This information is current as of May 17, 2016.

Highly Accurate Sequencing of Full-Length Immune Repertoire Amplicons Using Tn5-Enabled and Molecular Identifier–Guided Amplicon Assembly

Charles Cole, Roger Volden, Sumedha Dharmadhikari, Camille Scelfo-Dalbey and Christopher Vollmers

J Immunol published online 8 February 2016
http://www.jimmunol.org/content/early/2016/02/06/jimmunol.1502563

Supplementary Material **http://www.jimmunol.org/content/suppl/2016/02/06/jimmunol.1502563.DCSupplemental.html**

The Journal of Immunology is published twice each month by
The American Association of Immunologists, Inc.,
9650 Rockville Pike, Bethesda, MD 20814-3994.

Print ISSN: 0022-1767 Online ISSN: 1550-6606.

Published February 8, 2016, doi:10.4049/jimmunol.1502563

Highly Accurate Sequencing of Full-Length Immune Repertoire Amplicons Using Tn5-Enabled and Molecular Identifier–Guided Amplicon Assembly

Charles Cole, Roger Volden, Sumedha Dharmadhikari, Camille Scelfo-Dalbey, and Christopher Vollmers

Ab repertoire sequencing is a powerful tool to analyze the adaptive immune system. To sequence entire Ab repertoires, amplicons are created from Ab H chain (IgH) transcripts and sequenced on a high-throughput sequencer. The field of immune repertoire sequencing is growing rapidly and the protocols used are steadily improving; however, thus far, immune repertoire sequencing protocols have not been able to sequence full-length immune repertoires including the entire IgH V region and enough of the IgH C region to identify isotype subtypes. In this study, we present a method that combines Tn5 transposase and molecular identifiers for the highly accurate sequencing of amplicons >500 bp using Illumina short read paired-end sequencing. We then apply this method to Ab H chain amplicons to sequence the first, to our knowledge, highly accurate full-length immune repertoire. ***The Journal of Immunology*****, 2016, 196: 000–000.**

Antibodies are encoded by H chain (IgH) and L chain (IGκ/λ) loci that undergo somatic recombination during B cell differentiation. In the H chain, VDJ recombination creates a highly diverse CDR3 when randomly and imperfectly combining one each of ~40 V, ~30 D, and 6 J segments. H chain loci are further modified by somatic hypermutation and class switch recombination (1). Somatic hypermutation introduces mutations and indels that can affect the binding characteristics of an Ab. Class-switch recombination changes the Ab isotype by genomic rearrangement of the isotype-determining C regions (IgM, IgD, IgG1–4, IgA1, IgA2, IgE). The isotype of an Ab changes the characteristics of an Ab, such as the ability to activate the complement system, pass the placenta, or bind certain F_c receptors.

Together, VDJ recombination, somatic hypermutation, and class-switch recombination create a virtually unique IgH locus in every mature B cell clone. Furthermore, these unique B cell clones expand and increase IgH transcription in response to activation by an Ag.

Therefore, analyzing the repertoire of IgH transcripts in a blood sample for isotype distribution, mutation levels, and clonal expansion provides insight into the composition and state of the adaptive immune system. Immune repertoire sequencing has so far been used in both basic and translational research. In basic research it has been applied to estimate the absolute size of the B cell repertoire in humans, track the effect of aging on the immune system, investigate V, D, and J pairing, and infer haplotypes of the genomic IgH locus (2–4). On the translational side, it has been used to track immune response to diseases and vaccines, to track minimal residual disease in leukemia, and to determine rejection events following organ transplantation (5–8).

All of these studies rely on capturing the diversity of the IgH repertoire but are limited by current sequencing technologies and protocols. Therefore, they either 1) use long read platforms (454), which allow for the sequencing of the whole IgH V region plus partial C region (2, 5, 9) but are often limited by high cost, lower throughput, and high error rates, or 2) use short read sequencers (Illumina HiSeq, Illumina MiSeq, Ion Torrent PGM), which allow for higher throughput at lower cost and error rate but are limited by their short read length to sequence only part of the IgH V region, including the CDR3 (7, 10).

To provide complete information of an IgH repertoire, an ideal full-length IgH transcript amplicon would be ~530 bp in length, starting at the leader exon and ending at 100 bp into the C region. Starting in the leader exon, which does not encode for the final Ab protein, would ensure that no bases in the V region are masked by primers, which would ensure the accurate identification of all V segment alleles. Ending 100 bp into the C region would provide enough sequencing information to distinguish all isotypes and subtypes, which is essential for allergy research.

Although an Illumina MiSeq 2 × 300 sequencing run theoretically allows for the sequencing of this ideal 530-bp IgH amplicon, in practice declining base quality does not allow for the sequencing of an amplicon >450 bp.

Recently, several groups have employed approaches utilizing molecular identifiers to improve sequencing accuracy, which is essential to differentiate somatic hypermutation from PCR and sequencing errors (7, 11, 12). Furthermore, several groups have developed protocols utilizing short read sequencers to sequence individual molecules exceeding the current raw read length of these sequencers. These protocols rely on inefficient steps in li-

Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064

ORCID: 0000-0002-2898-5150 (R.V.).

Received for publication December 9, 2015. Accepted for publication January 11, 2016.

