nature medicine

In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire

Brandon J DeKosky[1], Takaaki Kojima[1,2], Alexa Rodin[1], Wissam Charab[1], Gregory C Ippolito[3], Andrew D Ellington[4] & George Georgiou[1,3,5,6]

High-throughput immune repertoire sequencing has emerged as a critical step in the understanding of adaptive responses following infection or vaccination or in autoimmunity. However, determination of native antibody variable heavy-light pairs (VH-VL pairs) remains a major challenge, and no technologies exist to adequately interrogate the >1 × 10^6 B cells in typical specimens. We developed a low-cost, single-cell, emulsion-based technology for sequencing antibody VH-VL repertoires from >2 × 10^6 B cells per experiment with demonstrated pairing precision >97%. A simple flow-focusing apparatus was used to sequester single B cells into emulsion droplets containing lysis buffer and magnetic beads for mRNA capture; subsequent emulsion RT-PCR generated VH-VL amplicons for next-generation sequencing. Massive VH-VL repertoire analyses of three human donors provided new immunological insights including (i) the identity, frequency and pairing propensity of shared, or 'public', VL genes, (ii) the detection of allelic inclusion (an implicated autoimmune mechanism) in healthy individuals and (iii) the occurrence of antibodies with features, in terms of gene usage and CDR3 length, associated with broadly neutralizing antibodies to rapidly evolving viruses such as HIV-1 and influenza.

The determination of immune receptor repertoires using high-throughput (next-generation) DNA sequencing is rapidly becoming an indispensable tool for the understanding of adaptive immunity, for antibody discovery and in clinical practice[1–3]. However, because the variable domains of antibody heavy and light chains (VH and VL, respectively) are encoded by different mRNA transcripts, until recently it was only possible to determine the VH and VL repertoires separately, or else paired VH-VL sequences for small numbers, and more recently, moderate numbers, of cells (10^4–10^5)[4], far fewer than the ~0.7 × 10^6 to 4 × 10^6 B cells contained in a typical 10-ml blood draw. Thus, a technology for the facile determination of the paired antibody VH-VL repertoire at great depth and for a variety of B cell subsets is of interest for clinical research[5], for antibody discovery[6,7] and for addressing a host of questions related to the shaping of the antibody repertoire[2,8–13].

Several techniques have been reported for detection or sequencing of genomic DNA or cDNA from single cells; however, all are limited by low efficiency or low cell throughput (<200–500 cells) and require fabrication and operation of complicated microfluidic devices[14–17]. Chudakov and coworkers recently reported the use of one-pot cell encapsulation within water-in-oil emulsions, achieving cell lysis by heating at 65 °C concomitant with reverse transcription of the genes encoding T cell receptor α (TCRα) and TCRβ and linking by overlap-extension PCR to determine TCRα-TCRβ pairings, albeit only for TCRβV7 and with a very low efficiency (approximately 700 TCRα-TCRβ pairs recovered from 8 × 10^6 peripheral blood mononuclear cells (PBMCs))[17]. This is probably because one-pot emulsions result in a high degree of droplet size dispersity, and because the reverse transcription reaction is inhibited in volumes <5 nl (ref. 15), only the small fraction of cells encapsulated within larger droplets yields cDNA for further manipulation. Incomplete cell lysis and RNA degradation during thermal cell lysis further reduce the yield of linked cDNA products achieved with this approach.

Inspired by methods for the production of highly monodisperse polymeric microspheres for drug delivery purposes[18,19], we developed a new technology that enables sequencing of the paired VH-VL repertoire from millions of B cells within a few hours of experimental effort and using equipment that can be built inexpensively by most laboratories. For validation, we expanded memory B cells isolated from human PBMCs *in vitro* to obtain a sample that contained multiple clones of individual B cells and showed that among aliquots (technical replicates), the accuracy of VH-VL pairing is >97%. We show that ultra-high throughput determination of the paired VH-VL repertoire provides immunological insights such as the discovery of human light chains detected in multiple individuals that pair with a wide range of VH genes, the quantitative analysis of allelic inclusion in humans, i.e., of B cells expressing two different antibodies, and estimates of the frequencies of antibodies in healthy human repertoires that display known features of broadly neutralizing antibodies to rapidly evolving pathogens.

RESULTS

Device construction

For facile high-throughput single-cell manipulation, we assembled a simple axisymmetric flow-focusing device comprising three

[1]Department of Chemical Engineering, University of Texas at Austin, Austin, Texas, USA. [2]Laboratory of Molecular Biotechnology, Graduate School of Bioagricultural Sciences, Nagoya University, Nagoya, Japan. [3]Department of Molecular Biosciences, University of Texas at Austin, Austin, Texas, USA. [4]Center for Systems and Synthetic Biology, University of Texas at Austin, Austin, Texas, USA. [5]Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas, USA. [6]Department of Biomedical Engineering, University of Texas at Austin, Austin, Texas, USA. Correspondence should be addressed to G.G. (gg@che.utexas.edu).

Received 24 February; accepted 8 October; published online 15 December 2014; doi:10.1038/nm.3743

