High-Throughput Sequencing of the Zebrafish Antibody Repertoire

Joshua A. Weinstein, *et al.*

Science **324**, 807 (2009);
DOI: 10.1126/science.1170020

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

7. L. Wang, W. C. Jackson, P. A. Steinbach, R. Y. Tsien, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16745 (2004).
8. A. J. Fischer, J. C. Lagarias, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 17334 (2004).
9. S. J. Davis, A. V. Vener, R. D. Vierstra, *Science* **286**, 2517 (1999).
10. J. W. Harris, R. W. Kellermeyer, *The Red Cell* (Harvard Univ. Press, Cambridge, MA, 1970).
11. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
12. J. R. Wagner *et al.*, *J. Biol. Chem.* **283**, 12212 (2008).
13. E. Giraud *et al.*, *J. Biol. Chem.* **280**, 32389 (2005).
14. X. Yang, J. Kuk, K. Moffat, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 14715 (2008).
15. N. C. Rockwell, Y. S. Su, J. C. Lagarias, *Annu. Rev. Plant Biol.* **57**, 837 (2006).
16. J. R. Wagner, J. S. Brunzelle, K. T. Forest, R. D. Vierstra, *Nature* **438**, 325 (2005).
17. Materials and methods are available as supporting material on *Science* Online.
18. N. C. Shaner, P. A. Steinbach, R. Y. Tsien, *Nat. Methods* **2**, 905 (2005).
19. A. Bellacosa, J. R. Testa, S. P. Staal, P. N. Tsichlis, *Science* **254**, 274 (1991).
20. S. N. Waddington *et al.*, *Cell* **132**, 397 (2008).
21. Six mice after BV injection were observed for 3 days (the maximum time that we could hold mice for imaging according to our university-approved animal protocol).
22. R. Ollinger *et al.*, *Antioxid. Redox Signal.* **9**, 2175 (2007).
23. A. Nakao *et al.*, *Gastroenterology* **127**, 595 (2004).
24. C. H. Contag, M. H. Bachmann, *Annu. Rev. Biomed. Eng.* **4**, 235 (2002).
25. V. Ntziachristos *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 12294 (2004).
26. Electrospray mass spectrometry of bacterially expressed IFP1.4 holoprotein in 5% acetonitrile, 0.05% trifluoroacetic acid reported an average relative molecular mass (M_r) of 36,342 daltons, within experimental error of the M_r (36,332 daltons) expected from covalent incorporation of BV (582.6 daltons) into the apoprotein (35,749.6 daltons) from Ala^2 to the C terminus.
27. N. G. Abraham, A. Kappas, *Pharmacol. Rev.* **60**, 79 (2008).
28. D. B. Rusch *et al.*, *PLoS Biol.* **5**, e77 (2007).

29. We thank J. C. Lagarias and S. Field for donation of cDNAs encoding HO-1 and AKT1's PH domain, respectively; J. M. Saathoff and G. Tran for help with plasmid purification; M. Timmers for help with tissue culture; S. Adams for help with light scattering measurements; L. Gross for mass spectrometry; and Q. Xiong for flow cytometry. This work was supported by NIGMS grant R01 GM086197 and the Howard Hughes Medical Institute.

Supporting Online Material

www.sciencemag.org/cgi/content/full/324/5928/804/DC1
Materials and Methods
SOM Text
Figs. S1 to S11
Table S1
References

18 November 2008; accepted 10 March 2009
10.1126/science.1168683

High-Throughput Sequencing of the Zebrafish Antibody Repertoire

Joshua A. Weinstein,[1]* Ning Jiang,[2]* Richard A. White III,[3] Daniel S. Fisher,[1,4,5] Stephen R. Quake[1,2,3,4]†

Despite tremendous progress in understanding the nature of the immune system, the full diversity of an organism's antibody repertoire is unknown. We used high-throughput sequencing of the variable domain of the antibody heavy chain from 14 zebrafish to analyze VDJ usage and antibody sequence. Zebrafish were found to use between 50 and 86% of all possible VDJ combinations and shared a similar frequency distribution, with some correlation of VDJ patterns between individuals. Zebrafish antibodies retained a few thousand unique heavy chains that also exhibited a shared frequency distribution. We found evidence of convergence, in which different individuals made the same antibody. This approach provides insight into the breadth of the expressed antibody repertoire and immunological diversity at the level of an individual organism.

