

**Manuscript ready for Production**

1  
2  
3  
4  
5  
6  
7  
8

**Editorial summary:** IgSeqR is a bioinformatic pipeline for de novo assembly and characterization of the tumor immunoglobulin variable and constant region transcripts from RNA sequencing data. Immunoglobulin analysis can inform the cell of origin and predict clinical outcomes in B-cell cancers.

**Referee Statement:** Nature Protocols thanks Yang Cao, David Angeletti and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Type	Number	Filename	Legend or Descriptive Caption
Supplementary Table	Supplementary Tables 1 - 4	Supplementary Tables.xlsx	<p>Each type of file (Table, Video, etc.) should be numbered from 1 onwards. Multiple files of the same type should be listed in sequence, i.e.: Supplementary Video 1, Supplementary Video 2, etc.</p> <p>Whole original file name including extension. i.e.: <i>Smith_Supplementary_Video_1.mov</i></p> <p>Describe the contents of the file</p> <p><b>Supplementary Table 1:</b> A comparison of the success rate of the Immunoglobulin heavy chain variable gene (IGHV) using IgSeqR tool vs Sanger sequencing in cases with varying tumour infiltrations</p> <p><b>Supplementary Table 2:</b> A comparison between RNA-seq based Immunoglobulin Gene analysis tools, IgSeqR MiXCR and TRUST4. Alignments were carried out using the IMGT Program version 3.6.1 and reference directory verison 202330-1 (F+ORF+ in-frame P), Indels were assessed and no nucleotide trimming was performed or Single Chain FV</p>

			<p>Fragment assessment.</p> <p><b>Supplementary Table 3:</b> A comparison of the runtimes of three RNA-seq based Immunoglobulin gene analysis tools IgSeqR, MiXCR and TRUST4 run on 18 chronic lymphocytic leukaemia samples with high tumor infiltration</p> <p><b>Supplementary Table 4:</b> A comparison of the runtimes of three RNA-seq based Immunoglobulin gene analysis tool IgSeqR on chronic lymphocytic leukaemia (CLL) samples with high tumor infiltration and diffuse large B cell lymphoma (DLBCL) with unconfirmed tumor infiltration</p>
--	--	--	---

1

2

3 **Identification, assembly, and characterization of tumor Immunoglobulin**  
4 **transcripts from RNA sequencing data using IgSeqR**

5

6 Dean Bryant\*,<sup>1</sup> Benjamin Sale\*,<sup>1,2</sup> Giorgia Chiodin,<sup>1</sup> Dylan Tatterton,<sup>1</sup> Benjamin  
7 Stevens<sup>1</sup>, Alyssa Adlaon<sup>1</sup>, Erin Snook<sup>1</sup>, James Batchelor,<sup>1,2</sup> Alberto Orfao,<sup>3</sup> Francesco  
8 Forconi<sup>1,4</sup>

9 <sup>1</sup>Cancer Sciences, Faculty of Medicine, University of Southampton, Southampton, UK.

10 <sup>2</sup>Clinical Informatics Research Unit, University of Southampton, Southampton, UK.

11 <sup>3</sup>Cancer Research Center (IBMCC, USAL-CSIC), Cytometry Service (NUCLEUS),  
12 Department of Medicine, Biomedical Research Institute of Salamanca (IBSAL),  
13 University of Salamanca, Salamanca, Spain. <sup>4</sup>Haematology Department, Cancer Care  
14 Directorate, University Hospital Southampton NHS Trust, Southampton, UK.

15 \*These authors have contributed equally.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38

**Correspondence:** Francesco Forconi, [f.forconi@soton.ac.uk](mailto:f.forconi@soton.ac.uk).

**Key points:**

- IgSeqR is a bioinformatic pipeline for de novo assembly, identification and characterization of B-cell receptor immunoglobulin (IG) transcripts from tumor RNA-sequencing data, which can reveal B-cell receptor structure and provide insight into clinical outcomes.
- This approach is faster and less labor-intensive than the traditional Sanger sequencing-based method for IG gene analysis and achieves longer IG transcripts in less time compared to other RNA-seq-based methods.

**Key reference:**

Chiodin, G. *et al. Blood* **138**, 1570-1582 (2021): <https://doi.org/10.1182/blood.2021012052>

**[H1] Abstract**

Immunoglobulin (IG) gene analysis provides fundamental insight into B-cell receptor structure and function. In B-cell tumors, it can inform the cell of origin and predict clinical outcomes. Its clinical value has been established in the two main types of chronic lymphocytic leukemia, which are distinguished by the expression of unmutated or mutated Immunoglobulin heavy-chain variable region (*IGHV*) genes, and is emerging in other B-cell tumors. The traditional PCR- and Sanger sequencing-based techniques for *IG* gene analysis are labor-intensive and rely on attaining either a dominant sequence or a small number of subclonal sequences. Extraction of the expressed tumor *IG* transcripts using high-throughput RNA sequencing (RNA-seq) can be faster and allow the collection of the tumor *IG* sequence, and match this with the rest of the RNA-seq data. Analytical tools are regularly sought to increase the accuracy, depth, and speed of acquisition of the *IG transcript* sequences and combine the *IG* characteristics with other tumor features. We provide here a user-friendly protocol for the rapid (~1 hr) *de novo* assembly, identification, and accurate characterization of the full (leader to constant region) tumor *IG* templated and non-templated transcript sequence from RNA-seq data (<https://github.com/ForconiLab/IgSeqR>). The derived amino acid sequences can be interrogated for their physico-chemical characteristics and, in certain lymphomas, predict tumor glycan types occupying acquired N-glycosylation sites. These features will then be available for association studies with the tumor transcriptome. The resulting information can also help refine diagnosis, prognosis, and potential therapeutic targeting in the most common lymphomas.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

## [H1] Introduction

The B-cell receptor (BCR) immunoglobulin (IG) glycoprotein is the defining functional feature of a mature B cell, and *IG* gene analysis can provide fundamental insight into the origin and behavior of a B-cell tumor.<sup>1,2</sup> The IG glycoprotein is a Y-shaped dimer of 2 identical heavy and light chains, with 2 main functional components: a variable region that confers diversity to recognize different antigens and is unique to each B cell, and a constant region with an effector function. *IG* diversity results from a series of genetic recombination events at the *IG* heavy (*IGH*) and kappa (*IGK*) or lambda (*IGL*) light chain loci during B cell development in the bone marrow before a naïve B cell exits to the periphery (**Box 1**). For the heavy chain, the recombination events are accompanied by non-templated nucleotide additions and deletions at the junctions of one of ~51 *IGHV*, ~21 *IGHD*, and ~6 *IGHJ* genes in the complementarity-determining region 3 (*CDR3*), forming the “fingerprint” of an individual B cell, or its clonal expansion as it happens in B-cell tumors. Further variability is conferred by the recombination of a *V* gene with a *J* gene at the *IGK* or *IGL* loci. Following antigen encounter, naïve B-cells undergo class-switch recombination from *IGHM/D* to one of the downstream *IGHG1-4*, *IGHA1-2*, or *IGHI* constant region genes at the *14q32* locus, and somatic hypermutation of the *IG* variable region, typically in the presence of activation-induced cytidine deaminase (AID), T cells, and cytokines, for affinity maturation in the germinal center (GC) and subsequent differentiation into memory B cells or plasma cells.<sup>3</sup> The GC reaction involves proliferation, which makes the B cells vulnerable to damage and transformation into tumors. Tumor B cells preserve the *IG* sequence of the cell of origin. Therefore, analysis of the *IG* sequences allows the identification of the stage of differentiation reached by a B-cell before tumor transformation.<sup>4-6</sup>

*IG* analysis of chronic lymphocytic leukemia (CLL) has revealed two major CLL types defined by *IGHV* mutational status.<sup>5</sup> The CLL type with unmutated *IGHV* (U-CLL) derives from pre-germinal center CD5<sup>+</sup> B cells, while the CLL type with mutated *IGHV* (M-CLL) appears to arise from post-follicular CD5<sup>+</sup> B cells.<sup>7,8</sup> Since the discovery that U-CLL has a worse prognosis than M-CLL,<sup>9,10</sup> subsequent studies have demonstrated that each CLL type has a distinctive cellular origin, biology, genetic and epigenetic signatures, clinical prognosis, and response to therapy.<sup>5,11</sup> *IGHV* gene analysis has become an essential part of the diagnostic workup for any patient with CLL.

*IG* analysis also informs key tumor-specific features in certain lymphomas, such as the presence of acquired N-glycosylation sites (AGS). In classic follicular lymphoma (FL), the tumor *IG* acquires these sites, defined by the asparagine-X-serine/threonine motif (where X is any amino acid except proline).<sup>12</sup> AGS in FL are typically located in the CDRs of the variable region by somatic hypermutation,<sup>13</sup> and are occupied by tumor-specific oligomannose-type glycans.<sup>14-16</sup> These atypical glycans are unique to the tumor B cell, are present on the entire FL clone, and persist during the entire clonal history of FL through transformation into diffuse large B-cell lymphoma (DLBCL), despite ongoing somatic hypermutation.<sup>14,17</sup>

The current gold standard for *IG* gene analysis is Sanger sequencing. This approach offers a highly accurate *IG* sequence but is time-consuming, labor-intensive, and requires a dedicated experimental and analytical workflow on samples with documented high tumor infiltration.<sup>18</sup> The increasing adoption of high-throughput whole transcriptome RNA sequencing (RNA-seq) methods allows many features to be

1 detected in a single experiment. Through the application of appropriate analytical  
2 pipelines, a single RNA-seq experiment can yield comprehensive information on gene  
3 expression, isoform expression, single nucleotide polymorphisms (SNPs), and larger  
4 structural variants.<sup>19</sup>

5 RNA-seq can therefore be an attractive alternative to Sanger sequencing in *IG* gene  
6 analysis. However, the intrinsic high variability of the non-templated CDR3 sequences  
7 has been a challenge to the identification of the tumor *IG* sequence with the current  
8 RNA-seq analytical workflows, which have involved mapping reads to a reference  
9 transcriptome.

