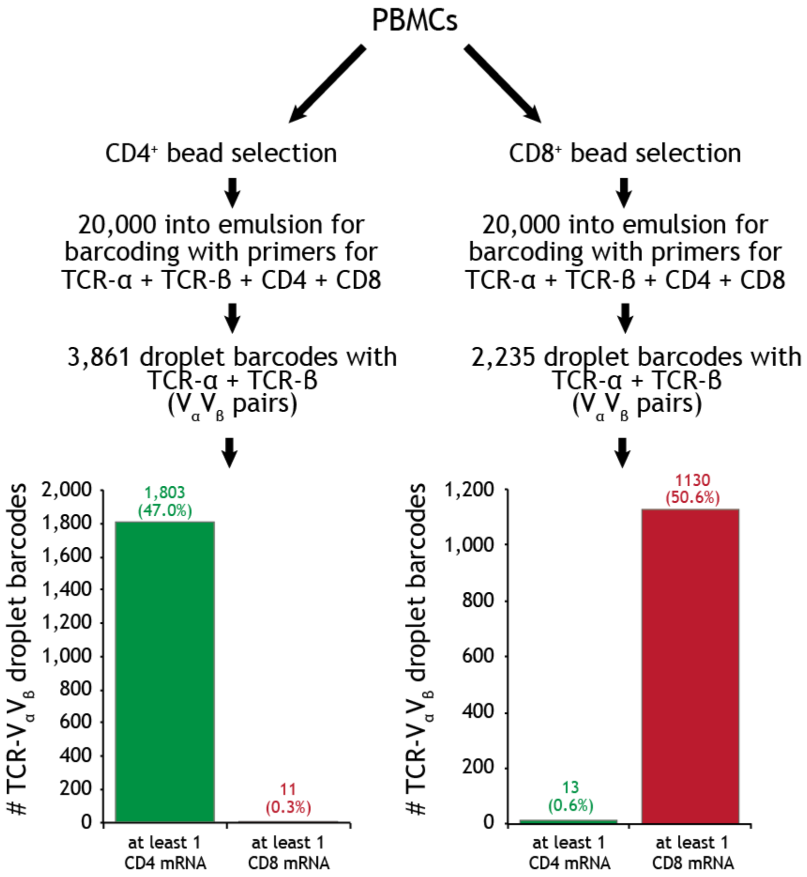
Non-overlapped droplets:	**326**
1-cell droplets:	**55**
2-cell droplets:	**2**
3-cell droplets:	**1**

average cells per droplet:
0.19

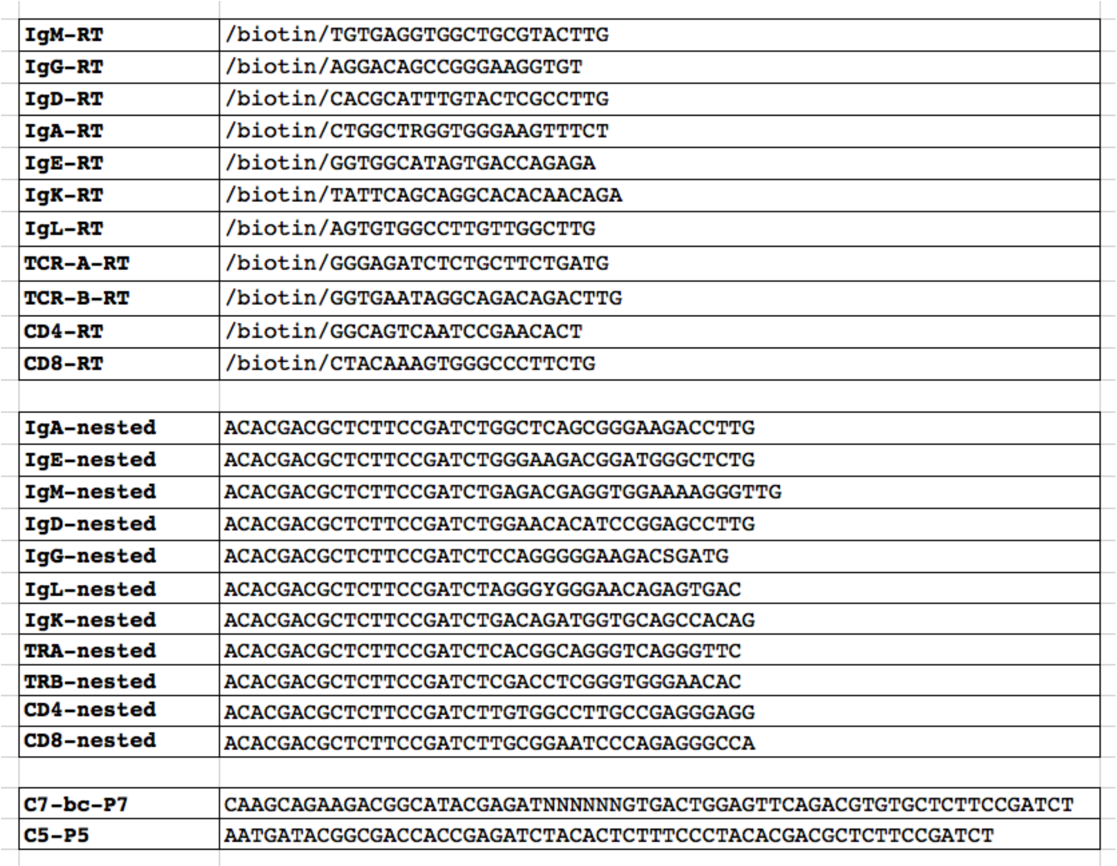
proportion of single-cell droplets:
94.8% (85.6-98.9 95%c.i.)