This work was supported by a special research grant awarded by the Committee on Research from the University of California Santa Cruz.

The raw data presented in this article have been submitted to the National Center for Biotechnology Information Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession number SRP061676.

Address correspondence and reprint requests to Prof. Christopher Vollmers, University of California Santa Cruz, Department of Biomolecular Engineering, Biomedical Sciences Building, 145/Thimann Receiving, Santa Cruz, CA 95064. E-mail address: vollmers@ucsc.edu

The online version of this article contains supplemental material.

Abbreviation used in this article: TMIseq, Tn5-enabled molecular identifier–guided amplicon sequencing.

0022-1767/16/$30.00

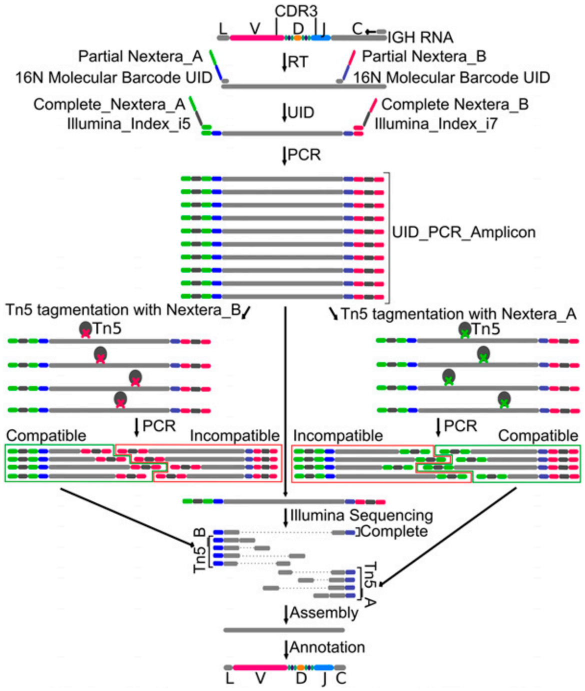

FIGURE 1. Schematic TMIseq library preparation and data analysis. IgH RNA is reverse transcribed and second- and third-strand cDNA is generated using 5′-modified primers. After PCR amplification the amplicons are tagmented using custom-loaded Tn5 enzymes. Three libraries per sample are sequenced and the resulting reads are grouped using molecular identifiers and assembled with a custom algorithm (AMPssembler).

brary preparation, including biotin pulldowns and intramolecular circulations (13–16).

To overcome read length, accuracy, and library preparation limitations, we developed Tn5-enabled molecular identifier–guided amplicon sequencing (TMIseq). TMIseq is based on a simple library preparation protocol utilizing molecular barcoding of individual molecules and Tn5 tagmentation (17), enabling the highly accurate and cost-effective sequencing of molecules exceeding Illumina read length (see Fig. 1).

Materials and Methods

PBMC extraction and RNA purification

All experiments were approved by the Internal Review Board at the University of California Santa Cruz and Stanford University. For sample I1, whole blood samples were collected from a healthy human adult volunteer by the University of California Santa Cruz Student Health Center. For sample I2, buffy coats were provided completely de-identified by the Stanford Blood Center. I1 and I2 samples were processed by Ficoll gradient (GE Healthcare) to extract PBMCs. PBMCs were lysed directly in RLT buffer and frozen at −80°C until RNA was extracted. RNA was extracted from 400,000 cells each using the RNeasy mini kit (Qiagen). Resulting RNA concentrations ranged from 20 to 50 ng/μl.

TMIseq library preparation

RNA (10 μl) was used for SuperScript II (Thermo Fisher Scientific) cDNA first-strand synthesis using a primer pool specific to all exons specific to the secreted isoform of all IgH isotypes (IgM, IgD, IgG1–4, IgA1, IgA2, IgE). In a two-cycle PCR reaction, second and third cDNA strands were synthesized using Phusion polymerase (Thermo Fisher Scientific) and two modified primer pools complementary to the beginning of the V leader exons and ∼100 bp into CH1 exons of all IgH isotypes and containing molecular identifiers and partial Nextera sequences. cDNA was purified and size selected twice with SPRI beads using a 0.7:1 beads/sample ratio corresponding to a cutoff discarding DNA <300 bp. In a 30-cycle PCR reaction, third cDNA strands were amplified using a pair of primers containing complete Nextera sequences as well as Illumina i5 and i7 indexes to index each individual sample. Samples with unique i5 and i7 indexes (i.e., each sample can be uniquely distinguished by either an i5 or i7 index; for example, sample 1, i5_1, i7_1; sample 2, i5_2, i7_2, and so forth) are pooled and split into three aliquots. To create Tn5_A libraries, aliquot 1 is tagmented using Tn5 enzyme (17) loaded with Nextera_A adapter and PCR amplified using a universal Nextera_B primer and a Nextera_A primer with an Illumina index not yet present in the library pool and purified and size selected for fragments >380 bp using 2% EX gels (Life Technologies). To create Tn5_B libraries, aliquot 2 is tagmented using Tn5 enzyme (17) loaded with Nextera_B adapter and PCR amplified using a universal Nextera_A primer and a Nextera_B primer with an Illumina index not yet present in the library pool and purified and size selected for fragments >380 bp using 2% EX gels (Life Technologies). Uncut (aliquot 3), Tn5_A, and Tn5_B libraries were pooled and sequenced according to standard Illumina protocols on an Illumina MiSeq (2 × 300 run) or HiSeq 3000 (2 × 150 run).

