Video Article

# VDJ-Seq: Deep Sequencing Analysis of Rearranged Immunoglobulin Heavy Chain Gene to Reveal Clonal Evolution Patterns of B Cell Lymphoma

Yanwen Jiang[1,2], Kui Nie[3], David Redmond[2], Ari M. Melnick[1], Wayne Tam[3], Olivier Elemento[2]

[1]Department of Medicine, Weill Cornell Medical College

[2]Institute for Computational Biomedicine, Weill Cornell Medical College

[3]Department of Pathology and Laboratory Medicine, Weill Cornell Medical College

Correspondence to: Yanwen Jiang at yaj2001@med.cornell.edu, Wayne Tam at wtam@med.cornell.edu, Olivier Elemento at ole2001@med.cornell.edu

URL: https://www.jove.com/video/53215
DOI: doi:10.3791/53215

Keywords: Medicine, Issue 106, Lymphoma relapse, Clonal heterogeneity, Clonal evolution, VDJ rearrangement, Somatic hypermutation, VDJ sequencing

Date Published: 12/28/2015

Citation: Jiang, Y., Nie, K., Redmond, D., Melnick, A.M., Tam, W., Elemento, O. VDJ-Seq: Deep Sequencing Analysis of Rearranged Immunoglobulin Heavy Chain Gene to Reveal Clonal Evolution Patterns of B Cell Lymphoma. *J. Vis. Exp.* (106), e53215, doi:10.3791/53215 (2015).

## Abstract

Understanding tumor clonality is critical to understanding the mechanisms involved in tumorigenesis and disease progression. In addition, understanding the clonal composition changes that occur within a tumor in response to certain micro-environment or treatments may lead to the design of more sophisticated and effective approaches to eradicate tumor cells. However, tracking tumor clonal sub-populations has been challenging due to the lack of distinguishable markers. To address this problem, a VDJ-seq protocol was created to trace the clonal evolution patterns of diffuse large B cell lymphoma (DLBCL) relapse by exploiting VDJ recombination and somatic hypermutation (SHM), two unique features of B cell lymphomas.

In this protocol, Next-Generation sequencing (NGS) libraries with indexing potential were constructed from amplified rearranged immunoglobulin heavy chain (IgH) VDJ region from pairs of primary diagnosis and relapse DLBCL samples. On average more than half million VDJ sequences per sample were obtained after sequencing, which contain both VDJ rearrangement and SHM information. In addition, customized bioinformatics pipelines were developed to fully utilize sequence information for the characterization of IgH-VDJ repertoire within these samples. Furthermore, the pipeline allows the reconstruction and comparison of the clonal architecture of individual tumors, which enables the examination of the clonal heterogeneity within the diagnosis tumors and deduction of clonal evolution patterns between diagnosis and relapse tumor pairs. When applying this analysis to several diagnosis-relapse pairs, we uncovered key evidence that multiple distinctive tumor evolutionary patterns could lead to DLBCL relapse. Additionally, this approach can be expanded into other clinical aspects, such as identification of minimal residual disease, monitoring relapse progress and treatment response, and investigation of immune repertoires in non-lymphoma contexts.

## Video Link

The video component of this article can be found at https://www.jove.com/video/53215/

## Introduction

Cancer is a clonal disease. Since thirty years ago when Peter C. Nowell proposed the cancer clonal evolution model[1], many studies have tried to dissect clonal populations within tumor samples and reconstruct clonal expansion and evolution patterns that underlie the tumorigenesis process[2]. Recently, whole-genome sequencing has enabled investigators to take a deep look at the clonal heterogeneity and evolution[3,4]. However, due to the lack of tractable markers in many cell types, it is difficult to infer the precise clonal architecture and evolutionary path. Fortunately there is a natural clonality marker in mature B cells from which many lymphoid malignancies, including DLBCL, originate. In response to antigen stimulation, each B-cell can form a single productive IgH VDJ sequence by joining a $V_H$ (variable), a D (diversity), and a $J_H$ (joining) segment together from a large pool of these segments. During this process, small portions of the original sequence may be deleted and additional non-templated nucleotides may be added to create a unique VDJ rearrangement. This specific VDJ rearrangement can be inherited in all the progeny of this B-cell, therefore tagging individual mature B-cell and its offspring[5]. Furthermore, SHM occurs on the recombined VDJ sequences in the subsequent germinal center (GC) reaction to introduce additional mutations for the expansion of the antibody pool and the enhancement of antibody affinity[6]. Therefore, by comparing and contrasting VDJ and SHM patterns of lymphoma samples that have undergone these processes, intra-tumor heterogeneity could be delineated and clonal evolution path of the disease may be deduced.

Previously, VDJ rearrangement and SHM could be identified by PCR amplifying the recombined regions, cloning the PCR products, and subsequently Sanger sequencing to obtain sequence information. This approach is low-throughput and low yield, retrieving only a very small portion of the entire recombined VDJ repertoire, and hindering the characterization of the overall representation of the clonal population within a given sample. A modified approach was created by generating NGS indexed sequencing libraries from VDJ PCR products and performing PE 2x150 bp sequencing to obtain more than half a million recombined VDJ sequences per sample. In addition, a custom pipeline was developed to perform quality control (QC), align, filter VDJ sequencing reads to identify rearrangements and SHMs of each read, and perform phylogenetic

 

analysis on the clonal architecture of each sample. In addition, a new approach has been established to further characterize the clonal evolution patterns for samples collected at various disease stages.

We have applied this technique to DLBCL patient samples. DLBCL is an aggressive form of non-Hodgkin lymphoma with frequent relapse in up to one third of the patients[7]. DLBCL relapses normally occur early, within 2 to 3 years of the initial diagnosis, although some do occur after 5 years[8]. Prognosis for relapsed patients is poor, with only 10% achieving 3 year progression-free survival due to limited treatment options. This is the basis to the urgent need for novel approaches to treat DLBCL relapse[9,10]. However, molecular mechanisms associated with DLBCL relapse are still largely unknown. Particularly, the role of clonal heterogeneity at diagnosis and clonal evolution during DLBCL relapse development are presently uncharacterized, making it difficult to define an accurate and useful biomarker to predict relapse. To address these questions, we applied our VDJ-sequencing approach on multiple pairs of matched primary diagnosis-relapse DLBCL sample pairs. Two distinct clonal evolutionary scenarios of relapse emerged from the comparison of the clonal architectures between the diagnosis and relapse samples that suggests multiple molecular mechanisms may be involved in DLBCL relapse.

