

High-resolution antibody dynamics of vaccine-induced immune responses

Uri Laserson^{a,b,c,1}, Francois Vigneault^{b,d,1}, Daniel Gadala-Maria^e, Gur Yaari^f, Mohamed Uduman^e, Jason A. Vander Heiden^e, William Kelton^g, Sang Taek Jung^g, Yi Liu^h, Jonathan Laserson^h, Raj Chari^b, Je-Hyuk Lee^b, Ido Bachelet^b, Brendan Hickeyⁱ, Erez Lieberman-Aiden^j, Bozena Hanczaruk^k, Birgitte B. Simen^k, Michael Egholm^k, Daphne Koller^h, George Georgiou^g, Steven H. Kleinstein^{e,f}, and George M. Church^{a,b,2}

^aHarvard-MIT Division of Health Sciences and Technology, Cambridge, MA 02139; ^bDepartment of Genetics, Harvard Medical School, Boston, MA 02115; ^cDepartment of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139; ^dAbViro, Boston, MA 02210; ^eInterdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven CT 06520; ^fDepartment of Pathology, Yale University School of Medicine, New Haven, CT 06520; ^gDepartment of Chemical Engineering, University of Texas at Austin, Austin, TX 78712; ^hDepartment of Computer Science, Stanford University, Stanford, CA 94305; ⁱDivision of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115; ^jHarvard Society of Fellows, Harvard University, Cambridge, MA 02138; and ^k454 Life Sciences, A Roche Company, Branford, CT 06405

Contributed by George M. Church, January 25, 2014 (sent for review May 3, 2013)

The adaptive immune system confers protection by generating a diverse repertoire of antibody receptors that are rapidly expanded and contracted in response to specific targets. Next-generation DNA sequencing now provides the opportunity to survey this complex and vast repertoire. In the present work, we describe a set of tools for the analysis of antibody repertoires and their application to elucidating the dynamics of the response to viral vaccination in human volunteers. By analyzing data from 38 separate blood samples across 2 y, we found that the use of the germ-line library of V and J segments is conserved between individuals over time. Surprisingly, there appeared to be no correlation between the use level of a particular VJ combination and degree of expansion. We found the antibody RNA repertoire in each volunteer to be highly dynamic, with each individual displaying qualitatively different response dynamics. By using combinatorial phage display, we screened selected VH genes paired with their corresponding VL library for affinity against the vaccine antigens. Altogether, this work presents an additional set of tools for profiling the human antibody repertoire and demonstrates characterization of the fast repertoire dynamics through time in multiple individuals responding to an immune challenge.

next-generation sequencing | influenza | immunology

The immune system is able to rapidly sense and respond to a vast array of invading organisms. Its arsenal contains systems that are immediately effective against commonly seen patterns (innate immunity) and systems that are capable of responding to novel invaders (adaptive immunity). Given the acute nature and diversity of infections, the immune system must be capable of rapid recognition of a pathogen, amplification of the response, and subsequent contraction of the response after the resolution of the infection. Adaptive immune responses rely on the continuous selection and amplification of specific clones from an enormous library of immune receptors (antibodies and T cell receptors). Specifically, stimulation of B-cell immunity results in the synthesis of antibodies that are secreted into the blood stream or into the mucosa as well as the programming of B memory cells that play a crucial role in the generation of rapid protective responses upon reinfection.

Currently, many immunology studies depend on characterizing lymphocyte subsets (e.g., assaying cell-surface receptors) and the ability to correlate them to encoded genetic information (1). Recent advances in next-generation sequencing (NGS) (2) have enabled any DNA-encodable assay to produce massive amounts of data. Indeed, NGS has enabled unprecedented views into the immune repertoire, as its immune receptor diversity is genetically encoded within a complex collection of lymphocytes (3–8).

The present study set out to dissect the rapid dynamics of the complete human peripheral antibody response against a controlled

immune challenge (vaccination), without the a priori notion of cell state markers or functions. We vaccinated three individuals with approved subunit vaccines designed to confer protection to viral antigens and banked blood samples at multiple time points before and after the vaccinations. The dynamic behavior of the immune repertoire in response to the vaccination was analyzed by RT-PCR of heavy chain V (VH) genes followed by DNA sequencing by using the 454 platform. We found that the VH gene repertoire is highly dynamic, and the response in each vaccination is qualitatively different. In contrast, we found that each individual uses the germ-line-encoded library of antibody components in very similar ways.

Because the immune system is shaped by selective pressures at the population and somatic levels (9), we observed statistical evidence that the immune system is equally likely to expend energy on expanding nondiversifying VJ combinations (innate action) as it is on expanding diversifying and class-switching VJ combinations (adaptive action). Finally, we synthesized a collection of the strongest-responding VH genes 1 wk after vaccination, paired them to their corresponding VL gene libraries by combinatorial phage display, and tested them for affinity against the vaccine antigens.

Results

Vaccination Time-Course Design. We characterized the antibody repertoire of three Personal Genome Project volunteers (three authors, G.M.C., I.B., and F.V.) in response to vaccination. In 2008, G.M.C. was vaccinated against seasonal influenza,

Significance

The immune system must constantly adapt to combat infections and other challenges. This is accomplished by continuously evolving the antibody repertoire, and by maintaining memory of prior challenges. By using next-generation DNA sequencing technology, we have examined the sheer amount of antibody made by individuals during a flu vaccination trial. We demonstrate one of the first characterizations of the fast antibody dynamics through time in multiple individuals responding to an immune challenge.

Author contributions: U.L., F.V., and G.M.C. designed research; U.L., F.V., W.K., S.T.J., J.-H.L., I.B., B. Hanczaruk, B.B.S., M.E., G.G., and G.M.C. performed research; Y.L., J.L., E.L.-A., and D.K. contributed new reagents/analytic tools; U.L., F.V., D.G.-M., G.Y., M.U., J.A.V.H., R.C., B. Hickey, S.H.K., and G.M.C. analyzed data; U.L., F.V., and G.M.C. wrote the paper, and M.E., D.K., G.G., S.H.K., and G.M.C. supervised work.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

¹U.L. and F.V. contributed equally to this work.

²To whom correspondence should be addressed. E-mail: gmc@harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1323862111/-DCSupplemental.

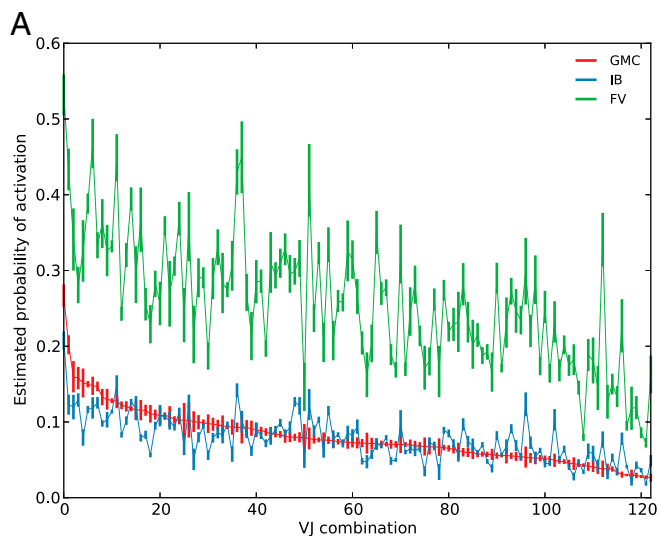
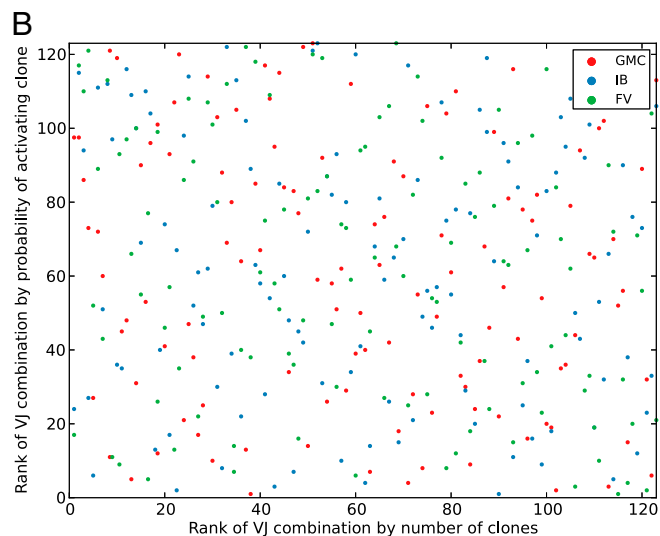


Fig. 2. Probability of clone activation by VJ use. (A) Each clone is labeled according to its VJ use and whether it is naïve (IgM-only with low mutation) or activated (IgG or IgA with high mutation). The probability of observing naïve or activated clones is estimated assuming a binomial distribution. The line plots the estimated probability of activation, and bars represent ± 1 SD. VJ combinations are ordered according to G.M.C. (B) For each VJ combination, we plot its VJ-use frequency rank against its rank of probability of activation. This is filtered on VJ combinations for which we obtain at least 100 clones that are classifiable as naïve or activated.

individuals (e.g., G.M.C. vs. F.V.) and also similar over time within an individual (e.g., G.M.C. in 2008 vs. 2009). For each time point, we compute a vector of use for the possible



combinations of the VJ gene segments. Indeed, the Spearman correlation between VJ-use vectors was consistently high across time points and individuals (*SI Appendix, Fig. S4*) and VJ-use