Control library preparation

Control libraries were generated as TMIseq libraries with the exceptions to the primer pools used for second- and third-strand cDNA synthesis. The FR1-specific primer pool was designed to bind 1–10 bp into the FR1 region, whereas the C-specific primer pool was designed to bind 20 bp into the CH1 exons of all IgH isotypes. The resulting library with an insert size of ∼ 400bp is sequenced on an Illumina MiSeq 2 × 300 bp run.

Raw data processing data assembly

Raw reads in fastq format are trimmed using Trimmomatic (18), discarding reads pairs containing adapters. For libraries sequenced on the MiSeq 2 × 300, reads were also cropped to 150 bp. TMIseq data were further processed according to the following pipeline. First, molecular identifiers are extracted from the trimmed fastq files. For Uncut libraries, the first 18 bases of read 1 represent molecular identifier 1 and the first 18 bases of read 2 represent molecular identifier 2. For Tn5_A libraries, the first 18 bases of read 2 represent molecular identifier 2. For Tn5_B libraries, the first 18 bases of read 1 represent molecular identifier 1. Second, reads of the Uncut library are grouped into molecular groups when their combined molecular identifiers differ by fewer than five mismatches. Third, reads with highly similar (fewer than two mismatches) molecular identifier

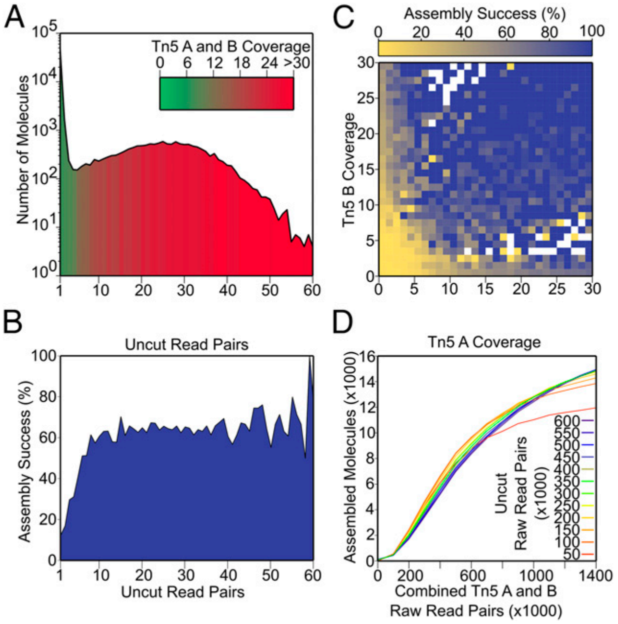

FIGURE 2. TMIseq subassembly coverage requirements. (**A**) Read pair coverage for IgH molecules in the I2 L1 Uncut library is shown as a histogram. Average combined Tn5_A and Tn5_B read coverage at increasing Uncut raw read coverage levels is shown as a color gradient. (**B**) Average assembly success at increasing I2 L1 Uncut read coverage levels is shown. (**C**) Heat map showing the correlation of assembly success and read coverage in I2 L1. Average success percentage for Tn5_A and Tn5_B coverage combinations is shown. (**D**) Number of I2 L1 IgH molecules successfully assembled from increasing numbers of subsampled Uncut raw read pairs (line colors) and combined Tn5_A and Tn5_B raw read pairs is plotted.

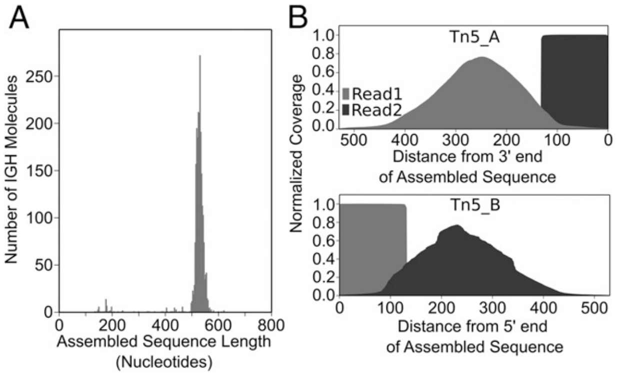

FIGURE 3. TMIseq assembles 530-bp IgH molecules. (**A**) Length distribution of I1 L1 IgH molecules assembled using TMIseq. (**B**) Trimmed Tn5_A and Tn5_B reads are mapped to assembled IgH molecules using BLAST. Mapped read coverage across IgH transcripts is shown as histograms.

