High-quality full-length immunoglobulin profiling with unique molecular barcoding

M A Turchaninova[1–4], A Davydov[2,4], O V Britanova[1,3,4], M Shugay[1,3,4], V Bikos[2,4], E S Egorov[1–3], V I Kirgizova[1], E M Merzlyak[1], D B Staroverov[1,3], D A Bolotin[1,3], I Z Mamedov[1,2], M Izraelson[1–3], M D Logacheva[3], O Kladova[3], K Plevova[2], S Pospisilova[2] & D M Chudakov[1–3]

[1]Shemiakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Science, Moscow, Russia. [2]Center of Molecular Medicine, CEITEC, Masaryk University, Brno, Czech Republic. [3]Pirogov Russian National Research Medical University, Moscow, Russia. [4]These authors contributed equally to this work. Correspondence should be addressed to D.M.C. (chudakovdm@mail.ru).

Published online 4 August 2016; doi:10.1038/nprot.2016.093

High-throughput sequencing analysis of hypermutating immunoglobulin (IG) repertoires remains a challenging task. Here we present a robust protocol for the full-length profiling of human and mouse IG repertoires. This protocol uses unique molecular identifiers (UMIs) introduced in the course of cDNA synthesis to control bottlenecks and to eliminate PCR and sequencing errors. Using asymmetric 400+100-nt paired-end Illumina sequencing and UMI-based assembly with the new version of the MIGEC software, the protocol allows up to 750-nt lengths to be sequenced in an almost error-free manner. This sequencing approach should also be applicable to various tasks beyond immune repertoire studies. In IG profiling, the achieved length of high-quality sequence covers the variable region of even the longest chains, along with the fragment of a constant region carrying information on the antibody isotype. The whole protocol, including preparation of cells and libraries, sequencing and data analysis, takes 5 to 6 d.

© 2016 Nature America, Inc. All rights reserved.

INTRODUCTION

High-throughput sequencing (HTS) enables thorough investigation of the diverse immune receptors that determine the specificity of T- and B-cell adaptive responses. Over the past few years, advanced methods have been developed for T-cell receptor (TCR)[1–4] and IG[5–9] HTS profiling. Application of these methods in basic and biomedical studies has become routine in immunology, and a multitude of works have already used HTS profiling of IG repertoires[5,6,10–16].

However, the more thoroughly we want to analyze immune repertoires, the more complicated it becomes to quantify low-frequency clonal variants and to distinguish true homologous variants from a plethora of accumulated PCR and sequencing errors. It becomes challenging to accurately characterize B- and T-cell functional subsets, and it raises the questions of appropriate data normalization and of the influence of artificial error-based diversity on repertoire analysis. Reliable profiling of IG repertoires is especially difficult, as the initial recombinatorial diversity is further increased by the process of hypermutation, which introduces nucleotide changes along the whole length of the variable region. Because of this complication, deep and error-free HTS analysis of full-length IG repertoires remained practically impossible until recently.

The situation changed with the introduction of unique molecular identifiers (UMIs), which are also called unique molecular barcodes[17,18]; the use of UMIs in adaptive immunity profiling has greatly improved quantification and allowed nearly error-free analysis of immune receptor repertoires[5,19–23].

Here we report a detailed UMI-based protocol for the full-length HTS profiling of IG heavy-chain (IGH) or IG light-chain (IGL) repertoires that improves sequencing quality and assists in error correction, including a step-by-step guide to the analysis of obtained sequencing data[20,24].

A

The protocol can be used for a wide range of basic and applied tasks that require deep full-length antibody profiling—e.g., the study of the mechanisms that underlie the development and shaping of naive and activated B-cell repertoires[25–27]; deduction of age-related trends in healthy antibody repertoires[28,29]; monitoring of immune responses during autoimmune[30,31] and infectious diseases[14,32], vaccination courses[6,33] and cancer[34]; identification and development of antigen-specific monoclonal antibodies[35,36]; and the development of optimal vaccination strategies[36,37].

In addition, the UMI-guided asymmetric 400+100-nt paired-end sequencing method described within the framework of the current protocol can be used in a wide range of tasks that require high-throughput sequencing of long (up to 750 bases) amplicons.

There are several limitations of the current protocol that should be considered to ensure robust and unbiased profiling of B-cell repertoires.

The first is inherent to all cDNA-based amplicon sequencing methods and arises from substantial differences in target gene (immunoglobulin) mRNA expression levels between functional subsets of B cells. Therefore, using unsorted bulk B cells as the mRNA source makes it challenging to deduce actual B-cell clonotype frequencies. We propose several ways to minimize the bias caused by the subset-to-subset and cell-to-cell variability of IG mRNA expression levels (Experimental design section). Using genomic DNA as a template can avoid biases caused by differential expression of IG mRNA and may be preferable for those experimental tasks in which quantification of clonal abundance is important. Thus, introducing UMIs at the genomic DNA (gDNA) level[17] would be a beneficial solution, given that sufficient efficiency is achieved in terms of successfully labeled and amplified IG gDNA molecules. However, although it should be technically possible, to date, no efficient technique for incorporating UMIs into IG gDNA—starting from several thousand plasma cells—has been published. We are working on such a protocol, but it is not yet ready. cDNA- and gDNA-based approaches could also

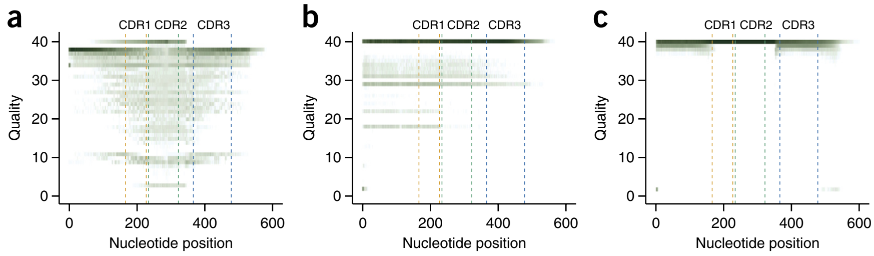

Figure 1 | Base quality score heatmaps across the template length for three data pre-processing strategies. (**a**–**c**) Sequencing quality of raw paired-end Illumina MiSeq reads as compared with that of UMI-based assembly of reads for 350+300-nt paired-end sequencing and asymmetric 400+100-nt paired-end sequencing strategies. (**a**) Sequencing-quality Phred scores are plotted for the merged raw paired-end reads obtained using the 350+300-nt paired-end sequencing strategy for 60,000 IGH sequencing reads. Human IGH libraries were prepared according to the present protocol. Approximate positioning of CDR regions is shown with dashed lines. (**b**) Consensus-quality scores are plotted following UMI-based assembly of merged raw paired-end reads, with a threshold of 5 reads per molecular barcode, obtained using the 350+300-nt paired-end sequencing strategy. (**c**) Consensus-quality scores are plotted following UMI-based assembly of merged paired-end reads performed for only the first mate (400-nt reads), with a threshold of 5 reads per molecular barcode, obtained using the 400+100-nt asymmetric paired-end sequencing strategy (see **Fig. 2** for the asymmetric sequencing scheme logic). Note that in the case of raw read pairing, the resulting quality is low in the middle of the IGH sequence, where CDR1 and CDR2 are located (**a**). By contrast, consensus-quality scores are stably higher across the length of the IGH sequence after UMI-based assembly (**b**) and are almost maximal after asymmetric assembly of the whole sequencing length (**c**). Relative base-pair density is shown with color: from white (low) to dark green (high).

© 2016 Nature America, Inc. All rights reserved.

be combined in order to obtain maximum information from a sample of interest. Both gDNA- and cDNA-based UMI-labeled library preparation methods should be compatible with the paired-end sequencing and data analysis proposed in the current protocol (starting from Step 18).

The second limitation is related to the error correction capabilities of the present protocol. UMI-based data analysis is able to eliminate nearly all errors and ambiguities, and it provides a highly accurate measurement of the clonal and intraclonal diversity of B cells. Yet it is worth mentioning that errors introduced in the reverse transcription step that occur before the incorporation of molecular identifiers escape UMI-based error correction. Detailed validation of error correction efficiency is presented below (Experimental design section).

Third, the general limitation of the UMI-based approach is the requirement to reach sufficient coverage of UMI-labeled cDNA or gDNA templates, providing high accuracy of clonotype identification at the cost of profiling depth. At least 4–5 reads per UMI-labeled template molecule is the requirement for efficient error correction (Experimental design section). In this respect, the use of the Illumina HiSeq2500 500-nt kit available for the rapid run mode, as well as future development of cheaper long-read sequencing options, could provide both desirable depth and quality for full-length IG profiling.

A

. Analysis of the full-length IGH or IGL cDNA variable region from the 5′UTR to the J gene segment (plus ~40 nt more in order to preserve information on the antibody isotype that is encoded in the C gene segment) requires a sequencing length of ~600–640 nt. This length is almost feasible with paired-end sequencing using the 600-nt MiSeq Illumina kit, which has enough reagents for up to 650 cycles (it is possible to increase the read length by 50 nt to achieve paired reads overlap). However, the quality of Illumina sequencing rapidly decreases along the read length, which makes it challenging to obtain reliable sequence information in the middle of the variable segment (. **1**). Grouping multiple sequencing reads from the same cDNA or gDNA starting molecule (i.e., carrying the same UMI) results in a high-quality consensus read, which can be used to obtain high-quality sequence throughout the whole IGH variable region (. **1**). For example, if the Illumina quality report states that there is a 1% chance that 'A' at position X is erroneous, but there are four such reads covering the same original starting molecule, the resulting error probability is only 0.000001%. Therefore, the reliability of each nucleotide position drastically increases.

Furthermore, our paired-end Illumina sequencing strategy with UMIs allows up to 750-nt-length fragments to be sequenced in an almost error-free manner (. **1** and **2**). To this end, the 500- or 600-nt kit is used for asymmetric 400+100-nt paired-end sequencing. Because of the 50/50 orientation of the PCR product after Illumina adaptor ligation, ~50% of the first 400-nt-length reads cover UMI-barcoded cDNA molecules from the 5′ end, whereas the other ~50% of the first reads cover the same molecules from the 3′ end. In this scenario, the second paired-end 100-nt read is used only to extract the UMI sequence. UMI-based assembly of the first reads results in high-quality 400-nt-length sequences covering the 3′- and the 5′-ends of the cDNA, separately. In the next analysis step, 400-nt 3′- and 5′-end sequences carrying the same UMIs are merged, resulting in high-quality 400+400-nt paired-end sequences of original cDNA molecules. This approach provides sufficient sequencing length to capture even extra-long[38,39] IG variants. Note that nonstrand-specific Illumina library preparation is a prerequisite for validity of the asymmetric data analysis strategy (. **2**). This means that Illumina adaptors should be either ligated or introduced during the course of PCR amplification as appropriate—e.g., using the primers introducing both A and B Illumina adaptors randomly from both sides of the library, resulting in both A–B and B–A orientations.