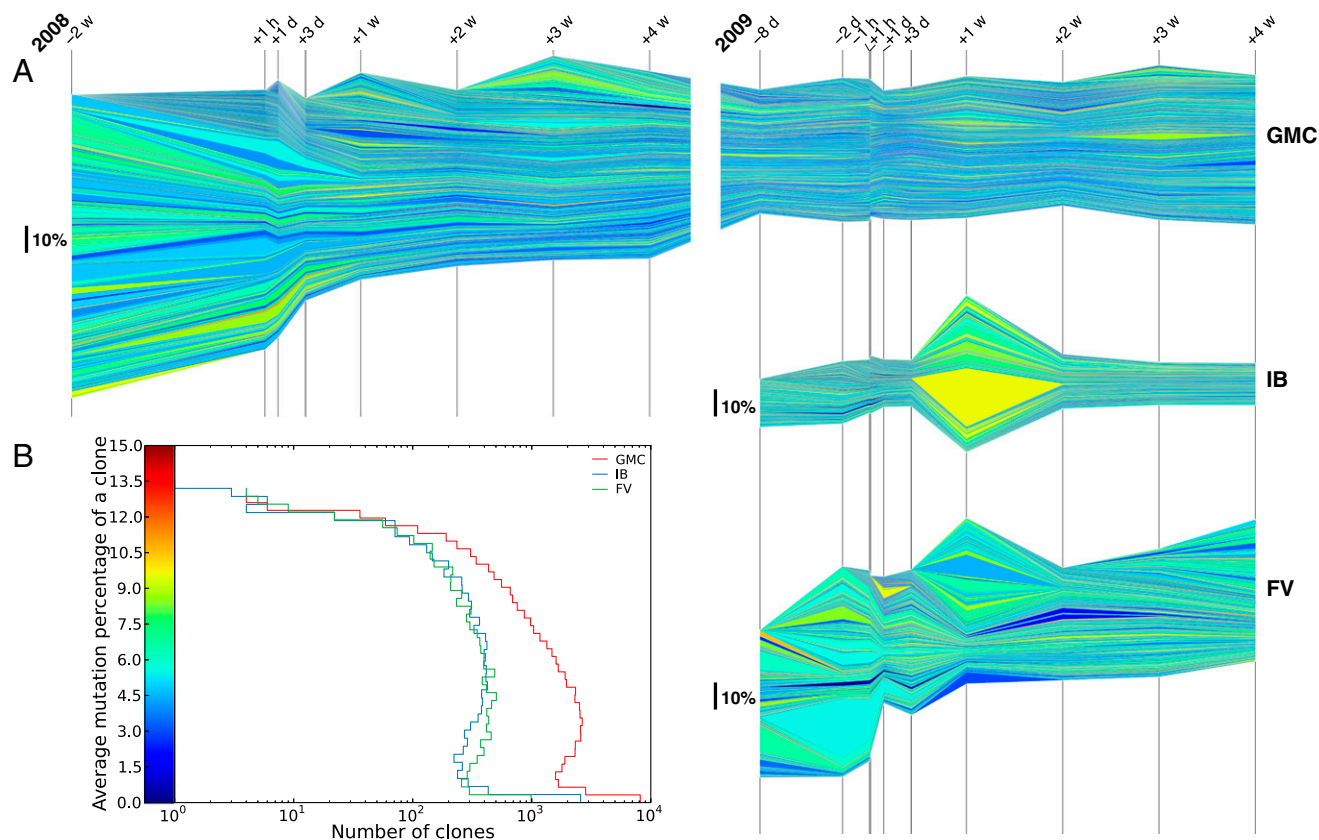


Fig. 3. Vaccination clone dynamics colored by mutation. (A) Each layer represents a clone. Time is shown by the grid lines on the x axis, and labeled relative to the two vaccination events. The thickness of each layer is proportional to the frequency of that clone at that time point. Each clone is colored based on the average mutation level of the corresponding reads (see color bar for B). Only clones seen in at least two time points are shown here. (B) Histogram of the average mutation level of all of the clones. Each clone is counted once (i.e., clones are not weighted by the number of corresponding reads). Stream graphs are stacked bar charts with moving baselines; if we did not filter out clones in only a single time point, the total thickness of the graph would add up to unity.

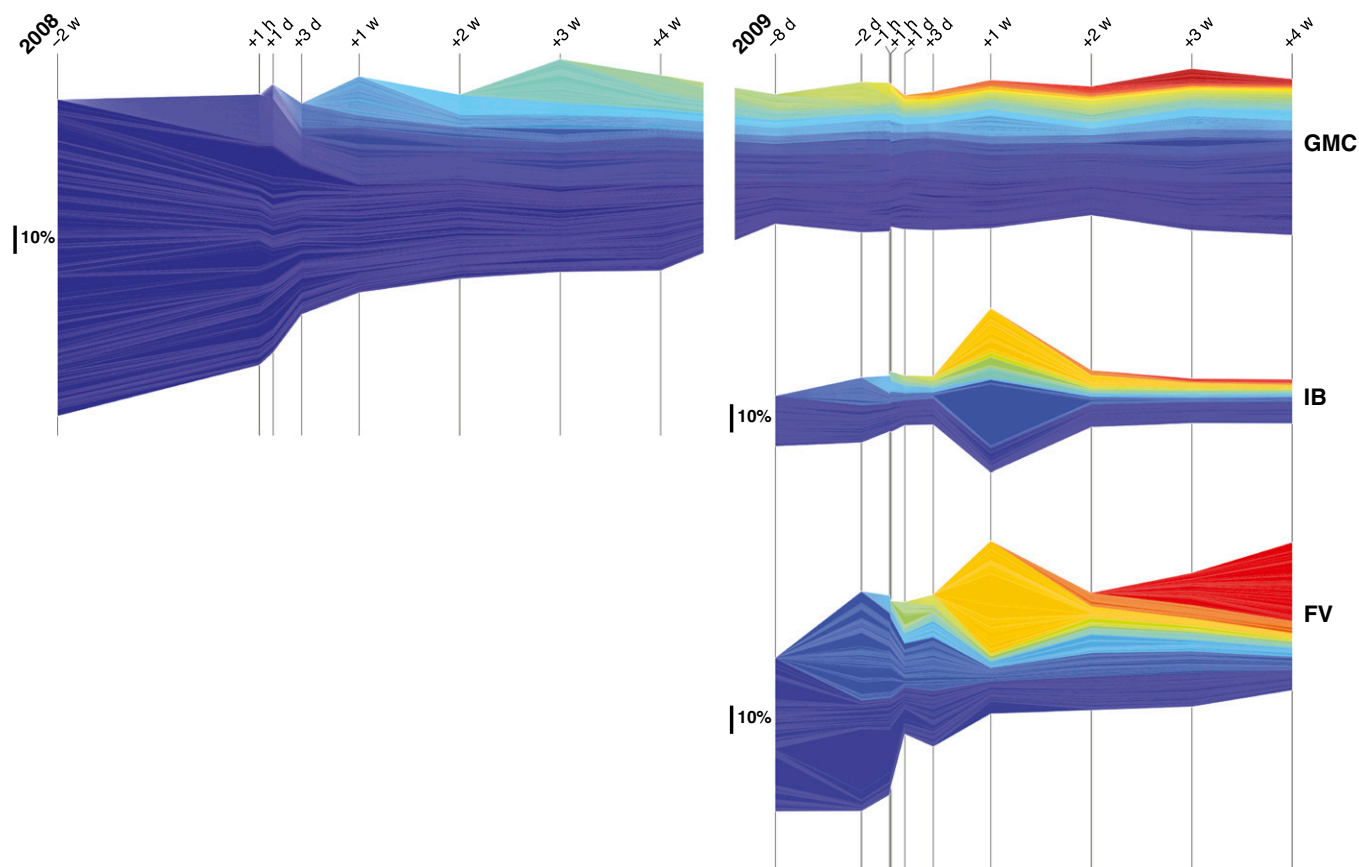


Fig. 4. Vaccination clone dynamics colored by onset time. Same as Fig. 3, except clones are colored by onset time. Onset times are ordered spectrally, so that all clones seen in the first time points are blue, followed by cyan, etc.

time series are remarkably stable as well (*SI Appendix, Fig. S5*). Finally, we built a neighbor-joining tree by using V-use vectors (as opposed to VJ-use) of each of the samples to see how V use is structured. For the most part, V use clustered first by individual, and then by isotype, implying that, even though V use is grossly similar across individuals, each individual still has a unique signature (Fig. 1).

For a majority of the reads, we were able to genetically discern the antibody isotype. We found that IgM antibody transcripts were the most abundant (43% of all reads), followed by IgA (27%), IgG (21%), IgD (9%), and IgE (0.01%; Table 1 and *SI Appendix, Fig. S1*). The level of IgA RNA is higher than the amount of IgA protein typically observed in serum. We believe this is observed because serum IgA are cleared more rapidly than IgG and also because the blood carries IgA cells in transit to mucosal sites. However, the isotype use varied significantly between time points (*SI Appendix, Fig. S6*).

Mutation levels were also measured across each of the reads. As expected, mutation rates were higher in the complementarity determining regions (CDRs) of the antibodies, and were much higher in IgG and IgA antibodies (*SI Appendix, Fig. S7*). We further processed our reads through the Bayesian estimation of antigen-driven selection (BASELINE) pipeline that estimates selection pressure on the antibodies (13). Framework regions (FWR) were universally negatively selected, whereas CDR regions showed neutral to slightly negative selection on average; however, CDR selection values were always more positive than FWR selection values (*SI Appendix, Fig. S8*).

The CDR3 length distribution we observed was consistent with IMGT/LIGM data (14) (*SI Appendix, Fig. S9*; CDR3 extraction is described in *Methods*). The 5th and 95th percentiles of the

observed CDR3 lengths are 36 nt and 75 nt, with median length 54 nt (with longest observed CDR3 at 140 nt).

B cells and their associated antibody transcripts can be present at vastly different quantities, depending on which lymphocyte subset they are derived from and the degree of expansion of particular clones. In particular, our method does not allow for functional differentiation between naïve, memory, or antibody-secreting cells (ASCs). Therefore, the variation in antibody expression may influence our view of the immune repertoire (e.g., ASCs, although rare in the periphery, exhibit very high expression of antibodies). Because the VDJ recombination process introduces so much diversity, the CDR3 sequence effectively functions as a natural barcode for a particular clone (15). To functionally define antibody clones, we perform clustering of the CDR3 sequences and define two reads as derived from the same clone if their CDR3 sequences are highly similar (same VJ combination and CDR3 within four nucleotide changes), as it is unlikely that two independent B cells will generate the same nucleotide sequence. In total, we observe >1.4 million clones across all of our data; however, only ~150,000 clones had more than three reads each and only 24 clones with >10,000 reads each. Summed over all three subjects, ~84,000 clones were seen in at least two separate time points, whereas only 257 heavy-chain clones were seen in every time point (*SI Appendix, Fig. S1*).