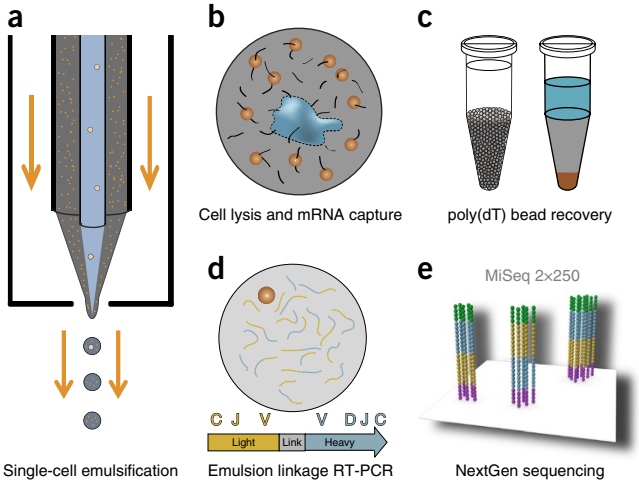

Figure 1 Technical workflow for ultra-high throughput VH-VL sequencing from single B cells. (**a**) An axisymmetric flow-focusing nozzle isolated single cells and poly(dT) magnetic beads into emulsions of predictable size distributions. An aqueous solution of cells in PBS (center, blue with pink circles) and cell lysis buffer with poly(dT) beads (gray with orange circles) exited an inner and outer needle and were surrounded by a rapidly moving annular oil phase (orange arrows). Aqueous streams focused into a thin jet, which coalesced into emulsion droplets of predictable sizes, and cells mixed with lysis buffer only at the point of droplet formation (**Supplementary Fig. 1**). (**b**) Single-cell VH and VL mRNAs annealed to poly(dT) beads within emulsion droplets (blue figure represents a lysed cell, orange circles depict magnetic beads and black lines depict mRNA strands). (**c**) poly(dT) beads with annealed mRNA were recovered by emulsion centrifugation to concentrate aqueous phase (left) followed by diethyl ether destabilization (right). (**d**) Recovered beads were emulsified for cDNA synthesis and linkage PCR to generate an ~850–base pair VH-VL cDNA product. (**e**) Next-generation sequencing of VH-VL amplicons was used to analyze the native heavy and light chain repertoire of input B cells.

concentric tubes: an inner needle carrying cells suspended in PBS, a middle tube carrying a lysis solution and magnetic poly(dT) beads for mRNA capture from lysed cells, and an external tube with a rapidly flowing annular oil phase, all of which passed through a 140-µm glass nozzle (**Fig. 1a**). The rapidly flowing outer annular oil phase focused the slower-moving aqueous phase into a thin, unstable jet that coalesced into droplets with a predictable size distribution; additionally, maintaining laminar flow regime within the apparatus prevented mixing of cells and lysis solution before droplet formation (**Supplementary Fig. 1**).

To evaluate cell encapsulation and droplet size distribution, we injected MOPC-21 immortalized B cells suspended in PBS through the inner needle at a rate of 250,000 cells per minute while we injected a solution of PBS containing the cell viability dye trypan blue (0.4% v/v) through the middle tubing so that dye mixed with cells at the point of droplet formation. The resulting emulsion droplets were 73 ± 20 µm in diameter (mean ± s.d.). Trypan blue exclusion revealed that, as expected, cells remained viable throughout the emulsification process (**Supplementary Fig. 2**). Replacing the trypan blue stream with cell lysis buffer containing lithium dodecyl sulfate (LiDS) and DTT to inactivate RNases resulted in complete cell lysis, as indicated by visual disappearance of cell membranes from emulsion droplets.

Single B cell VH-VL pairing: throughput and pairing accuracy

We isolated human $CD3^{-}CD19^{+}CD20^{+}CD27^{+}$ memory B cells from PBMCs from a healthy volunteer and expanded them for 4 d *in vitro* by stimulation with anti-CD40 antibody, interleukin-4 (IL-4), IL-10, IL-21 and CpG oligodeoxynucleotides. We performed *in vitro* expansion to create a cell population containing a sufficient number of clonal B cells so that we could assess the concordance of the VH-VL repertoire in two technical replicates. We divided 1,600,000 *in vitro*–expanded B cells into two aliquots and passed them through the flow-focusing nozzle at a rate of 50,000 cells per minute (i.e., 16-min emulsification for each replicate) and processed each aliquot as shown in **Figure 1**. We maintained the emulsion of lysed single cells with compartmentalized poly(dT) beads for 3 min at room temperature to allow specific mRNA hybridization onto poly(dT) magnetic beads (**Fig. 1b**); then, the emulsion was broken chemically (**Fig. 1c**), beads were reemulsified and overlap-extension RT-PCR was performed to generate linked VH-VL amplicons (**Fig. 1d**). We amplified the resulting cDNAs by nested PCR to generate an ~850-bp VH-VL product for high-throughput sequencing by Illumina MiSeq 2x250 or 2x300 (**Fig. 1e**). Owing to read-length limitations of current high-throughput sequencing technologies, we sequenced the FRH4-(CDR-H3)-FRH3-FRL3-(CDR-L3)-FRL4 antibody regions first to reveal the pairing of the VH and VL hypervariable loops. Each of these VH-VL pairs may also comprise one or more somatic variants containing mutations within the upstream portion of the VH and VL genes. We determine the complete set of somatic variants by separate MiSeq sequencing of the VH and VL portions of the paired 850-bp VH-VL amplicon followed by *in silico* gene assembly[4].

We processed sequence data by read-quality filtering, CDR-H3 clustering, VH-VL pairing and selection for paired VH-VLs with more than or equal to two reads in the data set. The clustering step resulted in high-confidence sequence data but with a lower-bound estimate of clonal diversity because clonally expanded or somatically mutated B cells with similar VH sequences collapse into a single CDR-H3 cluster. We observed 129,097 VH-VL clusters after separate analysis and clustering of replicates 1 and 2. Among these, we observed 37,995 CDR-H3 sequences in both replicates (and hence they must have originated from expanded B cells present in both technical replicates) with 36,468 paired with the same CDR-L3 across replicates, revealing a VH-VL pairing precision of 98.0% (**Fig. 2**, **Table 1**, **Supplementary Fig. 3** and Online Methods; VH-VL sequences are reported in **Supplementary Data Set 1**). The ratio of VH-VL clusters to input cells observed (typically between 1:10 and 1:15) is a reflection of the clonality of the memory B cell population (i.e., the presence of clonally related memory B cells), clustering threshold, RT-PCR efficiency and cell viability. For comparison, in our hands, sequencing the memory B cell VH repertoire by preparing amplicons directly by standard RT-PCR without pairing and using the same bioinformatic filters (sequences present at more than or equal to two reads, 96% clustering) resulted in a 1:6 ratio of VH clusters to input cells, which compares favorably to the yield of paired VH-VL clusters in **Table 1**. We also performed two additional pairing analyses of somewhat smaller B cell populations from different donors (**Table 1** and **Supplementary Figs. 4** and **5**; sequences provided in **Supplementary Data Set 1**). In a separate experiment designed to verify native VH-VL pairing accuracy, we transfected plasmids encoding 11 different known human antibodies separately into HEK293 cells. We mixed and processed aliquots containing comparable numbers of each of the transfected cells as described in **Figure 1**, and we identified native pairings for 11 of 11 antibodies (**Supplementary Table 1**). In yet another test, we mixed ~260 ARH-77 immortalized