10 Here we describe IgSeqR (pronounced I-G-Seeker), a protocol for the reference-free  
11 extraction, identification, and accurate determination of B-cell tumor-derived full *IGHV*-  
12 *IGHD-IGHJ* transcript sequence and *IGHC* (sub)class (*M*, *D*, *G1-4*, *A1-2*, or *E*) from  
13 whole transcriptome RNA-seq data.

#### 14 [H2] *Development of the protocol*

15 We first used IgSeqR to identify the tumor *IG* transcripts in a cohort of 489 DLBCL  
16 patient samples with publicly available RNA-seq data<sup>14</sup>. The data were deposited in  
17 the National Cancer Institute (NCI) Genomic Data Commons (accession  
18 phs001444.v1.p1).<sup>20,21</sup> The full *IGHV-IGHD-IGHJ* sequence rearrangements were  
19 identified from leader to constant region with high confidence in 339 (69%) samples,  
20 from which we could determine *IGHV*, *IGHD*, *IGHJ*, and *IGHC* use, homology to  
21 germline, and AGS presence and location. Since we were interested in those cases  
22 with N-glycosylation sites acquired by somatic hypermutation and no information was  
23 available on the tumor purity of these samples, we investigated only the 307 samples  
24 with mutated (<98% homology to germline) *IGHV* [14]. We found that the AGS were  
25 preferentially in the EZB (enriched for *EZH2* mutations and *BCL2* translocations)  
26 genetic subtype of the GC-B-cell-like (GCB) DLBCL. The majority of these AGS were  
27 located in the CDR, similarly to our observations in FL.<sup>12</sup> Following the generation of  
28 fragment antigen binding (Fab) glycoprotein from the tumor-derived *IG* heavy and light  
29 chain sequences, we documented that the glycan structure occupying the AGS was  
30 location-dependent and that the oligomannose-type glycans occupied the CDR-  
31 located sites only. We performed correlation studies with the transcriptome profile and  
32 defined genes and gene sets differentially expressed in samples with and without  
33 AGS. We performed correlations with the clinical characteristics of the DLBCL.  
34 Interestingly, we found that AGS in the EZB subtype conferred a poor prognosis,  
35 indicating that this approach for *IG* gene analysis could be adopted to predict both  
36 glycan structure and response to conventional therapies<sup>14</sup>. In this protocol, we present  
37 the IgSeqR script while further validating its accuracy in 18 primary CLL samples with  
38 matched *IG* heavy chain Sanger and bulk RNA-seq data (deposited in ArrayExpress,  
39 accession E-MTAB-12017).<sup>22</sup>

40

1 *[H2] Applications of the method*

2 IgSeqR is ideal for studies requiring high-quality base calls across the full sequence,  
3 including the non-templated CDR3 region, of the *IG* heavy and light chains of any  
4 mature B cell tumor. The protocol reduces the computational burden of *de novo*  
5 assembly by pre-filtering redundant data and allows the identification of the dominant  
6 nucleotide sequence of the *IG* heavy and light chains from leader to constant region  
7 from RNA-seq data. Through the alignment to the most updated *IG* sequence  
8 repertoires, currently the ImMunoGeneTics (IMGT) information system IMGT/V-  
9 QUEST reference directory 202349-3, program version 3.6.2 at <http://www.imgt.org>, it  
10 is possible to obtain insights into *IGHV*, *IGHD*, *IGHJ* heavy chain alleles, *IGKV*, *IGKJ*  
11 or *IGLV*, *IGLJ* light chain alleles, constant region class and subclass, homology to  
12 germline, CDR1-3 and FR1-4 characteristics. Due to the short time required for the  
13 identification of both the tumor *IG* heavy and light chains of multiple samples, we have  
14 already exploited IgSeqR to rapidly generate lymphoma-derived F(ab)s in large scale  
15 and use them for glycopeptide mass spectrometry and/or screening of antibodies  
16 specifically binding those F(ab)s.<sup>14</sup> These opportunities suggest that IgSeqR can also  
17 be exploited to improve screening tools for vaccine and anti-idiotypic antibody therapy  
18 development.<sup>23-25</sup>

19 Future work is planned to develop further the existing protocol and evaluate its efficacy  
20 for deriving smaller, less dominant, clonal populations. When the tumor *IG* sequence  
21 is already known, we will apply this approach for the determination of the minimal  
22 residual disease in repeat samples following anti-cancer therapy. We will also further  
23 investigate the protocol's potential use with RNA-seq data generated from FFPE  
24 material (for example the RNA preparation protocol – refer to the section  
25 “Experimental design”). There are intrinsic limitations of RNA-seq data quality from  
26 FFPE (for example RNA quantity and fragmentation), and areas where optimization or  
27 adaptation may be necessary will need to be identified. Future efforts will also focus  
28 on the annotation refinement of the *IG* constant region to facilitate the development of  
29 the protocol into a comprehensive bioinformatics tool for immunobiologists. The  
30 protocol will also be investigated for its use in any other genomic regions that are  
31 challenging to map to a reference genome, including the T-cell Receptor or specific  
32 fusion or deregulating gene rearrangements that are not represented in the reference  
33 transcriptome.<sup>52</sup>

34

35 *[H2] Comparison with other methods*

36 *[H3] Comparison with the PCR/Sanger sequencing approach for IG gene analysis*

37 IgSeqR was initially performed in 18 CLL samples where Sanger sequencing was  
38 applied from RT-PCR products using our standard procedures.<sup>26,27</sup> A comparison of  
39 these 2 approaches revealed that IgSeqR was fully concordant with Sanger  
40 sequencing for *IGHV*, *IGHD*, and *IGHJ* allele use and nucleotide sequence. IgSeqR  
41 maintained the same level of accuracy as Sanger sequencing. However, IgSeqR  
42 increased the length of the transcript containing the full *IGV(-IGD)-IGJ* rearrangements  
43 from leader to *IG* constant region (up to 2000 nucleotides), while Sanger sequencing  
44 of PCR product was limited to primers' location, generating sequences not longer than  
45 ~400-450 nucleotides and with limited information from the *IG* variable and constant  
46 region sequences. Obtaining the full transcript length will prevent missing AGSs, allow

1 structural prediction analyses and/or generation of F(ab)s for the variable region, or  
2 allow the identification of the IG (sub)class and/or unexpected mutations affecting Fc  
3 structure or function.<sup>28</sup> Also, IgSeqR will improve the chance of detecting a clonal  
4 sequence compared to a PCR-based approach, particularly in lymphoma samples. *IG*  
5 sequencing of lymphoma samples by Sanger is notoriously difficult from unpurified  
6 tumor samples, and demands significant amounts of equipment and time to identify  
7 the tumor *IG* in small cohorts of lymphoma patients, particularly if the tumor population  
8 is low and subcloning approaches are necessary.<sup>29-31</sup> In a cohort of 37 lymphomas  
9 with more than 10% tumor B cells of all the mononucleated cells in the test sample by  
10 flow cytometry, PCR/direct Sanger sequencing successfully identified a dominant *IG*  
11 rearrangement in only 11 (30%). By CIBERSORTx estimation of the frequencies of  
12 individual immune cells from RNAseq data,<sup>32</sup> 439 DLBCL samples from the NCI cohort  
13 had > 10% (tumor) B cells. IgSeqR identified the tumor *IG* rearrangement in 319  
14 (73%), a significantly superior frequency than Sanger ( $p < 0.0001$ ). IgSeqR was also  
15 successful in identifying the full *IGHV(-IGHD)-IGHJ* with their *IGHC* (sub)class  
16 sequence in 20 of the 44 (45%) samples with  $\leq 10\%$  B cell purity, although the success  
17 rate was lower compared to samples with  $> 10\%$  tumor B cell purity ( $p < 0.005$ ) (**Figure**  
18 **1** and **Supplementary Table 1**).

19 The experimental and analytical time to identify the sequences of such a large cohort  
20 by Sanger would be weeks, while it was days for the IgSeqR approach. This suggests  
21 that IgSeqR is more efficient, offering a higher success rate in a shorter experimental  
22 and analytical time compared to standard PCR and Sanger sequencing.

23 Although IgSeqR is currently not configured to build the *IGHC* sequence with contigs  
24 spanning from CDR3 to the 3' end of the constant region allele used, the recovered  
25 transcripts are generally sufficient to determine the *IGHC* class and subclass with high  
26 confidence. This is another advantage compared to the Sanger approach, where  
27 individual isotypes can only be identified using isotype-specific primers.

### 28 [H3] Comparison with other RNA-seq-based approaches

29 Several tools have been developed for *IG* analysis from bulk and single-cell RNA-seq  
30 (**Table 1**),<sup>33-41</sup> many of which preferentially rely on aligning RNA-seq reads to *IG*  
31 reference sequences.<sup>34,36-38</sup> MiXCR is widely adopted for immune profiling in both  
32 academic and industrial settings.<sup>34</sup> It primarily uses the N-regions at the *IGV(-IGD)-*  
33 *IGJ* junctions as a reference and identifies the *IG* repertoire by CDR3 diversity.  
34 However, it is less focused on producing full length transcripts, and highly mutated  
35 *IGV(-IGD)-IGJ* sequences may not be fully reconstructed. Instead, TRUST4 and IG\_ID  
36 tools utilize *de novo* transcriptome assembly. However, TRUST4 was initially designed  
37 for TCR, rather than BCR, repertoire analysis.<sup>40</sup> The IG\_ID tool can accurately produce  
38 full-length BCR transcripts comparable to Sanger sequencing, but has an extended  
39 processing time and generates large temporary files due to the *de novo* assembly of  
40 the whole transcriptome, limiting its use for large-scale analyses.<sup>33,34,36-38</sup>

41 We compared the performance of IgSeqR, MiXCR (v 4.3.2) and TRUST4 (v1.0.12)  
42 using the 18 CLL samples (**Table 2** and **Supplementary Table 2**) and a matching  
43 Sanger sequencing dataset.