1 (Tn5_B) or molecular identifier 2 (Tn5_A) to the Uncut molecular groups are added into these molecular groups. Third, the AMPssembler algorithm assembles IgH transcripts from each molecular group. The program sorts the raw reads into three categories: 1) reads derived from the 5′ end of the amplicon, 2) reads derived from the 3′ end, and 3) reads that are unanchored and likely to come from some place in the middle of the molecule. Then the program creates a high-quality consensus of the ends of the amplicon using a combination of quality and abundance of each nucleotide at each position. Finally, the program reduces all of the reads into k-mers and extends one of the ends until the program reaches the other end of the amplicon or the program runs out of extensions. It then reports the completed molecule in fastq format. When multiple extensions are possible, the program always selects the extension that results in the highest quality base being incorporated into the extension. Control data were processed as previously described (7). To analyze molecule coverage distribution in Fig. 3B, we aligned the raw reads of each molecule group to the assembled molecule using BLAST (19). Data were then converted to the fasta format and annotated using IgBLAST (20) with germline data retrieved from IMGT (21). For Fig. 6, IgH molecules were grouped into lineages across all samples analyzed using a single linkage clustering approach and a 90% CDR3 similarity cutoff. For Fig. 2D, reads were subsampled to the approximate target levels from the unprocessed fastq file pairs. The resulting subsampled files were then analyzed by the complete analysis pipeline. Further downstream analysis and visualization were done using Python/matplotlib (22).

Data access

The AMPssembler scripts used in the analyses of the data are available at GitHub at https://github.com/chkcole/AMPssembler. All other scripts are available upon request.

Raw data were uploaded to the National Center for Biotechnology Information Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession number SRP061676 (I1 data are identified by anonymized ID SHC1-3-1, and I2 data are identified by anonymized ID BB7).

Results

Overview of TMIseq

To assemble RNA molecules that exceed the sequencing length but not the cluster generation length of Illumina sequencers, TMIseq utilizes molecular identifiers and the unique characteristics of the Tn5 enzyme. We reverse transcribe RNA molecules into cDNA and then generate second- and third-strand copies of cDNA in a two-cycle amplification reaction using two primer pools. The primer pools we use for this two cycle reaction are (Supplemental Table II): 1) the V_Leader pool containing primers specific to the leader exons of all V segments, and 2) the C_long pool containing primers that bind 100 bp into the C regions of all isotypes (C_long). All primers in these pools feature modified 5′ ends to generate a single third-strand cDNA copy of each IgH RNA molecule tagged with 18-bp random molecular identifiers and partial Nextera (Illumina) sequences on both ends (Nextera_A for V_Leader, Nextera_B for C_long) (Fig. 1). We then amplify these uniquely tagged cDNAs using two primers specific to the partial Nextera A and B sequences, respectively. Both primers complete their respective Nextera sequence and add a sample index while preserving the molecular identifiers. This results in a dual-indexed ~530-bp amplicon library that is Illumina sequencing–ready. We then split the library into three aliquots.

The first and second aliquots (*Tn5_A* and *Tn5_B*) are tagmented with Tn5 enzyme loaded only with partial Nextera_A (*Tn5_A*) or Nextera_B (*Tn5_B*) oligonucleotides and PCR amplified to complete the Nextera_A (*Tn5_A*) or Nextera_B (*Tn5_B*) sequences, respectively (Fig. 1). At this point, *Tn5_A* and *Tn5_B* libraries are Illumina sequencing–ready, and there is no enrichment required, as Illumina chemistry only sequences molecules with both complete Nextera_A and Nextera_B sequences at their ends (Fig. 1). Therefore, *Tn5_A* and *Tn5_B* libraries exclusively produce raw read pairs in which one read is anchored by the V_Leader (*Tn5_A*) or C_long (*Tn5_B*) primers and contains one of the molecular identifiers associated with the original template molecule, whereas the other read is primed from the Nextera sequence that was introduced at a random location into the amplicon by Tn5.

The third aliquot (*Uncut*) is left unchanged and sequenced alongside the *Tn5_A* and *Tn5_B* libraries (Fig. 1) and exclusively

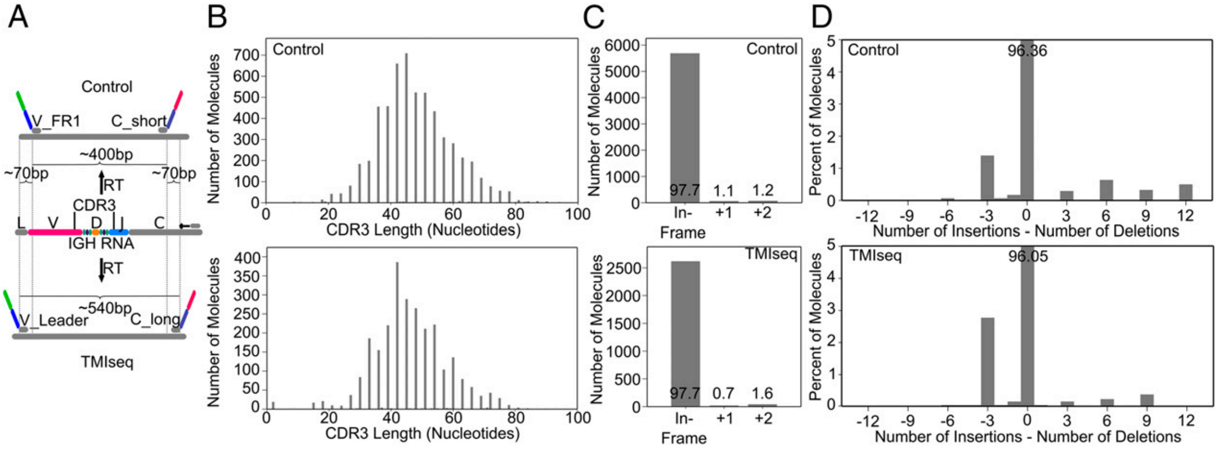

FIGURE 4. TMIseq mutations data equivalent to control libraries. (**A**) Schematic of primer positioning for second- and third-strand synthesis in TMIseq and control libraries. (**B–D**) Control and TMIseq sequences are compared for CDR3 length distribution (B), CDR3 translation frame (C), and shift in frame produced by indels (D).