## Protocol

# 1. VDJ Amplification

## 1.1) DNA Extraction from Tumor Samples

1. Extract DNA from thin sections (10-20 µm) of frozen O.C.T. embedded normal or malignant tissues.
    1. Digest 10-30 thin sections of embedded tissue cut by a cryostat microtome in 4 ml Nucleic Lysis Buffer (0.0075 M Tris HCl, pH 8.2; 0.3 M NaCl; 0.002 M $Na_2EDTA$) with Proteinase K (0.5 mg/ml, final concentration) and 0.625% SDS in a 15 ml centrifuge tube in a 37 °C water bath overnight.
    2. Add 1 ml of saturated NaCl (5 M) to the digestion mixture and shake vigorously for 15 sec.
    3. Centrifuge at 1,100 x g for 15 min at room temperature.
    4. Transfer the supernatant to a new 15 ml centrifuge tube and add two volumes of 100% Ethanol.
    5. Mix by inverting the tube 6-8 times. Centrifuge at 5,000 x g for 60 min at 4 °C to collect the precipitated DNA.
    6. Wash the DNA pellet twice with 70% ethanol. Centrifuge at 5,000 x g for 15 min each time to collect the pellet.
    7. Dissolve DNA in 100-400 µl TE buffer at room temperature on a shaker overnight. The DNA yield is 5 to 200 µg (final concentration between 50-500 ng/µl) depending on the size of the tissue
2. Extract DNA from thin sections (10-20 µm) of formalin-fixed, paraffin- embedded (FFPE) normal or malignant tissue.
    1. Incubate paraffin sections in 1 ml xylene twice at room temperature for 10 min each time in a 1.5 ml microcentrifuge tube to de-paraffinize. Collect the tissue sections by spinning at 13,000 x g for 5 min at room temperature.
    2. Incubate sections in 1 ml 100% ethanol twice at room temperature, 10 min each time to remove residue xylenes. Collect the tissue sections by spinning at 13,000 x g for 5 min at room temperature. The paraffin is completely dissolved at this point and only the tissue section is remaining.
    3. Air-dry the sections at room temperature for 10-15 min.
    4. Make a 0.5 mg/ml Proteinase K solution with 1x PCR buffer (diluted from 10x PCR buffer with nuclease-free water). Add the Proteinase K solution to the samples in a final volume of 50-100 µl and incubate overnight at 37 °C.
    5. Heat the samples at 95 °C for 10 min to inactivate Proteinase K.
    Note: At this step, sample DNA is dissolved in the PCR buffer and can be used in Steps 1.2 and 1.3 directly. The DNA yield is 0.5 to 20 µg (final concentration between 10-200 ng/µl) depending on the size of the tissue.

## 1.2) DNA Quality Assessment

1. Mix 0.25 µl Taq DNA polymerase with 45 µl of master mix from the commercial ladder kit in a PCR tube.
2. Add 5 µl DNA prepared from 1.1.1.7 or 1.1.2.5 into the PCR tube and mix well by pipetting up and down for at least 5 times.
3. Use the following conditions to amplify the DNA: 95 °C for 7 min; followed by 35 cycles of 45 sec at 95 °C, 45 sec at 60 °C, and 90 sec at 72 °C; then 72 °C for 10 min and hold at 15 °C.
4. Prepare a 2% agarose gel in TBE (Tris/Borate/EDTA).
5. Mix 20 µl PCR reaction with 4 µl 6x loading dye, and load onto a 2% agarose gel.
6. Stain the agarose gel with 0.5 µg/ml ethidium bromide solution and detect PCR products with a gel imagining system. Note: the samples that yield 5 PCR products at sizes of 100, 200, 300, 400, and 600 bp will be continued to generate VDJ amplicons.
CAUTION: Ethidium bromide is a potent mutagen. Please handle with extreme caution by wearing protective gears, *i.e.* gloves, and dispose into specific containers per institution's guidelines.

## 1.3) VDJ PCR

## 1.3.1) Amplify Recombined IgH VDJ Segment from Framework Region 1 (IgVHFR1)

1. Mix 45 µl master mix from the tube labeled as "Mix 2" of a commercial somatic IGH Hypermutation Assay for Gel Detection kit, 0.25 µl Taq DNA polymerase, and 5 µl sample DNA prepared from 1.1.1.7 or 1.1.2.5 in a PCR tube.
2. Use the following conditions to amplify the DNA: 95 °C for 7 min; followed by 35 cycles of 45 sec at 95 °C, 45 sec at 60 °C, and 90 sec at 72 °C; then 72 °C for 10 min and hold at 15 °C.

3. Resolve the entire PCR product in a 2% agarose gel by electrophoresis.
4. Stain the agarose gel with 0.5 µg/ml ethidium bromide solution and detect PCR products with a gel imaging system. A monoclonal amplicon is expected with the size range of 310-380 bp.
5. Excise the gel portion containing the monoclonal amplicon between 310-380 bp.
6. Purify DNA from excised gel using a standard Gel Extraction Kit according to manufacturer's protocol. Note: for samples that FR1 fragments could be obtained, there is no need to amply IgVHFR2 and IgVHFR3.