The nature of the immune system's antibody repertoire has been a subject of fascination for more than a century. This repertoire is highly plastic and can be directed to create antibodies with broad chemical diversity and high selectivity (*1*, *2*). There is also a good understanding of the potential diversity available and the mechanistic aspects of how this diversity is generated. Antibodies are composed of two types of chains (heavy and light), each containing a highly diversified antigen-binding domain (variable). The V, D, and J gene segments of the antibody heavy-chain variable genes go through a series of recombination events to generate a new heavy-chain gene (Fig. 1). Antibodies are formed by a mixture of recombination among gene segments, sequence diversification at the junctions of these segments, and point mutations throughout the gene (*3*). Estimates of immune diversity for antibodies or the related T cell receptors either have attempted to extrapolate from small samples to entire systems or have been limited by coarse resolution of immune receptor genes (*4*). However, certain very elementary questions have remained open more than a half-century after being posed (*1*, *5*, *6*): It is still unclear what fraction of the potential repertoire is expressed in an individual at any point in time and how similar repertoires are between individuals who have lived in similar environments. Moreover, because each individual's immune system is an independent experiment in evolution by natural selection, these questions about repertoire similarity also inform our understanding of evolutionary diversity and convergence.

Zebrafish are an ideal model system for studying the adaptive immune system because in evolutionary terms they have the earliest recognizable adaptive immune system whose features match the essential human elements (*7*, *8*). Like humans, zebrafish have a recombination activating gene (RAG) and a combinatorial rearrangement of V, D, and J gene segments to create antibodies. They also have junctional diversity during recombination and somatic hypermutation of antibodies to improve specificity, and the organization of their immunoglobulin (Ig) gene loci approximates that of human (*9*). In addition, the zebrafish immune system has only ~300,000 antibody-producing B cells, making it three orders of magnitude simpler than mouse and five orders simpler than human in this regard.

We developed an approach to characterize the antibody repertoire of zebrafish by analyzing complimentarity-determining region 3 (CDR3) of the heavy chain, which contains the vast majority of immunoglobulin diversity (*10*, *11*) and can be captured in a single sequencing read (Fig. 1). Using the 454 GS FLX high-throughput pyrosequencing technology allowed sequencing of 640 million bases of zebrafish antibody cDNA from 14 zebrafish in four families (Fig. 1B). Zebrafish were raised in separate aquaria for each family and were allowed to have normal interactions with the environment, including the development of natural internal flora. We chose to investigate the quiescent state of the immune system, a state where the zebrafish had sampled a complex but fairly innocuous environment and had established an equilibrium of normal immune function. mRNA was prepared from whole fish, and we synthesized cDNA using primers designed to capture the entire variable region.

Between 28,000 and 112,000 useful sequencing reads were obtained per fish, and we focused our analysis on CDR3 sequences. Each read was assigned V and J by alignment to a reference with a 99.6% success rate (table S3); failures were due to similarity in some of the V gene segments. D was determined for each read by applying a clustering algorithm to all of the reads within a given

[1]Biophysics Program, Stanford University, Stanford, CA 94305, USA. [2]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA. [3]Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305, USA. [4]Department of Applied Physics, Stanford University, Stanford, CA 94305, USA. [5]Department of Biology, Stanford University, Stanford, CA 94305, USA.

*These authors contributed equally to this work.
†To whom correspondence may be addressed. E-mail: quake@stanford.edu

VJ and then aligning the consensus sequence from each cluster to a reference. D was assigned to 69.6% of reads; many of the unassignable cases had D regions mostly deleted. Both the isotypes that are known to exist in zebrafish (IgM and IgZ) were found, and their relative abundance agrees with previous studies (*12*). Our analysis focused on IgM, which is the most abundant species; IgZ data are presented in figs. S3 and S4 (*13*).