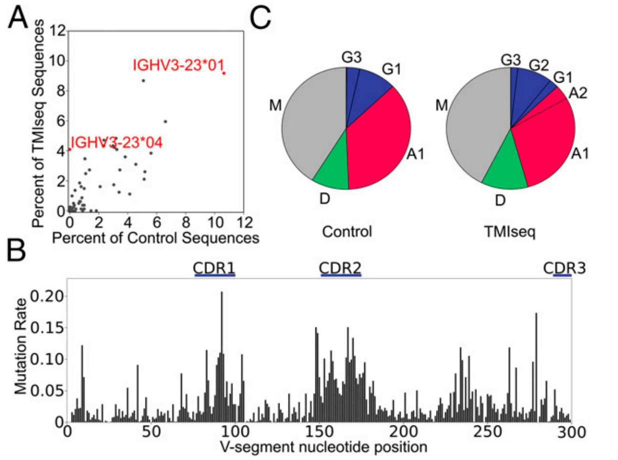

FIGURE 5. TMIseq can differentiate all V segments and isotypes. (**A** and **B**) Control and TMIseq sequences are compared for perfectly matched V segment usage in a scatter plot (A) and isotype usage in a pie chart (B). (**C**) Mutation rate across the entire V segment averaged across all IgH molecules using V segments of the IGHV3 family.

produces raw read pairs in which both reads are anchored by V_Leader or C_long primers and contain both molecular identifiers associated with the original template molecule (Fig. 1).

For analysis, after quality trimming and filtering, *Uncut, Tn5_A*, and *Tn5_B* read pairs containing highly similar molecular identifiers (Supplemental Fig. 1) in their anchored reads are combined into IgH molecule groups. IgH molecules are then assembled from each group using AMPssembler, a custom k-mer–based amplicon assembler that takes advantages of the known properties of the TMIseq protocol. Namely, the ends of the assembled sequences are defined by the anchored reads and there is only a single sequence to be assembled per IgH molecule group.

Application of TMIseq to the analysis of IgH transcript amplicons

To test the TMIseq protocol and data analysis, we created TMIseq libraries from two individuals (I1 and I2) from samples of PBMCs that contain B cells. For I1, we generated TMIseq *Uncut*, *Tn5_A*, and *Tn5_B* libraries for one sample (I1 L1), sequenced those libraries on a MiSeq 2 × 300 run, and truncated the resulting reads to 150 bp to model the shorter read length. The MiSeq run generated 125,200 raw reads for the I1 libraries, which yielded 120,104 quality trimmed reads. The trimmed read pairs were assembled by AMPssembler into 2779 IgH molecules. For I2, we generated TMIseq *Uncut*, *Tn5_A*, and *Tn5_B* libraries for eight samples (I2 L1–L8) and sequenced those libraries on a HiSeq 3000 (2 × 150 run). The HiSeq run generated 15,587,484 raw read pairs across the eight I2 samples, which yielded 10,577,945 quality trimmed read pairs. These trimmed read pairs were assembled by AMPssembler into 115,108 IgH molecules (11,075–16,985 IgH molecules per library) (Supplemental Table I).

TMIseq coverage requirements

We used I2 L1 data to determine the coverage requirements to assemble IgH molecules and enable future optimization of raw read depth. Successful assembly was strongly dependent on read coverage, which itself showed a strong positive correlation between the *Uncut* and *Tn5*_A/*Tn5_B* libraries (Fig. 2A, 2B). TMIseq assembly success increased from 15% for IgH molecules covered by only one *Uncut* read pair to 60–70% for reads covered by five or more *Uncut* read pairs (Fig. 2C). The assembly success of individual molecules was highly dependent on *Tn5_A* and *Tn5_B* coverage, reaching >90% for IgH molecules covered by >40 combined *Tn5_A* and *Tn5_B* read pairs (Fig. 2C).

Next, we performed rarefaction analysis to determine the ideal coverage levels required for effective assembly. Whereas subsampling of the *Tn5_A* and *Tn5_B* raw reads had a strong impact on the number of IgH molecules that were successfully assembled, subsampling of the *Uncut* raw reads had only minimal effect until the number of raw reads fell to <2- to 5-fold the maximum number of assembled IgH molecules (Fig. 2D). A good trade-off between assembled IgH molecules and raw read coverage therefore appears to be 5 *Uncut* raw reads and 30–40 raw reads each for *Tn5_A* and *Tn5_B* for every high-abundance IgH molecule in the *Uncut* library. In comparison with other approaches that enable the sequencing of molecules exceeding the Illumina read length limit, TMIseq already requires far less reads per assembled molecule (13). Furthermore, raw read requirements are likely to be lower when using a HiSeq 2500, as the HiSeq 3000 used in this study has a strong preference for short molecules, which resulted in ∼40% of *Tn5_A* and *Tn5_B* reads to be discarded in a quality filtering step because they were too short or contained adapter sequences (Supplemental Table I).