Supplementary Figure 3. Estimates of per-droplet cell occupancy for the emulsion containing 3 million healthy B-cells. Cells were initially entered into emulsion at an estimated ~0.2/droplet based on an input concentration of 3.1M/ml and an estimated droplet volume of 65pl. The mixture contains EvaGreen dye allowing visualization of the cell nuclei which are not destroyed by the detergent lysis conditions.



Supplementary Figure 4. Capture of additional phenotypic marker genes to annotate TCR alpha-beta pairs. Peripheral PBMCs from a healthy volunteer were separated into CD4+ and CD8+ populations using corresponding negative selection bead enrichment kits (EasySep, Stem Cell Technologies), before barcoding emulsion processing using primers mixes to simultaneously target TCR alpha and beta constant region and CD4 and CD8 transcripts. After sequencing (~5M reads on MiSeq), filtered droplet barcodes containing a TCR $V_\alpha V_\beta$ pair were annotated by their association with at least one CD4 or CD8 mRNA. Around half of each population could be linked to an mRNA of the expected phenotypic marker gene. Although the false negative rate is considerable (expected marker gene not found in a droplet barcode), the false positive rate is low ("wrong" marker gene found in a droplet barcode), a typical result for single-cell mRNA studies where despite presence of a particular protein, corresponding transcripts at any moment in time may be rare or absent.



IgM-RT	/biotin/TGTGAGGTGGCTGCGTACTTG
IgG-RT	/biotin/AGGACAGCCGGGAAGGTGT
IgD-RT	/biotin/CACGCATTTGTACTCGCCTTG
IgA-RT	/biotin/CTGGCTRGGTGGGAAGTTTCT
IgE-RT	/biotin/GGTGGCATAGTGACCAGAGA
IgK-RT	/biotin/TATTCAGCAGGCACACAACAGA
IgL-RT	/biotin/AGTGTGGCCTTGTTGGCTTG
TCR-A-RT	/biotin/GGGAGATCTCTGCTTCTGATG
TCR-B-RT	/biotin/GGTGAATAGGCAGACAGACTTG
CD4-RT	/biotin/GGCAGTCAATCCGAACACT
CD8-RT	/biotin/CTACAAAGTGGGCCCTTCTG

IgA-nested	ACACGACGCTCTTCCGATCTGGCTCAGCGGGAAGACCTTG
IgE-nested	ACACGACGCTCTTCCGATCTGGGAAGACGGATGGGCTCTG
IgM-nested	ACACGACGCTCTTCCGATCTGAGACGAGGTGGAAAAGGGTTG
IgD-nested	ACACGACGCTCTTCCGATCTGGAACACATCCGGAGCCTTG
IgG-nested	ACACGACGCTCTTCCGATCTCCAGGGGGAAGACSGATG
IgL-nested	ACACGACGCTCTTCCGATCTAGGGYGGGAACAGAGTGAC
IgK-nested	ACACGACGCTCTTCCGATCTGACAGATGGTGCAGCCACAG
TRA-nested	ACACGACGCTCTTCCGATCTCACGGCAGGGTCAGGGTTC
TRB-nested	ACACGACGCTCTTCCGATCTCGACCTCGGGTGGGAACAC
CD4-nested	ACACGACGCTCTTCCGATCTTGTGGCCTTGCCGAGGGAGG
CD8-nested	ACACGACGCTCTTCCGATCTTGCGGAATCCCAGAGGGCCA

C7-bc-P7	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
C5-P5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

Supplementary Table 1. Target-specific primers (BCR heavy+light, TCR alpha+beta, CD4 and CD8) used in the study for reverse transcription in emulsion and nested PCR post-emulsion. Nested primers carry 5' tails matching the Illumina P5 sequencing primer sequence. Illumina adaptor primers added in the final PCR step are shown.