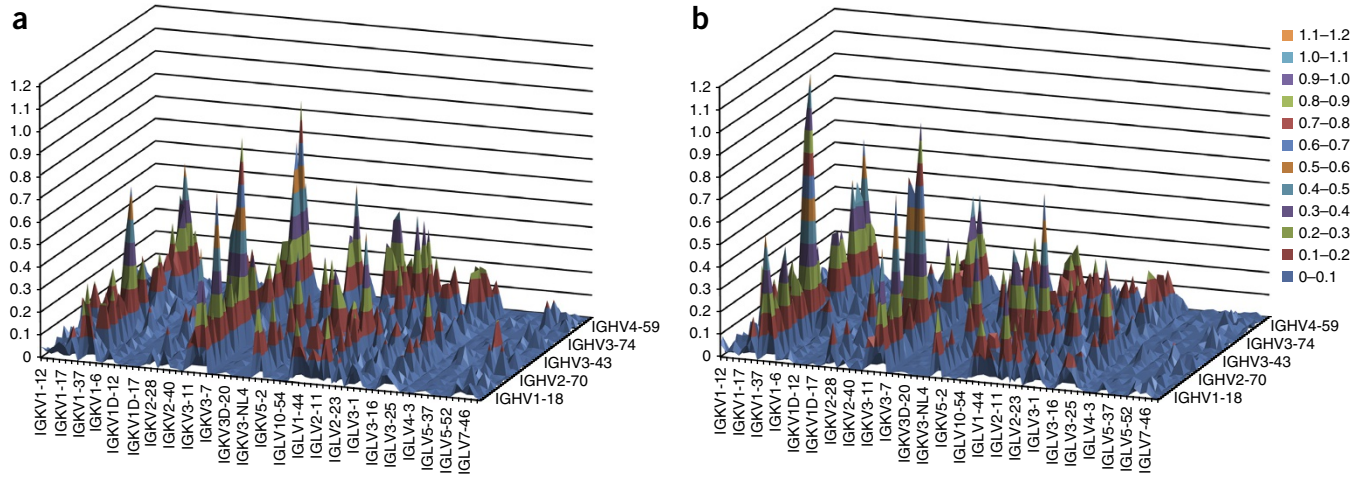

Figure 2 Heavy-light V-gene pairing landscape of CD3$^-$CD19$^+$CD20$^+$CD27$^+$ peripheral memory B cells in two healthy human donors. V genes are plotted in alphanumeric order; height indicates percentage representation among VH-VL clusters. (**a**) Donor 1 (n = 129,097 clusters). (**b**) Donor 2 (n = 53,679). VH-VL gene usage was highly correlated between donors 1 and 2 (Spearman rank correlation coefficient 0.757, $P < 1 \times 10^{-99}$). Additional heat maps are provided in **Supplementary Figures 3** and **4**; VH-VL sequences are reported in **Supplementary Data Set 1**.

human B cells[4] with 20,000 CD3$^-$CD19$^+$CD20$^+$CD27$^+$ expanded memory B cells (~100-fold excess). ARH-77 heavy and light chains were paired correctly, and the ratio of correctly paired ARH-77 VH-VL reads over the top correct VH–incorrect VL, a parameter that we denote signal to top VL noise, was 96.4:1 (2,604 correct ARH-77 VH-VL reads versus 27 reads for the top ARH-77 VH paired with an incorrect VL, **Supplementary Table 2**).

As discussed above, three sequencing reactions and *in silico* assembly are needed to determine the sequence of the complete linked VH-VL amplicon with Illumina MiSeq 2x250 or 2x300. Alternatively, the long-read Pacific Biosciences (PacBio) sequencing platform can be used to obtain the complete ~850-bp cDNA encoding linked VH-VL sequences. However, because of its substantially lower throughput and higher cost per read, we find that despite the need for three distinct MiSeq samples compared to only one for PacBio, MiSeq is currently much more cost effective for deep repertoire analyses. We found PacBio sequencing to be preferable only for certain specialized applications, for example in identifying VH-VL pairs in antibodies with extensive somatic hypermutation (SHM) such as broadly neutralizing antibodies that arise following persistent infection with rapidly evolving viruses, most notably HIV-1. For example, we used PacBio to sequence 15,000 VH-VL amplicons from an HIV elite neutralizer, donor CAP256 (ref. 7), and identified six variants of VRC26-class HIV broadly neutralizing antibodies within the VH-VL repertoire (**Supplementary Figs. 6** and **7**).

Promiscuous and public VL junctions

In contrast to heavy-chain rearrangements, light-chain rearrangements do not incorporate a diversity segment and exhibit restricted CDR-L3 lengths with low levels of nontemplated nucleotide addition (N addition) at the CDR-L3 junction. Light chains therefore have a much lower theoretical diversity than heavy chains, and the presence of light chain sequences paired with multiple heavy chains within a single donor, referred to as 'repeated' or 'promiscuous' light chains, is an expected result, especially for VL junctions that are mostly germline encoded and also derive from V and J genes with high prevalence in human immune repertoires[20]. However, the separate high-throughput sequencing of VH and VL repertoires, as has been practiced until now, cannot provide VH pairing information for a given VL and thus precludes identification and characterization of promiscuous light chains[21,22]. We observed thousands of heavy chains paired with promiscuous VL nucleotide junctions (34.9%, 29.4% and 19.6% of all heavy chains were paired with promiscuous VL junctions in donors 1, 2 and 3, respectively; a quantitative list of all promiscuous VL junctions is provided in **Supplementary Data Set 2**). We inspected high-frequency promiscuous light chains to see whether any promiscuous VL might be shared across individuals (i.e., a 'public' VL). We found that highly promiscuous VLs were nearly always public: for example, of the 50 highest-frequency promiscuous VL junctions in donor 1, 49/50 were also detected in donors 2 and 3. Promiscuous light chains showed an average of 0.04 nontemplated bases in the VL junction compared to an average of five nontemplated bases in nonpromiscuous light chains (i.e., VLs that paired with a single VH in a donor, see **Supplementary Fig. 8**, $P < 1 \times 10^{-10}$). The lack of nontemplated bases in promiscuous VL junctions indicated that promiscuity can be observed mainly in germline-encoded VL genes lacking SHM.