44 MiXCR generated *IGH* transcripts for all 18 of the samples. Seventeen (94%) of these  
45 spanned the full *IGHV-IGHD-IGHJ* rearrangement, and 14 (78%) had 100% identity  
46 with Sanger sequencing.

1 TRUST4 generated *IGH* transcripts from 17 (94%) of the samples, all of which were  
2 fully concordant with Sanger. However, TRUST4 failed to identify the only case that  
3 had a deletion of codon 66 of the *IGHV4-34* tumor sequence, possibly revealing a  
4 limitation of TRUST4 in identifying insertions or deletions.

5 IgSeqR produced the longest tumor transcripts, averaging a length of 2036  
6 nucleotides, compared to 589 and 769 nucleotides by MiXCR and TRUST4  
7 respectively. Notably, the majority (78%) of the IgSeqR transcripts were long enough  
8 to cover the full *IGH* region from leader to the membrane domain of the constant region  
9 with confidence, a feature not possible in the shorter transcripts generated by MiXCR  
10 or TRUST4 (**Figure 2**).

11 When efficiency was assessed, IgSeqR took on average 1.18 seconds per nucleotide  
12 assembled (s/nt) to complete the assembly, compared to 8.10 s/nt and 1.44 s/nt  
13 minutes by MiXCR and TRUST4, respectively (Supplementary **Table 3**).

14 Overall, IgSeqR produced longer transcripts, was more efficient per nucleotide  
15 assembled, and was more accurate than MiXCR and TRUST4.

16

## 17 [H2] Overview of the Protocol

18 The Procedure for using IgSeqR to assemble, quantify and annotate *IG* transcripts is  
19 divided into four key stages (**Figure 3**): (a) data pre-processing (Step 1 – 4), (b) *de*  
20 *novo* transcript assembly (Step 5), (c) *IG* transcript selection and quantification (Step  
21 6 – 10), and (d) *IG* transcript annotation and interpretation (Step 11 – 12).

## 22 [H3] Data pre-processing

23 IgSeqR can use RNA-seq data in either BAM (Binary Alignment Map) or FASTQ  
24 format. We assess the quality of the RNA-seq data using FastQC,<sup>42</sup> but alternative  
25 methods more familiar to the operator can be used. Alignment of the data to a  
26 reference transcriptome is performed using the HISAT2 alignment tool, which employs  
27 a hierarchical indexing strategy based on Burrows-Wheeler Transform.<sup>43</sup> If the input  
28 file has been previously aligned, FASTQ reads must first be extracted from the  
29 alignment file (Step 1-2) before being supplied to HISAT2 (Step 3). It is problematic to  
30 map *IG* variable genes, especially D and J genes, to a reference transcriptome using  
31 short-read RNA-seq data, which results in many *IG*-derived reads being unmapped  
32 following alignment.<sup>33</sup> Therefore, following alignment, the resultant BAM file is filtered  
33 to extract reads which align to specific *IG* associated genomic *loci* in addition to any  
34 unmapped reads (Step 4).

## 35 [H3] De Novo Assembly

36 The Trinity software is used for reference-free transcript reconstruction of the reads  
37 associated with *IG* sequences.<sup>44</sup> Trinity follows a three-step process using its modules  
38 Inchworm, Chrysalis, and Butterfly.<sup>44</sup> Inchworm builds initial contigs by assembling  
39 overlapping k-mers from the short reads. Chrysalis constructs a De Bruijn graph using  
40 the Inchworm contigs to represent connections between overlapping sequences and  
41 identifies alternative splicing events. Butterfly decomposes the De Bruijn graph into  
42 individual components representing distinct transcripts from the same gene. These  
43 components are refined and merged to generate complete transcript sequences. The  
44 filtered FASTQ files generated from the HISAT2 output are supplied to Trinity for *de*

1 *novo* transcript assembly, resulting in a FASTA file containing the assembled  
2 transcripts (Step 5).

### 3 **[H3] IG Transcript Selection and Quantification**

4 To remove any non-IG associated transcripts assembled by Trinity, the transcripts in  
5 the output FASTA file are aligned to reference IG databases using BLAST.<sup>45</sup>  
6 Reference FASTA IG sequences are concatenated to generate the databases (Step  
7 6). Transcripts that align with an IG reference sequence are retained (Step 7) and  
8 quantified using Kallisto,<sup>46</sup> a tool that quantifies transcript abundance from RNA-Seq  
9 data using pseudo-alignment instead of read alignment. A k-mer-based index is built  
10 (Step 8) for quantification of the filtered transcripts using the FASTQ reads (Step 9).

### 11 **[H3] IG Transcript Annotation and Interpretation**

12 The most abundant transcripts are identified using the transcript quantification outputs  
13 (Step 10) and passed through the IMGT/V-QUEST sequence alignment web tool,<sup>47</sup>  
14 benefiting from a comprehensive database of known germline IG alleles and  
15 polymorphisms for functional annotations (Step 11). V-QUEST identifies and  
16 annotates *IGHV-IGHD-IGHJ* and *IGKV-IGKJ* or *IGLV-IGLJ* rearrangements, detects  
17 nucleotide mutations and insertions/deletions, and indicates if the detected nucleotide  
18 sequence is functional. The annotated transcripts are then manually reviewed to  
19 identify the tumor transcript through a hierarchical filtering process (Step 12).

### 20 **[H2] Experimental design**

21 Our initial use of IgSeqR with RNA-seq data from a cohort with unknown tumor and  
22 residual normal B cell percentages demonstrated the utility of our protocol.<sup>14,20,21</sup> The  
23 CDR3 is the fingerprint of a B-cell clone. Therefore, it is identical in all the tumor B  
24 cells, different from any other residual non-tumor (polyclonal, hence with different  
25 CDR3) B cells, and quantitatively more abundant than any other CDR3 sequence in the  
26 sample. We designated an IG sequence as tumor-derived when their CDR3 was at  
27 least 5-fold more frequent than any other functional full IG transcript identified with a  
28 different CDR3. The full tumor *IGHV-IGHD-IGHJ* nucleotide sequences including the  
29 *IGHC* constant region isotype were identified in 339/489 (69%) samples with available  
30 RNA-seq data.<sup>14</sup> However, the probability of identifying the tumor sequence could be  
31 maximized by changing certain parameters, including the fold increase of the dominant  
32 sequence's frequency to the frequency of other sequences, or the desired length of  
33 the transcript. Although the success rate of identifying the tumor sequence was lower  
34 compared to samples with  $\geq 10\%$  estimated B cells, a full IG rearrangement could be  
35 identified in many samples with  $< 10\%$  B cells.

36 The sequencing chemistry employed for data generation can influence the outputs of  
37 the protocol. IgSeqR protocol has been designed and tested using RNA sequencing  
38 data derived from polyA-enriched library preparations from fresh-frozen samples.<sup>20,21</sup>  
39 This yielded the successful identification of the tumor IG transcript in 69% of samples,  
40 irrespective of tumor purity. RNA extracted from formalin-fixed paraffin-embedded  
41 (FFPE) tumor samples, which are commonly available in diagnostic settings, is often  
42 of lower quality<sup>48</sup>. We have had success using IgSeqR with ribodepleted total RNA  
43 library preparations in FFPE material, which have yielded IG transcript identification in  
44 54% of samples, unlike polyA-enriched FFPE preparations which did not work in a  
45 preliminary dataset of 10 samples. IgSeqR has not been tested on matched fresh  
46 frozen and FFPE samples, preventing a direct comparison of the procedures.

1 Nevertheless, the current data indicate that IgSeqR can be run both from fresh-frozen  
2 and FFPE material. However, they suggest that polyA-enriched preparations should  
3 not be used from FFPE material.

4 Additionally, paired-end sequencing is recommended for *de novo* assembly of RNA  
5 libraries.<sup>49</sup> IgSeqR was designed for use with whole-transcriptome RNA sequencing  
6 assays and relies on the presence of overlapping cDNA fragments to enable accurate  
7 *de novo* assembly of the entire *IG* transcript. Capture-based methods do not identify  
8 non-templated (CDR3) or heavily mutated fragments and will impair the chance of  
9 success of IgSeqR. Additionally, analytical pipelines that remove unmapped reads will  
10 severely limit *IG* transcript identification and should not be used upstream of IgSeqR.

#### 11 *[H2] Limitations of the method*

12 The main limitation of IgSeqR is accessibility to high-quality RNA and the experimental  
13 costs of RNA-seq.

14 A benchmarking comparison of 10 DLBCL samples demonstrated notably longer  
15 runtimes when compared to our CLL cohort, with average runtimes taking 247 minutes  
16 in DLBCL vs 33 minutes in CLL per sample (Supplementary **Table 4**). The cellular  
17 complexity and lower tumor purity (Supplementary **Table 1**) of a DLBCL tissue sample  
18 may contribute to these longer runtimes compared to CLL blood samples. However,  
19 this is likely to have been compounded by the higher number of starting reads in  
20 DLBCL cases (121.2 million on average) compared to CLL (71.1 million on average),  
21 which increases the processing requirements at each stage of the protocol.

22 Overall, sample characteristics, sequencing chemistry, and data quality may limit the  
23 efficacy of IgSeqR. The quality control assessments described above and in the  
24 “Procedure” should be performed and any necessary errors corrected, before using  
25 IgSeqR.

#### 26 *[H2] Expertise Required to implement the protocol*

27 To effectively implement IgSeqR, individuals must be familiar with computational  
28 biology and have basic expertise in navigating and running commands in a Linux  
29 command-line environment. Users will need to be comfortable installing bioinformatics  
30 tools using the conda package manager or from a repository using git. Familiarity with  
31 large-scale sequencing datasets and their data formats (**Table 3**) and the principles of  
32 immunogenetics, BCR structure and function, and B-cell biology in health and disease  
33 is expected for the interpretation and curation of the results  
34 (<https://www.imgt.org/IMGTEducation/>). While the protocol can be performed by a  
35 skilled graduate student or postdoctoral researcher with the necessary computational  
36 expertise, collaboration with a specialized core facility for sequencing analysis may be  
37 advantageous when generating and processing primary high-throughput sequencing  
38 data.