. PCR errors are abundantly accumulated in the course of library preparation. At the same time, true hypermutated IG subvariants with one or several natural mismatches may be present at low concentrations, being indistinguishable from artificial variants arising as a result of PCR errors in a major clonotype. Both somatic hypermutation and accumulation of PCR and sequencing errors are, from a statistical point of view, branching processes. It is extremely difficult to distinguish between these factors when analyzing a group of homologous IG variants. For these reasons, frequency-based error correction that is generally efficient in TCR profiling[40–42] (although it may also result in loss of homologous TCR variants resulting from convergent recombination) is prone to losing

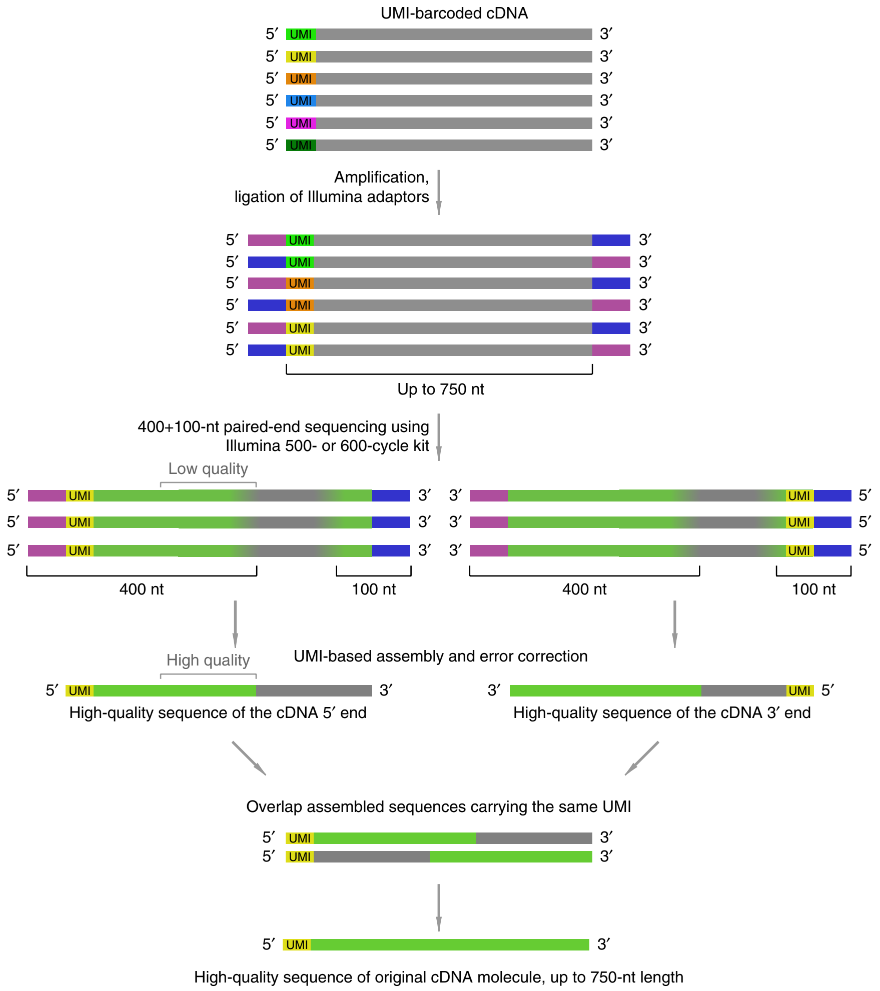
UMI-barcoded cDNA
5′
UMI
3′
Amplification,
ligation of Illumina adaptors
Up to 750 nt
400+100-nt paired-end sequencing using
Illumina 500- or 600-cycle kit
Low quality
400 nt
100 nt
High quality
UMI-based assembly and error correction
High-quality sequence of the cDNA 5′ end
High-quality sequence of the cDNA 3′ end
Overlap assembled sequences carrying the same UMI
High-quality sequence of original cDNA molecule, up to 750-nt length

Figure 2 | Logic of the asymmetric paired-end sequencing strategy and data analysis with UMIs. The asymmetric 400+100-nt-length paired-end sequencing strategy (Step 22B) with UMIs allows high-quality throughout sequencing of up to 750-nt-length fragments. It requires setting appropriate software parameters for data analysis, as indicated in Steps 25 and 26.

hypermutated subvariants of interest in sequencing of IG repertoires[20].

However, most of the errors occur during the later PCR rounds, when many molecules are being amplified. Such errors are therefore present in minor percentages of reads that cover the same starting DNA or cDNA molecule. UMI information allows reads to be grouped into molecular identifier groups (MIGs) and eliminates most PCR errors accumulated during library preparation via consensus assembly[17,20,22,23,43]. Following the same logic, sequencing errors are also efficiently corrected within MIGs. Importantly, even low-frequency homologous subvariants are preserved in such analysis, which is not possible using frequency-based error correction[20].

It should be noted, however, that some low-frequency erroneous subvariants survive UMI-based correction. To estimate the uncorrected error rate, we analyzed two control data sets. The first contained four samples of sequencing reads covering the CDR3 region of the IGH gene from the EHEB B cell line (data from Shugay *et al.*[20]). The second contained seven chronic lymphocytic leukemia (CLL) clonotypes used as spike-in controls, with sequencing reads spanning the full length of the IGH sequence of the CLL clone. For these two data sets, overall error rates after UMI-based correction were similar and constituted ~1 per 10,000 bp × UMI (**1**). Note that typical long-read MiSeq runs have relatively poor quality, with Phred scores of 20–25 or lower, leading to error rates of 1 per 100–300 bp × read or higher (along with the earlier accumulated PCR errors).

In the deep-sequencing experiment on EHEB CDR3, the error rate after UMI-based correction resulted in the ratio of a true clonotype sequence to the largest erroneous subvariant of ~4,000:1 (**1**). In the full-length IGH profiling of human plasma cells, the largest clonotypes were typically represented by up to 400 UMI-labeled cDNA molecules (MIGs) in our experiments (**1**). Correspondingly, for clonotypes of this size, the remaining uncorrected errors are few, and they should be mostly represented by one, or occasionally two, MIGs (i.e., by one or two uniquely labeled cDNA molecules), as confirmed by the examples of full-length profiling of control CLL clones (**2**).

The uncorrected errors may come from two sources: errors during cDNA synthesis[44] and errors during the early PCR cycles that happen to dominate within the MIGs[20]. As the UMI is introduced during cDNA synthesis, errors that occur during the synthesis itself cannot be corrected by UMIs. However, cDNA synthesis errors, as well as PCR errors during the first cycles, are relatively rare, as they happen at the stage with a minimal number of synthesized molecules. Therefore, exclusion of 'singletons'—an MIG consensus that is not confirmed by identical independent MIG consensus—efficiently filters out most remaining errors (shown in gray in **2**) and should be considered sufficient for medium- and small-sized clonotypes represented by up to a hundred UMI-labeled cDNA molecules in a sample. For large clonotypes represented by several hundred cDNA molecules, subvariants represented by only 2–3 UMI-labeled cDNA molecules (MIGs) should also be considered unreliable in typical analysis performed with the current protocol.

It should be noted here that such filtering can be deleterious for particular questions of interest. Analysis of immune repertoires for the naive B cells mostly produces cDNA 'singletons' (**1**), as the efficiency of the template-switch method is about or less than 1 molecule per naive B lymphocyte (our current estimations, see below), and naive clones are extremely small, with each usually being represented by one cell in a sample. Filtering cDNA singleton events would essentially deplete the observed diversity of naive B cells. In such cases, remaining rare errors may be overlooked to enable the general features of the naive repertoire to be analyzed.

. Between a starting sample of cells and analyzed sequencing reads, researchers

© 2016 Nature America, Inc. All rights reserved.

TABLE 1 | Error rates observed in model data sets and share of sequencing artifacts.

Data set description	Overall error rate, 1/bp/UMI	Artificial diversity, 1/bp/UMI	Top error ratio	Share of noncoding molecules, 1/bp/UMI	Share of noncanonical molecules, 1/UMI
EHEB spike-in control, deep CDR3 sequencing[20], *n* = 4	9.0 ± 0.3 × 10^{-5}	1.8 ± 0.1 × 10^{-5}	2.6 ± 0.2 × 10^{-4}	2.7 ± 0.1 × 10^{-5}	2.8 ± 0.1 × 10^{-4}
Healthy PBMC IGH, CDR3 sequencing[20], *n* = 4	NA	NA	NA	1.7 ± 0.2 × 10^{-4}	7.3 ± 0.3 × 10^{-2}
Control set of CLL clonotypes, full-length IGH sequencing (A.D., M.S., V.B., K.P., D.M.C., data not shown), *n* = 7[a]	1.0 ± 0.3 × 10^{-4}	8.0 ± 3.0 × 10^{-5}	4 ± 2 × 10^{-3}[b]	1.7 ± 1.0 × 10^{-5}	<10^{-3}[c]
Plasma cell samples, full-length IGH sequencing (M.A.T., A.D., O.V.B., M.S., D.M.C., data not shown), *n* = 2[d]	NA	NA	NA	5.3 ± 2.8 × 10^{-5}	1.6 ± 0.8 × 10^{-2}

Overall error rate was estimated as total abundance of erroneous molecules divided per base pair of parent variant sequence per UMI tag. Artificial diversity was estimated as total number of distinct erroneous variants per base pair per UMI tag. Top error ratio is the ratio between abundances of the most abundant distinct erroneous variant and its parent variant. This (the upper bound on the signal-to-noise ratio) is a highly relevant metric characterizing the extent of our ability to filter erroneous variants, as it directly shows the extent of uncorrected errors. Using estimates present in this table, for a typical experimental setup with 20,000 cDNA molecules (UMI tags) and an average IGH sequence length of 400 nt (excluding the conserved 'C' segment), we expect to observe approximately 800 erroneous cDNA molecules, each error mostly represented by a single cDNA. Noncoding molecules are molecules that carry stop codons or frameshifts. Noncanonical molecules are molecules with the first or last amino acid of the CDR3 region differing from the conserved Cys or Trp residue. These two types of errors result in some artifactual IGH sequences: notably, 30.1 ± 0.7% of errors result in noncoding variants and 6.6 ± 0.5% of errors result in noncanonical variants with changed conserved Cys and Trp residues, as seen from the control experiment with the EHEB cell line. However, the rate of noncoding and noncanonical variants was lower in EHEB and CLL controls as compared with PBMC and plasma cell samples, respectively, indicating that B cells carry a portion of noncoding IGH mRNA molecules that survive nonsense-mediated mRNA decay. Mean ± s.d. values across samples are reported. NA, not applicable.

[a] **Supplementary Table 2**. [b] Note that the top error estimate here is subject to bias due to relatively low depth, as the size of the most abundant distinct erroneous variant cannot be lower than 1 (single MIG), which can be illustrated as follows: consider 100 singleton errors per 10,000 UMIs that result in a 1/1,000 top error rate; randomly sampling 1,000 UMIs will result in a 1/100 top error rate. [c] A single noncanonical variant with a frequency of 7.4 × 10^{-4} was found in one of seven samples. [d] **Supplementary Table 1**.