Each clone is labeled as “activated” or “naïve” based on its isotype and mutation level. For each VJ combination, we estimated the probability that a clone is activated (assuming a binomial distribution) and found that it is reproducibly biased by VJ use (Fig. 24). The V regions most likely to become activated are dominated by IGHV4- and IGHV5-family genes, and the three individuals have highly correlated biases in the VJ-activation probabilities (Spearman correlation of ~0.7). Nevertheless,

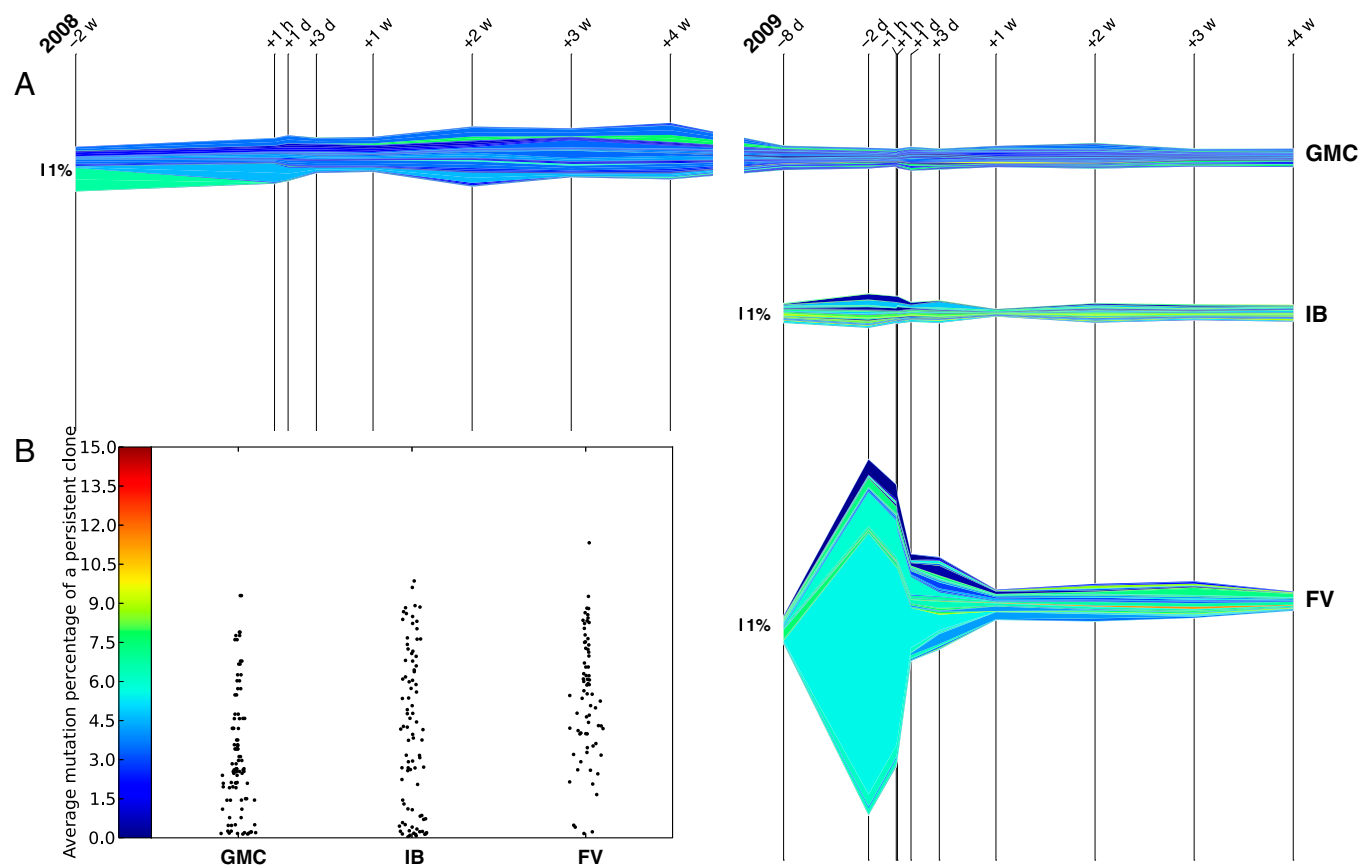


Fig. 5. Dynamics of persistent clones. (A) Stream graphs of only clones that are observed in every single time point for a given individual. They are colored as in Fig. 3. (B) Distribution of average mutation level of the persistent clones.

we find that there is virtually no correlation between whether a gene derived from a particular VJ germ-line combination is likely to become activated (i.e., subjected to somatic hypermutation and clonally expanded) and whether it is highly used (Fig. 2B). One hypothesis consistent with these results is that the immune system is being selected for naïve antibodies that confer protection. If, instead, the sole function of antibodies is in adaptive immunity, it seems wasteful to highly express VJ combinations that are less likely to become activated. However, we observed no correlation between the level of VJ combination use and activation. We interpret this as consistent with the hypothesis that the antibody repertoire is shaped by selective forces at population and somatic time scales. More precisely, use of the VDJ germ-line library may be optimized for naïve interactions with common pathogens at the population scale, and the propensity of any given germ-line gene to somatically mutate may be optimized for the evolvability of the target organisms. In the context of our hypothesis, the population of antibodies can be seen to occupy an innate-adaptive spectrum (9, 16).

Antibody Repertoire Dynamics. Each of the three volunteers was given the clinically-indicated seasonal flu vaccine, and G.M.C. was also given boosters to hepatitis A/B in 2008. None of the subjects were naïve to the antigens at the time of vaccination (through prior vaccination or infection). Each read was assigned to a clone and a time point, allowing us to compute time series. The clone frequencies were tracked across all 38 time points to produce >20 million clone-frequency measurements (although the matrix is sparse). In contrast to the relative stability of the VJ use, antibody clones were highly dynamic and variable across individuals (Figs. 3 and 4).

We used expression-level/burstiness as a proxy for immune response strength. Responses to each of the four vaccination events were qualitatively different: IB produced a “textbook” response with large proliferating clones 7 d after vaccination; FV displayed high-frequency clones just before vaccination as well as large clonal expansions 7 d postvaccination; the repertoire analysis suggested that G.M.C. showed weak response from his first vaccination, and his second vaccination appeared to have produced no high-frequency responses. Taken together, these data show that individuals can demonstrate a highly varying array of responses against an identical immune challenge, likely influenced by prior exposure, age, and other concurrent immune responses during the course of this experiment, among other factors.

We verified that samples that are closer in time share more unique clones. We computed the number of shared CDR3 sequences between all 703 possible pairs of samples across all 38 time points, and observed that closer time points within an individual indeed share a larger number of unique CDR3s (*SI Appendix*, Fig. S10). Consistent with this, interindividual comparisons between time points showed very little CDR3 overlap.

We also quantified the range of dynamic behavior of the clones, finding that clones generally fluctuate wildly (*SI Appendix*, Fig. S11). Interestingly, each individual had a number of clones that were present at every time point sampled, including the samples separated by more than 1 y (257 clones total with median clone frequency 1.1×10^{-4} ; Fig. 5A). It is possible that these clones correspond to a long-lived, expanded B-cell memory population or are chronically responding to antigens (foreign or auto-) that are always present; indeed, these clones include sequences that are highly mutated (Fig. 5B).

Supporting Information

Laserson et al.

SI Materials and Methods

Sample Collection. Blood samples were collected under the approval of the Personal Genome Project (1). Sample collection was coordinated with clinically indicated vaccinations for each individual. Total RNA was immediately extracted from each blood sample and stored at -80 until use.

Primer design. All oligonucleotides were ordered from Integrated DNA Technologies (IDT, Coralville, IA). For the design of the upstream variable-region oligonucleotides (IGHV-PCR), we extracted the L-PART1 and L-PART2 sequences from all IMGT/GENE-DB (2) reference segments annotated as “functional” or “ORF”. These two segments are spliced together *in vivo* to form the leader sequence. We positioned our primer sequence to cross the exon-exon boundary to ensure amplification from cDNA rather than gDNA. For the design of the downstream constant-region oligonucleotides (IGHC-RT and IGH-PCR), the first 100 nucleotides of the CH1 exon were extracted from the IMGT/GENE-DB. Oligonucleotides were then selected as close as possible to the 5' end of the C-region to take advantage of sequence conservation between different variants, and to ensure that isotypes would be distinguishable.

Sequencing library preparation. We reverse-transcribed the immunoglobulin heavy chain mRNA using a pool of 6 primers specific to the Ig constant regions and amplified the cDNA using 16 cycles of PCR with a pool of 46 V region-specific primers and 6 nested constant region primers. Following ligation of 454-compatible sequencing adapters, we purified the expected VH fragment using PAGE. Each sample derived from a given time-point was uniquely bar-coded during the ligation process, allowing subsequent mixing of all the time points into one common reaction sample (performed independently for each replicate run). Emulsion PCR and 454 GS FLX sequencing were performed directly at the 454 Life Sciences facility according to the manufacturer's standard protocols.

Data processing overview. Following data generation, the resulting reads were processed through an in-house software pipeline. The sequencing reads were filtered for proper fragment size and presence of a sample identity barcode. The reads were aligned to the reference IMGT database to identify the V, D, and J regions. We then partitioned the reads by VJ usage and

hierarchically clustered them using their CDR3 junction to define unique clones. This data was finally used for subsequent time series and statistical analyses, including selection estimation with BASELINE (13) and phylogeny inference with Immunitree (19).

VDJ alignment process. For each segment we performed a semiglobal dynamic programming alignment against each reference sequence, choosing the best match. To maximize the number of distinguishing nucleotides, we performed our alignment in order of decreasing segment length (V then J then D), and subsequently pruned off successfully aligned V or J regions before attempting alignment of the next segment. Since we know that the V and J segments must reside at the ends of the reads, we used a method that is similar to the Needleman-Wunsch algorithm (3). In contrast to the canonical algorithm, we used zero initial conditions to allow the start of the alignment to occur anywhere without penalty. The alignment is then reconstructed and scored by starting at the maximum value of the score matrix along the last row or last column, and backtracing. Finally, the identified V or J segments are removed before proceeding to the J or D alignment, respectively. For the D region alignment, we used the canonical Smith-Waterman local alignment algorithm (3), as we have no prior information as to where the D segment should reside. Finally, we compared the performance of our aligner against IMGT/V-QUEST (4) and generate ROC curves (Fig. S13).

Sequence clustering. We performed sequence clustering in order to group our sequences (reads) into unique clones. This process is primarily used to associate sequences that originated from the same cell/clone, while allowing minor variations attributable to sequencing errors. For most of our work, we chose to use single-linkage agglomerative hierarchical clustering with Levenshtein edit distance as the metric. To make the clustering process more tractable, we partitioned our reads based on VJ identity. Within each partition, we then performed sequence clustering using only the CDR3 junction nucleotide sequence. To account for sequencing errors, we examined the distribution of cophenetic distances observed in the linkage tree (Fig. S14), and determined the optimal distance to clip the tree at 4-5 edits (Fig. S14). This distance corresponded to an observable change in the distribution that we believe corresponds to the distinction between sequencing errors and real somatic mutation.

Mutation analysis pipeline. After removing the primers from both ends of each raw read, High V-Quest (5) was used to assign V and J genes and to align the sequences through the IMGT

unique numbering scheme. In this step most of the insertions/deletions were identified and corrected by either removing any insertion or adding “N” to replace any deletion.

Following this step, sequences that potentially had artificial mutations due to incorrect germline assignments were excluded. This was done by: 1) excluding nonfunctional sequences (due to the occurrence of a stop codon and/or due to a shift in the reading frame between the V and the J gene), 2) excluding sequences with more than 14% mutations, 3) excluding sequences with more than 7 mutations in any 12 nucleotide window. This final step was taken in order to account for the possibility of an insertion following a deletion event, which can be incorrectly viewed as several dense point mutations.

Clonality was determined using a two-step approach. First, the sequences were divided into groups based on equivalence of their V-gene assignment, J-gene assignment, and the number of nucleotides in their junction. Following this step, clones were then defined within each of these groups as a collection of sequences with junction regions that differ from one sequence to any of the others by no more than three point mutations. The threshold of three was determined after manual inspection of the mutation patterns in resulting clones identified through building phylogenetic trees.