39

### 40 **[H1] Materials**

#### 41 *[H2] Hardware*

42 The IgSeqR protocol is designed to be versatile, allowing compatibility with various  
43 computing resources, ranging from laptops to high-performance computing clusters,

1 and cloud computing platforms. All analyses, including those for comparison with  
2 MixCR and TRUST4, were conducted using the Iridis5 high-performance computing  
3 cluster at the University of Southampton, utilizing 8 x 2.0 GHz CPU (Central Processing  
4 Unit) cores and 32 GB RAM (Random Access Memory) to simulate a typical desktop  
5 workstation. Default settings were used for MixCR and TRUST4 following the RNA-  
6 seq from raw FASTQ files protocols from each tool's documentation.

7 However, the protocol can be run on less powerful hardware with longer expected  
8 runtimes. Before starting the protocol, users should carefully consider the exact  
9 resources available on their machine, including CPU cores and RAM (considering the  
10 RAM utilized by the operating system), to mitigate errors.

## 11 [H2] Software

- 12 • Operating system: Linux distribution (tested on Red Hat Enterprise v 7.9 and  
13 Ubuntu 16, 22 and 24 distributions)
- 14 • Conda package manager (<https://conda.io>) to install the IgSeqR environment.
- 15 • All dependencies of IgSeqR are documented in the environment file which can  
16 be found in the IgSeqR GitHub Repository  
17 (<https://github.com/ForconiLab/IgSeqR/releases/tag/v1.0.1>), eliminating the  
18 need for manual installation of individual tools and dependencies. The main  
19 software tools used in IgSeqR are listed below along with their versions as  
20 documented in the environment file:
  - 21 ○ BLAST (v 2.13.0)<sup>45</sup>
  - 22 ○ HISAT2 (v 2.2.1)<sup>43</sup>
  - 23 ○ Kallisto (v 0.48.0)<sup>46</sup>
  - 24 ○ Samtools (v 1.16.1)<sup>50</sup>
  - 25 ○ Trinity (v 2.13.2)<sup>44</sup>

26  
27 To create a conda environment from the command line, navigate to the  
28 directory containing the environment file and run the following command:

```
29  
30 $ conda env create -f environment.yml
```

31  
32 Replacing 'environment.yml' with the filepath of the environment file.

33  
34 Once the environment is created, it can be activated by running the following  
35 command:

```
36  
37 $ conda activate igseqr
```

38  
39 **CRITICAL** The [IgSeqR GitHub repository](#) provides the necessary resources  
40 to install and run IgSeqR from the Linux command line. Users are strongly  
41 advised to read the Procedure for an overview of the steps involved before  
42 running IgSeqR. Documentation on our GitHub repository outlines the  
43 installation of IgSeqR, and our guided Tutorial provides example scripts and an  
44 example dataset with which to test your IgSeqR install.

## 45 46 [H2] Data

47 In order to implement this protocol users will need:

- Paired-end RNA sequencing data in either FASTQ or BAM format
- Indexed reference transcriptome for HISAT2 alignment. The protocol was designed and tested using the HISAT2 pre-indexed GRCh38 reference which can be downloaded from the HISAT2 Repository using the command:

```
$ wget https://genome-idx.s3.amazonaws.com/hisat/grch38_snptran.tar.gz
```

Alternatively, custom indexed reference from a user provided reference transcriptome can be generated using the `hisat2-build` command, as described in the HISAT2 documentation (<https://daehwankimlab.github.io/hisat2/manual/>)

- Genomic coordinates associated with target regions for *de novo* transcript assembly. In this application, we have focused on *IG* heavy and light chains coordinates which are supplied in the Procedure section below.
- Reference sequences for *IG* heavy (*IGHV*, *IGHD*, *IGHJ*) and light (*IGKV*, *IGKJ*, *IGLV*, *IGLJ*) chain genes. The references used to develop this protocol can be found in the DATA directory of the IgSeqR GitHub repository (<https://github.com/ForconiLab/IgSeqR/releases/tag/v1.0.1>). However, the IMGT database is regularly updated online. Therefore, it is advised to download the latest individual gene reference files from IMGT (**Table 4**) and merge these into reference FASTA files for *IG* heavy and *IG* light chains before use of the pipeline (see Troubleshooting Step 6, Table 5).

## [H1] Procedure

CRITICAL The procedure below provides a detailed explanation of each command required for the operation of IgSeqR. This allows each command to be run independently and configured based on project and user requirements. Additionally, the procedure can be paused between each step and continued at a later time. However, the protocol has been developed to be run as a complete pipeline from a Linux command line.

### [H2] Data Pre-processing

#### [H3] Data pre-processing of newly generated sequencing data (Pre-pipeline)

1. The pipeline has been optimized on FASTQ files from Illumina sequencing platforms (Illumina, Hayward, CA, USA). For newly generated Illumina sequencing data in BCL format (Binary base Call format) follow Illumina protocols for converting data into fastq format. If starting with FASTQ files commence the pipeline at Step 3 (Genome Alignment).

CRITICAL STEP: It is important to perform quality control (QC) to ensure that the data is of sufficient quality for downstream analysis. A widely used QC tool is FastQC, which produces a detailed report of several quality metrics including per base sequence quality, per sequence quality scores, per base sequence content,

1 per sequence GC content, and sequence length distribution, among others detailed  
2 in the FastQC documentation.<sup>42</sup> If any issues are identified, corrective measures  
3 should be taken as per local procedures implemented by the users' bioinformatics  
4 core facility or general best practice.<sup>51</sup>

### 5 *[H3] Pre-processing published and existing sequencing data*

6 *Timing ~ 5 minutes*

7 **2.** FASTQ files are required for downstream steps in this pipeline, however published  
8 RNA-seq datasets often provide aligned or unaligned BAM files, in which case  
9 FASTQ records must first be extracted from these files, using the `fastq` command  
10 from Samtools.

11  
12  
13

14 Use the following example command to sort, and extract FASTQ records from a  
15 paired-end BAM file 'sample.bam'. This command uses 8 CPU threads for  
16 parallelization, and outputs compressed FASTQ files for read 1, read 2, and  
17 unpaired singleton reads to 'read1.raw.fastq.gz', 'read2.raw.fastq.gz',  
18 respectively:

19

```
20 $ samtools sort -n -@ 8 sample.bam -o sorted.bam
```

```
21 $ samtools fastq -@ 8 -n -c 6 sorted.bam \  
22 -1 read1.raw.fastq.gz \  
23 -2 read2.raw.fastq.gz \  
24 -0 /dev/null -s /dev/null
```

25

26 The `-n` parameter in `sort` is used to sort BAM files by name, `-@` parameter  
27 specifies the number of CPU threads to be used for parallelization of tasks. The `-n`  
28 option in `fastq` is used to leave the read names as they are provided. The `-c`  
29 option sets the compression level of the output files. 'sample.bam' specifies the  
30 path to the input BAM file. `-1` and `-2` specify the desired paths for the compressed  
31 FASTQ output files for read 1 and read 2, respectively. `-0 /dev/null` and  
32 `-s /dev/null` exclude any supplementary, secondary and singleton reads  
33 from the output fastq files.

34 **CRITICAL STEP:** The `fastq` command requires BAM files to first be sorted by  
35 name rather than using the default sorting by chromosomal coordinates, to ensure  
36 proper read pairing. Sorting by name can be achieved by running the Samtools  
37 `sort` command with the `-n` option.

### 38 *[H3] Genome Alignment*

39 *Timing ~ 10 minutes*

40 **3.** FASTQ reads are aligned to a reference genome using HISAT2, producing a SAM  
41 (Sequence Alignment Map) output file which is processed by Samtools. The below  
42 commands can be run as a pipeline to save computational resources. The HISAT2

1 SAM can be passed to Samtools `view` for conversion to BAM format which is then  
2 sorted using the Samtools `sort` command. Upon completion of Samtools `sort`,  
3 Samtools `index` is run to create an accompanying index file for the BAM.  
4  
5 Use the following example command to align FASTQ input files  
6 'read1.raw.fastq.gz' and 'read2.raw.fastq.gz' to the GRCh38 reference  
7 transcriptome using 8 CPU threads. The resulting HISAT2 aligned BAM file is  
8 output as 'hisat\_output.bam' and its corresponding index as  
9 'hisat\_output.bam.bai':  
10  
11

```

1  $ hisat2 -p 8 --phred33 -x grch38_snp_tran \
2  -1 read1.raw.fastq.gz -2 read1.raw.fastq.gz | \
3  samtools view -@8 -bS -0 - - | \
4  samtools sort -@8 - -o hisat_output.bam &&
5  samtools index -@8 hisat_output.bam -o hisat_output.bam.bai
6

```

7 The `-p` or `-@` parameter specifies the number of CPU threads to be used for
8 parallelization, while `--phred33` specifies the encoding format of the quality
9 scores. The `-x` parameter specifies the path and basename of the indexed
10 reference transcriptome files. The input FASTQ file paths are specified by `-1` and
11 `-2` for read 1 and read 2, respectively. The SAM is converted to BAM using `-bS`
12 with `-0` specifying no additional filtering or format conversions, and `-` signifying
13 the standard input from the previous command. The sorted output is written to a
14 file path specified by `-o hisat_output.bam` from which an index file is created
15 and written to the file path specified by `-o hisat_output.bam.bai`.
16

### 17 [H3] Read selection

18 *Timing <5 minutes*

- 19 4. Samtools is used to remove all reads except those that map to the *IG*-associated
20 loci and those that are unmapped from the HISAT2 aligned BAM file, ensuring that
21 highly variable *IG* regions that are difficult to map are retained.

22  
23

24 The `view` command is used to extract the *IG*-associated loci and unmapped reads
25 independently, before joining them using the `merge` command.