Supplementary Methods

Read processing and isotype assignment.
Illumina MiSeq reads were processed using custom pipelines built around the pRESTO package[1] to generate full length consensus sequences for mRNA molecules and droplets, annotated with IgBLAST [2] and IMGT/HighV-QUEST [3], and processed with custom scripts and the Change-O package [4] to generate statistics and figures. MiSeq reads were demultiplexed using Illumina software. Positions with less than Phred quality 5 were masked with Ns. Isotype-specific primers, droplet barcodes (DBs), and molecular barcodes (MBs) were identified in the amplicon and trimmed, using pRESTO MaskPrimers-cut with maximum error 0.2. A read 1 and read 2 consensus sequence was generated separately for each mRNA from reads grouped by unique molecular identifier (UMI, comprised by the DB and MB together), which are PCR replicates arising from the same original mRNA molecule of origin. UMI read groups were aligned with MUSCLE, and pRESTO was used to BuildConsensus with parameters “—maxdiv 0.1 –bf PRIMER –prfreq 0.6 –maxmiss 0.5 –q 5”, requiring >=60% of called PCR primer sequences agree for the read group, maximum nucleotide diversity of 0.1, using majority rule on indel positions, and masking alignment columns with low posterior (consensus) quality. Paired end consensus sequences were then stitched in two rounds. First, ungapped alignment of each read pair’s consensus sequence termini was optimized using a Z-score approximation and scored with a binomial p-value as implemented in pRESTO AssemblePairs-align “--minlen 8 --alpha 1e-5 --maxerr 0.3”. For read pairs failing to stitch this way, stitching was attempted using the human BCR and TCR germline V exons to scaffold each read prior to stitching or gapped read-joining, using pRESTO’s AssemblePairs-reference “--minident 0.5 --evalue 1e-5”.

V(D)J segment annotation and isotype confirmation
We used IgBLAST, Change-O, and custom scripts to identify the germline V(D)J genes of origin, trim mRNA sequences to V(D)J region, identify the CDR3 region, and calculate mutation from germline V nucleotide sequence. IgBLAST counts Ns as mismatches but mRNA sequences with more than 6 V-region Ns were filtered for mutation analyses and cross-fraction pairing precision analysis. For IG heavy chains, isotype identity was confirmed by matching non-primer C-regions (constant region exons) to expected sequences using pRESTO MaskPrimers-score “--start 0 --maxerr 0.2”. Amplicons with discordant primer/non-primer C-region calls were discarded, except for two primer/non-primer combinations where a specific primer crosstalk event was resolved by visual inspection.

Grouping V(D)J sequences into clonal lineages.
V(D)J sequences were grouped into clones using single-linkage clustering with a weighted intraclonal distance as previously described [5]. Clustering was performed with Change-O package DefineClones-bygroup “—model m1n –gene first –dist 4.0 –

norm none". First, all functional IGHV chains' droplet consensus sequences were binned into V-J-junction bins, such that sequences possibly arising from the same initial recombination event were binned together (based on matching best IGHV gene, IGHJ gene, and junction length as identified by IMGT/HighV-QUEST [3]. The intraclonal distance threshold was chosen by generating a histogram of nearest-neighbor distances within each IGHV bin using the distToNearest function of Change-O's shm package, and visually inspecting the histogram for a natural distance cutoff (in the trough of a bimodal histogram). Light chains' clonal clusters were defined using the same distance model and threshold.

Droplet filtering, pairing fidelity calculation

Heavy-light pairing confidence was assessed in two independent ways: using intra-droplet mRNA sequence agreement, and inter-replicate pair agreement. Intradroplet mRNA agreement was defined as mean pairwise nucleotide difference (Nei's pi [6] <0.02 of V(D)J sequences within a locus. mRNA sequences were trimmed down to V(D)J nucleotide coding sequences using IgBLAST annotations. Within each droplet all productive mRNA sequences were grouped by V locus; within each group sequences were multiple aligned using MUSCLE as implemented in pRESTO AlignSets using default parameters. Droplet consensus chains were built from multiple mRNAs per locus using the pRESTO parameters "BuildConsensus.py --maxdiv 0.2 --maxmiss 0.5". Randomly shuffled droplets were used to select the diversity cutoff pi<=0.02; in shuffled droplets less than 0.01% of heavy chain loci (<0.2% of light chain loci) met this criteria. Multi-cell or immune-receptor included droplets were separated for further precision analysis.