© 2016 Nature America, Inc. All rights reserved.

have essentially no ability to track the fate of the incoming nucleotide molecules. How many of them were lost during RNA/DNA purification? How many of them were successfully amplified? How uniform was this amplification? Amplification with a set of multiplex primers with differing efficiencies[3,41,45–47] and the preference of DNA polymerase for different templates[48] essentially bias the relative abundances of IG variants carrying particular V- and J-gene segments. However, when each starting molecule has been labeled with a unique identifier, researchers can directly count the number of successfully analyzed cDNA or genomic DNA molecules (MIGs) in the sequencing data. Thus, the bottlenecks (in terms of the number of template molecules that have been successfully amplified and sequenced) become traceable, and the number of analyzed molecules can be equalized in order to perform unbiased comparison of samples of interest[17,19].

and secrete large amounts of antibodies—IG mRNA expression levels in plasma B cells may exceed those in naive or memory B cells by more than 30-fold[52]. In peripheral blood, plasma B cells may constitute 1–5% of all B cells[53,54]. In RNA-based analysis of IG repertoire for a total peripheral blood mononuclear cell (PBMC) blood sample that includes 100,000 B cells, several plasma B cells can be confused with a clonal expansion that represents 1% of all B cells (in terms of the number of template IG cDNA molecules) because of high IG mRNA expression levels in plasma B cells. This difference in IG mRNA expression levels is clearly visible in comparative analysis of total PBMCs and purified memory B-cell biological replicates of the same individual (**. 3**), and it would lead to erroneous interpretation of clonal composition and dynamics of B-cell clones in peripheral blood. Moreover, e.g., in cerebral spinal fluid, plasma B cells may constitute ~30% of all

- -

A (1 2). As mentioned above, an important issue in RNA-based IG repertoire analysis is the difference in immunoglobulin mRNA expression levels between functional subtypes of B cells. In contrast to T cells, for which TCR mRNA expression levels were reported to be nearly identical among all quiescent T-cell subsets (naive, memory)[49,50] and to diverge approximately plus or minus twofold from clone to clone in effector T-cell subsets[49,51], in the case of B cells, researchers have to take into account the potential presence of plasma B cells in a sample of interest. Due to their function—to produce

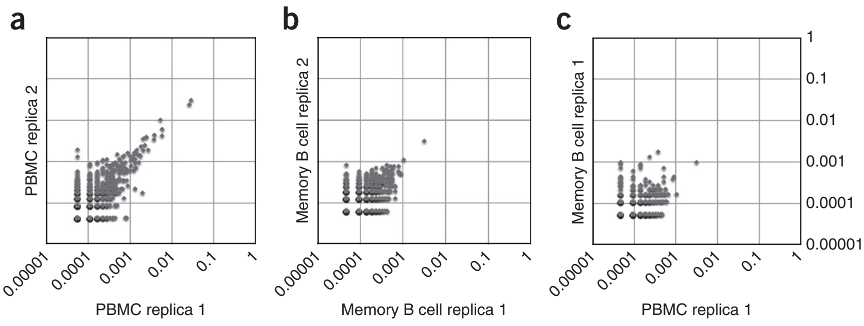

Figure 3 | Replicate sample analysis. Clonotype cDNA frequencies in IGH CDR3 profiling for replicate samples of memory B cells purified using magnetic separation (Memory B Cell Isolation Kit, Miltenyi), and PBMCs from the same individual. Each replicate included ~200,000 B cells. Each dot represents the frequency of an IGH clonotype shared between the two samples in each of these samples. (**a**) Comparison of two PMBC replicas. (**b**) Comparison of two memory B-cell replicas. (**c**) Comparison of one memory B-cell and one PBMC replica. Note a portion of seemingly large clones (frequency above 0.001) in PBMC samples resulting from the high IGH mRNA expression levels in plasma cells.

B cells[55], and RNA-based analysis would therefore mostly yield plasma B-cell but not memory B-cell IG repertoire. Therefore, in each study, it is preferable to focus on a functional B-cell subset of interest (e.g., naive, memory or plasma B cells isolated using FACS sorting or magnetic bead separation) in order to reduce the variance in IG mRNA expression levels.

Cell-to-cell variation in mRNA expression level may be pronounced between individual B cells even within the same functional subset. Therefore, it is beneficial to prepare each library from B cells taken from two replicate samples (**. 3**). Such libraries prepared from two or more independent samples can be used to estimate the variance in clonotype frequencies resulting from stochastic cell sampling and mRNA expression. Note that RNA[5]- or cDNA[56]-based replicates control only for the accuracy of library preparation method, not for the cell-sampling- and mRNA-expression-level biases. The most efficient way to improve clonal quantification is to use more cells per sample for RNA purification and cDNA synthesis. Importantly, in the latter case only a portion of obtained cDNA should be used in further library preparation in order to achieve sufficient oversequencing per cDNA molecule (see below).

(**3 20).** Here we report an IG library preparation protocol based on a RACE (Rapid Amplification of cDNA Ends) approach with template-switch effect[57]. This approach was first used for TCR sequencing by

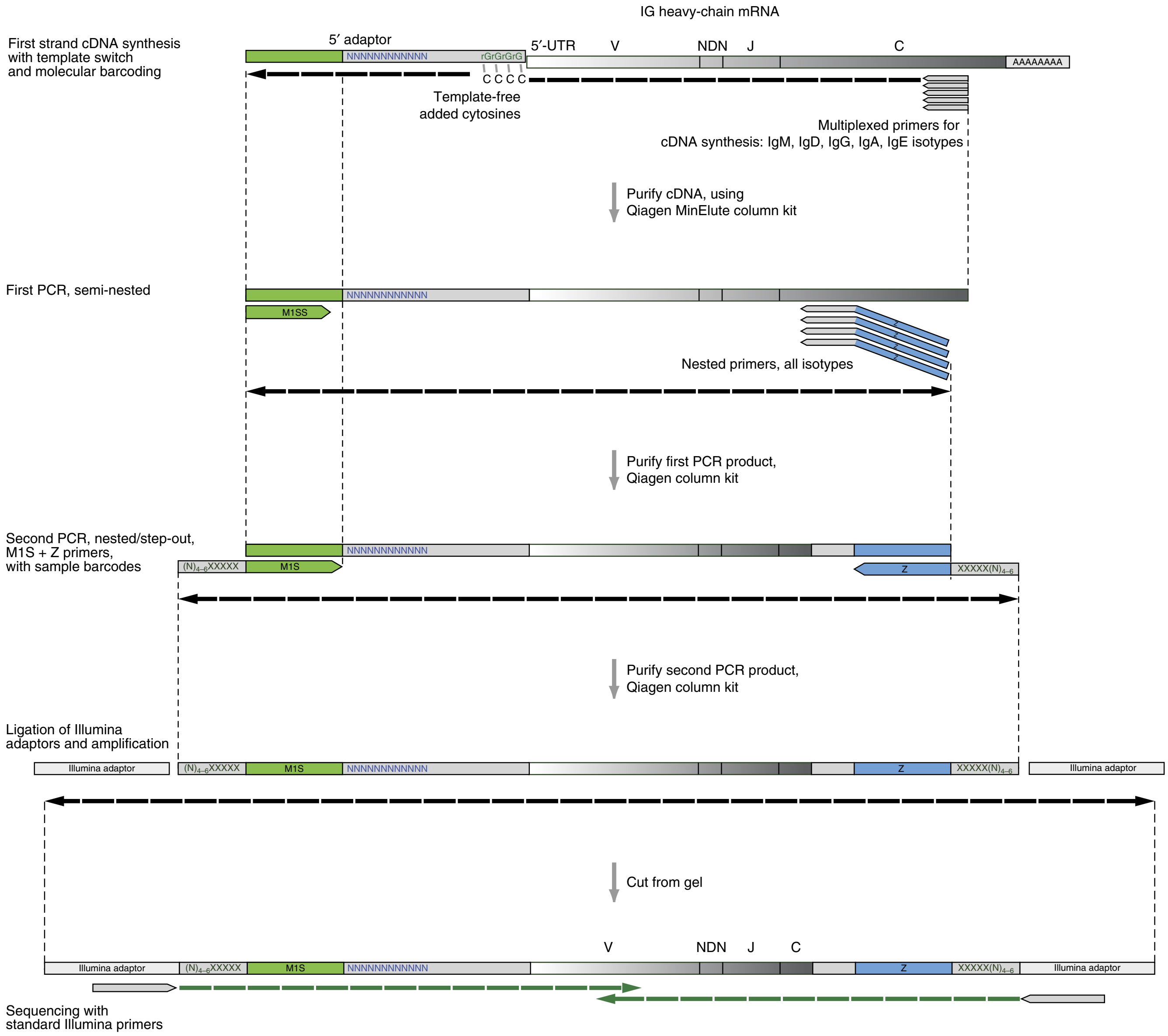
IG heavy-chain mRNA
First strand cDNA synthesis with template switch and molecular barcoding
5′ adaptor
NNNNNNNNNNNN
rGrGrGrG
5′-UTR
V
NDN
J
C
AAAAAAAA
CCCC
Template-free added cytosines
Multiplexed primers for cDNA synthesis: IgM, IgD, IgG, IgA, IgE isotypes
Purify cDNA, using Qiagen MinElute column kit
First PCR, semi-nested
NNNNNNNNNNNN
M1SS
Nested primers, all isotypes
Purify first PCR product, Qiagen column kit
Second PCR, nested/step-out, M1S + Z primers, with sample barcodes
NNNNNNNNNNNN
(N)4–6XXXXX
M1S
Z
XXXXX(N)4–6
Purify second PCR product, Qiagen column kit
Ligation of Illumina adaptors and amplification
Illumina adaptor
(N)4–6XXXXX
M1S
NNNNNNNNNNNN
Z
XXXXX(N)4–6
Illumina adaptor
Cut from gel
V
NDN
J
C
Illumina adaptor
(N)4–6XXXXX
M1S
NNNNNNNNNNNN
Z
XXXXX(N)4–6
Illumina adaptor
Sequencing with standard Illumina primers

Figure 4 | Scheme of the cDNA library preparation and sequencing (Steps 6–22). NNNNNNNNNNNN, unique molecular identifier; UMI; XXXXX, sample barcode; $(N)_{4–6}$, from 4 to 6 random nucleotides that are introduced at the ends of the library in order to (1) protect the sample barcode from damage and (2) generate diversity at the very beginning of sequencing reads for better cluster identification on the Illumina sequencing machine. The sequence corresponding to the annealing part of the M1S primer is shown in green, and the sequence corresponding to the annealing part of the Z primer is shown in blue. The whole variable region is covered by paired-end Illuimna MiSeq sequencing, e.g., 300+300-nt, 350+300-nt, 325+325-nt, 330+280-nt, or 400+100-nt sequencing. In the latter case, asymmetric assembly is used in data analysis, following the logic shown in **Figure 2**. AAAAAAAA, polyA; C, constant segment of antigen receptor; J, joining segment of antigen receptor; NDN, diversity segment of antigen receptor (D) with nucleotides (N) spanning V–D and D–J junctions; rG, riboguanosine; V, variable segment of antigen receptor.

© 2016 Nature America, Inc. All rights reserved.

TABLE 2 | Oligonucleotides.