TMIseq data quality

To assess TMIseq data quality and characteristics, we analyzed IgH molecules assembled from the I1 L1 library. The average length of the assembled IgH molecules was 530 bp (Fig. 3A), and trimmed Tn5_A and Tn5_B reads aligned to the assembled molecules in the pattern expected based on the library preparation protocols (Fig. 3B). Of the 2779 assembled IgH molecules, 98% were identified as H chain transcript and annotated by IgBLAST (20). We then compared these annotated IgH molecules to standard molecular identifier–based immune repertoire control data (I1 control) derived from a biological replicate and produced using a <400-bp amplicon and a 2 × 300 run on a MiSeq (Fig. 4A).

To assess base-exchange errors, we took advantage of IgD sequences that are thought to be expressed almost exclusively by naive B cells. The vast majority of sequenced IgD sequences should therefore be not mutated. Indeed, we found that most IgD

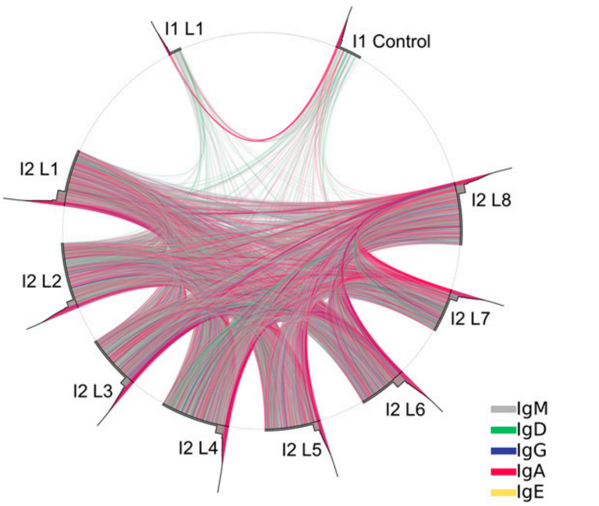

FIGURE 6. TMIseq data identify clonal IgH lineages. Visualization of IgH molecule lineages shared between samples is shown. IgH molecule lineages of each sample are plotted on the circumference of the circle, with the area representing abundance of the respective lineages (logarithmic) and the color representing isotype. IgH molecule lineages present in two time points are connected with lines colored according to their isotype.

sequences were not mutated: 95.16% of IgD sequences in the I1 L1 TMIseq library and 93.6% of IgD sequences in the I1 control library showed >99% identity to reference. Most importantly, the percentage of mutated IgD sequences was comparable between TMIseq and error-corrected control libraries.

Next, we tried to assess the rates of artificial insertions and deletions of the TMIseq protocol, which, as it relies on computational assembly of sequences, might be prone to generate these kinds of errors. First, we analyzed the observed CDR3 length and potential frame-shifts in the V region. Lengths of the CDR3s, which is the result of the random recombination of V, D, and J segments and the addition of quasirandom P and N nucleotides, are expected to occur in steps of three to maintain the reading frame of the Ab H chain transcript. Second, we analyzed indels occurring in the rest of the V region. Indels in the V region should occur in multiples of three to result in the addition or loss of whole amino acids while maintaining the reading frame of the transcripts. We found that the rates of out-of-frame CDR3 (Fig. 4B, 4C) and frame-shift events in the rest of V region (Fig. 4D) were very similar between I1 L1 TMIseq and I1 control libraries. Taken together, this confirmed that the error rate generated by the TMIseq is equivalent to the rates of the error-corrected control protocol (7)

V and C region coverage by IgH amplicon and TMIseq

The increased sequencing length made possible by TMIseq enabled us to create a longer amplicon by priming in the leader exon and 100 bp into the C region. Priming in the leader exon, which is not included in the final Ab protein, allowed us to read every base of the V region without it being covered by a possibly mismatching primer and therefore modified. This enabled us to uniquely identify all V segment alleles. In contrast to the I1 control library, the I1 L1 TMIseq library was able to identify the V segment allele IGHV3-23*04 that differs from the more common IGHV3-23*01 allele by a single base in the first 20 bp of the segment (Fig. 5A). Additionally, priming in the leader exon enables us to identify mutation hot spots in the entire V region, including potential hot spots in the first 20 bases of the IGHV3 segment family (Fig. 5B).

On the other end of the amplicon, priming 100 bp into the C regions creates an amplicon that contains enough distinct base positions to not only distinguish isotypes such as IgM and IgG, but isotype subtypes such as IgG1 and IgG3. Indeed, in contrast to the I1 control Library, the I1 L1 TMIseq library differentiates isotype subtypes, including IgG1, IgG2, and IgG3 as well as IgA1 and IgA2 (Fig. 5C). Although IgG4 and IgE, which are essential for allergy research, were detected at very low levels in the data, this is likely due to the low sequencing depth and their naturally low levels in a mix of IgH transcripts.

Finally, to test the data for obvious recurring assembly artifacts and contaminations, we compared IgH molecules derived from all I1 and I2 libraries. Similar to what we had previously shown for standard immune repertoire data (7), IgH molecule lineages derived from I1 and I2 samples were shared at high levels between the samples of an individual, yet only at very low levels between individuals (Fig. 6), which confirmed the absence of rampant cross-contamination and assembly artifacts.