There are 975 possible VDJ combinations in zebrafish (39 V × 5 D × 5 J = 975 VDJ). In any given fish, the VDJ combination coverage was at least 50% and in some cases at least 86% (Fig. 2). By using subsets of the full data set to perform rarefaction studies, we demonstrated that our sampling of the VDJ repertoire was asymptoting toward saturation (Fig. 3A). Any VDJ classes that may be missing from the data are occurring at frequencies below 10^{-4} to 10^{-5}. There was a commonality to the frequency distributions of VDJ usage that was independent of the specific VDJ repertoire for individual fish (Fig. 3B). Specifically, the majority of VDJ combinations in each fish were of low abundance, but a similarly small fraction—although different combinations for different fish—were found at high frequencies. This distribution could be used to constrain theoretical models of repertoire development.

We next asked whether there is a quantitative relationship between the VDJ usage of different fish. The VDJ repertoire is a vector in which each element records the number of reads that map to a particular VDJ class. The dot product between VDJ repertoire vectors measures the degree of correlation between different fish (table S5 and Fig. 3C) [control experiments are described in (*13*)].

Most fish were uncorrelated in their VDJ repertoires; however, some fish were highly correlated, and three pairs of fish had correlation coefficients in the range 0.62 to 0.75. Some of these correlations appear to derive from the largest VDJ class in the repertoire (table S5A and Fig. 3C). When the fish-fish VDJ correlations were computed in the absence of the largest VDJ class, we discovered that although the largest correlations disappeared, a new set of correlations appeared between a larger fraction of the fish (table S5B and Fig. 3D). These correlations were mostly weaker than the previous correlations but still well above the statistical noise.

We were surprised to find measurable correlation in antibody repertoires between independent organisms. We created a model for random VDJ repertoire assembly, using simulated VDJ distributions that replicated the actual measured distributions and coverage fractions. The correlations in these simulated VDJ repertoires are all near zero, and the probability of two fish having a highly correlated random repertoire is less than 10^{-6} (Fig. 3, C and D). Thus, even though the VDJ repertoire is believed to be generated by a series of random molecular events within independent individual cells, in zebrafish the VDJ repertoire appears substantially structured and nonrandom on a global scale. It is possible that the source of this structure is simply convergent evolution, that the fish see a similar enough environment that selection in their quiescent immune systems converges to correlated VDJ usage. It is also possible that this distribution reflects bias in the VDJ recombination mechanisms, which would have important implications for antibody diversity space and would suggest that the number of solutions to a given antigen recognition problem, or at least the number that are readily evolvable, may be much smaller than previously assumed.

Summarizing the VDJ repertoire with a simple count of the number of different VDJ combinations neglects the variation in abundance of different VDJ species. Ecologists have the same problem in characterizing species diversity; they refer to the counting approach as species richness and have developed other methods to characterize variation of abundance, which they term "heterogeneity" (*14*). The most popular approach to characterize heterogeneity is based on information theory, specifically the Shannon-Weaver entropy, which summarizes the frequency distribution in a single number (*14*). The VDJ repertoire entropies generally varied between 3.1 and 7.7 bits for individual fish. Exponentiating the entropy indicates the effective size of the VDJ repertoire, and this varied between 9 and 200 with an average of 105, or an average effective VDJ repertoire coverage of about 9%. This can be interpreted as the fraction of highly expressed VDJ classes.

Whereas the VDJ repertoire provides a coarse view of immunological diversity, each VDJ class can contain a large number of distinct individual antibodies that differ as a result of hypermutations and junctional changes. We characterized the antibody repertoire by using quality threshold clustering of Smith-Waterman alignments to group similar reads together; each cluster defines an antibody. Performing this analysis on control data with well-defined sequence clones allowed us to calibrate the clustering algorithm and separate true hypermutation diversity from sequencing errors. Many VDJ combinations included a