Analysis of selection pressures. Selection pressure analysis was carried out using BASELINE (Bayesian estimation of Antigen-driven SElectioN (13) based on the local test formalism (see (6)). The output of BASELINE is a full posterior probability distribution function for each sequence and for a collection of sequences. Here, we used the mean selection estimation for each sequence for the tree analysis. For Fig. S8, we have calculated a combined selection score (and 95% confidence intervals) for each combination of individual, time point, and isotype

Clone phylogeny inference. To determine the most likely phylogeny of a clone of reads, we use the Immunitree algorithm. Immunitree uses a probabilistic generative model that assigns a probability to each possible phylogeny. We apply MCMC to sample from this probability distribution of possible phylogenies, subject to the constraint that the phylogeny must generate the observed empirical data. MCMC generates an entire chain of samples of possible phylogenetic trees. Per MCMC iteration, we perform block gibbs on each of the parameters: phylogenetic tree structure, birth and death times of individual subclones, birth and death rates, mutation rates, read error rates, subclone consensus sequences, and assignment of reads to

subclones. Finally, we perform a brief optimization on each of the sampled trees, and select the best such optimized sample as the final output.

V-usage clustering. After assigning the sequences to clones, each clone is associated with one V-gene. A V-gene usage vector for each individual-isotype combination was created. Using a Euclidean distance metric for these vectors, a neighbor joining tree was created in Fig. 1.

Clone synthesis/affinity. We tested whether we could find antigen-specific clones by choosing the most highly expressed clones at the +7 day time points. We picked a subset of the largest clones from multiple time points before and after vaccination (-2 day, +7 days, +21 days) and synthesized them chemically. Because high-throughput technology to pair heavy and light chains from single cells were not yet available, we cloned the full light chain repertoires from the corresponding time points. The constructs were then paired in an scFv format and panned using phage display against the influenza antigens present in the vaccines. After three rounds of selection against hemagglutinin, we found only a single clone (GMC J-065) at day +7 from GMC-2009 that displayed significant affinity for H3N2 HA (5 nM).

Software tools. Processing of raw data was performed by python packages and is available here:

<https://github.com/laserson/vdj>

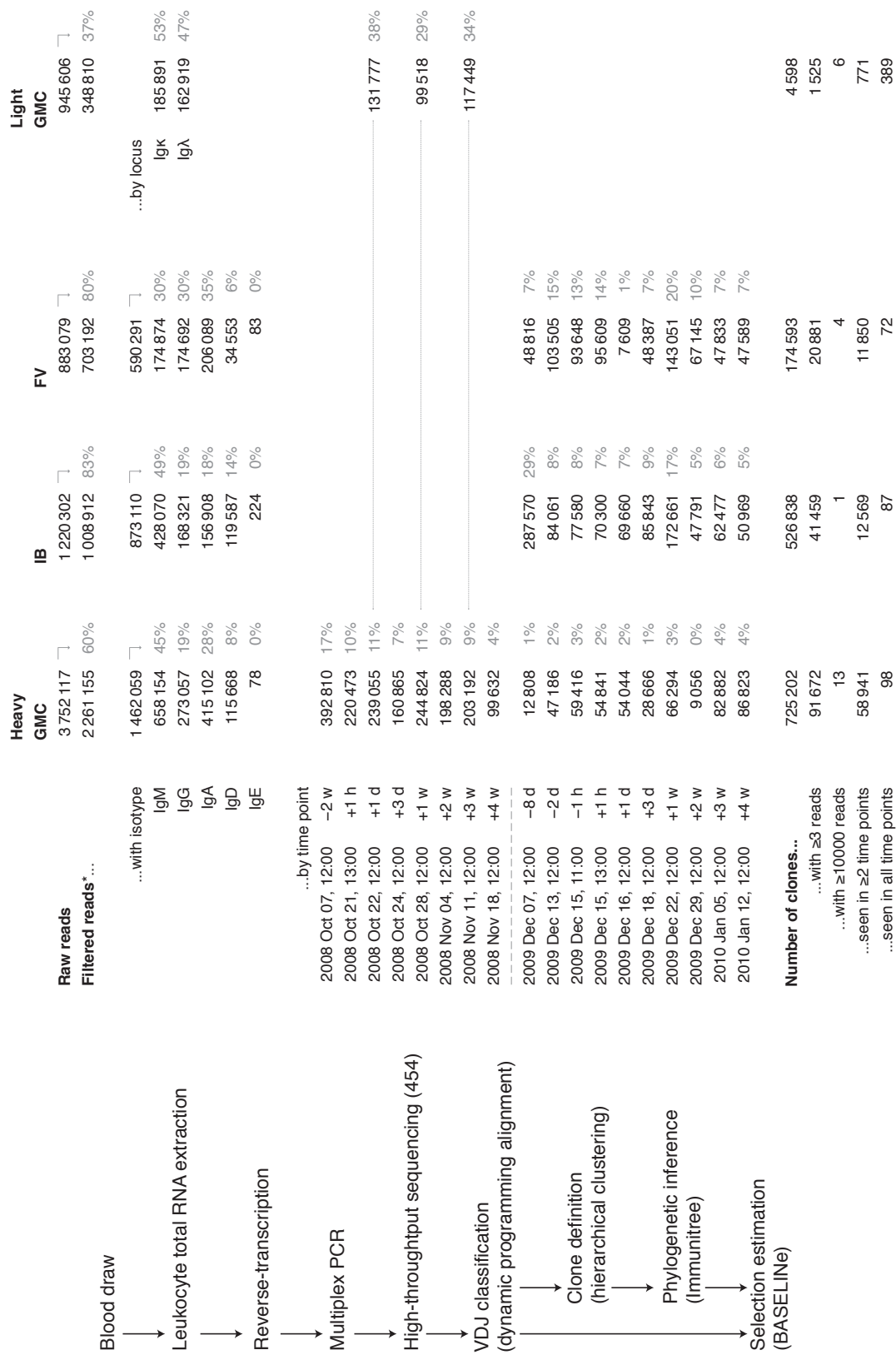
<https://github.com/laserson/pytools>

Figures were produced with matplotlib, R, and graphviz. Scripts for Figure preparation are available upon request.

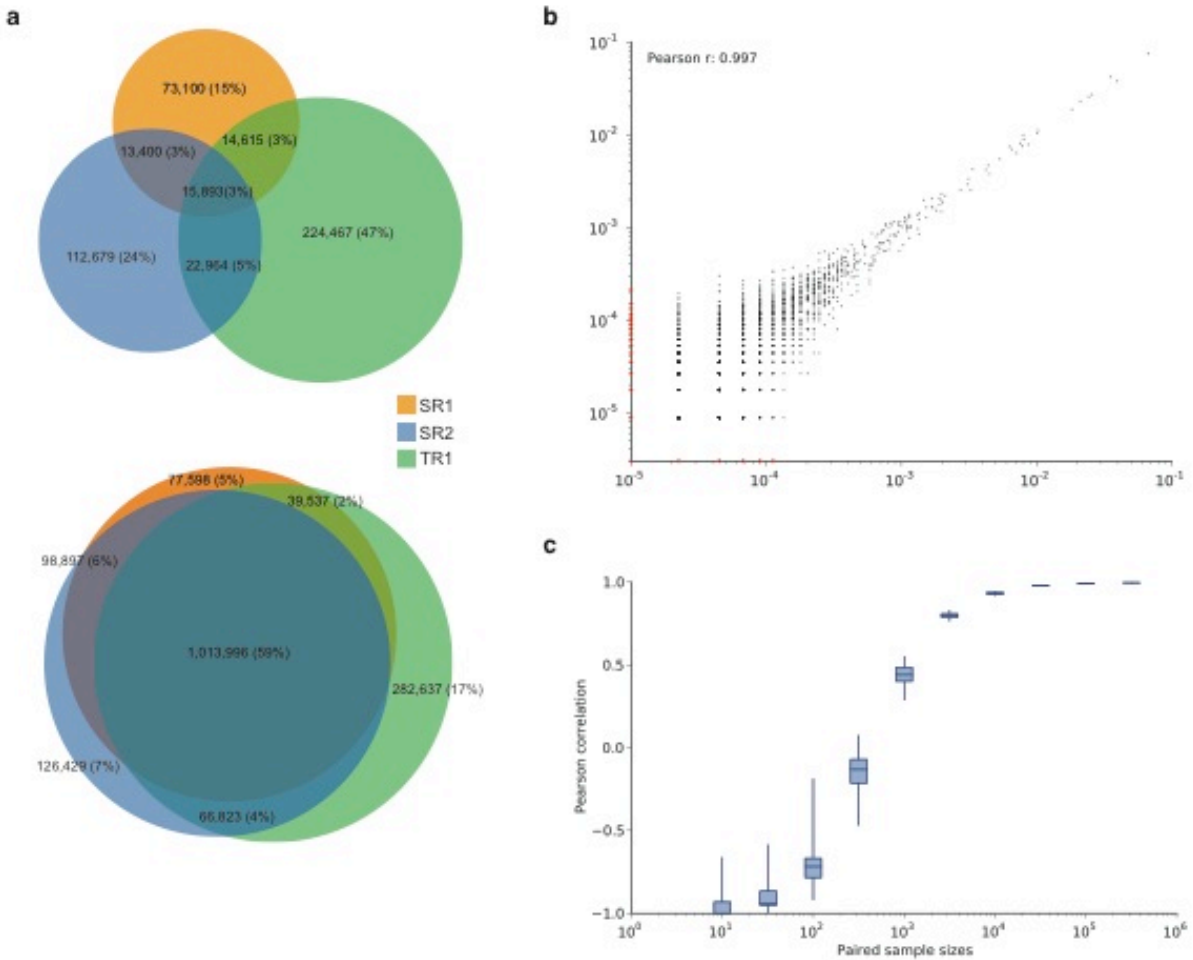
REFERENCES:

1. Church GM (2005) The Personal Genome Project. *Mol Syst Biol* 1:–.
2. Giudicelli VV, Chaume DD, Lefranc M-PM (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res* 5:D256–D261.
3. Durbin R, Eddy SR, Krogh A, Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press).
4. Brochet X, Lefranc M-P, Giudicelli V (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* 36:W503–8.
5. Lefranc M-P et al. (2009) IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res* 37:D1006–12.
6. Uduman M et al. (2011) Detecting selection in immunoglobulin sequences. *Nucleic Acids Res* 39:W499–504.