Discussion

In this study, we show that the TMIseq protocol enables the sequencing of amplicons that exceed Illumina read length but not cluster generation length. With the current state of Illumina technology, this includes amplicons 450–800 bp in length. We applied TMIseq to IgH amplicons to create an immune repertoire sequencing protocol that is unprecedented in its combination of sequencing accuracy and coverage of V and C regions. Indeed, there currently exists no other protocol to accomplish this combination. The high sequencing accuracy of TMIseq paired with the coverage of the complete V region enables the identification of all mutations in an Ab H chain, and increased coverage of the C region makes it possible to also identify the isotype subtype for every sequenced IgH molecule.

The increases in accuracy and coverage provided by TMIseq will be beneficial to the full range of current immune repertoire sequencing–based research efforts. First, we have shown (Fig. 5A) that accurate and complete V region coverage enables the identification of highly similar but distinct V segment alleles. This accurate distinction of V segment alleles will be essential to both define the true allelic diversity as well as haplotypes of the IG loci, which was previously attempted but was hampered by lower throughput and accuracy of the 454 sequencer (4, 23). Second, as complete V region coverage identifies all mutations in an IgH molecule, TMIseq will enable the construction of highly accurate clonal B cell lineage trees (2, 24), which can further be refined using the isotype subtype information that is provided through the increased C region coverage. These highly accurate and refined clonal lineage trees can then be used to gather information of unprecedented detail on clonal expansion, affinity maturation, and the class-switching process, which are of particular interest in allergy research (25). Third, TMIseq will make it possible to adapt immune repertoire sequencing for the pairing of H and L chains as was recently shown for T cells (26). Fourth, TMIseq will enable the cloning of Ab chains without the need to infer bases that were covered by primers or not sequenced. This could improve current high-throughput methods to pair Ab H and L chains (27), which only provide partial V region coverage.

The TMIseq protocol we present is extremely efficient and easily implemented. The protocol can be completed within a day, and our pooling strategy allows us to minimize the use of the Tn5 enzyme. The sequencing cost per TMIseq sample is lower than sequencing a shorter amplicon, which would omit either V or C region coverage, on a MiSeq 2 × 300 run using molecular identifiers for error correction.

Although we applied the TMIseq protocol to IgH amplicons, it will be easy to adapt the protocol to any other amplicon. TMIseq could, for example, be applied to immune repertoire amplicons prepared using 5′RACE (28). 5′RACE eliminates potential amplification bias introduced by V segment–specific primers, but when paired with C region primers that enable the differentiation of isotype subtypes, it produces amplicons too long for the current MiSeq read length. There are also several amplicon-based applications outside of immune repertoire sequencing for which TMIseq might be beneficial. These applications include, among others, the sequencing of 16S RNA and cancer amplicon panels.

Taken together, the data quality and coverage requirements shown prove that the TMIseq protocol is capable of sequencing full-length immune repertoires, or any other amplicon between 450 and 800 bp, highly accurately and at high throughput.

Acknowledgments

We thank Ed Malloy, Aisha Coons, and Jerrold Michaud of the Student Health Center at the University of California Santa Cruz for expert assistance with blood draws. We thank Rebecca DuBois and Lu Yin for help with expressing the Tn5 enzyme.

Disclosures

C.V. and C.C. have a patent SC2015-974 (482.43) 62/186,152: provisional application filed. The other authors have no financial conflicts of interest.