We examined in detail two representative promiscuous and public VL junctions that contained V and J genes with high prevalence in steady-state human immune repertoires (*KV1-39–KJ2*, 9–amino acid (aa) CDR-L3, *LV1-44–LJ3*, 11-aa CDR-L3, both observed at a

Table 1 High-throughput VH-VL sequence analysis of CD3$^-$CD19$^+$CD20$^+$CD27$^+$ *in vitro*–expanded human B cells

Human donor	V-region primer set	No. cells analyzed	Emulsification rate (cells per minute)	Observed VH-VL clusters	CDR-H3 detected in both replicates	CDR-H3–CDR-L3 clusters detected in both replicates	VH-VL pairing precision
Donor 1	Framework 1	1,600,000	50,000	129,097	37,995	36,468	98.0%
Donor 2	Framework 1	810,000	50,000	53,679	19,096	18,115	97.4%
Donor 3	Leader peptide	210,000	33,000	15,372	4,267	4,170	98.9%

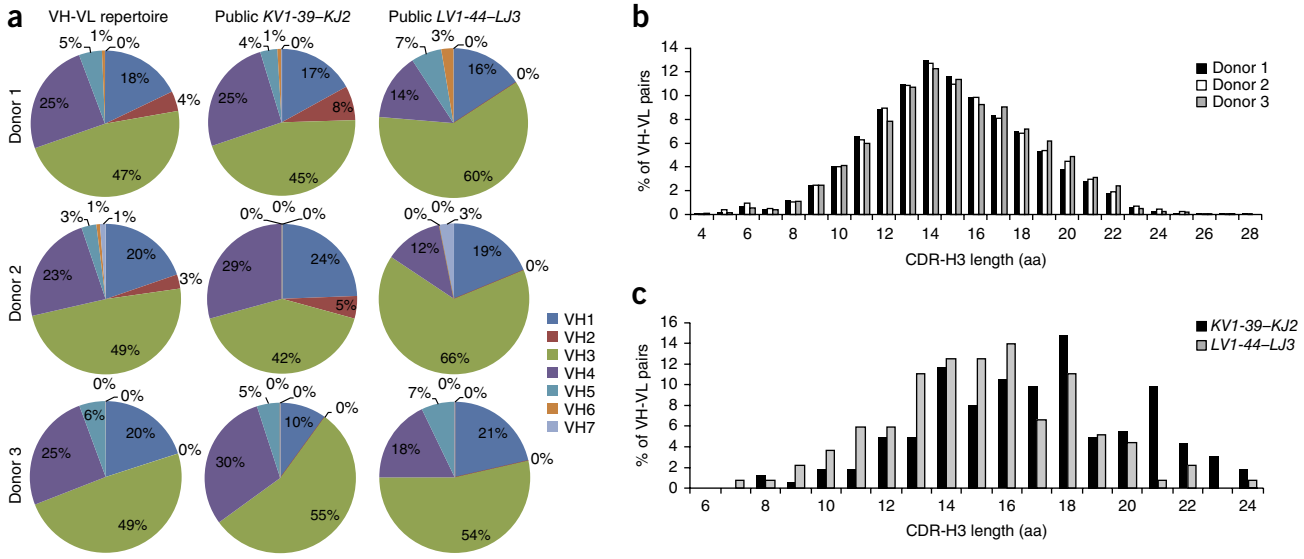

Figure 3 VH pairing statistics for representative promiscuous and public light chains. (**a**) VH gene family utilization in total paired VH-VL repertoires (left; donor 1 $n = 129{,}097$ clusters, donor 2 $n = 53{,}679$, donor 3 $n = 15{,}372$), heavy chains paired with a representative highly ranked public and promiscuous VL observed in all three donors (center; *KV1-39–KJ2* 9-aa CDR-L3, TGTCAACAGAGTTACAGTACCCCGTACACTTTT; donor 1 $n = 106$ clusters, donor 2 $n = 41$, donor 3 $n = 20$) and heavy chains paired with a different highly ranked public VL in all three donors (right; *LV1-44–LJ3* 11-aa CDR-L3, TGTGCAGCATGGGATGACAGCCTGAATGGTTGGGTGTTC; $n = 76$ clusters, $n = 32$ and $n = 28$, respectively). (**b**) CDR-H3 length distribution in VH-VL repertoires (donor 1 $n = 129{,}097$ clusters, donor 2 $n = 53{,}679$, donor 3 $n = 15{,}372$). (**c**) CDR-H3 length distribution for all antibodies containing the two representative public VL chains from **a**.

frequency of ~1 per 1,000 VH-VL clusters)[23,24] to check for biases in VH pairing of promiscuous VL chains. *KV1-39–KJ2* and *LV1-44–LJ3* both paired with VH genes of diverse germline lineages and CDR-H3 length that reflected the overall VH gene usage in the repertoire (**Fig. 3**, Spearman rank correlation coefficients: *KV1-39–KJ2* $\rho = 0.889$, $P < 1 \times 10^{-21}$; *LV1-44–LJ3* $\rho = 0.847$, $P < 1 \times 10^{-17}$), indicating that VL nucleotide sequence promiscuity arises mostly from distinct VL recombination events rather than B cell activation and subsequent clonal expansion. We note that no two donors shared more than two VH nucleotide sequences, and no VH sequence was detected in all three donors, consistent with previous reports showing that in contrast to VL junctions, the VH nucleotide repertoire is highly private[2,24].