## 1.3.2) Amplify Recombined IgH VDJ Segment from Framework Region 2 (IgVHFR2)

1. Mix 45 µl master mix from the tube labeled as "Tube B" of a commercial IGH Gene Clonality Assay for Gel Detection kit, 0.25 µl Taq DNA polymerase, and 5 µl sample DNA prepared from 1.1.1.7 or 1.1.2.5 in a PCR tube.
2. Use the following conditions to amplify the DNA: 95 °C for 7 min; followed by 35 cycles of 45 sec at 95 °C, 45 sec at 60 °C, and 90 sec at 72 °C; then 72 °C for 10 min and leave at 15°C.
3. Resolve the entire PCR product in a 2% agarose gel by electrophoresis.
4. Visualize PCR product(s) by 0.5 µg/ml ethidium bromide staining. A monoclonal amplicon may be observed within the size range of 250-295 bp.
5. Excise the gel portion containing the monoclonal amplicon between 250-295 bp.
6. Purify DNA from excised gel using a standard Gel Extraction Kit according to manufacturer's protocol.

## 1.4) Optimize VDJ PCR

1. Use the following modified PCR conditions to amplify IgVHFR1 and IgVHFR2 using suboptimal DNA: 95 °C for 7 min, 40 cycles of 95 °C for 60 sec, 60 °C for 60 sec, and 72 °C for 90 sec, final extension at 72 °C for 10 min.

# 2. VDJ Amplicon Library Preparation and Sequencing

## 2.1) Library Preparation

## 2.1.1) End-repair

1. Transfer VDJ PCR product from 1.3 into a PCR tube and add Resuspension Buffer from DNA Sample Preparation Kit to bring up the volume to 60 µl.
2. Add 40 µl of End Repair Mix from DNA Sample Preparation Kit and mix thoroughly.
3. Incubate the reaction at 30 °C for 30 min in a pre-heated thermocycler (30 °C) with a pre-heated lid at 100 °C.
4. Mix 136 µl magnetic beads and 24 µl PCR grade water first in a 1.5 ml tube, then transfer the entire End-repair reaction from 2.1.1.3 into the 1.5 ml tube and mix well with the beads solution.
5. Place tubes on a magnetic stand for 2 min to allow the separation of the beads from the solution. Aspirate the supernatant, and wash the beads with 80% fresh prepared EtOH twice while the tube is on the magnetic stand.
6. Aspirate the ethanol solution completely and allow the beads to air dry for 15 min at room temperature for 15 min.
7. Take tube off the magnetic stand and resuspend beads in 17.5 µl Resuspension Buffer.
8. Place tubes back onto the magnetic stand for 2 min to separate beads from the Resuspension Buffer. Remove the Resuspension Buffer (End-repair product is resuspended in the Resuspension buffer now) into a clean PCR tube.

## 2.1.2) A-tailing

1. Add 12.5 µl A-tailing mix into the PCR tube containing End-repair product and mix thoroughly.
2. Incubate the PCR tube at 37 °C for 30 min in a pre-heated thermocycler (37 °C) with a pre-heated lid at 100 °C.

## 2.1.3) Adaptor Ligation

1. Add 2.5 µl Resuspension Buffer, 2.5 µl Ligation Mix, and 2.5 µl DNA Adaptor Index into the A-tailing reaction.
2. Incubate the reaction at 30 °C for 30 minl in a pre-heated thermocycler (30 °C) with a pre-heated lid at 100 °C.
3. Add 5 µl Stop Ligation Buffer into each reaction and mix thoroughly.
4. Mix 42.5 µl well-mixed magnetic beads to clean the reaction following steps 2.1.1.5 to 2.1.1.8. Add 50 µl Resuspension Buffer to elute the adaptor-ligated product.
5. Clean the adaptor-ligated product for a second time by using 50 µl well-mixed AMPure XP beads, and elute the product in 25 µl Resuspension Buffer into a clean PCR tube.

## 2.1.4) Amplify DNA Fragments

1. Add 5 µl PCR Primer Cocktail and 25 µl PCR Master Mix to the PCR tube containing adaptor-ligated product and mix thoroughly.
2. Perform amplification in a pre-programmed thermocycler with the following conditions: 98 °C for 30 sec; 10 cycles of 98 °C for 10 sec, 60 °C for 30 sec and 72 °C for 30 sec; then 72 °C for 5 min and hold at 10 °C.
3. Clean up the reaction by using 50 µl well-mixed magnetic beads, and elute the final product in 30 µl Resuspension Buffer.

## 2.1.5) Library Validation

1. Quantify the final product with a fluorometer using manufacturer's protocol. Final library concentration is between 2.5 to 20 ng/µl.
2. Assess the quality of the final product by using an analytical instrument with High Sensitivity DNA chip according to manufacturer's protocol. Expect a single band with the expected size for either IGVHFR1, IGVHFR2 or IGVHFR3. The expected size of the library product equals the size of the original VDJ amplicon plus the size of the adaptors (~125 bp).

## 2.2) VDJ-PCR Library Pooling and Sequencing

1. Calculate the molarity of each library fraction using the following formula: nM = [(ng/1,000)/bp*660] x $10^9$ where ng is the concentration of the VDJ-PCR library measured in step 2.1.5.1 and bp is the peak size of the VDJ-PCR library measured in step 2.1.5.2.
2. Dilute the library to 2 nM (10 µl total) with DNase-free water.
3. Combine the 10 µl of the diluted 2 nM libraries together into a 1.5 ml tube.
4. Add an equal volume of PhiX spike-in control into the 1.5 ml tube.
5. Load the pool at 7 pM concentration onto a flowcell.
6. Sequence the libraries using a paired-end 150 cycle sequencer according to manufacturer's protocol.

## 3. Data Analysis

Note: A summary of the bioinformatics scripts used in this section can be found as a Supplementary Code File.

## 3.1) Alignment and QC

1. Map paired-end sequencing reads against a human IGH V, D and J region database downloaded from the IMGT website[11] using an adapted nucleotide blast algorithm based upon 'blastn' available from NCBI with gap open penalty of 2, gap extension penalty of 1, word length of 7, and e-value threshold of $10^{-4}$. Discard the read-pair does not map to both an IGH V and a J region.
2. Count the remaining read-pairs that have IGH V and J mappings to obtain the frequencies of matching VJ combinations across all read pairs obtained from the sequencing run on the sample. Count a read-pair if it is mapped to both an IGH V and J gene, and then add its allele to the corresponding count for the particular VJ combination.
3. Rank the counts for each recombined VDJ region obtained from 3.1.2 from the highest to the lowest. The VDJ region has the highest reads count is defined as major rearrangement combination.
4. Discard aligned sequences that cover less than 35% of the domain major rearrangement identified in 3.1.3.