Primer	Application	Sequence
First-strand cDNA synthesis[a]		
SmartNNNa	5′ – template-switch adaptor. U = dU	AAGCAGUGGTAUCAACGCAGAGUNNNNUNNNNUNNNNU CTT(rG)$_4$
Human IGH cDNA synthesis primer mix		
hIGG_r1	Primer for cDNA synthesis, human IgG heavy-chain mRNA	GAAGTAGTCCTTGACCAGGCA
hIGM_r1	Primer for cDNA synthesis, human IgM heavy-chain mRNA	GTGATGGAGTCGGGAAGGAAG
hIGA_r1	Primer for cDNA synthesis, human IgA heavy-chain mRNA	GCGACGACCACGTTCCCATCT
hIGD_r1	Primer for cDNA synthesis, human IgD heavy-chain mRNA	GGACCACAGGGCTGTTATC
hIGE_r1	Primer for cDNA synthesis, human IgE heavy-chain mRNA	AGTCACGGAGGTGGCATTG
Human IGL cDNA synthesis primer mix		
hIGLC_r1	Primer for cDNA synthesis, human IgL light-chain mRNA	GCTCCCGGGTAGAAGT
hIGKC_r1	Primer for cDNA synthesis, human IgK light-chain mRNA	GCGTTATCCACCTTCC
Mouse IGH cDNA synthesis primer mix		
mIGG12_r1	Primer for cDNA synthesis, mouse IgG1/IgG2 heavy-chain mRNA	KKACAGTCACTGAGCTGCT
mIGG3_r1	Primer for cDNA synthesis, mouse IgG3 heavy-chain mRNA	GTACAGTCACCAAGCTGCT
mIGA_r1	Primer for cDNA synthesis, mouse IgA heavy-chain mRNA	CCAGGTCACATTCATCGTG
mIGM_r1	Primer for cDNA synthesis, mouse IgM heavy-chain mRNA	CTGGATGACTTCAGTGTTGT
mIGD_r1	Primer for cDNA synthesis, mouse IgD heavy-chain mRNA	GCCATTTCTCATTTCAGAGG
mIGE_r1	Primer for cDNA synthesis, mouse IgE heavy-chain mRNA	GTTCACAGTGCTCATGTTC
Mouse IGL cDNA synthesis primer mix		
mIGLC_r1	Primer for cDNA synthesis, mouse IgL light-chain mRNA	TGTACCATYTGCCTTCCAG
mIGKC_r1	Primer for cDNA synthesis, mouse IgK light-chain mRNA	ACTGCCATCAATCTTCCAC
First PCR amplification[b]		
M1SS	Step-out primer 1, anneals on the switch adaptor	AAGCAGTGGTATCAACGCA
Human IGH reverse primer mix		
hIGGE_r2	Nested primer with Z adaptor, human IgG/IgE heavy-chain cDNA	ATTGGGCAGCCCTGATTARGGGGAAGACSGATG
hIGA_r2	Nested primer with Z adaptor, human IgA heavy-chain cDNA	ATTGGGCAGCCCTGATTCAGCGGGAAGACCTTG
hIGM_r2	Nested primer with Z adaptor, human IgM heavy-chain cDNA	ATTGGGCAGCCCTGATTAGGGGGAAAAGGGTTG
hIGD_r2	Nested primer with Z adaptor, human IgD heavy-chain cDNA	ATTGGGCAGCCCTGATTATATGATGGGGAACAC
Human IGL reverse primer mix		
hIGLC_r2	Nested primer with Z adaptor, human IgL light-chain cDNA	ATTGGGCAGCCCTGATTGYGGGAACAGAGTGAC
hIGKC_r2	Nested primer with Z adaptor, human IgK light-chain cDNA	ATTGGGCAGCCCTGATTGATGGTGCAGCCACAG

(continued)

© 2016 Nature America, Inc. All rights reserved.

TABLE 2 | Oligonucleotides (continued).

Primer	Application	Sequence
Mouse IGH reverse primer mix		
MIGG12_r2	Nested primer with Z adaptor, mouse IgG1/IgG2 heavy-chain cDNA	ATTGGGCAGCCCTGATTAGTGGATAGACMGATG
mIGG3_r2	Nested primer with Z adaptor, mouse IgG3 heavy-chain cDNA	ATTGGGCAGCCCTGATTAAGGGATAGACAGATG
mIGA_r2	Nested primer with Z adaptor, mouse IgA heavy-chain cDNA	ATTGGGCAGCCCTGATTTCAGTGGGTAGATGGTG
mIGM_r2	Nested primer with Z adaptor, mouse IgM heavy-chain cDNA	ATTGGGCAGCCCTGATTGGGGGAAGACATTTGG
mIGD_r2	Nested primer with Z adaptor, mouse IgD heavy-chain cDNA	ATTGGGCAGCCCTGATTCTCTGAGAGGAGGAAC
mIGE_r2	Nested primer with Z adaptor, mouse IgE heavy-chain cDNA	ATTGGGCAGCCCTGATTAAGGGGTAGAGCTGAG
Mouse IGL reverse primer mix		
mIGLC_r2	Nested primer with Z adaptor, mouse IgL light-chain cDNA	ATTGGGCAGCCCTGATTAGRGGAAGGTGGAAAC
mIGKC_r2	Nested primer with Z adaptor, mouse IgK light-chain cDNA	ATTGGGCAGCCCTGATTGGATGGTGGGAAGATG
Second PCR amplification		
M1S	Nested primer, anneals on M1SS[c]	$(N)_{4-6}$(XXXXX)CAGTGGTATCAACGCAGAG
Z	Step-out primer[c]	$(N)_{4-6}$(XXXXX)ATTGGGCAGCCCTGATT

[a]Simultaneous first-strand cDNA synthesis of all heavy-chain isotypes (IgA, IgM, IgG, IgD and IgE for human or IgA, IgM, IgG1, IgG2, IgG3, IgE, IgD for mouse) is possible. Simultaneous first-strand cDNA synthesis of all light-chain variants (IgK and IgL for human or for mouse) is possible. [b]Simultaneous amplification of all heavy-chain isotypes (IgA, IgM, IgG, IgD and IgE for human or IgA, IgM, IgG1, IgG2, IgG3, IgE, IgD for mouse) is possible. Simultaneous amplification of all light-chain variants (IgK and IgL for human or for mouse) is possible. [c]$(N)_{4-6}$—random nucleotides introduced at the 5′ ends of the library for better cluster differentiation by the Illumina sequencer. dU, deoxyuridine; XXXXX, sample barcode.

© 2016 Nature America, Inc. All rights reserved.

Douek *et al.*[58]. In 2011, Kivioja *et al.*[18] suggested incorporating a unique molecular identifier (UMI) within the template-switch adaptor; we have successfully used this method in TCR[19,22,59,60] and IG[20] profiling.

The current protocol, which is schematically shown in **4**, was specifically designed for IG repertoires. Briefly, cDNA synthesis starts from a set of C-region-specific primers and, with the template-switch effect, incorporates the universal adaptor at the 5′ end of the library, simultaneously labeling each cDNA molecule with a UMI. The cDNA library is further amplified in a two-stage PCR. The first PCR is multiplex and seminested on the 3′ end, approaching the J segment but preserving ~20 nt of the C segment, which allows information on antibody isotype to be retained. A universal 3′-adaptor is introduced during the first PCR. This avoids the need for multiplexing in the second PCR amplification, thus allowing the reaction to be performed with two sample-barcoded primers and generating IG libraries with sample barcodes on both the 5′ and 3′ ends. Such double-ended barcoding of each sample essentially protects against cross-sample contamination in further analysis. Oligonucleotides optimized for human and mouse IGH and IGL libraries are summarized in **2.** Illumina adaptors can be ligated separately to each library (in our experience, this is the best strategy to protect against cross-sample contamination) or to pooled multiplex PCR products (which is faster and cheaper, yet protects against cross-sample contamination with almost the same efficiency as ligating adaptors separately to each library).

. The prerequisite for efficient UMI-based correction of PCR and sequencing errors, as well as for the general improvement of sequence quality along >>300-nt-long reads (**. 1**), is that each uniquely labeled cDNA molecule is sequenced multiple times. In this respect, the fewer the cDNA molecules analyzed, the higher the resulting sequencing quality. At the same time, the aim of HTS analysis is to look deep into the repertoire complexity—i.e., to analyze more molecules from the starting library. Compromise between these two issues defines rather a narrow window for the optimal number of IG cDNA molecules analyzed per Illumina MiSeq sequencing run.

The number of target IG paired-end sequences that pass all quality filters may vary from 10 to 15 million for a successful MiSeq run. Taking into account desirable 10-fold coverage of each starting molecule and unevenness of the coverage, the optimal number of starting cDNA molecules per MiSeq run is in the range of ~200,000–250,000.

Among B cells, plasma cells show the highest IG expression levels, followed by memory cells with moderate expression, and naive B cells having the lowest IG expression. According to our current estimations, the approximate ratio of IGH mRNA levels for human B cells is 500:5:2 for plasma, memory and naive B cells, respectively. On the basis of FACS cell counts for input samples, we estimate the efficiency of the template-switch RACE protocol for the human IGH libraries as ~150, 1.5 and 0.6 UMI-tagged cDNA molecules per cell for these three subsets, respectively. Therefore, if, for example, six samples of interest are combined on one MiSeq run, we recommend using at least 50,000 naive, 100,000 memory and 5,000 plasma FACS-sorted human B cells per sample. Preferably, all RNA should be used for RNA purification and cDNA synthesis. For 50,000 naive B cells, all cDNA can be used for the library preparation. Importantly, however, only a portion of cDNA equivalent to ~20,000 memory B cells (1/5th of the cDNA obtained from 100,000 cells) or 200 plasma B cells (1/25th of the cDNA obtained from 5,000 cells) should be used.

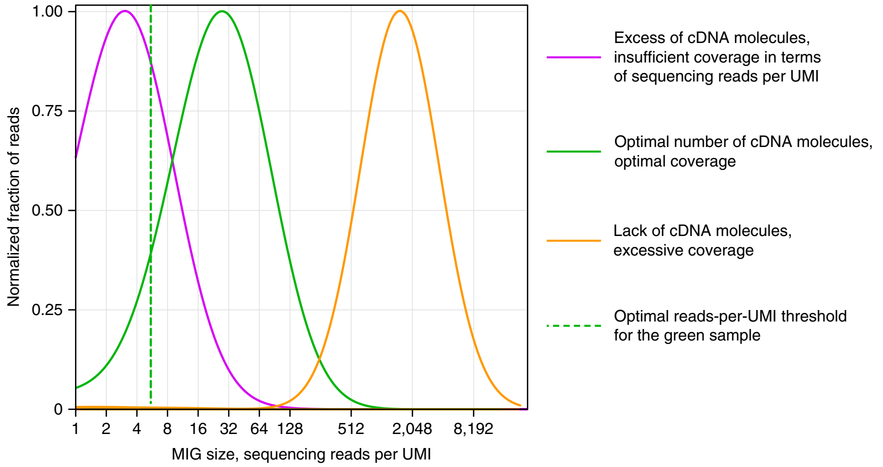

Figure 5 | Required cDNA sequencing coverage. This figure schematically shows examples of UMI coverage distributions. *y* axis: fraction of sequencing reads in all MIGs with a given coverage (*x* axis), normalized to 1. Optimal average sequencing coverage (empirically estimated to be ~30 sequencing reads per UMI[20]) and optimal reads-per-UMI threshold for this scenario are shown in green. The threshold is empirically selected for filtering of UMI tags that have low coverage and/or are erroneous subvariants of high-coverage UMI tags. The threshold is selected by the following formula: 2^(log2(peak position)/2).

Such an approach provides better quantification, as the influence of cell sampling bias is decreased with the increase of sampled cell numbers, and mRNA expression levels are averaged between many cells. At the same time, sufficient coverage in terms of sequencing reads per cDNA molecule is achieved. Appropriate selection of the amount of cDNA used for library preparation is critical for further data analysis, as an excess of starting cDNA molecules will result in insufficient sequence coverage of each cDNA event, whereas an insufficient number of cDNA molecules will lead to superficial IG profiling (.5).