Quantifying allelic inclusion in human memory B cells

Clonal selection theory postulates that each lymphocyte expresses one antibody. However, studies in mice have confirmed that this is not always the case. Allelic inclusion, the phenomenon whereby one B cell expresses two B cell receptors (BCRs), overwhelmingly one VH gene with two different VLs, has been well documented in mice and has been proposed to be particularly important in autoimmunity because the expression of a second BCR can dilute a preexisting auto-reactive BCRs and limit the expansion of autoreactive B cells. Similarly, allelic inclusion can also provide a mechanism for autoreactive antibodies to evade central tolerance[25–29]. Almost 20 years ago, Lanzavecchia and coworkers used FACS sorting of cells expressing both κ and λ immunoglobulin proteins on their cell surface (sIgκ$^+$sIgλ$^+$, denoting surface-expression of both Igκ and Igλ) followed by Epstein-Barr virus immortalization to show that sIgκ$^+$sIgλ$^+$ allelic inclusion occurs in 0.2–0.5% of human memory B cells[30]. However, the inability to sort dual sIgκ$^+$ and dual sIgλ$^+$ human B cells and the absence of methods for the determination of the VH-VL repertoire at sufficient depth (because the frequency of allelic inclusion is low) have precluded more comprehensive determination of allelic inclusion in humans. We detected VL allelic inclusion at a rate of approximately 0.4% of VH clusters for donors 1 and 2, with dual κ- and λ-transcribing B cells in approximately equal proportions to dual κ- and κ- and λ- and λ-transcribing B cell clones (**Fig. 4**; a list of allelically included clones provided in **Supplementary Data Set 3**). These heavy chains paired only with their two allelically included light chains (exact nucleotide match) in two technical replicates, and we observed that approximately 80% of these antibodies displayed somatic mutations. The somatic mutation frequency detected in allelically included VH-VL pairs was comparable to previous reports by Lanzavecchia and coworkers for allelically included sIgκ$^+$sIgλ$^+$ cells (three of five Epstein-Barr virus–immortalized clones in ref. 30). Also consistent with the earlier study, we observed

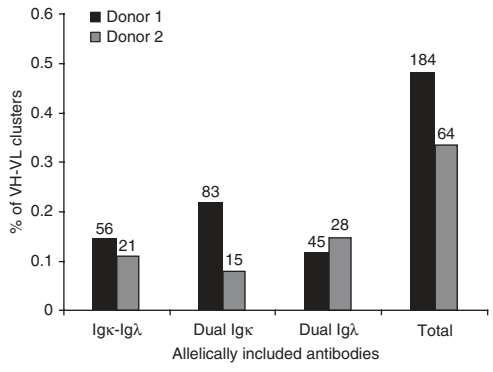

Figure 4 Frequency of VL transcript allelic inclusion in two donors ($n = 184$ and $n = 64$ allelically included antibodies from $n = 37{,}995$ and $n = 19{,}096$ VH-VL clusters detected across replicates in donor 1 and donor 2, respectively). Fourteen allelically included antibodies were detected in donor 3 (eight dual κ-λ, two dual κ-κ, two dual λ-λ, $n = 4{,}267$ VH-VL clusters detected across replicates). Numbers above each category indicate the absolute number of observed allelically included antibodies.

stop codons resulting from somatic mutation that inactivated a subset of allelically included VL transcripts[30] (sequences provided in **Supplementary Data Set 4**). For the ~20% of allelically included VHs that do not display SHM, we cannot rule out the possibility that these clones were derived from pre-B expansion.

Gene signatures of antiviral broadly neutralizing antibodies

High-resolution sequence descriptions of the immune repertoire can inform on B cell trajectories for the emergence of broadly neutralizing antibodies (bNAbs) to rapidly evolving pathogens[7,8,31,32]. Many bNAbs display highly unusual features including very long CDR-H3 and short CDR-L3 sequences[7,31,33,34], and these properties have raised the question as to whether antibodies with similar features are normally found in the repertoire of healthy donors and thus could evolve following stimulation by infection or vaccination to yield neutralizing antibodies. We found ~1:6,000 VH-VL clusters exhibited general characteristics of known VRC01-class anti-HIV antibodies (22, 9 and 0 for donors 1, 2 and 3, respectively; germline *VH1-02*, a very short ≤5-aa CDR-L3, and CDR-H3 length between 11 and 18 aa[34]), whereas antibodies with genetic characteristics of the anti-influenza antibody clone FI6 occurred in approximately 2×10^4 to 5×10^4 memory B cells (six and one antibodies detected in donors 1 and 2, respectively; *VH3-30*, *KV4-1*, 22-aa CDR-H3, 9-aa CDR-L3 (ref. 31)).

DISCUSSION

We have developed an easy-to-implement, ultra-high-throughput technology for sequencing the VH-VL repertoire at relatively low cost and with high pairing accuracy. The workflow presented here permits sequence analysis of the entire population of human B cells contained in a 10-ml blood draw, or, if needed, even in a unit of blood (450 ml) in a single-day experiment, an improvement of orders of magnitude relative to what is feasible using robotic single-cell RT-PCR[35]. As many as 6 million B cells (or, alternatively, as few as 1,000 B cells) can be analyzed per operator in a single day. Of note, the number of antibody sequences reported here (~200,000) dwarfs the entire set of <19,000 human VH-VL sequences that had been deposited in the International Nucleotide Sequence Database Collaboration over the past 25 years (in addition to the ~5,000 human VH-VL pairs we reported previously[4]).

The determination of the paired antibody repertoire at great depth can provide insights on a number of medically and immunologically important issues. For example, we used high-throughput single-cell VH-VL sequencing to detect highly used promiscuous and germline-encoded VL junctions that are observed in multiple donors, to identify antibodies with bNAb-like features in subjects with HIV-1 (**Supplementary Figs. 6** and **7**, ref. 7) and to quantify the frequency of bNAb-like V gene rearrangements in healthy donors, as described above. The latter is a key factor in determining whether a vaccine immunogen might be able to elicit protective immunity[33,34]. High-throughput VH-VL sequencing can also be used to search for public antibody VH-VL clonotypes[36,37] and to identify antibodies having specific features determined by computational or structural biology analyses or with relevance to pathogen neutralization[7,34,38–40]. In autoimmunity, high-throughput VH-VL sequencing can reveal an individual's repertoire of allelically included B cells and the presence of B cell clones expressing antibodies containing hallmark autoimmune signatures (with respect to paratope net charge, CDR-H3 and CDR-L3 lengths, etc.) as well as other attributes of potential diagnostic and therapeutic utility[26,28,29]. As DNA sequencing technologies continue to progress, low-cost high-throughput single-cell antibody sequencing can enable paired antibody repertoire analysis at great depth in large study cohorts and clinical patients and in turn provide insights into humoral responses associated with vaccine development, autoimmunity, infectious diseases and other human disease states.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. DNA sequences can be downloaded from the NCBI Short Read Archive (SRA) under study accession number SRP047462, and computer source code is available in GitHub repository NMED-NT69968.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank C. Berkland, M. Singh and N. Dormer for critical help and advice. We thank B. Iverson and D. Derryberry for insightful feedback, M. Ronalter, C. Das, M. Wirth and Y. Wine for help with experiments, J. Lavinder for reviewing the manuscript, O. Lungu for help with data analysis, B. Briney for sharing an unpublished primer sequence, L. Morris (National Institute for Communicable Diseases of the National Health Laboratory Service, South Africa) and P. Kwong (Vaccine Research Center, NIAID, USA) for sharing CAP256 samples, C. McHenry for PacBio sequencing, and J. Wheeler and S. Hunicke-Smith for Illumina MiSeq sequencing. This work was funded by fellowships to B.J.D. from the Hertz Foundation, the University of Texas Donald D. Harrington Foundation and the National Science Foundation, and by the US Defense Threat Reduction Agency (DTRA) HDTRA1-12-C-0105 (G.G.). A.D.E. would like to acknowledge funding from US National Security Science and Engineering Faculty Fellowship (FA9550-10-1-0169) and grants from the DTRA (HDTRA1-12-C-0007) and the Welch Foundation (F-1654). The content is solely the responsibility of the authors and do not necessarily represent the official views of the sponsors.