## 3.2) SHM Profile Identification

1. Count all SHM patterns across the reads from step 3.1.3. Define each SHM pattern as a subclone.
2. Count the number of the subclones that are corresponding to different unique SHM patterns, and the number of reads that are mapped to individual subclones.

## 3.3) Graphical Representation of Results

1. Perform the phylogenetic analysis on the subclones using their corresponding SHM patterns using a neighborhood joining method from R package "ape" according to manufacturer's protocol. Calculate a distance matrix from the individual distinct alignments to recreate the subclone phylogeny rooted to the germline sequence of the VJ combination in question for each paired sample set.
2. Use the nucleotide sequence of each subclone as character vector and calculate the approximate string distance between all subclones in the major VJ rearrangement for both diagnosis and corresponding relapse samples using a Levenshtein distance measure where mathematically the distance between two alignments is given by $d_{x,y}(i,j)$ where $d_{x,y}(i,j) = 1$ if $i \neq j$ and 0 otherwise, and then sum this function over the length of the string corresponding to nucleotide sequences i and j.
3. Graphically display the resulting matrix of subclone distances and their corresponding subclone counts in the following two ways:
    1. Use R package "MASS" according to manufacturer's protocol to apply multidimensional scaling to the subclone distance matrix and generate a two dimensional principle coordinates map, which is then plotted along with the logarithm of subclonal counts as the radius of the circle.
    2. Construct an undirected graph based upon the Levenshtein distance, where the vertices correspond to each distinct subclone arising from the major VJ rearrangement.
    Note: If the distance between two distinct clones is equal to one then there exists an edge between these two vertices. If the distance is greater than one, then there is no edge connecting them.
    3. Plot the graph of 3.3.3.2 using the tkplot function in the "iGraph" R package with the Kamada Kawai layout according to manufacturer's protocol. The radius of the vertices is proportional to the cubed root of the total number of clones mapped to it and the shading uses a red and blue range to denote the proportion of clones that either belong to the diagnosis (blue) or relapse (red) samples.

## 3.4) Heterogeneity (Entropy) Measurement

1. Examine the numbers and subclone counts of the individual patterns of somatic hypermutation occurring in the major VJ rearrangement for each paired set of samples.

2. Calculate the empirical entropy and residual empirical entropy adjusted for the number of distinct clones in each sample using the standard histogram based entropy estimation.

## Representative Results

The overall procedure of VDJ sequencing (VDJ-seq), including DNA extraction, recombined VDJ region amplification and purification, sequencing library construction, reads processing, and phylogenetic analysis, is represented in **Figure 1**. Routinely 5-200 µg DNA can be retrieved from frozen solid tissue sections or 0.5-20 µg DNA from formalin-fixed paraffin-embedded tissue sections. Depending on the quality, rearrangement pattern, and SHM degree of individual samples, a variety of VDJ PCR products from different samples could be obtained (**Figure 2**), including IGVHFR1 (310-380 bp), FR2 (250-295 bp), and FR3 (70-120 bp). Samples from which only FR3 fragment can be obtained from VDJ PCR were excluded from the remaining study due to the short product that does not provide sufficient amount of sequence information for downstream data analysis purpose. Only samples of which a major FR1 or FR2 PCR product can be visualized on the agarose gel are processed to generate sequencing libraries. The yield of the major PCR product after gel purification ranges from 10 to 50 ng in total.

The entire purified PCR product was used for the subsequent library construction for sequencing library. After adaptor ligation, each sample obtains its own unique index. During library amplification, 10 cycles of PCR were performed which could yield more than 30 ng as final library products. The library quality was accessed by a bioanalyzer. As shown in **Figure 3**, there is one single product in the library sample with a size shift of ~125 bp form the original VDJ PCR amplicon product indicating the additional of adaptor. For each sequencing run, 5-10 libraries were pooled together at 7 µM. In addition, due to the high sequence similarity between the VDJ PCR products, the library pool was mixed with 50% PhiX spike-in to ensure the complexity of the run and correct identification of base additional after each sequencing cycle. Library DNA fragments were sequenced for 150 bp on both ends, which allows the retrieval of sufficient amount of sequence information spanning VD and DJ junctions for the FR1 fragments, and SHM mutation patterns for phylogenetic analysis.

Previously, we performed VDJ sequencing on 32 DLBCL samples collected at diagnosis and relapse from 14 patients (several patients had multiple samples collected at either diagnosis or relapse stage) using the condition described above[12]. By performing the phylogenetic analysis of the SHM profiles of the major $V_HDJ_H$ rearrangements between the paired samples, an “Early-divergent” and a “Late-divergent” mode of clonal evolution associated with DLBCL relapse were identified[12]. Herein, the same approach was applied to a newly collected DLBCL diagnosis and relapse pair. The FR1 regions were obtained in both samples after PCR amplification. A major PCR product with the size around 344 bp (measured by bioanalyzer) was obtained from both the diagnosis and the relapse samples. After sequencing, a total of 0.70 million paired-end reads were obtained for the diagnosis sample and 0.73 million paired-end reads for the relapse sample, which were within the range of the number of reads obtained per sample in the previous study (0.38-1.42 million, average 0.75 ± 0.26 million)[12]. After mapping paired-end reads against IMGT database, reads that do not have hits in all three V, D and J regions were discarded. $V_HDJ_H$ arrangement was assigned to each paired-end read. A total of 0.64 million and 0.67 million $V_HDJ_H$ junctions were identified in the diagnosis and relapse samples respectively, representing greater than 90% alignment rate. By counting how many times each combination of $V_HDJ_H$ was found in individual samples, 808 distinct $V_HDJ_H$ junctions were found in the diagnosis sample and 581 distinct $V_HDJ_H$ junctions in the relapse sample. The dominant $V_HDJ_H$ junction in both sample is IGHV4-34*2 IGHD5-5*01 IGHJ2*01, confirming that they are clonally related. This dominant $V_HDJ_H$ junction accounted for 97% of the entire mapped $V_HDJ_H$ junction reads in both the diagnosis sample and the relapse sample.