In general, this oversequencing requirement limits the depth and thus the quantitative power of the analysis. Memory B-cell clones are relatively small; the largest clones typically occupy <<0.5% of all memory B cells in peripheral blood (.3). Therefore, in a random sample of 20,000 analyzed cDNA molecules, even a large clone may be represented by only a dozen IG cDNA molecules, which may be sufficient to capture but insufficient to reliably quantify such clones, because of both cell- and cDNA-sampling biases[22]. As a feasible option for performing deeper UMI-based IG sequencing, the rapid run mode of the Illumina HiSeq2500 now allows 400+100-nt paired-end analysis to be performed. Because of the higher number of produced sequencing reads, HiSeq should allow more cells and more cDNA to be used and/or more samples to be analyzed per sequencing run.

MATERIALS

A

- ▲ **CRITICAL** The materials listed here are those needed for the example application of preparing an immunoglobulin library from a freshly isolated cell sample. For library preparation from frozen samples, some materials will need to be added—for example, cell culture media (RPMI-1640, Sigma-Aldrich, cat. no. R8758) and human serum (First Link, cat. no. 20-00-810).
- Blood collection tubes (e.g., BD Plastic Whole Blood tube with spray-coated K$_2$EDTA, cat. no. 367844)
- Ficoll Paque (GE Healthcare, cat. no. 17-1440-02)
- PBS (Sigma-Aldrich, cat. no. D8537)
- Trizol Reagent (Invitrogen, cat. no. 15596-026) or Qiazol Lysis Reagent (Qiagen, cat. no. 79306) ! **CAUTION** Trizol Reagent contains phenol (toxic and corrosive) and guanidine isothiocyanate (an irritant). Trizol Reagent should be handled under a fume hood while wearing a lab coat, gloves and safety glasses.
- Chloroform (Sigma-Aldrich, cat. no. C2432-25ML) ! **CAUTION** Chloroform should be handled under a fume hood.
- Isopropyl alcohol (Sigma-Aldrich, cat. no. I9516-25ML) ! **CAUTION** Isopropyl alcohol is flammable and toxic. Isopropyl alcohol should be handled under a fume hood and only while using personal protective equipment.
- Co-precipitant (e.g., Merck Millipore Pellet Paint, cat. no. 69049)
- RNAse-free water
- Oligonucleotides (e.g., Evrogen JSC, see 2 for sequences) ▲ **CRITICAL** The 5′-template-switch adaptor (SmartNNNa) must be HPLC-purified.
- SMARTscribe reverse transcriptase (Clontech, cat. no. 639536) or Mint reverse transcriptase (Evrogen, cat. no. SK005) ▲ **CRITICAL** Reverse transcriptase with high template switching activity[57] must be used.
- Biological samples that can be used for isolation of B cells: whole-blood samples from human or mouse; tissue samples (e.g. spleen, bone marrow) ! **CAUTION** All human samples must be obtained after written informed consent from the donor and ethical approval from the applicable regulating body. All mouse samples must be obtained in accordance with institutional, local and national laws and standards for animal work.
- dNTPs (Life Technologies, cat. no. R0181)
- RNAse inhibitor (Promega, cat. no. N2111)
- Uracil DNA glycosylase (New England BioLabs, cat. no. M0280S)
- Q5 Hot-start High-Fidelity DNA Polymerase (New England BioLabs, cat. no. M0493S)
- QIAquick PCR Purification Kit (Qiagen, cat. no. 28104)
- MinElute PCR Purification Kit (Qiagen, cat. no. 28004)
- TAE buffer (Sigma-Aldrich, cat. no. T8280)
- Agarose (Sigma-Aldrich, cat. no. A0169)
- Ethidium bromide (Invitrogen, cat. no. 15585011) or any alternative nucleic acid gel stain (e.g., SYBR Gold, Invitrogen, cat. no. S11494) ! **CAUTION** Nucleic acid stains are usually mutagenic. Use personal protective equipment when handling nucleic acid gel stains and dispose of waste according to institutional regulations.
- 1-kB DNA ladder (Sybenzyme, cat. no. M11)
- 100-bp DNA ladder (Sybenzyme, cat. no. M15)
- NEBNext DNA Library Prep Master Mix Set for Illumina (New England BioLabs, cat. no. E6040)
- NEBNext Multiplex Oligos for Illumina (Index Primers Sets 1 and 2, cat. no. E7335, E7500)
- Agencourt AMPure XP (Beckman Coulter, cat. no. A63881)
- 96% (vol/vol) ethanol (Sigma-Aldrich, cat. no. E7148) ! **CAUTION** Ethanol is flammable and toxic. Ethanol should be handled under a fume hood and only with personal protective equipment.
- RNAeasy Micro Kit (Qiagen, cat. no. 74004)

- Centrifuge with a swinging bucket rotor (e.g., Eppendorf 5810R or equivalent system)
- Falcon tubes (15 ml, 50 ml, conical, screw cap, sterile, disposable; BD Falcon, cat. nos. 352096, 352070)
- Refrigerated benchtop centrifuge (e.g., Eppendorf 5415R or equivalent system)

© 2016 Nature America, Inc. All rights reserved.

- PCR laminar flow cabinet (e.g., Lamsystems, cat. no. 620.100)
- Programmable thermostat (e.g., DNA-Technology, cat. no. Î-TT2-EU or equivalent system)
- PCR thermocycler with a heated lid (e.g., Applied Biosystems, cat. no. 4314878)
- QuBit Fluorometer (Life Technologies, cat. no. Q33216)
- MiSeq desktop sequencer (Illumina)
- 2100 Electrophoresis Bioanalyzer Instrument (Agilent Technologies, cat. no. G2939AA)
- Safe-Imager Blue-Light Transilluminator (Thermo Fischer Scientific, cat. no. G6600UK)
- Mini-Sub Cell GT horizontal electrophoresis system (Bio-Rad, cat. no. 1704406)
- Electrophoresis power supply, PowerPac Basic (Bio-Rad, cat. no. 1645050)
- Vortex mixer (e.g., Biosan, cat. no. BS-010201-AAA)
- Single-channel pipettes (volume range: 0.5–10 µl, 2–20 µl, 20–200 µl, 200–1,000 µl; e.g., Gilson, cat. nos. F144801, F123600, F123615, F123602, respectively)
- RNase- and DNase-free filtered pipette tips (10 µl, 20 µl, 200 µl, 1,000 µl; e.g., SSI Vertex, cat. nos. 4117NSFS, 4237NAFS, 4237NSFS, 4337NSFS, respectively)
- Plastic tubes (0.2-, 0.5- and 1.5-ml; e.g., SSI, cat. no. 3225; Eppendorf LoBind tubes, cat. nos. 022431005 and 022431021)

:

MiXCR (http://mixcr.milaboratory.com/) or MiGMAP (https://github.com/mikessh/migmap)
MiTools (https://github.com/milaboratory/mitools)
MIGEC (https://github.com/mikessh/migec)

A

80% (/) Prepare from 96% (vol/vol) ethanol by adding water.

75% (/) Prepare from 96% (vol/vol) ethanol by adding water.

Java Runtime Environment needs to be installed in order to run the pipeline described here (http://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html).

PROCEDURE

Preparation of starting material—cell purification ● TIMING 4–8 h

▲ CRITICAL Perform cell isolation, RNA purification, cDNA synthesis and first PCR preparation steps in separate clean workplaces. General recommendations to lower the risk of RNA degradation and contamination should be followed: use labcoats, gloves, tips with aerosol filters and certified RNAse/DNAse-free reagents, and perform nontemplate control reaction.

1| Perform isolation of mononuclear cells from whole blood using Ficoll–Paque density-gradient centrifugation[61].

2| Purify the B-cell subset of interest using magnetic separation or fluorescence-activated cell sorting[62,63].

▲ CRITICAL STEP An IG cDNA library can be generated starting from RNA isolated from total leukocytes, PBMCs or any tissue containing B cells. However, due to differences in IG mRNA expression levels, isolation of the specific cell subset of interest is recommended.

▲ CRITICAL STEP Control for the counts of isolated B cells of interest is desirable in order to manage the amount of input cells and RNA material (see Experimental design section).

▲ CRITICAL STEP It is preferable to use a freshly isolated cell sample for RNA purification. In the case of frozen samples, culture thawed cells overnight in RPMI-1640 supplemented with 10% (vol/vol) human serum.

▲ CRITICAL STEP If you are using FACS for purifying a limited number of B cells, direct sorting into lysis buffer from RNA isolation kit is recommended. The cells should be lysed immediately in the collection tube and the mRNA should be protected from degradation. In this case, cellular lysate in RLT buffer (Qiagen) can be stored at −70 °C for a month.

Total RNA extraction ● TIMING 1 h

3| For a small number of cells <100,000), use a column-based RNA isolation method—for example, an RNeasy Micro Kit (Qiagen) or other high-quality column-based kit, according to the manufacturer's instructions. For larger numbers of cells, use either column-based RNA isolation or any commercially available reagents for RNA extraction based on the acid guanidinium thiocyanate–phenol–chloroform method, such as Trizol (Invitrogen) or QIAzol (Qiagen), or other analogous products. For large numbers of cells, when using a column-based RNA extraction method, DNAse treatment is strongly recommended. When using a phenol–chloroform-based RNA extraction method, ensure that you use the correct amount of lysis reagent—for example, use 1 ml of Trizol for up to 10^7 cells. Add co-precipitant as a carrier to the aqueous phase, as recommended by the manufacturer.

▲ CRITICAL STEP An improper ratio of Trizol reagent:sample may lead to insufficient cell lysis, and it may affect the RNA yield. Use only fresh solutions of 70% (vol/vol) ethanol and 80% (vol/vol) ethanol for RNA extraction.

▲ CRITICAL STEP Co-precipitant is recommended, as it allows easy identification of the colored pellet and does not inhibit subsequent first-strand cDNA synthesis or PCR.

4| Verify the RNA quantity and quality, e.g., using an Agilent Bioanalyzer or gel electrophoresis. An RNA integrity number >7, or a correct 28S rRNA:18S rRNA ratio (~1.5–2.5:1) and a low number of shadow bands above and below the 18S band are indicative of high-quality RNA.

▲ CRITICAL STEP High RNA quality is critical for efficient IG library preparation.

? TROUBLESHOOTING

■ PAUSE POINT: RNA can be stored in 75% (vol/vol) ethanol for at least 1 year at −20 °C, or at least 1 week at 4 °C. For small RNA amounts (less than 100 ng), it is better to start cDNA synthesis immediately after RNA extraction.

© 2016 Nature America, Inc. All rights reserved.

cDNA synthesis with template switch ● TIMING 2 h

5| In a sterile thin-walled 0.2-ml reaction tube, mix the following reagents for a final volume of 8 μl (mix 1).

Component	Amount (μl)	Final amount/concentration[a]
RNA from Step 4	2–6	Up to 700 ng
cDNA synthesis primer mix (10 μM each)[b]	2	1 μM for each primer
RNAse-free water	0–4	

[a] Final concentration refers to concentration in 20 μl following the addition of mix 2 (in Step 8). [b] See **Table 2** for human and mouse primer mixes.

▲ CRITICAL STEP Preferably, use all RNA extracted from the B-cell sample of interest. At a later stage (Step 11), a portion of the obtained cDNA may be used in order to achieve desirable oversequencing per cDNA molecule.

6| Place the reaction tube(s) into a thermal cycler and incubate for 2 min at 70 °C and then decrease the incubation temperature to 42 °C to anneal the synthesis primers (1–3 min). Keep the tubes in the thermal cycler at 42 °C while preparing the mix 2 (up to 15 min). Use a heated lid.

7| While incubating, mix the following components in a separate tube for a final volume of 12 μl (mix 2).