Supplementary Figure 1

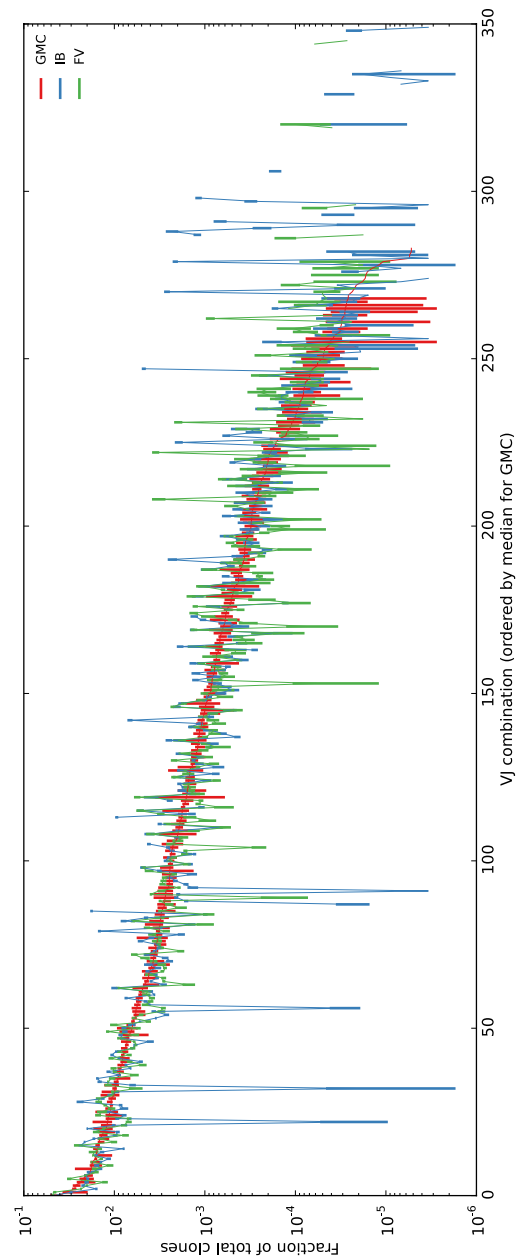


*Size-selected, VJ-filtered



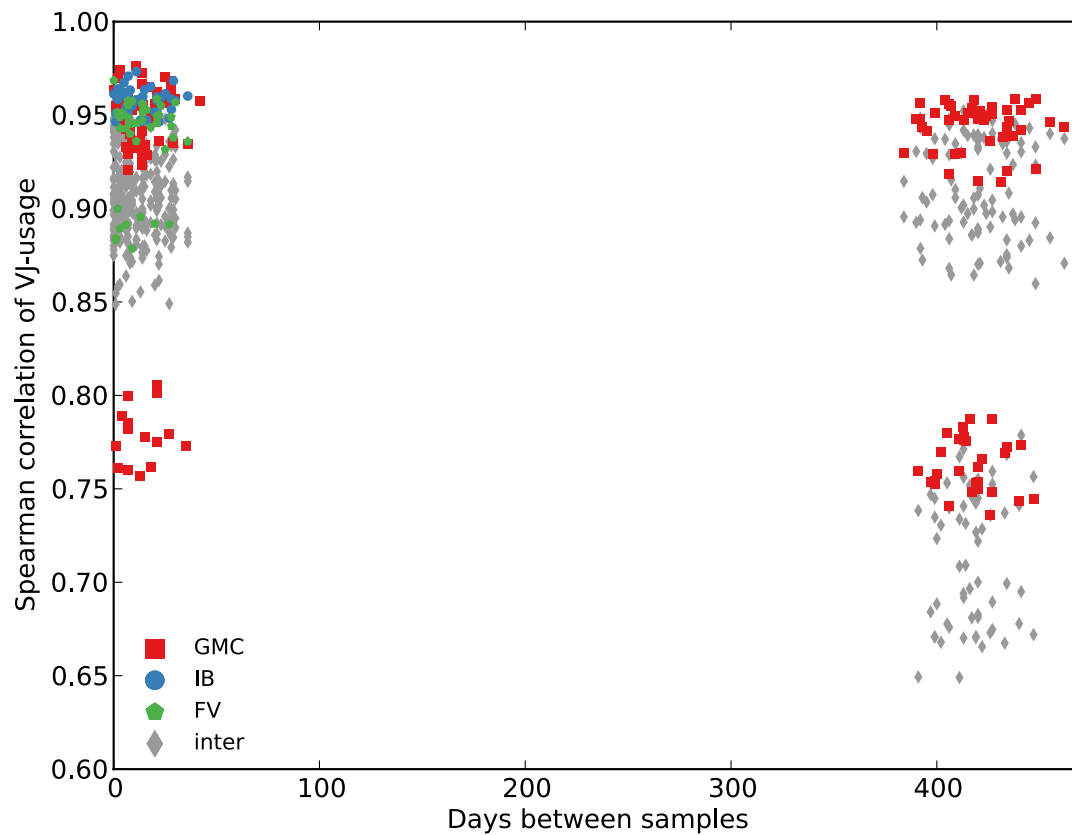
Supplementary Figure 2

Reproducibility in vaccination experiment. (a) Venn diagrams showing overlapping clones (top) and the same overlaps weighted by number of reads (bottom) for replicates. SR1 and SR2 are sequencing replicates of the same library and TR1 is a technical replicate of an independent sequencing library from the same RNA sample. (b) Correlation between technical replicates. Axis scales are clone frequencies; red points are zero-valued on that axis. (c) Correlation between paired random samples of reads of the indicated size.



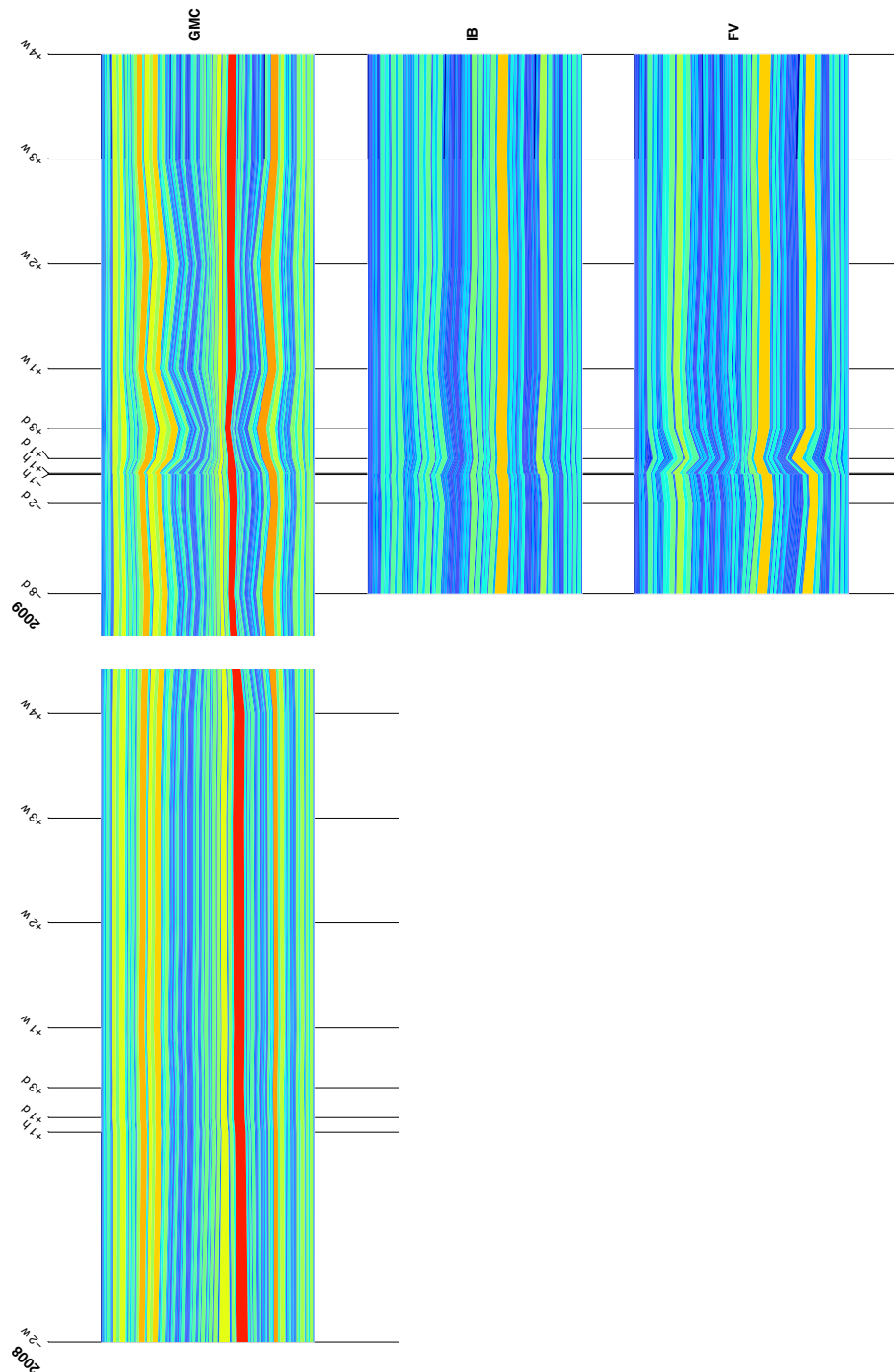
Supplementary Figure 3

VJ usage. For each sample from each individual, the number of clones with a particular VJ combination was computed. Reads for a given clone are collapsed so that each count represent a single recombination event. For each possible VJ combination, a distribution of frequencies was computed for each individual across all samples. Each bar represents the 25th--75th percentile value across the different sample, while the line tracks the median values. The VJ combinations are ordered by median for GMC.