AUTHOR CONTRIBUTIONS

B.J.D. and G.G. developed the methodology and wrote the manuscript; B.J.D., T.K., G.C.I., A.D.E. and G.G. designed the experiments; B.J.D., T.K., A.R. and W.C. performed the experiments; B.J.D. carried out the bioinformatic analysis; and B.J.D. and T.K. analyzed the data.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Warren, E.H., Matsen, F.A. & Chou, J. High-throughput sequencing of B- and T-lymphocyte antigen receptors in hematology. *Blood* **122**, 19–22 (2013).
2. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32**, 158–168 (2014).
3. Logan, A.C. *et al.* High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc. Natl. Acad. Sci. USA* **108**, 21194–21199 (2011).
4. DeKosky, B.J. *et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* **31**, 166–169 (2013).
5. Sasaki, S. *et al.* Limited efficacy of inactivated influenza vaccine in elderly individuals is associated with decreased production of vaccine-specific antibodies. *J. Clin. Invest.* **121**, 3109–3119 (2011).
6. Smith, K. *et al.* Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen. *Nat. Protoc.* **4**, 372–384 (2009).
7. Doria-Rose, N.A. *et al.* Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* **509**, 55–62 (2014).
8. Wu, X. *et al.* Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* **333**, 1593–1602 (2011).
9. Fischer, N. Sequencing antibody repertoires: the next generation. *MAbs* **3**, 17–20 (2011).
10. Wilson, P.C. & Andrews, S.F. Tools to therapeutically harness the human antibody response. *Nat. Rev. Immunol.* **12**, 709–719 (2012).

11. Finn, J.A. & Crowe, J.E. Jr. Impact of new sequencing technologies on studies of the human B cell repertoire. *Curr. Opin. Immunol.* **25**, 613–618 (2013).
12. Finco, O. & Rappuoli, R. Designing vaccines for the twenty-first century society. *Front. Immunol.* **5**, 12 (2014).
13. Newell, E.W. & Davis, M.M. Beyond model antigens: high-dimensional methods for the analysis of antigen-specific T cells. *Nat. Biotechnol.* **32**, 149–157 (2014).
14. Marcus, J.S., Anderson, W.F. & Quake, S.R. Microfluidic single-cell mRNA isolation and analysis. *Anal. Chem.* **78**, 3084–3089 (2006).
15. White, A.K. *et al.* High-throughput microfluidic single-cell RT-qPCR. *Proc. Natl. Acad. Sci. USA* **108**, 13999–14004 (2011).
16. Furutani, S., Nagai, H., Takamura, Y., Aoyama, Y. & Kubo, I. Detection of expressed gene in isolated single cells in microchambers by a novel hot cell-direct RT-PCR method. *Analyst* **137**, 2951–2957 (2012).
17. Turchaninova, M.A. *et al.* Pairing of T-cell receptor chains via emulsion PCR. *Eur. J. Immunol.* **43**, 2507–2515 (2013).
18. Berkland, C., Kim, K.K. & Pack, D.W. Fabrication of PLG microspheres with precisely controlled and monodisperse size distributions. *J. Control. Release* **73**, 59–74 (2001).
19. Berkland, C., Pollauf, E., Pack, D.W. & Kim, K. Uniform double-walled polymer microspheres of controllable shell thickness. *J. Control. Release* **96**, 101–111 (2004).
20. Jackson, K.J.L., Kidd, M.J., Wang, Y. & Collins, A.M. The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor. *Front. Immunol.* **4**, 263 (2013).
21. Jackson, K.J.L. *et al.* Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells. *Immunogenetics* **64**, 3–14 (2012).
22. Hoi, K.H. & Ippolito, G.C. Intrinsic bias and public rearrangements in the human immunoglobulin Vλ light chain repertoire. *Genes Immun.* **14**, 271–276 (2013).
23. Ippolito, G.C. *et al.* Antibody repertoires in humanized NOD-scid-IL2R γnull mice and human b cells reveals human-like diversification and tolerance checkpoints in the mouse. *PLoS ONE* **7**, e35497 (2012).
24. Glanville, J. *et al.* Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci. USA* **108**, 20066–20071 (2011).
25. Pelanda, R. Dual immunoglobulin light chain B cells: trojan horses of autoimmunity? *Curr. Opin. Immunol.* **27**, 53–59 (2014).
26. Liu, S. *et al.* Receptor editing can lead to allelic inclusion and development of B cells that retain antibodies reacting with high avidity autoantigens. *J. Immunol.* **175**, 5067–5076 (2005).
27. Rezanka, L.J., Kenny, J.J. & Longo, D.L. Dual isotype expressing B cells [κ^+/λ^+] arise during the ontogeny of B cells in the bone marrow of normal nontransgenic mice. *Cell. Immunol.* **238**, 38–48 (2005).
28. Casellas, R. *et al.* Igκ allelic inclusion is a consequence of receptor editing. *J. Exp. Med.* **204**, 153–160 (2007).
29. Andrews, S.F. *et al.* Global analysis of B cell selection using an immunoglobulin light chain–mediated model of autoreactivity. *J. Exp. Med.* **210**, 125–142 (2013).
30. Giachino, C., Padovan, E. & Lanzavecchia, A. $\kappa^+\lambda^+$ dual receptor B cells are present in the human peripheral repertoire. *J. Exp. Med.* **181**, 1245–1250 (1995).
31. Corti, D. *et al.* A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science* **333**, 850–856 (2011).
32. Wrammert, J. *et al.* Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. *J. Exp. Med.* **208**, 181–193 (2011).
33. Jardine, J. *et al.* Rational HIV immunogen design to target specific germline B cell receptors. *Science* **340**, 711–716 (2013).
34. Zhou, T. *et al.* Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* **39**, 245–258 (2013).
35. Busse, C.E., Czogiel, I., Braun, P., Arndt, P.F. & Wardemann, H. Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *Eur. J. Immunol.* **44**, 597–603 (2014).
36. Parameswaran, P. *et al.* Convergent antibody signatures in human dengue. *Cell Host Microbe* **13**, 691–700 (2013).
37. Jackson, K.J.L. *et al.* Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host Microbe* **16**, 105–114 (2014).
38. Lavinder, J.J. *et al.* Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc. Natl. Acad. Sci. USA* **111**, 2259–2264 (2014).
39. Wine, Y. *et al.* Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response. *Proc. Natl. Acad. Sci. USA* **110**, 2993–2998 (2013).
40. Boutz, D.R. *et al.* Proteomic identification of monoclonal antibodies from serum. *Anal. Chem.* **86**, 4758–4766 (2014).