Component	Amount (μl)	Final concentration[a]
First-strand buffer (5×, Evrogen or Clontech)	4	1×
DTT (20 mM, from SMARTscribe or Mint cDNA synthesis kit)	2	2 mM
5′-Template switch adaptor (10 μM)[b]	2	1 μM
dNTP solution (10 mM each)	2	1 mM each
SMARTScribe Reverse Transcriptase (10×, Clontech) or Mint Reverse Transcriptase (10×, Evrogen)	2	10 U/μl

[a] Final concentration in 20 μl after adding mix 2 to mix 1. [b] See **Table 2** for sequences of primers for cDNA synthesis.

▲ CRITICAL STEP Optionally, 20 units of RNAse inhibitor (e.g., RNAsin, Promega) can be added to mix 2 to prevent RNA degradation during cDNA synthesis.

8| Add mix 2 to mix 1, gently pipette the reaction mix and incubate it for 60 min at 42 °C.

9| Add 1 μl of uracil DNA glycosylase (5 U/μl) and incubate the mixture for 40 min at 37 °C.
▲ CRITICAL STEP Uracil DNA glycosylase treatment removes the residual template switch adaptor, which is critical for the accurate labeling of starting cDNA molecules.

10| Purify cDNA using the MinElute PCR Purification Kit (Qiagen). During purification, wash twice with PE buffer.
▲ CRITICAL STEP Residual quantities of oligonucleotides and enzymes used in the reverse transcription reaction negatively affect subsequent PCR. Removal of these components allows pure final bands to be obtained and enables the subsequent first PCR to be carried out in a smaller reaction volume.
■ PAUSE POINT At this time, purified cDNA can be stored at 4 °C overnight. For extended storage (1 month), it is recommended that the cDNA product be stored at 20 °C. However, it is safer to proceed with the first PCR the same or next day without freezing, in order to avoid losing the material.

© 2016 Nature America, Inc. All rights reserved.

© 2016 Nature America, Inc. All rights reserved.

First PCR amplification ● TIMING 2 h

11| In a sterile thin-walled 0.2-ml reaction tube, mix the following reagents in a final volume of 50 μl:

Component	Amount (μl)	Final concentration
First-strand cDNA from Step 10	2.5	—
Q5 polymerase buffer (5×, NEB)	10	1×
dNTP mix (10 mM each)	1	0.2 mM each
Primer M1SS (10 μM)[a]	1	0.2 μM
First PCR reverse primer mix[a]	1	0.2 μM each in heavy-chain mix, 0.2 μM each in light-chain mix
Q5 polymerase (NEB)	0.5	0.02 U/μl
Nuclease-free water	34	—

[a]See **Table 2** for human and mouse primer mixes.

▲ CRITICAL STEP Use a portion of cDNA equivalent to ~50,000 naive B cells, ~20,000 memory B cells or ~200 plasma B cells for the preparation of a library planned as 1/6 of a MiSeq run. Note that the initial amount of cells used should be as high as possible (the higher the better), in order to minimize cell sampling bias and to average IG mRNA expression levels between many cells.

▲ CRITICAL STEP During optimization of the PCR amplification step, we have tested several polymerases from different suppliers. The best result was obtained with Q5 polymerase (NEB). Alternatively, other polymerases with high fidelity and processivity may be used.

12| Perform PCR using the following parameters:

Cycle	Denature	Anneal	Extend
1	95 °C for 1 min 30 s	—	—
18	95 °C for 10 s	60 °C for 20 s	72 °C for 40 s
1	—	—	72 °C for 4 min

13| Combine PCR products that were obtained starting from the same cDNA synthesis and purify them using the QIAquick PCR Purification Kit (or other purification system). During purification, wash the sample twice with PE buffer, supplied in the kit. Elute the purified PCR product in 30 μl of elution buffer.

■ PAUSE POINT At this time, the purified product of the first PCR can be stored at 4 °C overnight. For extended storage (1–2 months), it is recommended that the PCR product be stored at 20 °C.

Second PCR amplification ● TIMING 2 h

14| In a sterile thin-walled 0.2-ml reaction tube, mix the following reagents for a final volume of 50 μl:

Component	Amount (μl)	Final concentration
Purified first PCR product from Step 13	2	—
Q5 polymerase buffer (5×, NEB)	10	1×
dNTP mix (10 mM each)	1	0.2 mM each
M1S primer with sample barcode (10 μM)[a]	1	0.2 μM
Z primer with sample barcode (10 μM)[a]	1	0.2 μM
Q5 polymerase (NEB)	0.5	0.02 U/μl
Nuclease-free water	34.5	

[a]See **Table 2** for second PCR primers.

▲ CRITICAL STEP On a single MiSeq run, combine M1S and Z primers with different numbers of random nucleotides at the 5′ end for different samples (4, 5 or 6 nt; see **Table 2**). This provides better diversity generation, which is critical for cluster differentiation by the Illumina sequencer.

15| Perform PCR using the following parameters:

Cycle	Denature	Anneal	Extend
1	95 °C for 1 min 30 s	—	—
10–15	95 °C for 10 s	60 °C for 20 s	72 °C for 40 s
1	—	—	72 °C for 4 min

16| Verify the quality and concentration of the obtained PCR product by analyzing an aliquot of the sample alongside the DNA ladder on an agarose gel or by using an Agilent Bioanalyzer.
▲ CRITICAL STEP A visible band (~2 ng/µl) should usually be obtained within 11–12 cycles of the second PCR. An excessive amount of PCR product produced after 10 cycles of the second PCR may indicate that too many IG cDNA molecules were used at the start of the first PCR, which may result in insufficient oversequencing. The absence of a visible band by 17 cycles of second PCR may indicate that less than 1,000 cDNA molecules have entered the first PCR, which would result in insufficient depth of profiling.
? TROUBLESHOOTING

17| Purify the PCR product using the QIAquick PCR Purification Kit (or another column-based purification system). During purification, wash the product twice.
▲ CRITICAL STEP It is important to purify the products of the second PCR within an hour after amplification. Otherwise, residual enzyme activities may damage the ends of the library that carry sample barcodes that are required for the data demultiplexing in further sequence analysis. Preferably store the products at +4 °C in the meantime.
■ PAUSE POINT At this time, purified libraries can be stored at 4 °C overnight. For extended storage (up to 1 month), it is recommended that the PCR product be stored at 20 °C before adaptor ligation.

Sequencing library preparation ● TIMING 1 d
18| For each of the obtained libraries, determine the concentration using the QuBit Fluorometer.

19| Process libraries for sequencing by pooling before adaptor ligation (option A) or after ligation of adaptors (option B):
(A) Pooled adaptor ligation
(i) For a MiSeq run, prepare a pool of your libraries by combining equal molar or equal volume portions of each individual sample. The resulting amount of pooled PCR products should be at least 300 ng.
▲ CRITICAL STEP We recommend generating parallel libraries of similar content (e.g., six samples of 5,000 plasma B cells each, sorted from three individuals in two replicates) using exactly the same protocol and number of PCR cycles, and mixing the obtained libraries in equal volume proportions. This allows homogeneous coverage (in terms of sequencing reads per UMI) to be obtained, which is optimal for subsequent comparative bioinformatic analysis. For example, a library that started from 30,000 cDNA molecules may produce more PCR product than one that started from 5,000 cDNA molecules after the same number of PCR cycles. However, the former library would also carry proportionally more IG cDNA molecules, and thus it requires more sequencing reads to achieve comparable reads-per-UMI sequencing coverage.
(ii) Use pooled PCR products from the previous step to prepare a sequencing library. Use the NEBNext Ultra DNA Library Prep Kit for Illumina sequencing and apply a standard protocol according to the manufacturer's recommendations (https://www.neb.com/protocols/2014/05/22/protocol-for-use-with-nebnext-ultra-dna-library-prep-kit-for-illumina-e7370).
(B) Separate adaptor ligation
(i) Alternatively, prepare separate libraries for each sample, using the NEBNext Ultra DNA Library Prep Kit for Illumina sequencing, as described in Step 19A(ii). Use at least 300 ng of each PCR product.
▲ CRITICAL STEP Double-end sample barcoding in the second PCR amplification efficiently protects from cross-sample contaminations during coamplification of joined PCR products after adaptor ligation. Nevertheless, some minimal cross-sample contamination may still occur. To provide 100% protection from cross-sample contamination, ligation of Illumina adaptors separately to each library is recommended.
(ii) For a MiSeq run, pool libraries of interest with ligated Illumina adaptors in equal volumes.

20| Purify the target library on an agarose gel. Prepare a 2% agarose gel with 1× TBE buffer. Mix the target library with 5× loading buffer and apply 20 µl per gel well. Run the agarose gel electrophoresis at 120 V for ~60 min. Next, excise

© 2016 Nature America, Inc. All rights reserved.

the bands of interest from the gel containing amplified libraries in the range of ~600–800 nt (the size of IGH amplicons with Illumina adaptors is sample-dependent).
▲ CRITICAL STEP Even minor amounts of short nonspecific products may essentially reduce counts of target sequencing reads, as short fragments are much more efficient in solid-phase bridge amplification.
▲ CRITICAL STEP Cut from the gel widely, so that the shortest and longest variants are not lost.
▲ CRITICAL STEP Cut from the gel quickly, or preferably use Safe Imager and Sybr Green staining, to avoid damaging the library.

Sequencing ● TIMING 2 d

21| Spike the library with 30% of the PhiX library.

TABLE 3 | An example of barcodes.txt file.

Sample ID	Master barcode sequence	Slave barcode sequence
S1	ACAATcagtggtatcaacgcagagt NNNNtNNNNtNNNNtct	CGTAAattgggcagccctgatt
S2	GCGGAcagtggtatcaacgcagag tNNNNtNNNNtNNNNtct	TCATTattgggcagccctgatt
S3	AGGGAcagtggtatcaacgcagag tNNNNtNNNNtNNNNtct	AGATAattgggcagccctgatt

First five upper-case characters match sample barcode and *N* characters are recognized by MIGEC as UMI. The master barcode sequence corresponds to the 5′ end of the library molecule and contains sample barcode and UMI sequence. The slave barcode can be found at the 3′ end and contains only sample barcode.

22| Analyze the resulting library using paired-end Illumina MiSeq sequencing, standard Illumina sequencing primers and MiSeq Reagent Kit v3 (600 cycles). To cover the whole IG length, increase the sequencing length by changing the cycle number during MiSeq sample sheet creation. Use either symmetric (option A) or asymmetric (option B) sequencing.

(A) Symmetric sequencing.

(i) Use 325+325-nt paired-end sequencing. Ignore software objections concerning the available number of cycles.

(B) Asymmetric sequencing.

(i) Use 400+100-nt paired-end sequencing in order to proceed with the asymmetric data analysis protocol shown in **Figure 2**.

Data analysis ● TIMING 5 h

▲ CRITICAL Input data must be in `fastq` format. All required software can work with compressed fastq (`fastq.gz`).

23| Prepare a tab-delimited file named `barcodes.txt` containing information about sample barcodes and UMI position. This tab-delimited text file should have the structure shown in **Table 3**.