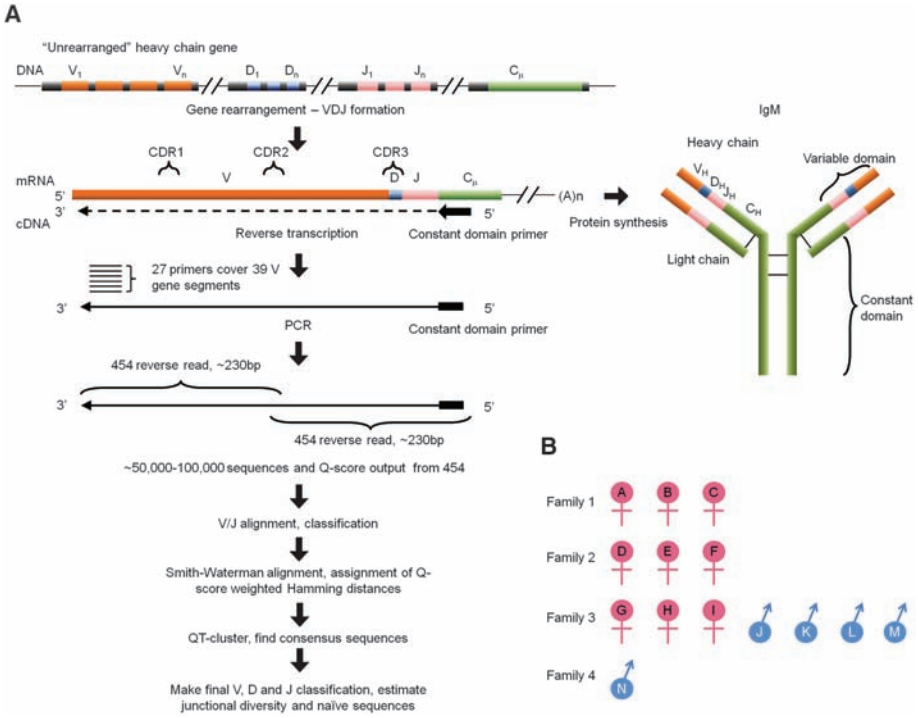

Fig. 1. (**A**) Schematic drawing of the VDJ recombination of an antibody heavy-chain gene, the cDNA amplicon library construction, and the infomatics pipeline. The heavy-chain VDJ segment of an antibody is created by recombination, junctional diversity, and hypermutation. We designed primer sets to amplify the expressed heavy-chain mRNA, which were then sequenced and analyzed as outlined. High-throughput sequencing allows determination of the identity of nearly all heavy-chain sequences. (**B**) Gender and family information for the 14 sequenced zebrafish.