ONLINE METHODS

Flow-focusing apparatus. An axisymmetric flow focusing emulsification apparatus was constructed by inserting a 26-gauge needle within 19-gauge hypodermic tubing (Hamilton Company, Reno, NV, USA), and the needle was adjusted so that the needle tip was nearly flush with the end of the hypodermic tubing (**Supplementary Fig. 1**). The concentric needles were placed inside 3/8 inch OD glass tubing (Wale Apparatus, Hellertown, PA, USA) with a 140-µm orifice such that the needle exit was approximately 2 mm from the nozzle orifice. A syringe pump (KD Scientific Legato 200, Holliston, MA, USA) was used to control aqueous flow rates, and a gear pump (M-50, Valco Instruments, Houston, TX, USA) was used to control oil flow rates. The flow focusing nozzle and supply lines were cleaned using 70% EtOH followed by PBS before all experiments.

Single-cell emulsification, cell viability and lysis analyses. MOPC-21 cells were resuspended at a concentration of 500,000 cells/mL in PBS, and the resulting cell solution was injected through the needle at 500 µL/min. A PBS/0.4% trypan blue solution (Sigma-Aldrich, St. Louis, MO, USA) was injected through the 19-gauge hypodermic tubing at 500 µL/min and oil phase (molecular biology grade mineral oil with 4.5% Span-80, 0.4% Tween 80, 0.05% Triton X-100, v/v%, Sigma Aldrich Corp.) passed through the glass tubing at 3 mL/min. The resulting emulsion was analyzed via light microscopy (**Supplementary Fig. 2**); ImageJ post-processing was used to measure droplet diameters. Next, the PBS/0.4% trypan blue solution was replaced with a solution of poly(dT) magnetic beads (1.0 µm diameter, New England Biosciences, Ipswich, MA, USA) pelleted and resuspended in cell lysis/binding buffer (100 mM Tris pH 7.5, 500 mM LiCl, 10 mM EDTA, 1% lithium dodecyl sulfate, 5 mM DTT) at a concentration of 45 µL magnetic bead stock/mL lysis/binding buffer to verify cell lysis. Upon emulsification no cell membranes could be observed, indicating total cell lysis (data not shown).

VH-VL pairing of technical replicates of expanded memory B cells. PBMCs were isolated from donated human whole blood after informed consent had been obtained (Gulf Coast Regional Blood Center, Houston, TX), and non-B cells were depleted via magnetic bead sorting (Miltenyi Biotec, Auburn, CA). B cells were stained with anti-CD20-FITC (clone 2H7, 1:12.5 dilution, BD Biosciences, Franklin Lakes, NJ, USA), anti-CD3-PerCP (HIT3a, 1:50, BioLegend, San Diego, CA, USA), anti-CD19-v450 (HIB19, 1:25, BD), and anti-CD27-APC (M-T271, 1:12.5, BD). $CD3^{-}CD19^{+}CD20^{+}CD27^{+}$ memory B cells were incubated four days in the presence of RPMI-1640 supplemented with 10% FBS, 1 × GlutaMAX, 1 × non-essential amino acids, 1 × sodium pyruvate and 1 × penicillin/streptomycin (Life Technologies) along with 10 µg/mL anti-CD40 antibody (5C3, BioLegend), 1 µg/mL CpG ODN 2006 (Invivogen, San Diego, CA, USA), 100 units/mL IL-4, 100 units/mL IL-10, and 50 ng/mL IL-21 (ref. 41) (PeproTech, Rocky Hill, NJ, USA, **Supplementary Table 3**). Memory B cells were then resuspended in PBS at a concentration of 100k/mL and passed through the innermost, 26-gauge needle of the flow focusing device at 500 µL/min. poly(dT) magnetic beads (1.0 µm diameter, New England Biosciences, Ipswich, MA, USA) were pelleted and resuspended in cell lysis/binding buffer (100 mM Tris pH 7.5, 500 mM LiCl, 10 mM EDTA, 1% lithium dodecyl sulfate, 5 mM DTT) at a concentration of 45 µL magnetic bead stock/mL lysis/binding buffer. The cell lysis/beads mixture was passed through the 19-gauge hypodermic tubing at 500 µL/min while oil phase (molecular biology grade mineral oil with 4.5% Span-80, 0.4% Tween 80, 0.05% Triton X-100, v/v%, Sigma Aldrich Corp.) was passed through the outermost glass tubing at 3 mL/min. The emulsified stream was collected into a series of 2-mL Eppendorf tubes, and each tube was maintained at room temperature for 3 min before being placed on ice for a minimum of twenty minutes. Tubes were centrifuged at 16,000*g* for 5 min at 4 °C, and the upper mineral oil layer was discarded. 200 µL cold water-saturated diethyl ether was added to break the emulsion in each tube, and the tubes were centrifuged again at 16,000*g* for 5 min at 4 °C to pellet the beads; for larger emulsion volumes (donor 1) emulsion collected in Eppendorf tubes was pooled into a 50-mL conical for centrifugation at 5,000 rpm for 7 min at 4 °C, and the emulsion was broken with an equal volume of cold diethyl ether after removal of the upper mineral oil layer. Magnetic beads were withdrawn using a pipette, pelleted, washed once in cold wash buffer (100 mM Tris pH 7.5, 500 mM LiCl, 1 mM EDTA) and then resuspended in 2 mL cold lysis/binding buffer (100 mM Tris pH 7.5, 500 mM LiCl, 10 mM EDTA, 1% LiDS, 5 mM DTT). Beads were washed and resuspended in OE RT-PCR mixture as in previous reports[4]. The OE RT-PCR mixture bead suspension was emulsified and thermally cycled, cDNA was extracted, and a nested PCR was performed as reported previously. Nested PCR product was electrophoresed to purify ~850 bp linked transcripts and sequenced via Illumina 2 × 250 bp sequencing. VH-VL analysis for donor 3 was performed using the same procedure with 0.5% LiDS in cell lysis buffer (as opposed to 1% LiDS as above), and leader peptide primers were used to test VH-VL pairing precision with a different human primer set (**Supplementary Table 4**)[42]. Analysis with a known ARH-77 VH-VL control cell spike was also performed using sorted and *in vitro* expanded memory B cells from one human donor, as described above. This study was approved by the University of Texas at Austin Institutional Review Board (2012-08-0031) and Institutional Biosafety Committee (2010-06-0084), and blood was drawn after informed consent had been obtained. Cells were expanded *in vitro* as described above and ~260 ARH-77 immortalized cells were spiked into a sample of ~20,000 expanded memory B cells in 3 mL PBS; single-cell VH-VL analysis was then performed as above using 1% LiDS and Framework 1 primers.