26

27 Use the following example command to filter the HISAT2 aligned BAM file
28 'hisat2\_output.bam', retaining reads mapping to the *IGH*, *IGK*, and *IGL* loci
29 and unmapped reads, using 8 threads for parallelization. The resulting filtered BAM
30 file is output as 'IG\_filtered.bam'. Process substitution can be applied when
31 using a supported Unix shell to avoid the generation of temporary files:

32

```

33  $ samtools merge -f IG_filtered.bam \
34  <(samtools view -@ 8 -b -f 4 hisat2_output.bam) \
35  <(samtools view -@ 8 -b hisat2_output.bam 14:105550000-
36  106900000 2:87000000-92000000 22:20500000-24500000)
37

```

38 Where `-@` specifies the number of CPU threads to be utilized for parallelization,
39 `-b` specifies the output format as BAM, `-f 4` returns sequences which have the
40 unmapped Samtools flag, `hisat2_output.bam` is the full input HISAT2 aligned
41 BAM file. The *IG* coordinates `14:105550000-106900000`, `2:87000000-92000000`
42 and `22:20500000-24500000` for *IGH*, *IGK* and *IGL*, respectively,
43 are specified in the format `chr:start-end` where `chr` is the chromosome
44 number, `start` is the numerical position of the first nucleotide in the locus and
45 `end` is the numerical position of the last nucleotide.

1 CRITICAL STEP: The format of the *IG* coordinates will depend on the reference  
2 transcriptome used to generate the aligned BAM file. The HISAT2 indexed  
3 GRCh38 reference uses numerical values for chromosomes (e.g., 14 for  
4 chromosome 14). However, other references may also include a 'chr' prefix (e.g.,  
5 chr14). Additionally, if a different reference transcriptome build is used (e.g.,  
6 GRCh37), the coordinates should be converted accordingly.

## 8 [H2] De Novo Transcript Assembly

9 *Timing ~ 15 minutes*

10 5. Trinity accepts FASTQ input files which must first be extracted from the  
11 'IG\_filtered.bam' BAM file using the Samtools `fastq` command (as described  
12 in Step 2):

```
13 $ samtools sort -n -@ 8 IG_filtered.bam \  
14 -o IG_filtered_sorted.bam \  
15 $ samtools fastq -@ 8 -n -c 6 IG_filtered_sorted.bam \  
16 -1 IG_filtered_read1.fastq \  
17 -2 IG_filtered_read2.fastq \  
18 -0 /dev/null -s /dev/null
```

19  
20  
21 Then, use the following example command to perform Trinity *de novo* assembly  
22 with the input filtered FASTQ files 'IG\_filtered\_read1.fastq' and  
23 'IG\_filtered\_read2.fastq', using 8 threads for parallelization and 32Gb  
24 RAM. The resulting transcriptome FASTA file is output as  
25 'trinity\_transcripts.fasta':

```
26  
27 $ Trinity --CPU 8 --max_memory 32G --seqType fq \  
28 --left IG_filtered_read1.fastq \  
29 --right IG_filtered_read2.fastq \  
30 --output trinity_transcripts \  
31 --no_normalize_reads \  
32 --min_contig_length 500 \  
33 --full_cleanup
```

34  
35 Where `--CPU` specifies the number of CPU threads to be utilized for  
36 parallelization, `--max_memory` specifies the maximum memory to be utilized, `--`  
37 `seqType fq` specifies that the input files are in FASTQ format, `--left` and `-`  
38 `right` are the filtered input FASTQ files for read 1 and read 2, respectively, and  
39 `-output <output>` is the basename of the output files.

40  
41 CRITICAL STEP: Read normalization aims to reduce bias in assembly by down-  
42 sampling highly expressed reads. Input data will be enriched for *IG* transcripts.  
43 Read normalization here can lead to a reduction of reads for low-abundance  
44 transcripts, resulting in incomplete assembly or loss of rare transcripts, and should  
45 be disabled using `--no_normalize_reads`.

1 CRITICAL STEP: Short contigs may represent partial or fragmented *IG* transcripts,  
2 which can affect downstream analysis and interpretation. Using a minimum contig  
3 length of 500 with `-min_contig_length` ensures that most assembled  
4 transcripts will contain the full *IGV-(IGD)-IGJ* recombination.  
5  
6

1 [H2] *IG* transcript Selection and Quantification

2 [H3] *IG* Transcript Selection

3 *Timing < 5 mins*

4 CRITICAL The protocol permits the detection and quantification of *IG* heavy and/or  
5 light chains. The steps below provide examples of *IG* heavy chain transcript extraction,  
6 but can be adapted to extract the *IG* light chain transcript.

7 6. To extract putative *IG* sequences from the Trinity assembly, the transcriptome  
8 FASTA file containing the assembled contigs (from Step 5) is used to search  
9 against a reference sequence using BLAST.

10

11 Individual BLAST databases should be generated using the reference sequences  
12 for *IG* heavy (*IGHV,IGHD,IGHJ*) and light (*IGKV,IGKJ,IGLV,IGLJ*) chains as  
13 required using the `makeblastdb` command.

14

15 Use the following example command to generate a BLAST database from the *IG*  
16 heavy reference FASTA sequences '`IGH_reference.fasta`':

17

```
18 $ makeblastdb -in IGH_reference.fasta -parse_seqids -dbtype  
19 nucl
```

20

21 Where `-in` specifies the input FASTA file containing reference sequences, `-`  
22 `parse_seqids` allows the FASTA headers to be parsed along with their  
23 sequence, and `-dbtype nucl` specifies the sequence content to be nucleotides.

24 TROUBLESHOOTING

25

26

27 7. The assembled transcripts should then be compared against the reference  
28 database(s) generated in step 6 using the `BLASTN` command. This produces a  
29 tabular output that can be passed to the `cut` and `uniq` commands to obtain a  
30 unique list of *IG* transcript IDs, which are used by `samtools faidx` to extract  
31 the corresponding sequences from the assembled transcripts FASTA file.

32

33 Use the following example command to select *IG* transcripts covering reference *IG*  
34 heavy FASTA sequences in the '`IGH_reference.fasta`' file from the  
35 assembled transcripts '`trinity_transcripts.fasta`', to produce the filtered  
36 FASTA file '`IGH_transcripts.fasta`':

37

```
38 $ blastn -db IGH_reference.fasta \  
39 -query trinity_transcripts.fasta -outfmt 6 | \  
40 cut -f1 | uniq | xargs -n 1 samtools faidx  
41 trinity_transcripts.fasta > IGH_transcripts.fasta
```

42

43 Where, `-db` specifies the path to the FASTA file used to generate the reference  
44 database for either *IG* heavy or light sequences, `-query` specifies the path to the  
45 FASTA file containing the Trinity assembled transcripts, `-outfmt 6` sets the  
46 output format to be tabular, `cut -f1` selects the transcript ID (first) column in the

1 tabular BLASTN output, `uniq` removes duplicate transcript IDs, `xargs -n 1`  
2 reads the IDs from the output of `uniq` (one ID per line) and passes them to  
3 `samtools faidx` as separate arguments.

### 4 *[H3] Transcript Quantification*

5 *Timing < 5 minutes*

6 **8.** Abundance of selected transcripts is quantified using the Kallisto pseudoalignment  
7 tool which first requires a Kallisto index to be built from the input FASTA file using  
8 the `index` command.

9

10 Use the following example command to generate a Kallisto index file  
11 'kallisto.index' for the *IGH* heavy chain filtered transcript FASTA sequence file  
12 'IGH\_transcripts.fasta':

13

```
14 kallisto index -i kallisto.index IGH_transcripts.fasta
```

15

16 Where `-i` specifies the filename of the Kallisto index to be constructed and  
17 'IGH\_transcripts.fasta' is the path to the filtered *IGH* transcripts FASTA  
18 sequences.

19

20 **9.** The generated index is used in the `quant` command, along with FASTQ files used  
21 to assemble the transcripts to quantify the abundance of the *IGH* filtered transcripts.

22

23 Use the following example command can be used to quantify the abundance of  
24 transcripts in the *IGH* filtered transcript FASTQ files (generated in step 5)  
25 'IG\_filtered\_read1.fastq' and 'IG\_filtered\_read2.fastq', using 8  
26 threads:

27

```
28 kallisto quant -i kallisto.index -t 8 \  
29 IG_filtered_read1.fastq IG_filtered_read2.fastq
```

30

31 Where `-i` specifies the filename of the Kallisto index, `-t` specifies the number  
32 of CPU threads to be utilized for parallelization, and  
33 'IG\_filtered\_read1.fastq' and 'IG\_filtered\_read2.fastq' are the *IGH*  
34 filtered FASTQ files for read 1 and read 2, respectively.

35

### 36 *[H3] Dominant IGH Transcript Selection*

37

38 **10.** The five most abundant transcript IDs are identified based on their transcript per  
39 million (TPM) value by passing the Kallisto output through the `tail`, `sort`, `head`  
40 and `cut` commands, and their corresponding FASTA sequences are extracted  
41 using the `samtools faidx` command.

42

43 Use the following example command to identify the five most abundant transcript  
44 IDs from the Kallisto output 'abundance.tsv', extract their corresponding  
45 transcript sequences from 'IGH\_transcripts.fasta' and write the results to  
46 an output FASTA file called 'IGH\_TPM\_filtered.fasta':

47

```
1 $ tail -n +2 abundance.tsv | \  
2 sort -t $'\t' -k5,5nr | head -5 | cut -f1 | \  
3 xargs -n 1 samtools faidx IGH_transcripts.fasta >  
4 IGH_TPM_filtered.fasta
```

5  
6 Where `-n +2` selects all rows except the first (header) from the Kallisto quantification output, `-t $'\t'` specifies the delimiter of the input as tab, `-k5,5nr` sorts the remaining lines by the fifth column (TPM) in reverse numerical order, `head -5` outputs the first 5 lines of the sorted file and `cut -f1` extracts the first column (IDs) from the output. The IDs are read (one ID per line) using `xargs -n 1` which then passes them to `samtools faidx` as separate arguments.