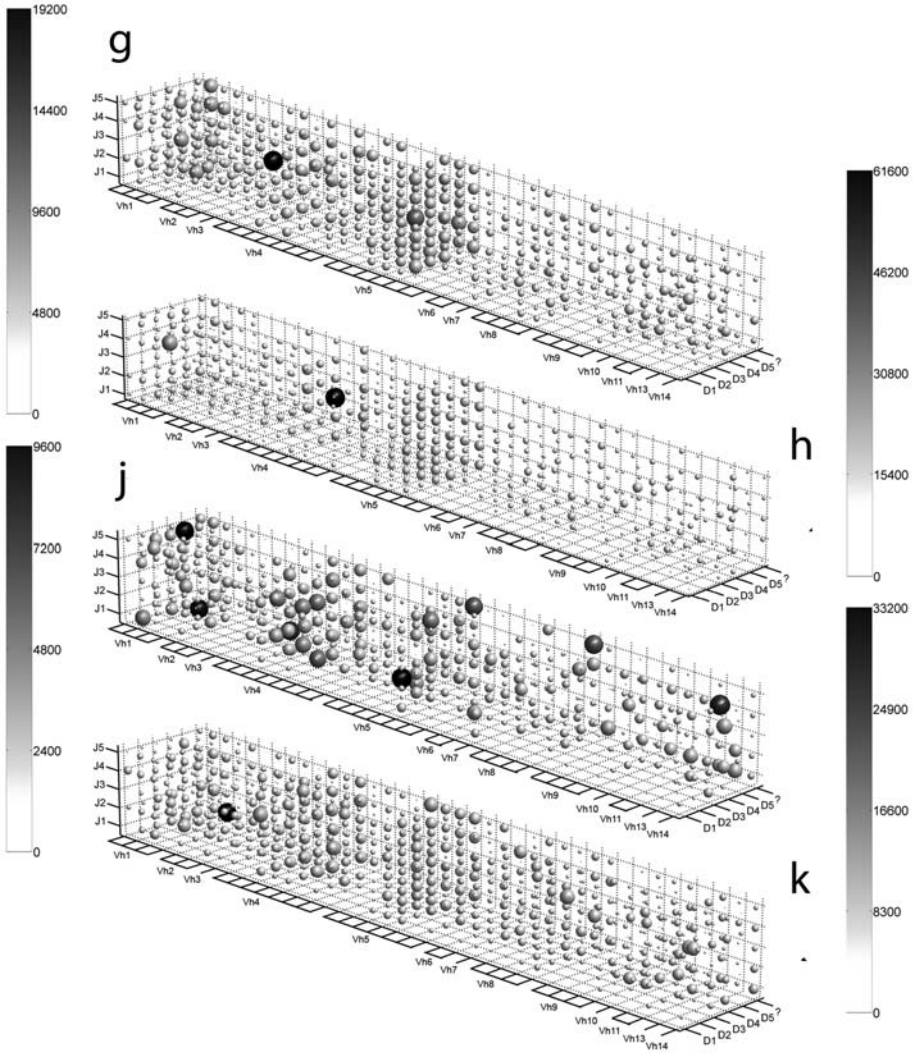

Fig. 2. The entire expressed VDJ repertoires for individual fish g, h, j, and k (top to bottom). The three axes enumerate all possible V, D, and J values, so each point in three-space is a unique VDJ combination. Both the size of the sphere at each point and the intensity correspond to the number of reads matching that particular VDJ combination. Gray scale is plotted on a linear scale, and the dot size is plotted on a log scale. The upper limits of the scales are set to the most populated VDJ combination for each fish, with PCR bias factored out.

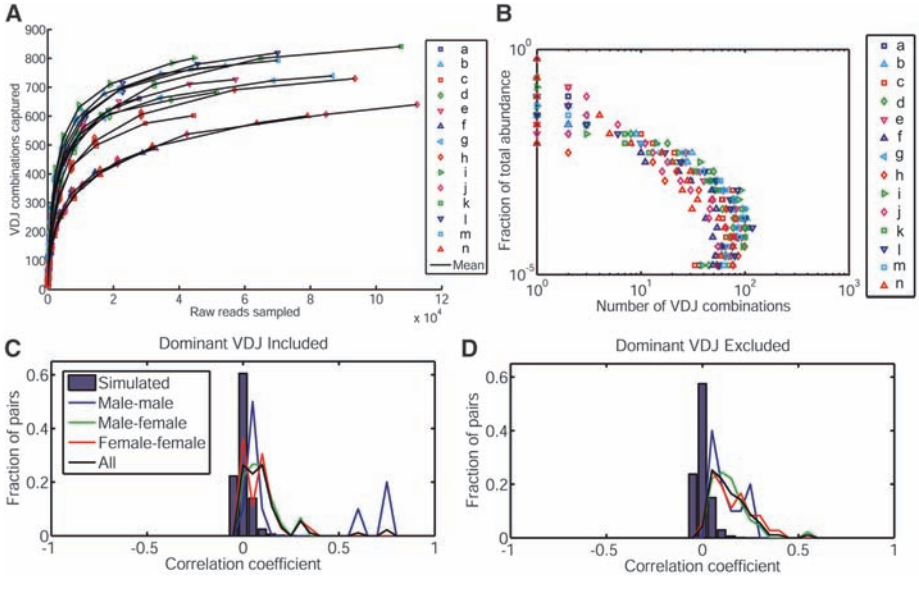

Fig. 3. VDJ repertoire analysis for all 14 fish. (**A**) Abundance distribution for each VDJ combination. A small number of VDJ combinations are highly represented in each fish, and most VDJ combinations are represented only at low abundance. The shape of the distribution is common among all of the fish sampled. This histogram is oriented sideways (from left to right) to emphasize that a small number of VDJ combinations are highly abundant, with a distribution that falls off rapidly. (**B**) Rarefaction analysis of VDJ diversity demonstrates that as one sequences more deeply into a fish, the number of new VDJ classes discovered saturates. (**C**) Histogram of correlations between VDJ repertoires. The data are collected as histograms and compared to simulated fish which have random VDJ repertoires. The simulated fish have no significant correlations, whereas some of the real fish have high correlations, representing 5 SD outliers of the random model. The highest correlations are from males in the same family (table S5A). (**D**) When the largest VDJ class in each fish is eliminated, the correlations are reduced and there is a larger proportion of moderate female correlations.