Pairing precision was calculated based on observation of the same clone-pair across multiple "replicates" (separate emulsion experiments), similar to previous methods [7], focusing on those VDJ clusters likely containing only a single lineage, that is, arising from a single V(D)J and VJ rearrangement followed by expansion. Similar VDJ rearrangements can arise within an individual multiple independent times, leading to the same heavy chain V(D)J rearrangement natively paired with multiple different light chain VJ rearrangements. Because rare V(D)J rearrangements would provide a more accurate measure of our technical precision, we focused on long heavy CDR3s (HCDR3) for this analysis (as a proxy for rarer V(D)J rearrangements). We also removed sequences with >6Ns to increase clonal assignment confidence. Pairing precision increased with HCDR3 length to over 96% for the longest quartile of clones observed across fractions (2,604 clones with junction length>=54nt). Because the probability of clone-pair agreements is the joint probability of true pairs in two independent experiments, pairing precision is estimated as the square root of the pairing agreement across replicates, calculated as follows:

d_{hl}^{f} is the number of droplet barcodes *d* with paired heavy clone *h* and light clone *l,* and found in physical fraction *f*. Mean (squared) pairing precision for each experiment is estimated by averaging, over heavy clones *h* and all pairs of fractions (*f,g*), the agreement of paired light clones (*l,k):*

$$\langle precision^2 \rangle \ = mean(P_f P_g)$$

$$= \frac{consistent\ heavy\ light\ pairs\ across\ fractions}{total\ pairs\ where\ heavy\ clone\ seen\ across\ fractions}$$

$$\frac{consistent\ heavy\ light\ pairs}{consistent\ pairs + inconsistent\ pairs} = \frac{\sum_h (\sum_{l=k}^{f\neq g} d_{hl}^f \cdot d_{hk}^g)}{\sum_h (\sum_{l=k}^{f\neq g} d_{hl}^f \cdot d_{hk}^g + \sum_{l\neq k}^{f\neq g} d_{hl}^f \cdot d_{hk}^g)}$$

$$\langle precision^2 \rangle \ = \frac{33157}{35922}$$

Therefore the mean precision of each experiment, (to within the variance in precision between experiments) is 96.1%.

HIV phylogenetic analysis

New broadly-neutralizing antibodies (bNAbs) to HIV were discovered by mining our high-throughput paired antibody processed sequences for similarity to known bNAbs. Previously known bNAbs from PGT-donor and other donors were mined from the literature [8,9]. All HIV IGH mRNAs recovered from emulsion were scored for similarity to known CDR3 amino acid sequences via tblastx [10]. Using IGH mRNA sequences from a healthy donor to generate a background distribution of sequence similarities, a bit score cutoff of 27 was used to segregate candidate bNAb-like CDR3s for further analysis. V(D)J sequences of candidate sequences were aligned to known bNAb's using MUSCLE [11] with default parameters, and in particular to PGT-donor lineages using default parameters except "--gapopen -15". Trees were generated with PhyML [12] default parameters, manipulated and visualized with Newick Utils [13] and Dendroscope[14] and manually inspected to select immunoglobulin heavy chain sequences interspersing with known bNAbs sequences. Consensus sequences for each droplet were built as previously described with manual inspection of alignments of any within-droplet amino acid conflicts using in JALVIEW[15]. Eight heavy chain sequences and their natively paired light chain antibody sequences were selected for synthesis, cloning expression. and neutralization assays.

Data analysis and plotting

Plots were generated using the dplyr and ggplot2 R packages[16,17]. Data was randomly downsampled and/or jittered with R for visualization purposes only in scatter plot figures. Downsampling minimum was 20,000 droplets per isotype or as otherwise noted on plot. Points were jittered by adding vertical and horizontal noise drawn from the same uniform probability distribution, with maxima <=0.2 for mRNA units and <=0.6 % for mutation.

Supplementary References

1. Vander Heiden, J. A. *et al.* pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinforma. Oxf. Engl.* 30, 1930–1932 (2014).
2. 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).
3. Giudicelli, V. *et al.* IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res.* 34, D781–784 (2006).
4. N Gupta *et al.* Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinforma. Oxf. Engl.* in press
5. Stern, J. N. H. *et al.* B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Transl. Med.* 6, 248ra107 (2014).
6. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* 76, 5269–5273 (1979).
7. DeKosky, B. J. *et al.* In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* 21, 86–91 (2015).
8. Walker, L. M. *et al.* Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* 477, 466–470 (2011).
9. Eroshkin, A. M. *et al.* bNAber: database of broadly neutralizing HIV antibodies. *Nucleic Acids Res.* 42, D1133–1139 (2014).
10. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402 (1997).
11. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797 (2004).
12. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321 (2010).
13. Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinforma. Oxf. Engl.* 26, 1669–1670 (2010).
14. Huson, D. H. & Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* 61, 1061–1067 (2012).
15. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinforma. Oxf. Engl.* 25, 1189–1191 (2009).
16. Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer New York, 2009). at <http://had.co.nz/ggplot2/book>
17. Wickham, H. & Francois, R. *dplyr: A Grammar of Data Manipulation*. (2015). at <http://CRAN.R-project.org/package=dplyr>