14 [H2] *IG* Transcript Annotation and Interpretation

16 [H3] *Dominant IG* Transcript Annotation

17 *Timing ~ 15 minutes*

18  
19 **11.** Submit the top 5 most abundant transcripts identified in step 10 to the IMGT/V-  
20 QUEST tool ([https://imgt.org/IMGT\\_vquest/input](https://imgt.org/IMGT_vquest/input)) for sequence analysis and  
21 annotation. Provide the top 5 transcript sequences to the sequence submission  
22 section of the IMGT/V-QUEST tool, either by copying and pasting the sequences  
23 from the FASTA file or by directly uploading the FASTA file from Step 10. Set the  
24 parameters 'Species' and 'Receptor type or locus' to 'Homo sapiens (human)' and  
25 'IG', respectively. Finally, set the output format to 'C.Excel file'. The IMGT/V-  
26 QUEST tool will annotate and analyze the submitted sequences for their  
27 corresponding *IGV*, *IGHD* (for the heavy chain only) and *IGJ* genes, their junction  
28 at the *CDR3* region, and other related features.

30 [H3] *Transcript Interpretation*

31  
32 **12.** Use the outputs of the Kallisto quantification (Step 10) and IMGT/V-QUEST results  
33 (Step 11) to identify the dominant tumor *IG* transcript present within the RNA-seq  
34 dataset. This process may require manual interpretation but follows the following  
35 hierarchical filtering criteria:

- 36 i. Presence of a full transcript sequence (from codon 1 in framework region  
37 1 to codon 129 in framework region 4 included), identified by  
38 IMGT/VQUEST.
- 39 ii. Presence of 'productive' V-domain functionality call by IMGT/V-QUEST
- 40 iii. The highest estimated read count (est. count) determined by Kallisto.
- 41 iv. The est. count is more than 5-fold higher than the est. count for any of  
42 the other 4 most abundant transcripts if their sequences are different.
- 43 v. The ability to determine the *IG* constant region class and subclass.

44 TROUBLESHOOTING

45

46

1 **[H1] Troubleshooting**

2 Troubleshooting advice can be found in Table 5.

3 Table 5. Troubleshooting.

Step	Problem	Possible Reason	Possible Solution
6	Unable to make BLAST database	Invalid characters included in header identifiers	Ensure that the sequence headers in the FASTA file are free of any special characters, especially the pipe ( ) character. The pipe character is used as a delimiter by IMG_T for its references, but it is also a reserved character for the ID parser, which can cause errors. Pre-built BLAST databases and a script to automate the creation of the latest versions, <code>make_imgt_blast_db.sh</code> , are available in the data/IMG_T section of the GitHub repository.
12	No transcripts remain after the filtering is applied	Unmapped reads unavailable in previously aligned files	Unmapped reads are essential for high quality recovery of the non-templated IG regions and highly mutated IG transcripts. Previously aligned sequencing data may have had these reads removed by their processing pipelines. Use <code>\$ samtools view -c -f 4 input_file.bam</code> to check for the number of unmapped reads in a BAM file; if 0, it is unlikely that IgSeqR will be successful with these data.
		Insufficient transcripts retained in step 10	The number of most abundant transcripts to take forward has been suggested as 5. This has been found to strike a good balance between analytical efficiency and identification of the dominant tumor transcript. In instances where full-length, productive transcripts are not obtained within the top 5 transcripts due to tumor purity or sequencing quality, users may wish to increase the number of transcripts to take forward for analysis in step 10
		Est. counts fold change threshold is too high	In cases with a low tumor purity, there is a higher proportion of background non-tumor IG transcripts. A reduction of the fold change threshold may increase the probability of identifying a “dominant” sequence.
		Low tumour purity	If the tumor purity was not established prior to RNA sequencing and IgSeqR is unable to identify immunoglobulin transcripts, cellular deconvolution tools such as CIBERSORTx <sup>32</sup>

			can be run on RNAseq data to diagnose problems with low tumour purity. Detailed tutorials for the implementation of CIBERSORTx are available on the author's website ( <a href="https://cibersortx.stanford.edu/tutorial.php">https://cibersortx.stanford.edu/tutorial.php</a> ).
		Non-full-length transcripts belonging to the same transcript (i.e. transcripts sharing the same CDR3)	Some samples recover non-full-length transcripts belonging to the same transcript (i.e. transcripts sharing the same CDR3), which may lead to an artificial decrease in fold change. In these cases, the estimated counts in transcripts with identical CDR3s should be summed to establish a more reliable fold change compared to the next unrelated transcript.

1

2

### 3 [H1] Timing

4 Benchmarking was conducted using the computational hardware described in the  
5 Materials section. The dominant *IG* heavy and light chain transcripts were extracted  
6 from FASTQ files generated from high-purity CLL samples with an average starting  
7 read count of 71.1 million following initial HISAT2 alignment. In similar conditions, the  
8 full pipeline can be expected to take less than 1 hour per sample. Specific timings can  
9 be found in the procedure section headers for each stage of the analytical pipeline.  
10 The duration of each stage may vary depending on the input file type (BAM files require  
11 additional pre-processing), hardware used to run the pipeline, heterogeneity of B-cell  
12 populations, and number of starting sequencing reads generated from the samples.

### 13 [H1] Anticipated results

14 Upon successful completion of the IgSeqR protocol, users will have generated the  
15 following output files for *IG* heavy and/or light chain transcripts:

- 16 • The five most abundant assembled *IG* transcripts in FASTA format
- 17 • Table of quantifications for these *IG* transcripts in tab-separated value (tsv)  
18 format
- 19 • Annotations for the top five *IG* transcripts generated by IMGT/V-QUEST.

20 For further insights, users can refer to our previous publication <sup>14</sup>, which includes  
21 results and examples of downstream analysis.

### 22 [H1] Data Availability

23 The datasets referenced in the protocol are available in the following repositories:

- 24 • NCI DLBCL dataset: The data IgSeqR was developed to analyze have been  
25 deposited in dbGaP under accession code [phs001444.v1.p1](https://dbgap.ncbi.nlm.nih.gov/oa/GET.cgi?acc=phs001444.v1.p1) and detailed

1 results of IgSeqR analysis can be found in the publication by Chiodin et al.  
2 (2021) Blood<sup>14</sup>  
3 • CLL dataset: The data used to benchmark IgSeqR have been deposited in  
4 ArrayExpress under accession number [E-MTAB-12017](#).

5 These datasets are publicly available and can be accessed by following the respective  
6 repository guidelines. For further information, please refer to the repository links  
7 provided.

### 8 **[H1] Code Availability**

9 The code for IgSeqR is available on GitHub at  
10 <https://github.com/ForconiLab/IgSeqR/releases/tag/v1.0.1> This repository contains  
11 the source code, installation instructions, and usage examples to facilitate the use of  
12 IgSeqR. For additional information on the implementation and usage of IgSeqR,  
13 please refer to the documentation provided in the repository.

### 14 **Supplementary information**

- 15 • Supplementary Tables.xlsx

16

### 17 **[H1] Author contributions**

18 D.B. and B.J.S. developed the IgSeqR bioinformatic pipeline, analyzed and interpreted  
19 data, and wrote the manuscript. D.T. and G.C analyzed, interpreted data, and  
20 contributed to the immunoglobulin gene analysis pipeline validation. B.S., A.A, E.S.,  
21 A.O. and J.B contributed to the analysis and interpretation of the data. F.F. designed  
22 the study, supervised research, interpreted data and wrote the manuscript. All authors  
23 reviewed and approved the manuscript.

24

25

26

1 **[H1] Acknowledgments**

2 The authors are grateful to the Faculty of Medicine Tissue Bank (Cancer Sciences,  
3 University of Southampton) for the processing and storage of the primary lymphoma  
4 specimens. This work was supported by Cancer Research UK (ECRIN-M3 accelerator  
5 award C42023/A29370, and BTERP project C36811/A29101), Leukaemia UK Pioneer  
6 Award. B.S. was also supported by the Cancer Sciences Talent Fund. D.T. was funded  
7 by the Eyles Cancer Immunology PhD scholarship. G.C. was funded by the Leukaemia  
8 UK John Goldman Fellowship, the Wessex immunology the Eyles Cancer Immunology  
9 Fellowship and the Southampton Cancer Immunology Centre Pump- priming award  
10 2021). Genetic data for the IgSeqR protocol were obtained via the National Cancer  
11 Institute Genomic Data Commons for Genotypes and Phenotypes (accession  
12 phs001444.v1.p1). The authors acknowledge the use of the IRIDIS High Performance  
13 Computing Facility, and associated support services at the University of Southampton,  
14 in the completion of this work.