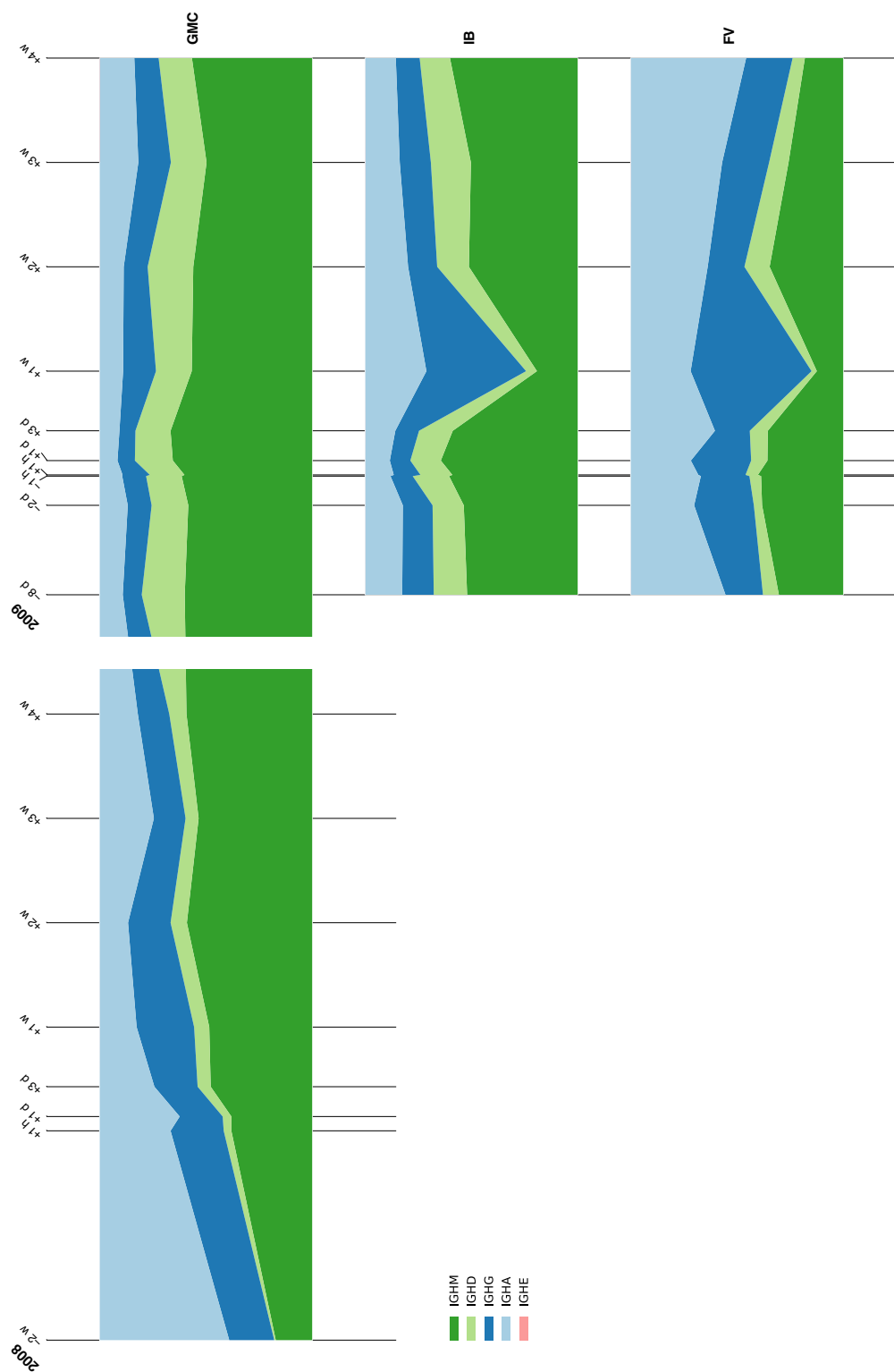
Supplementary Figure 4

For each sample, a VJ-usage vector is formed. The Spearman correlation is computed between every pair of vectors; intra-individual comparisons are shown with the indicated color, while inter-individual comparisons are shown in gray. Note that intra- and inter-individual comparisons both show comparable correlations. The multimodality of the correlation coefficients has two causes: low sequencing coverage in some cases, and individual differences in comparisons between individuals, where GMC and FV appear more highly correlated to each other than IB.



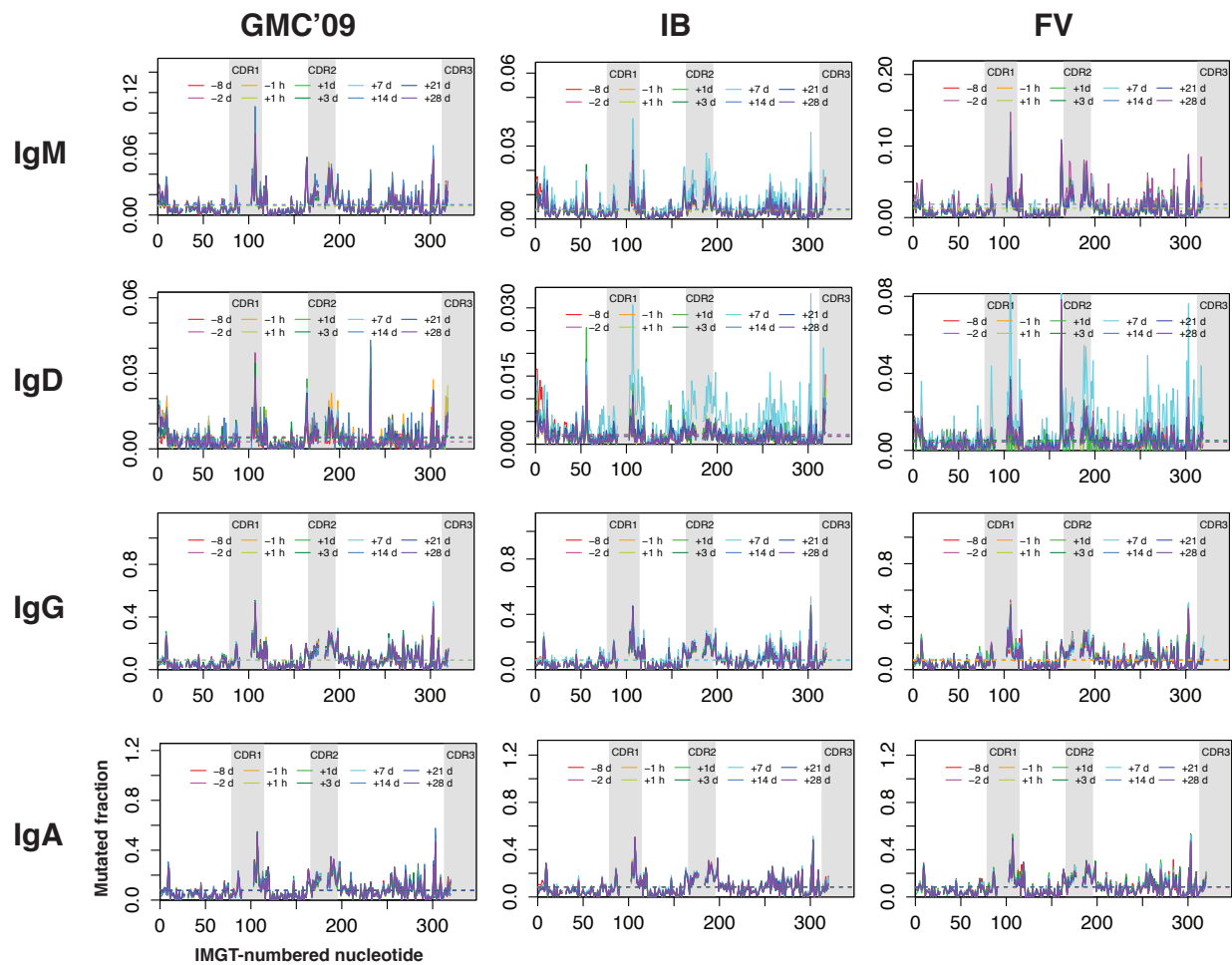
Supplementary Figure 5

VJ-usage dynamics. Streamgraph of VJ-usage for each individual. Time is listed on the x-axis, with time points relative to vaccinations marked at the grid lines. Each stream/layer corresponds to a particular VJ combination, and its thickness at a given time point is proportional to its frequency at that time point. All streams add up to 100% usage at each time point.



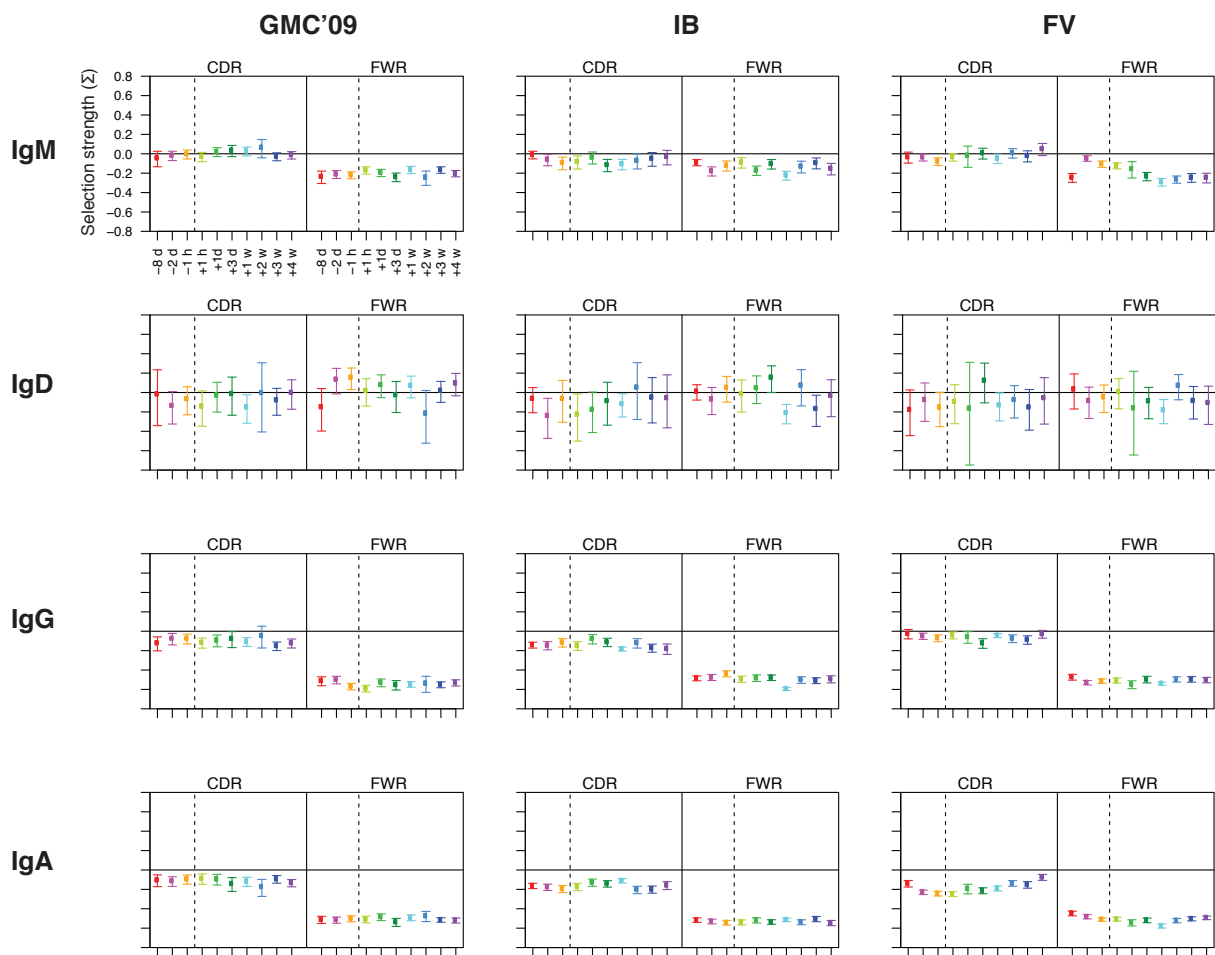
Supplementary Figure 6

Isotype dynamics. Streamgraph of isotype usage at each timepoint for each individual.