To trace the clonal evolution of this DLBCL relapse case, phylogenetic analysis of the SHM profiles of the major $V_HDJ_H$ rearrangements between this pair of diagnosis and relapse samples was performed. We observed that the clonal evolution pattern of this pair was following the “Late-divergent” path (**Figure 4**), that the subclones from the diagnosis and relapse tumors clustered together on the same branch of the phylogenetic tree, and the dominant diagnosis subclone and the dominant relapse subclones shared similar SHM pattern with a few additional mutations occurred in the relapse subclone. These results suggest that potentially there was a subclone in the diagnosis tumor that was either chemo-resistant or escaped from the treatment, and then eventually developed into the relapse tumor.

To further characterize and visualize clonal architecture of each samples and clonal evolution patterns during relapse process, two new analyses have been developed. In these analyses, the entire set of subclones of the major VJ rearrangement in each sample were used to calculate the pairwise string distance between all subclones as defined by their patterns of SHM. Then by using multi-dimensional scaling (**Figure 5A, 5B**) or by building an undirected graph (**Figure 5C, 5D**) as described in Data Analysis, plots representing the distribution and heterogeneity of subclones were created. As illustrated in **Figure 4**, the MDS plot exhibits far less sequence diversity between major subclones in the late divergent cases (**Figure 5A**) than that in the early divergent cases (**Figure 5B**). Furthermore, the sub-clones of the diagnosis sample and the sub-clones of the relapse sample in the early divergent case separate completely from each other, indicating the lack of similarity of the SHM patterns between these tumors even though they carry the same VDJ rearrangement. Likewise, the undirected graphs in (**Figure 5C**) and (**Figure 5D**) show that the distribution of subclone counts in the late divergent case (**Figure 5C**) differs from the early divergent (**Figure 5D**). The latter has more prominent but genetically diverse major clones from diagnosis to relapse. In addition, the lack of edges and smaller subclones connections between the major diagnosis and relapse clones in the early divergent case (**Figure 5D**), demonstrate the sequence diversity nature of the early divergent relapse case.

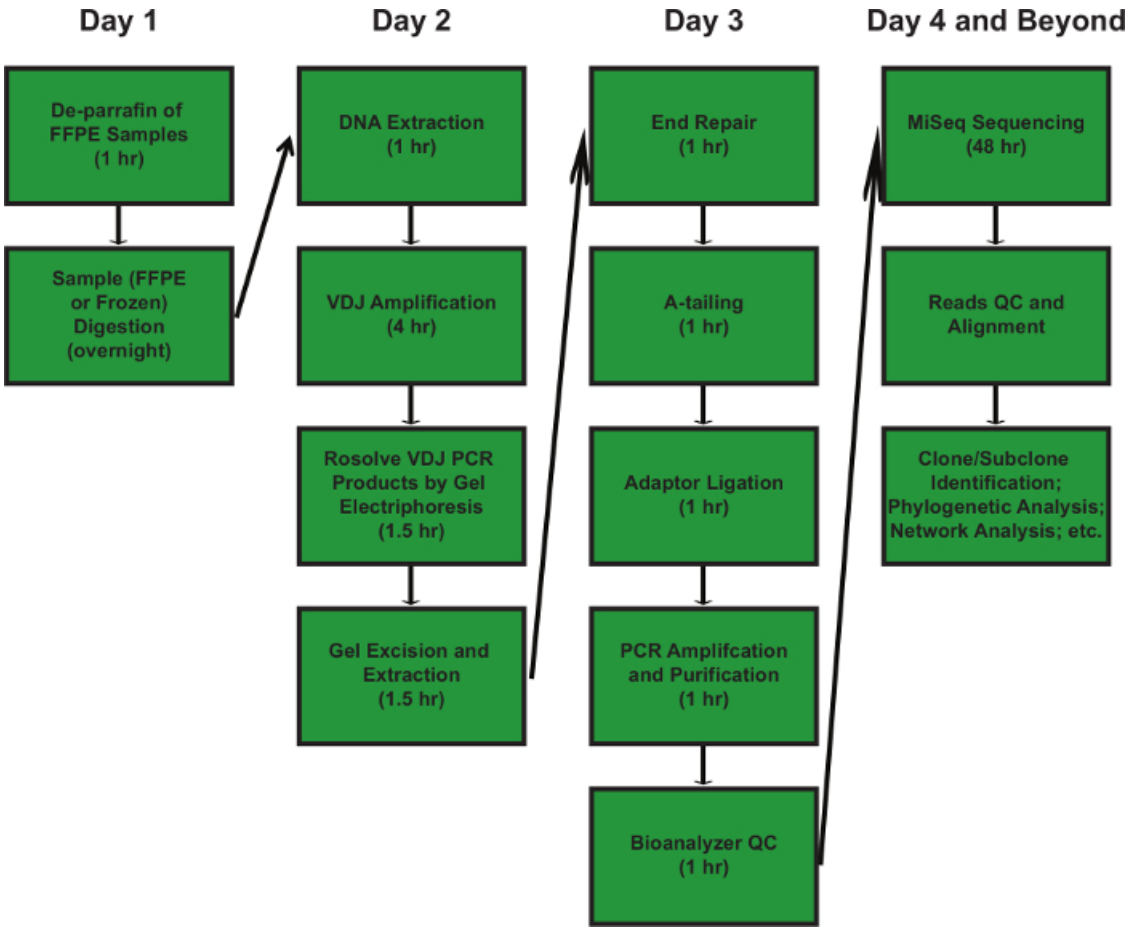


**Figure 1: Step by step flow chart of the VDJ-seq approach.**The overall procedure of VDJ sequencing is shown. Please click here to view a larger version of this figure.