15 **[H1] Competing interests**

16 The authors declare no potential conflicts of interest.

17

18

## 1 [H1] References

- 2 1 Lam, K. P., Kuhn, R. & Rajewsky, K. In vivo ablation of surface immunoglobulin on  
3 mature B cells by inducible gene targeting results in rapid cell death. *Cell* **90**, 1073-  
4 1083, doi:10.1016/s0092-8674(00)80373-6 (1997).
- 5 2 Stevenson, F. K. *et al.* The occurrence and significance of V gene mutations in B cell-  
6 derived human malignancy. *Adv Cancer Res* **83**, 81-116, doi:10.1016/s0065-  
7 230x(01)83004-9 (2001).
- 8 3 Victora, G. D. & Nussenzweig, M. C. Germinal Centers. *Annual Review of Immunology*  
9 **40**, 413-442, doi:10.1146/annurev-immunol-120419-022408 (2022).
- 10 4 Forconi, F., Lanham, S. A. & Chiodin, G. Biological and Clinical Insight from Analysis  
11 of the Tumor B-Cell Receptor Structure and Function in Chronic Lymphocytic  
12 Leukemia. *Cancers (Basel)* **14**, doi:10.3390/cancers14030663 (2022).
- 13 5 Stevenson, F. K., Forconi, F. & Kipps, T. J. Exploring the pathways to chronic  
14 lymphocytic leukemia. *Blood* **138**, 827-835, doi:10.1182/blood.2020010029 (2021).
- 15 6 Efremov, D. G., Turkalj, S. & Laurenti, L. Mechanisms of B Cell Receptor Activation  
16 and Responses to B Cell Receptor Inhibitors in B Cell Malignancies. *Cancers* **12**, 1396  
17 (2020).
- 18 7 Forconi, F. *et al.* The normal IGHV1-69-derived B-cell repertoire contains stereotypic  
19 patterns characteristic of unmutated CLL. *Blood* **115**, 71-77, doi:10.1182/blood-2009-  
20 06-225813 (2010).
- 21 8 Seifert, M. *et al.* Cellular origin and pathophysiology of chronic lymphocytic leukemia.  
22 *J Exp Med* **209**, 2183-2198, doi:10.1084/jem.20120833 (2012).
- 23 9 Damle, R. N. *et al.* Ig V gene mutation status and CD38 expression as novel prognostic  
24 indicators in chronic lymphocytic leukemia. *Blood* **94**, 1840-1847 (1999).
- 25 10 Hamblin, T. J., Davis, Z., Gardiner, A., Oscier, D. G. & Stevenson, F. K. Unmutated Ig  
26 V(H) genes are associated with a more aggressive form of chronic lymphocytic  
27 leukemia. *Blood* **94**, 1848-1854 (1999).
- 28 11 Niemann, C. U. *et al.* Fixed-duration ibrutinib–venetoclax versus chlorambucil–  
29 obinutuzumab in previously untreated chronic lymphocytic leukaemia (GLOW): 4-year  
30 follow-up from a multicentre, open-label, randomised, phase 3 trial. *The Lancet*  
31 *Oncology*, doi:[https://doi.org/10.1016/S1470-2045\(23\)00452-7](https://doi.org/10.1016/S1470-2045(23)00452-7) (2023).
- 32 12 Stevenson, F. K. & Forconi, F. The essential microenvironmental role of  
33 oligomannoses inserted into the antigen-binding sites of lymphoma cells. *Blood*,  
34 doi:10.1182/blood.2023022703 (2023).
- 35 13 Zhu, D. *et al.* Acquisition of potential N-glycosylation sites in the immunoglobulin  
36 variable region by somatic mutation is a distinctive feature of follicular lymphoma.  
37 *Blood* **99**, 2562-2568, doi:10.1182/blood.V99.7.2562 (2002).
- 38 14 Chiodin, G. *et al.* Insertion of atypical glycans into the tumor antigen-binding site  
39 identifies DLBCLs with distinct origin and behavior. *Blood* **138**, 1570-1582,  
40 doi:10.1182/blood.2021012052 (2021).
- 41 15 Coelho, V. *et al.* Glycosylation of surface Ig creates a functional bridge between human  
42 follicular lymphoma and microenvironmental lectins. *Proc Natl Acad Sci U S A* **107**,  
43 18587-18592, doi:10.1073/pnas.1009388107 (2010).
- 44 16 Linley, A. *et al.* Lectin binding to surface Ig variable regions provides a universal  
45 persistent activating signal for follicular lymphoma cells. *Blood* **126**, 1902-1910,  
46 doi:10.1182/blood-2015-04-640805 (2015).
- 47 17 Odabashian, M. *et al.* IGHV sequencing reveals acquired N-glycosylation sites as a  
48 clonal and stable event during follicular lymphoma evolution. *Blood* **135**, 834-844,  
49 doi:10.1182/blood.2019002279 (2020).
- 50 18 Sutton, L. A. *et al.* Immunoglobulin genes in chronic lymphocytic leukemia: key to  
51 understanding the disease and improving risk stratification. *Haematologica* **102**, 968-  
52 971, doi:10.3324/haematol.2017.165605 (2017).
- 53 19 Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics.  
54 *Nat Rev Genet* **10**, 57-63, doi:10.1038/nrg2484 (2009).

1 20 Schmitz, R. *et al.* Genetics and pathogenesis of diffuse large B-cell lymphoma. *New*  
2 *England Journal of Medicine* **378**, 1396-1407 (2018).

3 21 Wright, G. W. *et al.* A Probabilistic Classification Tool for Genetic Subtypes of Diffuse  
4 Large B Cell Lymphoma with Therapeutic Implications. *Cancer Cell* **37**, 551-568.e514,  
5 doi:<https://doi.org/10.1016/j.ccell.2020.03.015> (2020).

6 22 Bryant, D. *et al.* Network analysis reveals a major role for 14q32 cluster miRNAs in  
7 determining transcriptional differences between IGHV-mutated and unmutated CLL.  
8 *Leukemia* **37**, 1454-1463, doi:10.1038/s41375-023-01918-9 (2023).

9 23 Forconi, F. *et al.* Insight into the potential for DNA idiotypic fusion vaccines designed  
10 for patients by analysing xenogeneic anti-idiotypic antibody responses. *Immunology*  
11 **107**, 39-45, doi:10.1046/j.1365-2567.2002.01452.x (2002).

12 24 Stevenson, G. T., Elliott, E. V. & Stevenson, F. K. Idiotypic determinants on the surface  
13 immunoglobulin of neoplastic lymphocytes: a therapeutic target. *Fed Proc* **36**, 2268-  
14 2271 (1977).

15 25 Hawkins, R. E. *et al.* Idiotypic vaccination against human B-cell lymphoma. Rescue of  
16 variable region gene sequences from biopsy material for assembly as single-chain Fv  
17 personal vaccines. *Blood* **83**, 3279-3288 (1994).

18 26 Forconi, F. *et al.* Tumor cells of hairy cell leukemia express multiple clonally related  
19 immunoglobulin isotypes via RNA splicing. *Blood* **98**, 1174-1181,  
20 doi:10.1182/blood.v98.4.1174 (2001).

21 27 D'Avola, A. *et al.* Surface IgM expression and function are associated with clinical  
22 behavior, genetic abnormalities, and DNA methylation in CLL. *Blood* **128**, 816-826,  
23 doi:10.1182/blood-2016-03-707786 (2016).

24 28 Wasim, L. *et al.* Mutations in the IgG B cell receptor associated with class-switched B  
25 cell lymphomas. *bioRxiv*, 2024.2004.2012.585865, doi:10.1101/2024.04.12.585865  
26 (2024).

27 29 McCann, K., Sahota, S. S., Stevenson, F. K. & Ottensmeier, C. H. Idiotypic gene rescue  
28 in follicular lymphoma. *Methods Mol Med* **115**, 145-171, doi:10.1385/1-59259-936-  
29 2:145 (2005).

30 30 Ottensmeier, C. H. & Stevenson, F. K. Isotype switch variants reveal clonally related  
31 subpopulations in diffuse large B-cell lymphoma. *Blood* **96**, 2550-2556 (2000).

32 31 Ottensmeier, C. H. *et al.* Analysis of VH genes in follicular and diffuse lymphoma shows  
33 ongoing somatic mutation and multiple isotype transcripts in early disease with  
34 changes during disease progression. *Blood* **91**, 4292-4299 (1998).

35 32 Newman, A.M., Steen, C.B., Liu, C.L. *et al.* Determining cell type abundance and  
36 expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**, 773-782  
37 (2019). <https://doi.org/10.1038/s41587-019-0114-2>

38 33 Blachly, J. S. *et al.* Immunoglobulin transcript sequence and somatic hypermutation  
39 computation from unselected RNA-seq reads in chronic lymphocytic leukemia. *Proc*  
40 *Natl Acad Sci U S A* **112**, 4322-4327, doi:10.1073/pnas.1503587112 (2015).

41 34 Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling.  
42 *Nat Methods* **12**, 380-381, doi:10.1038/nmeth.3364 (2015).

43 35 Canzar, S., Neu, K. E., Tang, Q., Wilson, P. C. & Khan, A. A. BASIC: BCR assembly  
44 from single cells. *Bioinformatics* **33**, 425-427, doi:10.1093/bioinformatics/btw631  
45 (2017).

46 36 Kuchenbecker, L. *et al.* IMSEQ--a fast and error aware approach to immunogenetic  
47 sequence analysis. *Bioinformatics* **31**, 2963-2971, doi:10.1093/bioinformatics/btv309  
48 (2015).

49 37 Mandric, I. *et al.* Profiling immunoglobulin repertoires across multiple human tissues  
50 using RNA sequencing. *Nat Commun* **11**, 3126, doi:10.1038/s41467-020-16857-7  
51 (2020).

52 38 Mose, L. E. *et al.* Assembly-based inference of B-cell receptor repertoires from short  
53 read RNA sequencing data with V'DJer. *Bioinformatics* **32**, 3729-3734,  
54 doi:10.1093/bioinformatics/btw526 (2016).

1 39 Rizzetto, S. *et al.* B-cell receptor reconstruction from single-cell RNA-seq with  
2 VDJ-Puzzle. *Bioinformatics* **34**, 2846-2847, doi:10.1093/bioinformatics/bty203 (2018).

3 40 Song, L. *et al.* TRUST4: immune repertoire reconstruction from bulk and single-cell  
4 RNA-seq data. *Nat Methods* **18**, 627-630, doi:10.1038/s41592-021-01142-2 (2021).

5 41 Upadhyay, A. A. *et al.* BALDR: a computational pipeline for paired heavy and light  
6 chain immunoglobulin reconstruction in single-cell RNA-seq data. *Genome Med* **10**,  
7 20, doi:10.1186/s13073-018-0528-3 (2018).

8 42 Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data.  
9 *Babraham Bioinformatics* (2010).

10 43 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome  
11 alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*  
12 **37**, 907-915, doi:10.1038/s41587-019-0201-4 (2019).

13 44 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without  
14 a reference genome. *Nat Biotechnol* **29**, 644-652, doi:10.1038/nbt.1883 (2011).

15 45 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment  
16 search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/s0022-2836(05)80360-2 (1990).

17 46 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq  
18 quantification. *Nat Biotechnol* **34**, 525-527, doi:10.1038/nbt.3519 (2016).