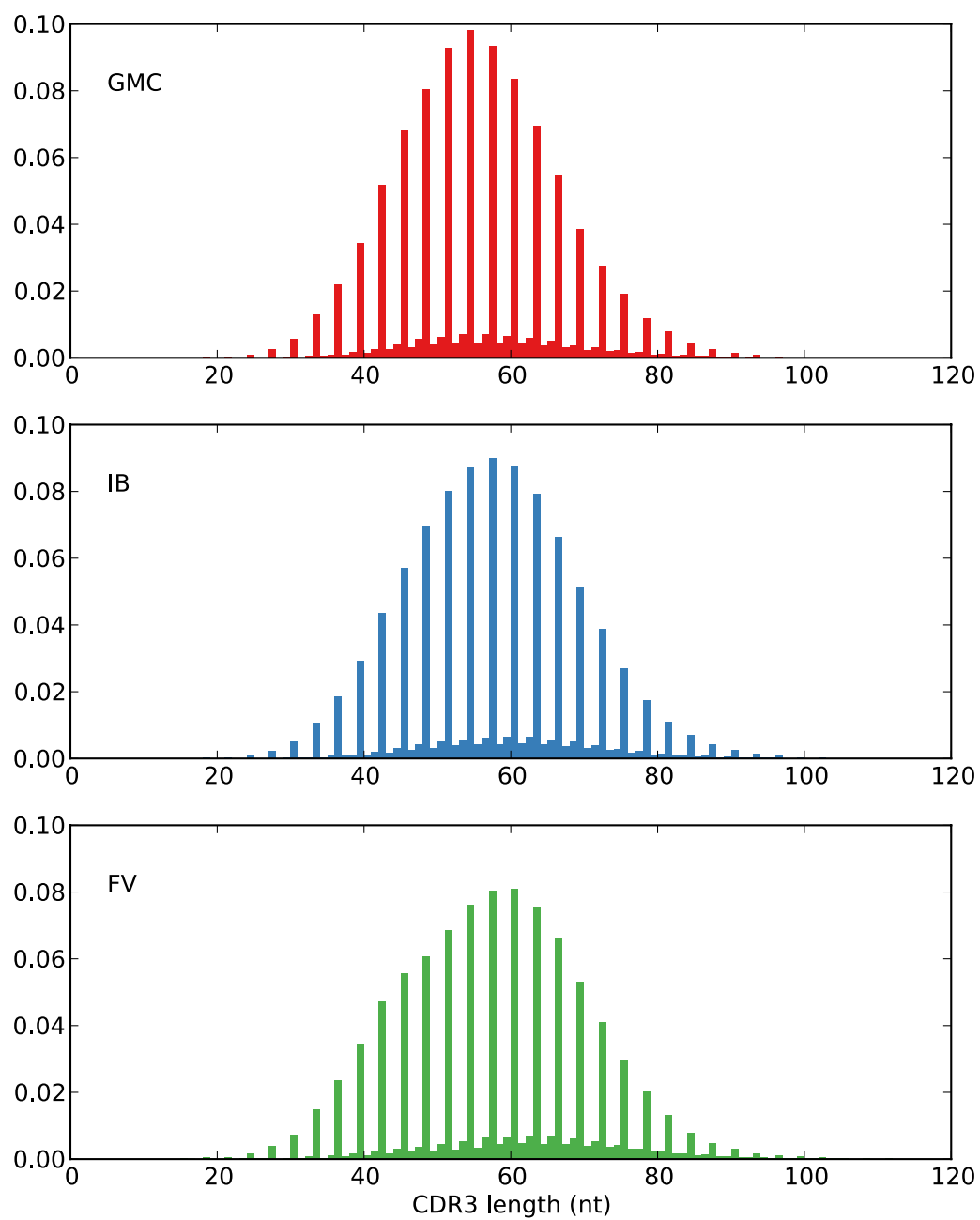
Supplementary Figure 7

Antibody mutation patterns. The mutation density for the indicated subset of reads is computed along the length of the gene. Note that different scales are used for different subplots.



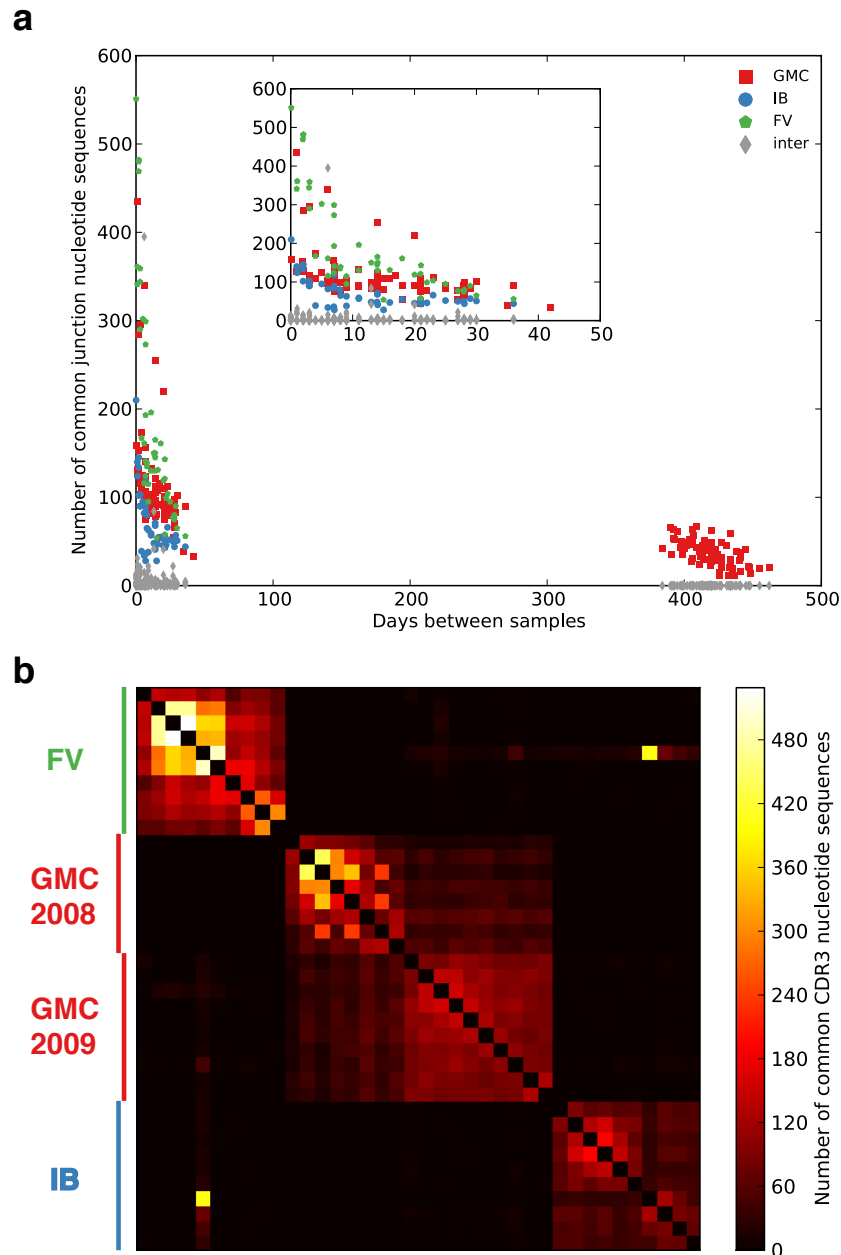
Supplementary Figure 8

Antibody selection estimation. For each set of antibody sequences, the selection pressure has been estimated with BASELINE (11).



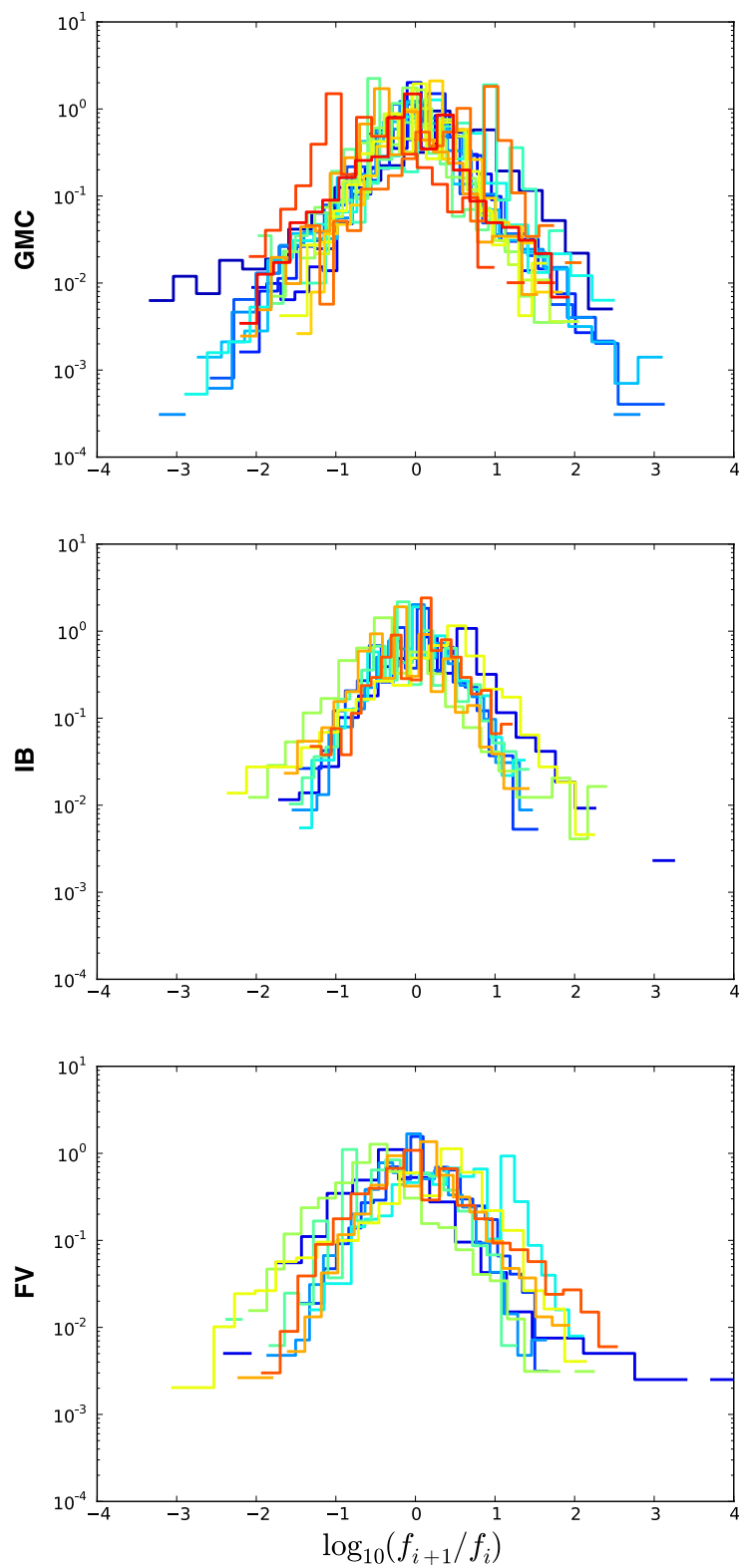
Supplementary Figure 9

CDR3 length distributions. The CDR3 is defined according to the IMGT numbering scheme as the segment spanning the second conserved cysteine through the conserved tryptophan.