24| Perform demultiplexing and unique molecular identifier extraction. During this step, fastq files are demultiplexed by sample barcodes and UMI sequences are extracted and recorded in the headers of the new fastq files. Use the latest version of the MIGEC software (https://github.com/mikessh/migec/releases) and run the following command:

```
java -jar migec.jar Checkout -cute barcodes.txt <R1.fastq.gz>
<R2.fastq.gz> <checkout_output_folder>
```

where `-cute` is a list of four parameters (c,u,t,e): c—tells MIGEC to produce the output in compressed form (`fastq.gz`), u—output UMI region to the read headers of the resulting fastq file, t—trim barcode sequences from the reads, e—remove template-switching trace, <R1.fastq.gz> — path to first read fastq file, <R2.fastq.gz> — path to second read fastq file, <checkout_output_folder> — the path where MIGEC will save demultiplexed files. This command will generate separate fastq files for each barcoded sample. It is expected that at the demultiplexing stage the software will report successful barcode extraction for more than 50% of sequencing reads if 30% of the PhiX library has been spiked in before sequencing.

25| Estimate the molecular identifier group (MIG, a set of reads tagged with the same UMI) size distribution—i.e., the coverage of individual sequenced molecules—using the MIGEC utility *Histogram*. An example command is as follows:

```
java -jar migec.jar Histogram --only-first-read <checkout_output_folder>
<histogram_output_folder>
```

This command will generate several files containing MIG statistics in a `histogram_output_folder`. Important files for further data processing are `overseq.txt` and `estimates.txt`. Each row of the first file contains information about the number of MIGs of a given size for each demultiplexed sample. The `estimates.txt` file contains information on estimated optimal threshold for MIG size.
▲ CRITICAL STEP The `--only-first-read` parameter allows only the first read to be used if you are carrying out asymmetric sequencing (400 + 100 nt, as described in Step 22B) to obtain correct statistical information for subsequent assembly of MIG groups (see **Fig. 2** for the logic of asymmetric data analysis).

© 2016 Nature America, Inc. All rights reserved.

26| Once the threshold for the minimally allowed MIG size is defined for each sample, perform assembly of the reads carrying the same UMI using the MIGEC utility AssembleBatch using the following command:

```
java -jar migec.jar AssembleBatch --force-collision-filter --force-
overseq X --only-first-read <checkout_output_folder>
<histogram_output_folder> <assemble_output_folder>
```

▲ **CRITICAL STEP** The `--force-overseq` parameter sets the reads per UMI threshold (MIG threshold). Efficient error correction preferably requires >5 reads per UMI. The MIG threshold should not be set below three reads per UMI in order to provide error correction. Setting the threshold below five reads per UMI is possible in order to obtain information on more IG cDNA molecules, but it will be associated with a higher error rate.
▲ **CRITICAL STEP** The `--force-collision-filter` parameter tells MIGEC to remove erroneous UMI variants that differ from a larger 'parent' UMI variant by a single mismatch. This parameter thus protects from artificial UMI variant accumulation that can remain after MIG size thresholding.
▲ **CRITICAL STEP** As the library preparation and sequencing (Steps 1–22) generate undirectional libraries, the first read fastq file contains sequences of both ends of the libraries. Using the AssembleBatch utility with the optional parameter `--only-first-read` will select only the read that was sequenced first according to raw data, which is recommended if you are using asymmetric 400+100-nt sequencing (Step 22B). This allows high-quality reads covering both the 5′ and 3′ parts of the IG sequence to be selected from files produced by Checkout, thus greatly improving consensus assembly efficiency. As a result, two fastq files are produced for each sample that contain consensus sequences for each assembled MIG, information about the number of reads corresponding to each MIG and the resulting consensus sequence quality for each MIG.

27| Perform merging of the 5′-end and 3′-end reads obtained after MIG assembly using the MiTools merge utility:

```
java -jar mitools.jar merge -ss -s 0.7 <R1.fastq>
<R2.fastq> <merge_output_folder/merged_R12.fastq>
```

where the `-ss` parameter tells the MiTools merge utility that reads are on the same strand and `-s` sets the minimum similarity of reads overlapping parts. This step produces fastq files that are ready for mapping to IGH reference sequences and for clonotype assembly.

28| Map sequences to references and assemble clonotypes using either the MiXCR software (option A) or the MiGMAP software (option B).

(A) Mapping and clonotype assembly using MiXCR

(i) Perform alignment of IGH sequences in the resulting MIG `fastq` files using the MiXCR software. An example command is as follows:

```
mixcr align --loci IGH -s hsa -
OvParameters.geneFeatureToAlign=VTranscript <input_fastq_file>
alignments.vdcja
```

This command will produce a `vdjca` file, which contains alignments of all data. The `--loci` parameter specifies target immunological loci (here it is the immunoglobulin heavy chain gene). If you are using RNA as a starting material, the parameter `-OvParameters.geneFeatureToAlign` should be set to `VTranscript` to increase sequencing information utilization from the 5′ end and thus the accuracy of V gene segment identification. The `-s` parameter allows the organism to be defined (`hsa` — Homo Sapiens or `mmu` — Mus Musculus).

(ii) Create an assembly of clonotypes using alignments obtained at the previous step. An example command is as follows:

```
mixcr assemble -OassemblingFeatures=VDJRegion -
OclusteringFilter.specificMutationProbability=1E-4 alignments.vdjca
clones.clns
```

▲ **CRITICAL STEP** The `-OclusteringFilter.specificMutationProbability` parameter sets the probability of PCR or sequencing error in clonal sequence for the frequency-based error correction. As an option for

© 2016 Nature America, Inc. All rights reserved.

analysis of data having a high MIG coverage threshold, it is possible to turn off MiXCR clustering (error correction) during the assemble stage using the -OcloneClusteringParameters=null parameter and relying on only MIG-guided error correction. Setting the -OassemblingFeatures parameter to VDJRegion is used for analyzing full-length sequences, and consists of framework 1, CDR1, framework 2, CDR2, framework 3, CDR3 and framework 4 (by default, clonotypes will be assembled using CDR3 sequences only). To assemble full-length clonotypes but without FR4, set the parameter -OassemblingFeatures to '{FR1Begin:CDR3End}':

```
mixcr assembly -OclusteringFilter.specificMutationProbability=1E-4 -
OassemblingFeatures='{FR1Begin:CDR3End}' alignments.vdjca clones.clns
```

In case of moderate MIG consensus-quality values (20–25), one can increase the number of extracted clonotypes by lowering the minimal quality threshold by setting the -ObadQualityThreshold parameter to, e.g., 15:

```
mixcr assembly -OclusteringFilter.specificMutationProbability=1E-4 -
OassemblingFeatures=VDJRegion -ObadQualityThreshold=15 alignments.vdjca
clones.clns
```

(iii) Export the results from the binary file (.clns) as a tab-delimited table. An example command is as follows:

```
mixcr exportClones clones.clns clones.txt
```

The resulting tab-delimited text file will contain columns with clonotype count (in terms of unique analyzed cDNA molecules); clonotype fraction; aligned clonotype sequences; MIGEC consensus qualities; best hits for V, D, J and C genes; and nucleotide and amino acid sequences of gene regions extracted from consensus sequences. It is possible to customize the output table by removing or adding necessary fields. For example, if one needs information about added nucleotides in V-(D)-J junctions, specifying the following nFeature (for nucleotide sequences) and aFeature (for amino acid sequences) parameters will add three columns to your final table (all available Gene Features and other options for output customization can be found in the documentation for the MiXCR software (http://mixcr.readthedocs.org/en/latest/)):

```
mixcr exportClones --preset full -nFeature VDJunction -nFeature
DJJunction -nFeature VJJunction clones.txt
```

© 2016 Nature America, Inc. All rights reserved.

(B) Mapping and clonotype assembly using MiGMAP

(i) Run the MiGMAP software with fastq files obtained after merging with the MiTools merge utility (Step 27). Before running the software, install stand-alone igBLAST software (ftp://ftp.ncbi.nih.gov/blast/executables/igblast/release/).

An example command is as follows:

```
java -jar migmap.jar --blast-dir <path to igblastn and makeblastdb>
--details all -R IGH -S human <input.fastq> <output.txt>
```

where --blast-dir indicates the path to the folder that contains the igblastn and makeblastdb binaries if they are not in the $PATH variable, -R specifies the receptor gene and chain, -S allows the organism to be set and --details specifies that additional fields will be added to the output. It can be nucleotide or amino acid sequences of frameworks and CDRs; for example, --details fr1nt, cdr1nt, fr1aa, cdr1aa will add columns with the nucleotide and amino acid sequences of the framework 1 and CDR1 regions. Value all will add all allowed fields (the nucleotide and amino acid sequences of all frameworks, CDRs and the clone consensus sequence).

By default, MiGMAP filters most ambiguous clonotype cases; this can be disabled by specifying --allow-incomplete, which will add clonotypes that CDR3 sequences cannot read completely; --allow-noncoding, which adds noncoding clonotypes; --allow-noncanonical, which adds clonotypes that have CDR3 regions not starting with a conserved Cys residue or ending with a conserved Phe/Trp; and --allow-no-cdr3, which will output clonotypes for which no CDR3 sequence was extracted. In addition, MiGMAP uses an alternative quality-filtering strategy that considers only the quality of mutations in germline sequence and N-regions of CDR3; the quality threshold can be adjusted with the -q parameter.

An example command is as follows:

```
java -jar migmap.jar -R IGH -S human -q 30 --allow-noncoding --allow-
noncanonical <input.fastq> <output.txt>
```

will extract and build clonotypes from sequences mapped to IGH, including ones that have noncoding and noncanonical CDR3 sequences with a quality threshold of 30.

Obtained information can be used for a wide range of downstream analysis tasks; for example, it can be exported to VDJtools software (https://github.com/mikessh/vdjtools)[60].

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 4**.

TABLE 4 | Troubleshooting table.

Step	Problem	Possible reason	Possible solution
4	gDNA contamination (when using phenol–chloroform extraction method)	Interphase pipetted up with aqueous phase	Do not draw off the entire aqueous phase after phase separation. Perform DNAse treatment
	Low RNA yield	Final RNA pellet was incompletely dissolved	Do not allow the RNA to dry completely after the final wash; the pellet can lose solubility
	RNA is degraded	Samples were stored too long before processing	Try to process the sample immediately after collection
		Isolated RNA was stored at an incorrect temperature	Store RNA samples at −20 °C and in 75% (vol/vol) ethanol
		RNAse contamination	Prepare new solutions of reagents and treat the equipment with RNAse decontamination solution (e.g., RNAseZap, Ambion)
16	Agarose gel electrophoresis reveals a low concentration of the PCR product or no product	The RNA may contain impurities that inhibit cDNA synthesis	In some cases, ethanol precipitation of RNA can remove impurities. If this does not help, reisolate the RNA
		Too few PCR cycles	Subject the samples to two or three additional PCR cycles (plus one extra final extension cycle) and recheck the products
		cDNA synthesis or PCR-kit-related problems	Use high-quality control RNA extracted from large amounts of B cells or white blood cells to verify kit performance
	Bands and background smear are very intense	Too many PCR cycles	Repeat the PCR amplification, using two or three fewer PCR cycles
	Background smear is intense or short-length fragments are visible		Purify the target library using AMPure XP beads or agarose gel purification

● TIMING

Steps 1 and 2, cell sample preparation: 4–8 h
Steps 3 and 4, total RNA extraction: 1 h
Steps 5–10, cDNA synthesis with template switch: 2 h
Steps 11–13, first PCR amplification: 2 h
Steps 14–17, second PCR amplification: 2 h
Steps 18–20, sequencing library preparation: 1 d
Steps 21 and 22, sequencing: 2 d
Steps 23–28, data analysis: 5 h

ANTICIPATED RESULTS

Using the protocol provided will typically produce a pure PCR band after 18 cycles of the first PCR and 11–15 cycles of the second PCR amplification, depending on the quality, IG mRNA content and amounts of starting mRNA. The number of

© 2016 Nature America, Inc. All rights reserved.

valid paired-end sequencing reads covering the IG variable region should be in the range of 10–14 million for a MiSeq run. Average coverage (in terms of sequencing reads per UMI) should be in the range between 20 and 40 sequencing reads per UMI. The number of successfully analyzed full-length IG cDNA molecules should be in the range between 70,000 and 200,000 for a MiSeq run (M.A.T., A.D., O.V.B., M.S. and D.M.C., data not shown), raw data deposited in SRA: PRJNA297771.