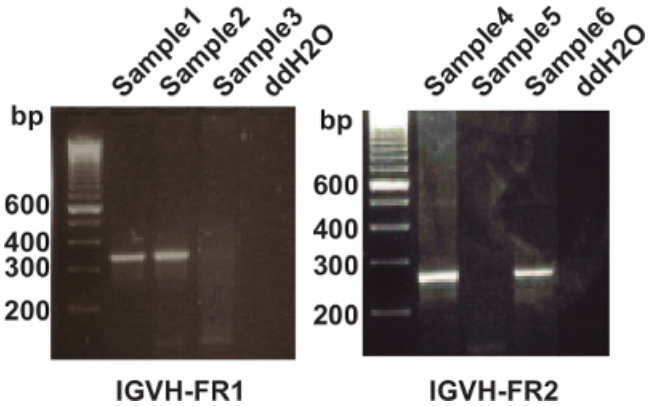


**Figure 2: Representative images of VDJ PCR amplicon products**. Gel images showing the representative IGVH-FR1 (left panel) and IGVH-FR2 (right panel) PCR products from DNA extracted from lymphoma patient samples. A PCR product could not be obtained in Sample 5 probably due to either the low sample DNA quality or SHM at primer sites that impaired the PCR reaction. Please click here to view a larger version of this figure.

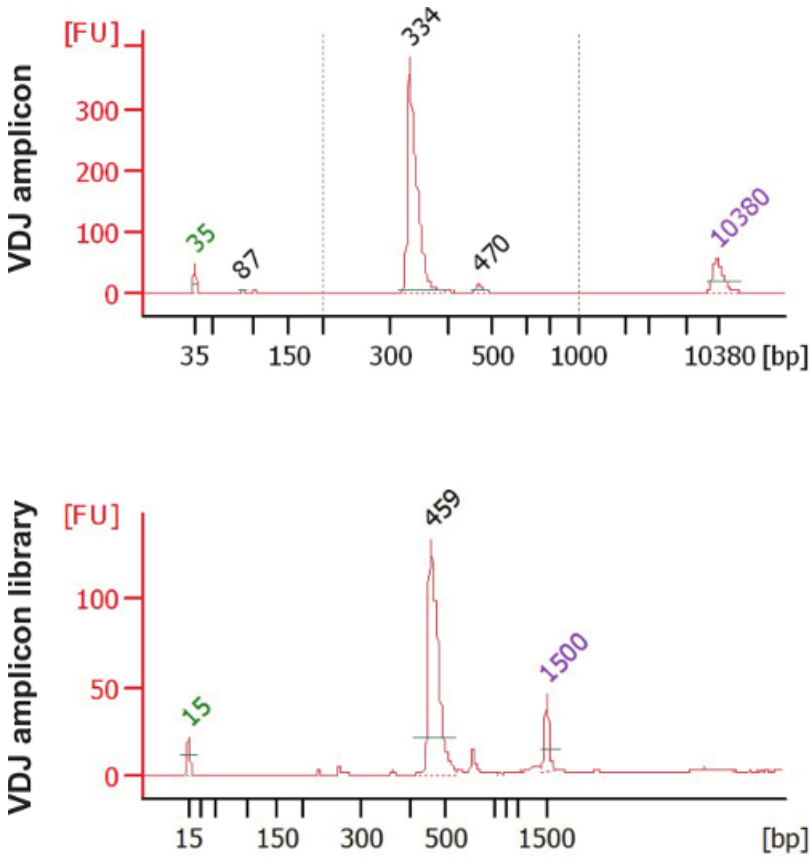


**Figure 3: QC of VDJ PCR amplicon and subsequent sequencing library by bioanalyzer.** Representative Bioanalyzer traces of the same VDJ amplicon before library construction (top panel) and after library construction (bottom panel). Note the size shift from 334 bp to 459 bp indicating the addition of the adaptor. Please click here to view a larger version of this figure.

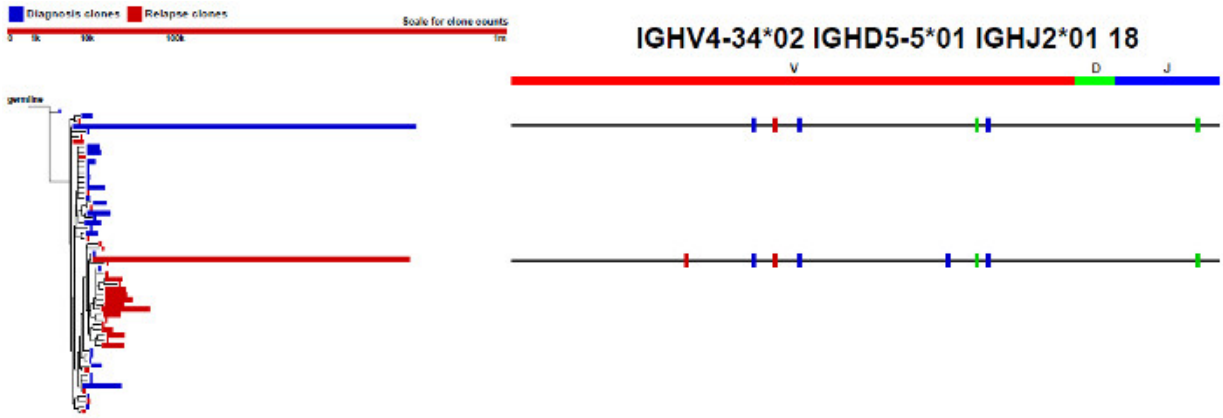


**Figure 4: Clonal evolution analysis of a pair of DLBCL diagnosis and relapse samples.** Phylogenetic analysis of a diagnosis-relapse DLBCL sample pair showing the "Late-divergent" clonal evolution mode. The color tickers represent the mutation status. Please click here to view a larger version of this figure.