Pacific Biosciences single-molecule real-time sequencing was performed with VH-VL amplicons derived from week 119 $CD27^{+}$ peripheral B cells of donor CAP256 (ref. 7) using Framework 1 region primers (**Supplementary Data Set 5**). HEK293 cells were transfected with known antibodies as reported previously[4] (**Supplementary Table 1**, antibody sequences reported in **Supplementary Data Set 6**).

Bioinformatic analysis. Raw Illumina sequences were quality-filtered, mapped to V, D and J genes and CDR3s extracted using the International Immunogenetics Information System (IMGT)[43]. Sequence data were filtered for in-frame V(D)J junctions, and productive VH and $V_{\kappa,\lambda}$ sequences were paired by Illumina read ID and compiled by exact CDR3 nucleotide and V(D)J gene usage match. CDR-H3 nucleotide sequences were extracted and clustered to 96% nt identity with terminal gaps ignored (USEARCH v5.2.32 (ref. 44)), and resulting VH-VL pairs with ≥2 reads comprised the list of VH-VL clusters for each data set. To determine the complete VH-VL sequence of the entire approx. 850 nt amplicon by Illumina Miseq 2x250 or 2x300, the separate VH and VL amplicons were sequenced as separate samples, and CDR3 junction regions were used as anchors for consensus sequences of each VH-VL pair[4].

Pacific Biosciences sequencing data were pre-processed with the PacBio SMRT Analysis suite followed by IMGT analysis of paired VH and VL regions. Complete VRC26-class antibody variable region sequences were identified from PacBio circular consensus sequences via BLAST search (**Supplementary Figs. 6** and **7** and **Supplementary Data Set 5**). Allelically included sequences for donors 1, 2, and 3 were detected using an iterative loop that identified heavy chain clusters paired with multiple light chains which were encoded by distinct Vκλ and Jκλ genes, and pairings were cross-confirmed in both replicates.

Randomization was performed by random number generation and statistical comparisons performed using a single-factor analysis of variance followed by a Tukey's HSD *post hoc* test with two-tailed *P* values (IBM SPSS v.20, **Supplementary Fig. 8**).

Precision, *P*, (also called positive predictive value, *PPV*) is calculated from the number of true positives (*TP*) and false positives (*FP*)[45]:

$$P = \frac{TP}{TP + FP}$$

True VH-VL pairs cannot be known for an entire human antibody repertoire, but we can approximate true positives and false positives with matched versus mismatched VH-VL pairs in technical replicates. We assume that a VH paired with the same VL in both replicates is a true positive in replicates 1 and 2 (TP_{1and2}); conversely, we assume any VH paired with a different VL in the two replicates is a false positive in at least one replicate (FP_{1or2}). Thus, the aggregate precision of two independently analyzed replicates 1 and 2 is as follows:

$$P_{1and2} = \frac{TP_{1and2}}{(TP_{1and2} + FP_{1or2})}$$

doi:10.1038/nm.3743

Probability theory dictates that the joint probability of independent events is equal to the product of the independent event probabilities; additionally, pairing precision of technical replicates 1 and 2 (P_1 and P_2, respectively) are assumed to be equal as a property of technical replicates.

$$P_{1and2} = P_1 \times P_2 = P^2$$

We can combine the previous two equations and solve for P, the VH-VL pairing precision of a single analysis.

$$P_{1and2} = P^2 \frac{TP_{1and2}}{(TP_{1and2} + FP_{1or2})}$$

$$P = \sqrt{\frac{TP_{1and2}}{TP_{1and2} + FP_{1or2}}}$$

For a hypothetical experiment: if 950 VH sequences were observed with matching VL in both groups ($TP_{1and2} = 950$) while 50 VH sequences displayed disparate VL in the two groups, VH-VL pairing precision P is estimated by $P = (950/1{,}000)^{0.5} = 97.5\%$.

41. Recher, M. *et al.* IL-21 is the primary common γ chain-binding cytokine required for human B-cell differentiation *in vivo*. *Blood* **118**, 6824–6835 (2011).
42. Wardemann, H. *et al.* Predominant autoantibody production by early human B cell precursors. *Science* **301**, 1374–1377 (2003).
43. Brochet, X., Lefranc, M.-P. & Giudicelli, V. 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–W508 (2008).
44. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
45. Saha, S. & Raghava, G.P.S. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.* **34**, W202–W209 (2006).