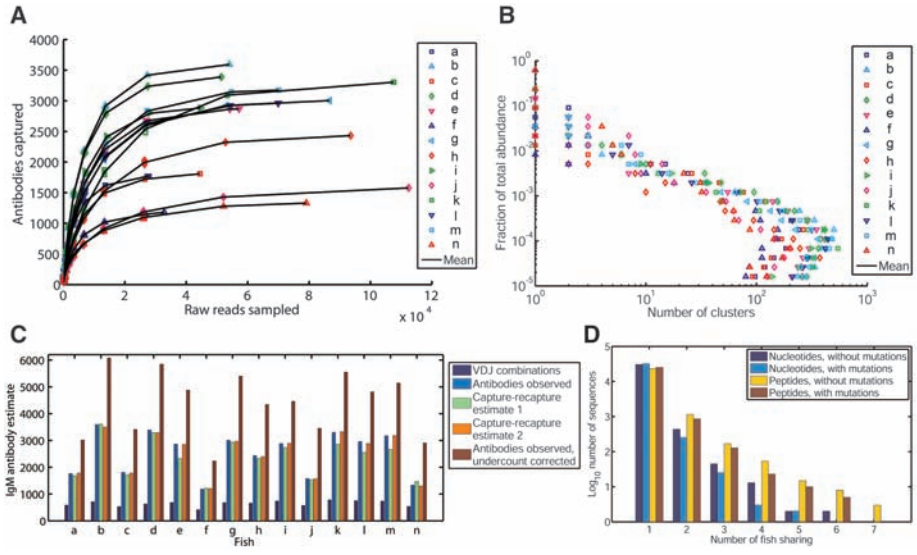

Fig. 4. Antibody heavy-chain repertoire diversity estimates for IgM in all 14 fish. (**A**) Rarefaction analysis of heavy-chain diversity demonstrates that as one sequences more deeply into a fish, the number of new antibodies discovered saturates at a few thousand. (**B**) Antibody abundance distributions for each fish. This histogram is oriented sideways (from left to right) to emphasize that a small number of antibodies (clusters) are highly abundant, with a distribution that falls off rapidly as a power law. The shape of the distribution is universal among all of the fish sampled. The bend at small abundance is caused by variability in the total reads sampled per fish and is not significant. (**C**) Total antibody diversity estimates for IgM using different criteria. VDJ diversity is the number of VDJ classes per fish, as described in Fig. 3A. Antibodies observed (>2 reads; VDJ classes composed only of antibody clusters with two or fewer reads are counted as one) is the number of unique antibodies per fish described in Fig. 4A. Capture-recapture estimate 1 refers to an estimate based on observed antibody abundances (*13*). Capture-recapture estimate 2 refers to an estimate using equal probability of all antibodies. Antibodies observed, undercount corrected refers to the upper bound. (**D**) Histogram of number of fish with shared IgM sequences. Hundreds of sequences are shared between pairs of fish, while a few tens of sequences are shared between three fish. Five sequences are shared between four or more fish, and none are shared among all fourteen fish. Sequence comparisons without mutations incorporate differences at the V/D and D/J junctions alone. Convergence on the amino acid level is also plotted.

large number of distinct antibodies. We found that the overall distribution of the abundances of the antibodies followed an apparent power law with scaling parameter 2.2, and this was consistent among all fish over two decades (Fig. 4B). This behavior may represent an important signature of the underlying dynamics of the adaptive immune system. It was not observed for either the control data or the VDJ distributions, and thus we ruled out the possibility that it is an artifact of polymerase chain reaction (PCR) bias.