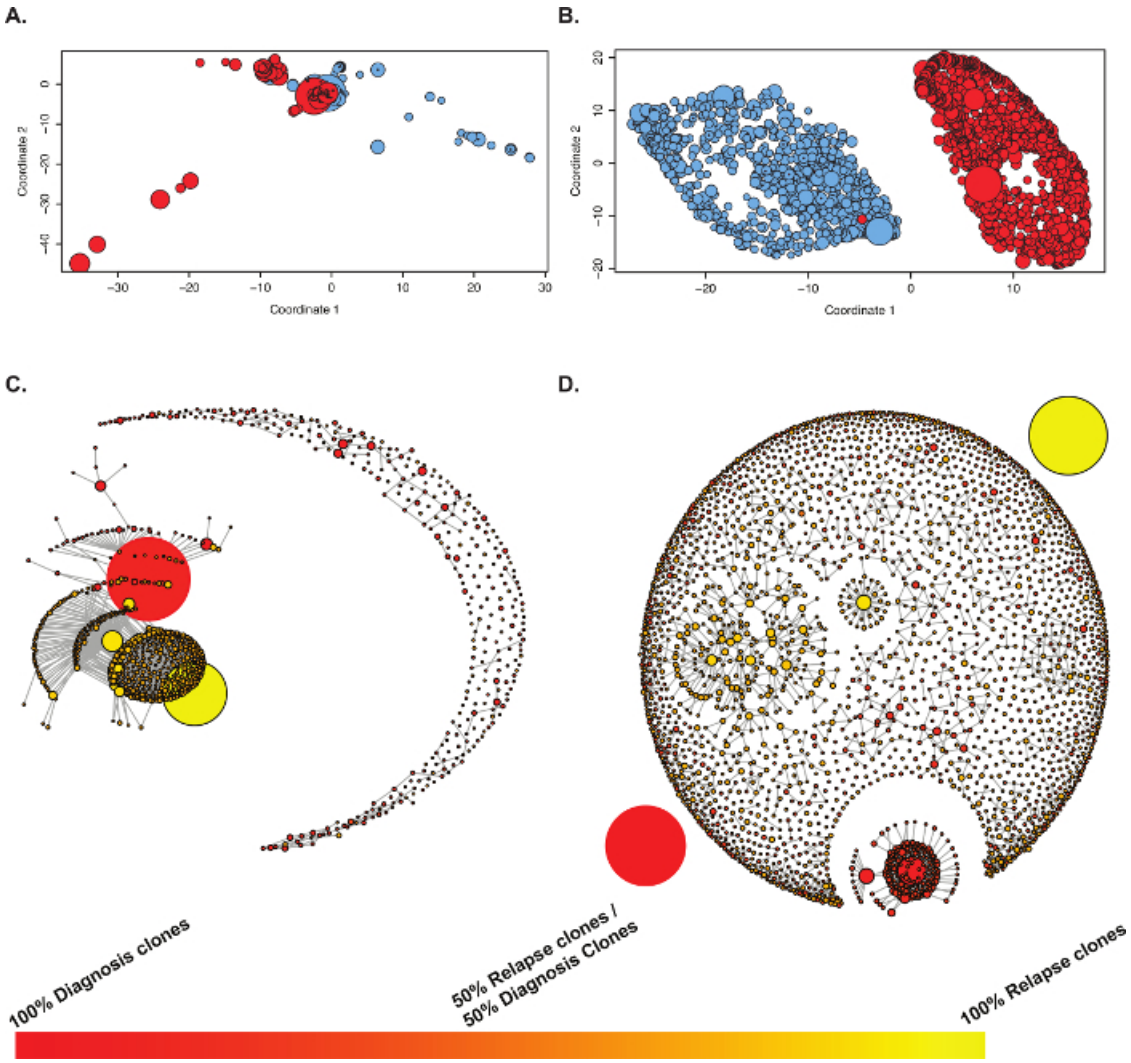


**Figure 5: New graphical methods to visualize clonal evolution patterns.** (**A**, **B**) Multidimensional scaling plots integrating SHM profiles and subclone counts of sample pair A showing the late-divergent relapse mode (**A**) and sample pair B showing early-divergent relapse mode (**B**). The radius of the circles indicate the count of subclones corresponding to a particular SHM profile and colors corresponding to the diagnosis (blue) sample or relapse (red) sample. The X- and Y-axes represent the top 2 components of MDA. (**C**, **D**) Undirected graphs of sample pair A showing the late-divergent relapse mode (**C**) and sample pair B showing early-divergent relapse mode (**D**). Edges correspond to two SHM profiles having a string distance of one letter separation and graduated shading (color scale bar) indicating the proportion of sequences mapping to a particular SHM profile corresponding to the diagnosis (red) sample or relapse (yellow) sample. Please click here to view a larger version of this figure.

## Discussion

Because of the nearly unlimited number of iterations of sequence information coded by VDJ rearrangement and SHM at the IGH locus of human B cells, examination of the entire IGH repertoire by high-throughput deep-sequencing proved to be an efficient and comprehensive way to delineate clonal and sub-clonal B-cell populations. Furthermore, this strategy can be used to study the clonal evolution path of B cell tumor development, remission, and relapse by comparing the clonal and sub-clonal architectures of patient samples collected along various stages of the disease. Although amplification of rearranged VDJ sequences from primary sample DNA is readily performed in a laboratory setting using commercially available primers, the efficiency of amplification is largely dependent on the quality of sample DNA. In particular, DNA extracted from FFPE samples exhibits low amplification efficiency, and often only the small FR3 fragment can be successfully amplified from these samples, making them unsuitable for the subsequent analysis. Since our study described herein was designed to understand the clonality of B cell lymphoma, which is a clonal disease with one or several major clones dominating the entire tumor, we only collected and sequenced the major VDJ-PCR product. For studies interested in cataloging the entire B cell clonal population of, for example, peripheral blood, a larger portion of the gel that contains multiple VDJ-PCR products (multiple bands or a smear on the gel) may be excised and sequenced. However, it is important to point out that it is not possible to capture the entire rearranged IGH repertoire because certain SHM may occur at where primers target and impair the amplification of these VDJs. Moreover, recently, a commercial assay directly coupling IGH rearrangement amplification and MiSeq sequencing has been introduced. This assay streamlines sample preparation process. However, since not every tumor samples would be able to generate the FR1 fragment, we still recommend including the gel electrophoresis step to check the product of the PCR reaction before pooling samples for sequencing.