References

1. Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature* 302: 575–581.
2. Jiang, N., J. He, J. A. Weinstein, L. Penland, S. Sasaki, X.-S. He, C. L. Dekker, N.-Y. Zheng, M. Huang, M. Sullivan, P. C. Wilson, H. B. Greenberg, M. M. Davis, D. S. Fisher, and S. R. Quake. 2013. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* 5: 171ra19.
3. Arnaout, R., W. Lee, P. Cahill, T. Honan, T. Sparrow, M. Weiand, C. Nusbaum, K. Rajewsky, and S. B. Koralov. 2011. High-resolution description of antibody heavy-chain repertoires in humans. *PLoS One* 6: e22365.
4. Kidd, M. J., Z. Chen, Y. Wang, K. J. Jackson, L. Zhang, S. D. Boyd, A. Z. Fire, M. M. Tanaka, B. A. Gaëta, and A. M. Collins. 2012. The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. *J. Immunol.* 188: 1333–1340.
5. Boyd, S. D., E. L. Marshall, J. D. Merker, J. M. Maniar, L. N. Zhang, B. Sahaf, C. D. Jones, B. B. Simen, B. Hanczaruk, K. D. Nguyen, et al. 2009. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Transl. Med.* 1: 12ra23.
6. Logan, A. C., H. Gao, C. Wang, B. Sahaf, C. D. Jones, E. L. Marshall, I. Buño, R. Armstrong, A. Z. Fire, K. I. Weinberg, et al. 2011. 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.
7. Vollmers, C., R. V. Sit, J. A. Weinstein, C. L. Dekker, and S. R. Quake. 2013. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl. Acad. Sci. USA* 110: 13463–13468.
8. Vollmers, C., I. De Vlaminck, H. A. Valantine, L. Penland, H. Luikart, C. Strehl, G. Cohen, K. K. Khush, and S. R. Quake. 2015. Monitoring pharmacologically induced immunosuppression by immune repertoire sequencing to detect acute allograft rejection in heart transplant patients: a proof-of-concept diagnostic accuracy study. *PLoS Med.* 12: e1001890.
9. Glanville, J., T. C. Kuo, H.-C. von Büdingen, L. Guey, J. Berka, P. D. Sundar, G. Huerta, G. R. Mehta, J. R. Oksenberg, S. L. Hauser, et al. 2011. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci. USA* 108: 20066–20071.
10. Meyer, E. H., A. R. Hsu, J. Liliental, A. Löhr, M. Florek, J. L. Zehnder, S. Strober, P. Lavori, D. B. Miklos, D. S. Johnson, and R. S. Negrin. 2013. A distinct evolution of the T-cell repertoire categorizes treatment refractory gastrointestinal acute graft-versus-host disease. *Blood* 121: 4955–4962.
11. Shugay, M., O. V. Britanova, E. M. Merzlyak, M. A. Turchaninova, I. Z. Mamedov, T. R. Tuganbaev, D. A. Bolotin, D. B. Staroverov, E. V. Putintseva, K. Plevova, et al. 2014. Towards error-free profiling of immune repertoires. *Nat. Methods* 11: 653–655.
12. He, L., D. Sok, P. Azadnia, J. Hsueh, E. Landais, M. Simek, W. C. Koff, P. Poignard, D. R. Burton, and J. Zhu. 2014. 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.
13. Hong, L. Z., S. Hong, H. T. Wong, P. P. K. Aw, Y. Cheng, A. Wilm, P. F. de Sessions, S. G. Lim, N. Nagarajan, M. L. Hibberd, et al. 2014. BAsE-Seq: a method for obtaining long viral haplotypes from short sequence reads. *Genome Biol.* 15: 517.
14. Hiatt, J. B., R. P. Patwardhan, E. H. Turner, C. Lee, and J. Shendure. 2010. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods* 7: 119–122.
15. Wu, N. C., J. De La Cruz, L. Q. Al-Mawsawi, C. A. Olson, H. Qi, H. H. Luan, N. Nguyen, Y. Du, S. Le, T.-T. Wu, et al. 2014. HIV-1 quasispecies delineation by tag linkage deep sequencing. *PLoS One* 9: e97505.
16. Lundin, S., J. Gruselius, B. Nystedt, P. Lexow, M. Käller, and J. Lundeberg. 2013. Hierarchical molecular tagging to resolve long continuous sequences by massively parallel sequencing. *Sci. Rep.* 3: 1186.
17. Picelli, S., A. K. Björklund, B. Reinius, S. Sagasser, G. Winberg, and R. Sandberg. 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* 24: 2033–2040.
18. Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.
19. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215: 403–410.
20. Ye, J., N. Ma, T. L. Madden, and J. M. Ostell. 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* 41: W34–W40.
21. Lefranc, M.-P., V. Giudicelli, C. Ginestoux, N. Bosc, G. Folch, D. Guiraudou, J. Jabado-Michaloud, S. Magris, D. Scaviner, V. Thouvenin, et al. 2004. IMGT-ONTOLOGY for immunogenetics and immunoinformatics. *In Silico Biol. (Gedrukt)* 4: 17–29.
22. Hunter, J. D. 2007. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* 9: 90–95.
23. Boyd, S. D., B. A. Gaëta, K. J. Jackson, A. Z. Fire, E. L. Marshall, J. D. Merker, J. M. Maniar, L. N. Zhang, B. Sahaf, C. D. Jones, et al. 2010. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J. Immunol.* 184: 6986–6992.
24. Gupta, N. T., J. A. Vander Heiden, M. Uduman, D. Gadala-Maria, G. Yaari, and S. H. Kleinstein. 2015. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* 31: 3356–3358.
25. Looney, T. J., J.-Y. Lee, K. M. Roskin, R. A. Hoh, J. King, J. Glanville, Y. Liu, T. D. Pham, C. L. Dekker, M. M. Davis, and S. D. Boyd. 2015. Human B-cell isotype switching origins of IgE. *J. Allergy Clin. Immunol.* doi: 10.1016/j.jaci.2015.07.014.
26. Howie, B., A. M. Sherwood, A. D. Berkebile, J. Berka, R. O. Emerson, D. W. Williamson, I. Kirsch, M. Vignali, M. J. Rieder, C. S. Carlson, and H. S. Robins. 2015. High-throughput pairing of T cell receptor α and β sequences. *Sci. Transl. Med.* 7: 301ra131.
27. DeKosky, B. J., T. Kojima, A. Rodin, W. Charab, G. C. Ippolito, A. D. Ellington, and G. Georgiou. 2015. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* 21: 86–91.
28. Tan, Y.-C., S. Kongpachith, L. K. Blum, C.-H. Ju, L. J. Lahey, D. R. Lu, X. Cai, C. A. Wagner, T. M. Lindstrom, J. Sokolove, and W. H. Robinson. 2014. Barcode-enabled sequencing of plasmablast antibody repertoires in rheumatoid arthritis. *Arthritis Rheumatol.* 66: 2706–2715.