There are several ways to use this data to estimate the number of unique antibodies per fish. The first is to perform rarefaction studies and determine whether the number of independent clusters tends to saturate. We did this and found that the saturation occurs at between ~1200 and 3500 unique antibodies per fish (Fig. 4A). Another way is by applying approaches used in ecology to estimate population sizes and diversity—sample and resample techniques (*15*). This yielded an estimate of between 1200 and 3700 unique antibodies per fish, whether applied blindly or using knowledge of the antibody abundance distributions (Fig. 4C). Both approaches are lower bounds on the true antibody diversity because antibodies that differ by only one or two mutations will be incorporated into the same cluster. We corrected for this effect by reanalyzing the data within each cluster with zero error tolerance, only matching exact reads. The largest clusters each had several subclusters with more than two reads each, and the control sequence data indicated that probably half of those clusters are real while the other half are artifacts due to sequencing error. By combining this stringent method of finding small differences in common sequences with the more permissive method of clustering rare sequences with less similarity together (thereby having tolerance to sequencing errors on rare transcripts), we estimated that the upper limit of heavy-chain antibody diversity is within 50% of the lower bound estimates, or between 5000 and 6000 antibodies in an individual fish.

To see how often repertoires converged to the same antibody, we searched for sequences that are shared between fish. Although there were no antibodies common to all fish, some antibodies were shared between smaller groups of fish (Fig. 4D). These cases of convergent evolution were more frequent than one would expect from a random usage model, with P values as low as 10^{-15}. Unexpectedly, different individuals shared heavy chains that were identical in the region we sequenced, even up to hypermutation. Specifically, there were 254 unique sequences shared between two fish and 2 unique sequences shared between five fish. These data illustrate the powerful forces of selection and perhaps can be used to estimate evolutionary dynamics in this system.

In conclusion, we have performed a comprehensive measurement of the heavy-chain antibody repertoire of zebrafish. We discovered that the abundance distributions of both the VDJ repertoire and antibody heavy-chain diversity were similar between individuals, that VDJ usage is not uniform, that individuals can have highly correlated VDJ repertoires, and that convergent evolution of identical heavy-chain sequences is unexpectedly common. With the rapid advance of sequencing throughput, it will soon be possible to make similar measurements on mice and humans. These organisms use the same molecular mechanisms for repertoire generation as fish; thus, we predict that they may also show similar distributions of antibody frequencies.

References and Notes

1. T. J. Kindt, J. D. Capra, *The Antibody Enigma* (Plenum Press, New York, 1984).
2. S. P. Singh, *Z. Allg. Mikrobiol.* **18**, 111 (1978).
3. C. A. Janeway, M. J. Shlomchik, M. Walport, P. Travers, *Immunobiology* (Garland Science Publishing, New York, ed. 6, 2004).
4. P. Boudinot *et al.*, *Mol. Immunol.* **45**, 2437 (2008).
5. E. S. Golub, *Cell* **48**, 723 (1987).
6. D. W. Talmage, *Science* **129**, 1643 (1959).
7. N. S. Trede, D. M. Langenau, D. Traver, A. T. Look, L. I. Zon, *Immunity* **20**, 367 (2004).
8. J. A. Yoder, M. E. Nielsen, C. T. Amemiya, G. W. Litman, *Microbes Infect.* **4**, 1469 (2002).
9. G. W. Litman, J. P. Cannon, L. J. Dishaw, *Nat. Rev. Immunol.* **5**, 866 (2005).
10. J. L. Xu, M. M. Davis, *Immunity* **13**, 37 (2000).
11. E. P. Rock, P. R. Sibbald, M. M. Davis, Y. H. Chien, *J. Exp. Med.* **179**, 323 (1994).
12. N. Danilova, J. Bussmann, K. Jekosch, L. A. Steiner, *Nat. Immunol.* **6**, 295 (2005).
13. Materials and methods are available as supporting material on *Science* Online.
14. R. K. Peet, *Annu. Rev. Ecol. Syst.* **5**, 285 (1974).
15. J. Bunge, M. Fitzpatrick, *J. Am. Stat. Assoc.* **88**, 364 (1993).

16. We thank W. Talbot for useful conversations and the generous loan of equipment and N. Neff for assistance with sequencing. IgM and IgZ sequence and quality-score files are available on the NIH short-read archive, with accession number SRA008134. This research was supported by the NIH Director's Pioneer award (S.R.Q.), the Arthritis Foundation Postdoctoral Fellowship (N.J.), and an NSF graduate fellowship (J.A.W.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/324/5928/807/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 to S6
References

19 December 2008; accepted 18 March 2009
10.1126/science.1170020