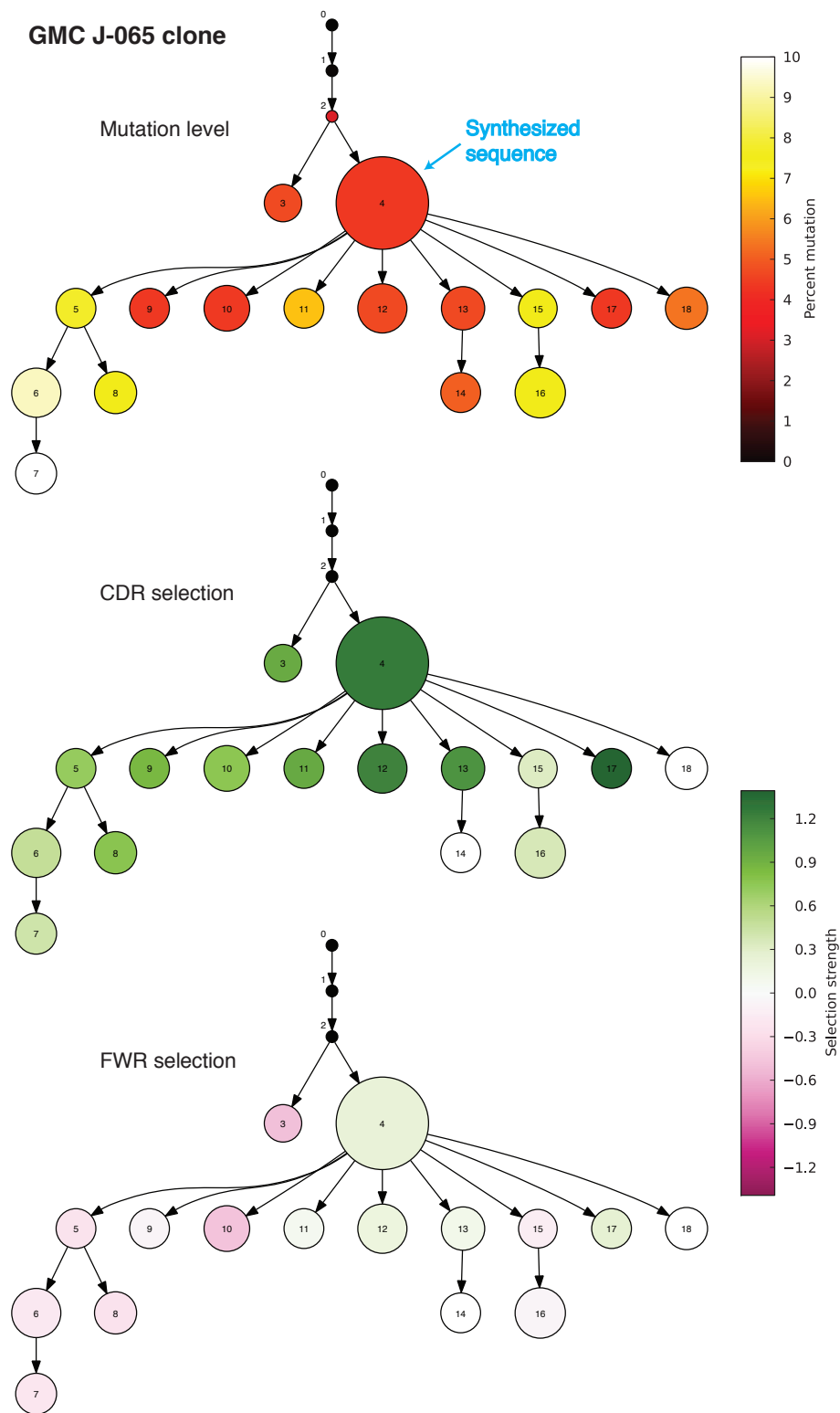
Supplementary Figure 10

Inter-sample CDR3 overlaps. (a) Subsampled CDR3s from each time point/individual are compared for common sequences. Comparisons of time points that are closer in time show higher levels of overlap. Inter-individual comparisons show very little overlap, as expected. (b) Overlap between each sample is plotted showing the three individual blocks. The strong overlap between an IB and an FV sample is likely due to some sample cross-contamination.



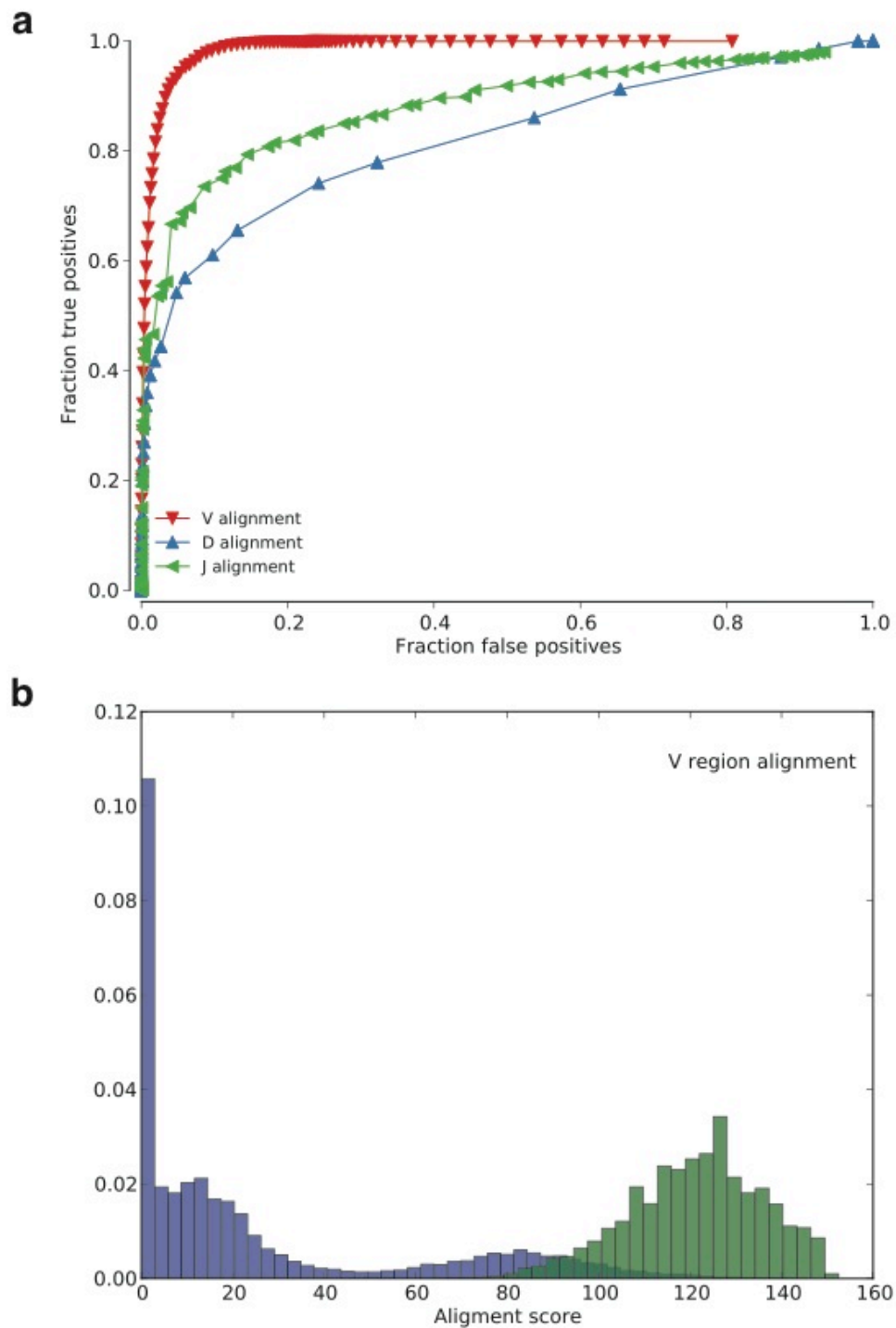
Supplementary Figure 11

Distribution of frequency changes. For each adjacent time point, the $\log_{10}$ ratio of frequencies (f_i) for each clone is histogrammed, when finite. Time points are plotted with different colors, arranged chronologically and spectrally (blue to red).



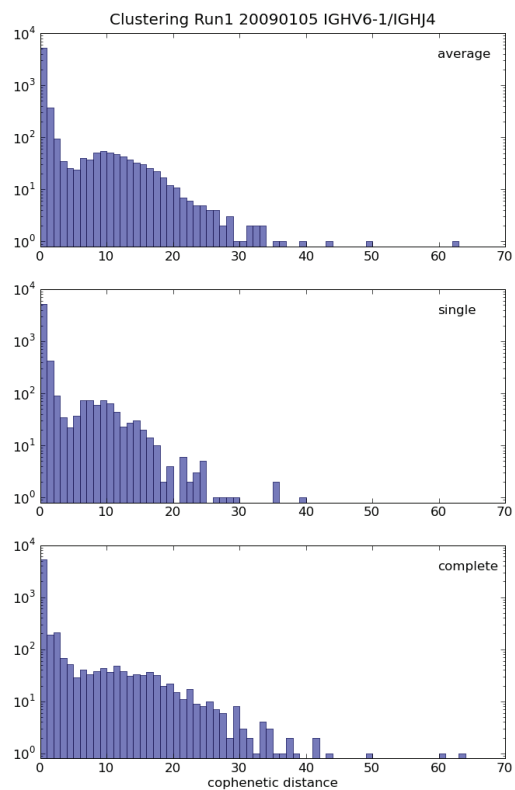
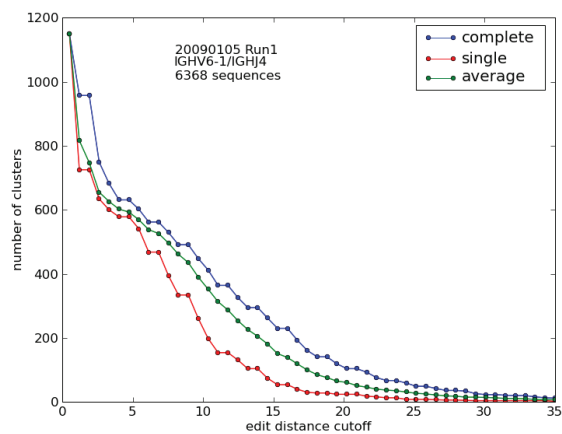
Supplementary Figure 12

GMC J-065 clonal phylogeny. A phylogenetic tree for GMC clone J-065 was constructed with Immunitree and overlaid with sequence mutation data and CDR/FWR selection estimates.



Supplementary Figure 13

VDJ aligner calibration. (a) ROC curves comparing our VDJ aligner to IMGT/V-QUEST as gold-standard. (b) V alignment scores for "correct" alignment versus incorrect alignments.

a**b**

Supplementary Figure 14

Clustering calibration. (a) Distribution of cophenetic distances for single, complete, and average linkage clustering. (b) Number of clusters as a function of clipping threshold