Sequencing 150 bp of both ends of the VDJ fragment (total of 300 bp sequence information) has the following advantages: 1) 300 bp sequencing can cover the majority of the FR1 and the entire FR2 fragment, providing ample sequence information for identifying VDJ rearrangement and somatic hyper-mutations; 2) The platform provides fast sequencing turn-around time that completes the entire 300 cycles of VDJ-sequencing under 48 hr; 3) Each run allows multiplexing up to 12 samples and still achieves around one million paired-end reads per sample output. Because the high sequence similarity between clones carrying the same VDJ rearrangement, especially in B cell lymphoma samples, it is critical to spike-in 20-50% PhiX during the sequencing run to allow accurately define base calling matrix. Recently, the MiSeq software has been updated to address this question. It is recommended to use as little as 5% PhiX spike-in for low complexity library sequencing. However, our sequencing facility has experienced inconsistent concentration of the PhiX spike-in control DNA provided by the manufacturer that would compromise the outcome of low complexity sequencing. Therefore, in current practice, we still use 10-20% PhiX spike-in for the VDJ sequencing. For sequencing of the entire peripheral blood B-cell repertoire, which normally contains a large number of different VDJ arrangements, it is possible to reduce the amount of PhiX spike-in. Although the number of clones and subclones identified of each sample is positively correlated with the number of reads sequenced, half to one million of paired-end reads per sample is sufficient to compare clonal structures between samples and deduce clonal evolution patterns between samples collected at different disease stages of the same patient[12]. It is possible that PE 150 bp sequencing might not achieve sufficient coverage of those long FR1 fragments (*i.e.* 350-380 bp). In some cases, sequence information of the D segment might not be collected or there might not be enough sequence information on the V segment to differentiate subclones by SHMs. However, with the advance of the sequencing technologies, it has become possible to sequence longer and cover the entire FR1 amplicon.

Through the sequencing of the IGH repertoire of B-cells it was possible to track the expansion of individual B cell clones and infer clonal evolution over the course from diagnosis to relapse. With the use of analysis on the populations of subclones generated by SHM patterns, we were able to differentiate two modes of cancer progression to relapse by examining their phylogenies. The latest graphical representations using multidimensional scaling plots and undirected graphs based on the Levenshtein distance further allow us to understand 1) the subclonal composition of each tumor, 2) how individual subclones are related to one another, and 3) how each subclone evolves during disease progression. Furthermore, statistical analysis showed that these phylogenetic structures were clearly different and that the subclonal populations differed as regards to their heterogeneity. Finally, the clonal evolution trajectory obtained by VDJ-sequencing can be well correlated with genetic evolution and characterized by either exome or targeted resequencing[12]. Therefore, combining these two approaches has the potential to identify driver mutations during lymphomagenesis and lymphoma relapse

In addition to defining the B cell IGH repertoire and trace clonal evolution of lymphoma progression, the VDJ-seq approach can be potentially used as a highly sensitive method for tracking minimal residue disease, monitoring tumor response to therapies, and potentially identifying the presence of drug-resistant clones. A similar approach can also be applied to T cells to map T cell receptor repertoire, and may be used to probe immune surveillance response to cancer. Moreover, VDJ-seq can also be performed in murine models using primers available from Hanna *et al.*[13] It is foreseeable that information collected through this approach in the next few years can result in a better understanding of the dynamics of tumor development as well as expand our knowledge of the immune system and its roles in defending our body from tumor invasion.

## Disclosures

The authors have no conflicts of interest to disclose.

## Acknowledgements

The authors would like to thank Dr Rita Shaknovich and members of Elemento lab, Melnick lab, and Tam lab for thoughtful discussions. We would also like to thank the Genomics Resources Core Facility at Weill Cornell Medical College for performing the VDJ-sequencing. YJ was supported by ASH Scholar Award. WT and OE are supported by Weill Cornell Cancer Center Pilot Grant. OE is supported by the NSF CAREER award, the Starr Cancer Consortium and the Hirschl Trust. We would also like to thank Katherine Benesch, JD, MPH for her generous support to this project.

## References

1. Nowell, P. C. The clonal evolution of tumor cell populations. *Science.* **194** (4260), 23-28 (1976).
2. Greaves, M., & Maley, C. C. Clonal evolution in cancer. *Nature.* **481** (7381), 306-313 (2012).
3. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature.* **481** (7382), 506-510 (2012).
4. Jiao, W., Vembu, S., Deshwar, A. G., Stein, L., & Morris, Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics.* **15** 35 (2014).
5. Schatz, D. G., & Ji, Y. Recombination centres and the orchestration of V(D)J recombination. *Nature reviews. Immunology.* **11** (4), 251-263 (2011).
6. Jacob, J., Kelsoe, G., Rajewsky, K., & Weiss, U. Intraclonal generation of antibody mutants in germinal centres. *Nature.* **354** (6352), 389-392 (1991).
7. Coiffier, B. *et al.* CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. *N Engl J Med.* **346** (4), 235-242 (2002).
8. Larouche, J. F. *et al.* Lymphoma recurrence 5 years or later following diffuse large B-cell lymphoma: clinical characteristics and outcome. *J Clin Oncol.* **28** (12), 2094-2100 (2010).
9. Friedberg, J. W. Relapsed/refractory diffuse large B-cell lymphoma. *Hematology Am Soc Hematol Educ Program.* **2011** 498-505 (2011).
10. Gisselbrecht, C. *et al.* Salvage regimens with autologous transplantation for relapsed large B-cell lymphoma in the rituximab era. *J Clin Oncol.* **28** (27), 4184-4190 (2010).

11. Lefranc, M. P. IMGT, the International ImMunoGeneTics Information System for Immunoinformatics : methods for querying IMGT databases, tools, and web resources in the context of immunoinformatics. *Molecular biotechnology.* **40** (1), 101-111 (2008).
12. Jiang, Y. *et al.* Deep sequencing reveals clonal evolution patterns and mutation events associated with relapse in B-cell lymphomas. *Genome biology.* **15** (8), 432 (2014).
13. Hanna, J. *et al.* Direct reprogramming of terminally differentiated mature B lymphocytes to pluripotency. *Cell.* **133** (2), 250-264 (2008).

 