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

ACKNOWLEDGMENTS This work was supported by Russian Science Foundation project no. 14-14-00533. The work was carried out in part using equipment provided by the Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry Core Facility (CKP IBCH). M.S. is supported by an individual fellowship (mol-a-dk RFBR grant 16-34-60179). A.D., V.B., K.P. and S.P. are supported by the Ministry of Education, Youth and Sports of the Czech Republic (CEITEC 2020, LQ1601). K.P., V.B. and S.P. are supported by Ministry of Health, Czech Republic (AZV 15-30015A).

AUTHOR CONTRIBUTIONS M.A.T., O.V.B., V.B., V.I.K., E.M.M., I.Z.M. and K.P. prepared the cDNA libraries and worked on the protocol. E.S.E., D.B.S. and O.K. worked on cell sample preparation. A.D. and M.D.L. worked on sequencing. M.A.T., O.V.B. and D.M.C. designed the experiments. A.D., M.S., D.A.B., M.I., S.P. and D.M.C. worked on data analysis and manuscript preparation.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Robins, H.S. *et al.* Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* **114**, 4099–4107 (2009).
2. Freeman, J.D., Warren, R.L., Webb, J.R., Nelson, B.H. & Holt, R.A. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res.* **19**, 1817–1824 (2009).
3. Mamedov, I.Z. *et al.* Quantitative tracking of T cell clones after haematopoietic stem cell transplantation. *EMBO Mol. Med.* **3**, 201–207 (2011).
4. Warren, R.L. *et al.* Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res.* **21**, 790–797 (2011).
5. Vollmers, C., Sit, R.V., Weinstein, J.A., Dekker, C.L. & Quake, S.R. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl. Acad. Sci. USA* **110**, 13463–13468 (2013).
6. Jiang, N. *et al.* Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* **5**, 171ra119 (2013).
7. Laserson, U. *et al.* High-resolution antibody dynamics of vaccine-induced immune responses. *Proc. Natl. Acad. Sci. USA* **111**, 4928–4933 (2014).
8. Kaplinsky, J. *et al.* Antibody repertoire deep sequencing reveals antigen-independent selection in maturing B cells. *Proc. Natl. Acad. Sci. USA* **111**, E2622–2629 (2014).
9. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32**, 158–168 (2014).
10. Weinstein, J.A., Jiang, N., White, R.A., Fischer, D.S. & Quake, S.R. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807–810 (2009).
11. Mora, T., Walczak, A.M., Bialek, W. & Callan, C.G. Jr. Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. USA* **107**, 5405–5410 (2010).
12. Jiang, N. *et al.* Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proc. Natl. Acad. Sci. USA* **108**, 5348–5353 (2011).
13. Rubelt, F. *et al.* Onset of immune senescence defined by unbiased pyrosequencing of human immunoglobulin mRNA repertoires. *PLoS One* **7**, e49774 (2012).
14. Parameswaran, P. *et al.* Convergent antibody signatures in human dengue. *Cell Host Microbe* **13**, 691–700 (2013).
15. Tan, Y.C. *et al.* High-throughput sequencing of natively paired antibody chains provides evidence for original antigenic sin shaping the antibody response to influenza vaccination. *Clin. Immunol.* **151**, 55–65 (2014).
16. Galson, J.D. *et al.* BCR repertoire sequencing: different patterns of B-cell activation after two Meningococcal vaccines. *Immunol. Cell Biol.* **93**, 885–895 (2015).
17. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. USA* **108**, 9530–9535 (2011).
18. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).
19. Britanova, O.V. *et al.* Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J. Immunol.* **192**, 2689–2698 (2014).
20. Shugay, M. *et al.* Towards error-free profiling of immune repertoires. *Nat. Methods* **11**, 653–655 (2014).
21. He, L. *et al.* Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. *Sci. Rep.* **4**, 6778 (2014).
22. Egorov, E.S. *et al.* Quantitative profiling of immune repertoires for minor lymphocyte counts using unique molecular identifiers. *J. Immunol.* **194**, 6155–6163 (2015).
23. Khan, T.A. *et al.* Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci. Adv.* **2**, e1501371 (2016).
24. Bolotin, D.A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).
25. Briney, B.S., Willis, J.R., McKinney, B.A. & Crowe, J.E. Jr. High-throughput antibody sequencing reveals genetic evidence of global regulation of the naive and memory repertoires that extends across individuals. *Genes Immun.* **13**, 469–473 (2012).
26. Larimore, K., McCormick, M.W., Robins, H.S. & Greenberg, P.D. Shaping of human germline IgH repertoires revealed by deep sequencing. *J. Immunol.* **189**, 3221–3230 (2012).
27. Wu, Y.C. *et al.* High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* **116**, 1070–1078 (2010).
28. Siegrist, C.A. & Aspinall, R. B-cell responses to vaccination at the extremes of age. *Nat. Rev. Immunol.* **9**, 185–194 (2009).
29. Wang, C. *et al.* Effects of aging, cytomegalovirus infection, and EBV infection on human B cell repertoires. *J. Immunol.* **192**, 603–611 (2014).
30. Tan, Y.C. *et al.* Barcode-enabled sequencing of plasmablast antibody repertoires in rheumatoid arthritis. *Arthritis Rheumatol.* **66**, 2706–2715 (2014).
31. Doorenspleet, M.E. *et al.* Rheumatoid arthritis synovial tissue harbours dominant B-cell and plasma-cell clones associated with autoreactivity. *Ann. Rheum. Dis.* **73**, 756–762 (2014).
32. Racanelli, V. *et al.* Antibody V(h) repertoire differences between resolving and chronically evolving hepatitis C virus infections. *PLoS One* **6**, e25606 (2011).
33. Ademokun, A. *et al.* Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell* **10**, 922–930 (2011).
34. Tschumper, R.C. *et al.* Comprehensive assessment of potential multiple myeloma immunoglobulin heavy chain V-D-J intraclonal variation using massively parallel pyrosequencing. *Oncotarget* **3**, 502–513 (2012).
35. Fridy, P.C. *et al.* A robust pipeline for rapid production of versatile nanobody repertoires. *Nat. Methods* **11**, 1253–1260 (2014).
36. Lu, D.R. *et al.* Identifying functional anti-*Staphylococcus aureus* antibodies by sequencing antibody repertoires of patient plasmablasts. *Clin. Immunol.* **152**, 77–89 (2014).
37. 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).
38. Briney, B.S., Willis, J.R. & Crowe, J.E. Jr. Human peripheral blood antibodies with long HCDR3s are established primarily at original recombination using a limited subset of germline genes. *PLoS One* **7**, e36750 (2012).

© 2016 Nature America, Inc. All rights reserved.

39. Yu, L. & Guan, Y. Immunologic basis for long HCDR3s in broadly neutralizing antibodies against HIV-1. *Front. Immunol.* **5**, 250 (2014).
40. Nguyen, P. *et al.* Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics* **12**, 106 (2011).
41. Bolotin, D.A. *et al.* Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur. J. Immunol.* **42**, 3073–3083 (2012).
42. Brodin, J. *et al.* PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. *PLoS One* **8**, e70388 (2013).
43. Brodin, J. *et al.* Challenges with using primer IDs to improve accuracy of next generation sequencing. *PLoS One* **10**, e0119123 (2015).
44. Yaari, G. & Kleinstein, S.H. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.* **7**, 121 (2015).
45. Elnifro, E.M., Ashshi, A.M., Cooper, R.J. & Klapper, P.E. Multiplex PCR: optimization and application in diagnostic virology. *Clin. Microbiol. Rev.* **13**, 559–570 (2000).
46. Markoulatos, P., Siafakas, N. & Moncany, M. Multiplex polymerase chain reaction: a practical approach. *J. Clin. Lab. Anal.* **16**, 47–51 (2002).
47. Carlson, C.S. *et al.* Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* **4**, 2680 (2013).
48. van Dijk, E.L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.* **322**, 12–20 (2014).
49. Schrum, A.G., Turka, L.A. & Palmer, E. Surface T-cell antigen receptor expression and availability for long-term antigenic signaling. *Immunol. Rev.* **196**, 7–24 (2003).
50. Cho, B.K., Wang, C., Sugawa, S., Eisen, H.N. & Chen, J. Functional differences between memory and naive CD8 T cells. *Proc. Natl. Acad. Sci. USA* **96**, 2976–2981 (1999).
51. Schrum, A.G., Wells, A.D. & Turka, L.A. Enhanced surface TCR replenishment mediated by CD28 leads to greater TCR engagement during primary stimulation. *Int. Immunol.* **12**, 833–842 (2000).
52. Shi, W. *et al.* Transcriptional profiling of mouse B cell terminal differentiation defines a signature for antibody-secreting plasma cells. *Nat. Immunol.* **16**, 663–673 (2015).
53. Wrammert, J. *et al.* Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* **453**, 667–671 (2008).
54. Franz, B., May, K.F. Jr., Dranoff, G. & Wucherpfennig, K. *Ex vivo* characterization and isolation of rare memory B cells with antigen tetramers. *Blood* **118**, 348–357 (2011).
55. Kuenz, B. *et al.* Cerebrospinal fluid B cells correlate with early brain inflammation in multiple sclerosis. *PLoS One* **3**, e2559 (2008).
56. Greiff, V. *et al.* Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunology* **15**, 40 (2014).
57. Matz, M. *et al.* Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Res.* **27**, 1558–1560 (1999).
58. Douek, D.C. *et al.* A novel approach to the analysis of specificity, clonality, and frequency of HIV-specific T cell responses reveals a potential mechanism for control of viral escape. *J. Immunol.* **168**, 3099–3104 (2002).
59. Feng, Y. *et al.* A mechanism for expansion of regulatory T-cell repertoire and its role in self-tolerance. *Nature* **528**, 132–136 (2015).
60. Shugay, M. *et al.* VDJtools: unifying post-analysis of T cell receptor repertoires. *PLoS Comput. Biol.* **11**, e1004503 (2015).
61. Jaatinen, T. & Laine, J. Isolation of mononuclear cells from human cord blood by Ficoll-Paque density gradient. *Curr. Protoc. Stem Cell Biol.* Chapter 2 Unit 2A 1, http://dx.doi.org/10.1002/9780470151808.sc02a01s1 (2007).
62. Sims, G.P. & Lipsky, P.E. Isolation of human B cell populations. *Curr. Protoc. Immunol.* Chapter 7 Unit 7 5, http://dx.doi.org/10.1002/0471142735.im0705s75 (2006).
63. Kjeldsen, M.K. *et al.* Multiparametric flow cytometry for identification and fluorescence activated cell sorting of five distinct B-cell subpopulations in normal tonsil tissue. *Am. J. Clin. Pathol.* **136**, 960–969 (2011).

© 2016 Nature America, Inc. All rights reserved.