19 47 Brochet, X., Lefranc, M. P. & Giudicelli, V. IMGT/V-QUEST: the highly customized and  
20 integrated system for IG and TR standardized V-J and V-D-J sequence analysis.  
21 *Nucleic Acids Res* **36**, W503-508, doi:10.1093/nar/gkn316 (2008).

22 48 Cazzato, G. *et al.* Formalin-Fixed and Paraffin-Embedded Samples for Next  
23 Generation Sequencing: Problems and Solutions. *Genes (Basel)* **12**,  
24 doi:10.3390/genes12101472 (2021).

25 49 Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the  
26 Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494-1512,  
27 doi:10.1038/nprot.2013.084 (2013).

28 50 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,  
29 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

30 51 Hesketh, A. R. RNA Sequencing Best Practices: Experimental Protocol and Data  
31 Analysis in *Yeast Systems Biology: Methods and Protocols*. (eds Stephen G. Oliver &  
32 Juan I. Castrillo) 113-129 (Springer New York, 2019).

33 52 Haas, B. J. *et al.* Accuracy assessment of fusion transcript detection via read-mapping  
34 and de novo fusion transcript assembly-based methods. *Genome Biology* **20**, 213,  
35 doi:10.1186/s13059-019-1842-9 (2019).

36

37

1 **Table 1. Published tools for IG analysis from bulk and single-cell RNA sequencing (RNA-seq)**

Tool	Description	Receptor	RNA sequencing data	Reference
IG_ID	<i>De Novo</i> assembly of BCR transcripts from bulk RNA-seq data	BCR	Bulk	Blachly et al 2015 [30]
MiXCR	Analysis of raw T- or B- cell receptor repertoire sequencing data	BCR and TCR	Bulk and Single Cell	Bolotin et al 2015 [31]
BASIC	Bayesian inference of immunoglobulin sequences. BASIC offers functionalities for V(D)J gene identification, clonotype analysis, and mutation profiling.	BCR	Single Cell	Canzar et al 2017 [32]
IMSEQ	Provides functionalities for the identification and quantification of IG genes, as well as the detection of somatic hypermutations	BCR and TCR	Bulk	Kuchenbecker et al 2015 [33]
ImReP	Extracts of receptor reads from sequencing data and assembles clonotypes, detects corresponding V(D)J recombination events and corrects PCR sequencing errors	BCR and TCR	Bulk	Mandric et al 2020 [34]
V'DJer	Customized read extraction, assembly and V(D)J rearrangement detection and filtering to produce contigs representing the most abundant portions of the BCR repertoire	BCR	Bulk	Mose et al 2016 [35]
VDJPuzzle	Provides a user-friendly interface for the identification of V(D)J rearrangements, clonotype analysis, and visualization of TCR and BCR repertoires.	BCR and TCR	Single Cell	Rizzetto et al 2018 [36]
TRUST4	Performs de novo assembly on V, J, C genes including the hypervariable complementarity-determining region 3 (CDR3) and reports consensus of BCR/TCR sequences	BCR/TCR	Bulk and Single Cell	Song et al 2021 [37]
BALDR	Infers the clonal structure of B-cell repertoires, providing information on clonal abundance, V(D)J gene usage, and somatic hypermutations	BCR	Single Cell	Upadhyay et al 2018 [38]

2

1 Table 2. A comparison between RNA-seq based Immunoglobulin Gene analysis tools, IgSeqR  
 2 MiXCR and TRUST4.

Property	IgSeqR	MixCR	TRUST4
<b>Recognized as IG (by IMGT/VQUEST)</b>	18 (100 %)	18 (100 %)	17 (94.44 %)
<b>Productive Sequence*</b>	18 (100 %)	18 (100 %)	17 (94.44 %)
<b>Complete VDJ</b>	18 (100 %)	17 (94.44 %)	17 (94.44 %)
<b>IGHV Gene match</b>	18 (100 %)	17 (94.44 %)	17 (94.44 %)
<b>IGHV Seq match</b>	18 (100 %)	14 (77.78 %)	17 (94.44 %)
<b>IGHD Gene Allele match</b>	18 (100 %)	18 (100 %)	17 (94.44 %)
<b>IGHD Seq match</b>	18 (100 %)	18 (100 %)	17 (94.44 %)
<b>IGHJ Gene Allele match</b>	18 (100 %)	18 (100 %)	17 (94.44 %)
<b>IGHJ Seq match</b>	18 (100 %)	18 (100 %)	17 (94.44 %)
<b>CDR3 Seq match</b>	18 (100 %)	18 (100 %)	17 (94.44 %)
<b>Full Sanger VDJ Concordance</b>	18 (100 %)	14 (77.78 %)	17 (94.44 %)
<b>Average Length</b>	2036	589	768
<b>Assembly efficiency (Seconds/Nucleotide)</b>	1.18	8.10	1.44

3 \*According to IMGT, a productive sequence is a rearranged immunoglobulin (IG) that has an open  
 4 reading frame (ORF) without any stop codons.

5  
 6

1 **Table 3. List of acronyms used in the IgSeqR protocol**

<b>Acronyms</b>	<b>Name</b>	<b>Description</b>
BAM	Binary Alignment Map	The BAM format is a binary, compressed representation of sequence alignment data in SAM format (see SAM acronym). It is commonly used in genomics to efficiently store the results of sequence alignment algorithms.
CPU	Central Processing Unit	The CPU is the most important processor in a given computer, responsible for performing basic arithmetic, logic, controlling, and input/output (I/O) operations specified by the instructions in a program
CSV	Comma-Separated Values	CSV is a simple file format used to store tabular data, such as a spreadsheet or database. Each line of the file represents a row of the table, and the values are separated by commas. CSV files are widely supported by spreadsheet and database software, making them suitable for easy import and export of data.
FASTA	FASTA Sequence Format	The FASTA format is a text-based format for representing nucleotide or protein sequences. It consists of a single-line description followed by lines of sequence data. The format is widely used in bioinformatics for storing and exchanging sequence data.

FASTQ	FASTQ Sequence Format	The FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. It is widely used to represent raw sequencing data from high-throughput sequencing platforms.
TSV	Tab-Separated Values	TSV is a file format similar to CSV, but with tab characters as the field separator instead of commas. TSV files are commonly used for storing and exchanging tabular data, especially when the data may contain commas or other special characters.
RAM	Random Access Memory	A temporary memory bank in a computer where data which requires quick access is stored. It keeps data easily accessible so a computer's processor can quickly find it without having to go into long-term storage to complete immediate processing tasks.
SAM	Sequence Alignment Map	The SAM format is a text based representation of sequence alignment data. SAM files are human-readable and contain both the sequence information and the alignment details.

1

2

1 **Table 4. Links to the most recent IMGT/V-QUEST reference immunoglobulin heavy and light**  
 2 **chain FASTA sequences**

Chain	Gene	IMGT Link
Heavy	IGHV	<a href="http://imgt.org/download/V-QUEST/IMGT_V-QUEST_reference_directory/Homo_sapiens/IG/IGHV.fasta">imgt.org/download/V-QUEST/IMGT_V-QUEST_reference_directory/Homo_sapiens/IG/IGHV.fasta</a>
	IGHD	<a href="http://imgt.org/download/V-QUEST/IMGT_V-QUEST_reference_directory/Homo_sapiens/IG/IGHD.fasta">imgt.org/download/V-QUEST/IMGT_V-QUEST_reference_directory/Homo_sapiens/IG/IGHD.fasta</a>
	IGHJ	<a href="http://imgt.org/download/V-QUEST/IMGT_V-QUEST_reference_directory/Homo_sapiens/IG/IGHJ.fasta">imgt.org/download/V-QUEST/IMGT_V-QUEST_reference_directory/Homo_sapiens/IG/IGHJ.fasta</a>
Light	IGKV	<a href="http://imgt.org/download/V-QUEST/IMGT_V-QUEST_reference_directory/Homo_sapiens/IG/IGKV.fasta">imgt.org/download/V-QUEST/IMGT_V-QUEST_reference_directory/Homo_sapiens/IG/IGKV.fasta</a>
	IGKJ	<a href="http://imgt.org/download/V-QUEST/IMGT_V-QUEST_reference_directory/Homo_sapiens/IG/IGKJ.fasta">imgt.org/download/V-QUEST/IMGT_V-QUEST_reference_directory/Homo_sapiens/IG/IGKJ.fasta</a>
	IGLV	<a href="http://imgt.org/download/V-QUEST/IMGT_V-QUEST_reference_directory/Homo_sapiens/IG/IGLV.fasta">imgt.org/download/V-QUEST/IMGT_V-QUEST_reference_directory/Homo_sapiens/IG/IGLV.fasta</a>
	IGLJ	<a href="http://imgt.org/download/V-QUEST/IMGT_V-QUEST_reference_directory/Homo_sapiens/IG/IGLJ.fasta">imgt.org/download/V-QUEST/IMGT_V-QUEST_reference_directory/Homo_sapiens/IG/IGLJ.fasta</a>

3

4

5

1 **[b1] Box 1. Immunoglobulin gene analysis provides insight into the cell of origin and**  
2 **behavior of B-Cell malignancies.** The immunoglobulin heavy-chain gene repertoire  
3 comprises ~51 functional variable (*IGHV*), ~21 diversity (*IGHD*), and ~7 joining (*IGHJ*) genes  
4 at the *14q32* locus. As shown in the figure, in the bone marrow, progenitor B cells (Pro-B cells)  
5 undergo an *IGHD-IGHJ* gene rearrangement. If successful, a complete *IGHV-IGHD-IGHJ*  
6 rearrangement occurs at the precursor B cell (Pre-B), which expresses a precursor B-cell  
7 receptor (pre-BCR) containing a surrogate VpreB1 light chain and a full IG heavy chain. The  
8 pre-BCR will promote rearrangement of *IGKV-IGKJ* at the *2p11.2* locus, and, if this is non-  
9 functional in both alleles, the rearrangement of *IGLV-IGLJ* will occur at the *22q11.22* locus in  
10 immature B cells. A successful rearrangement of the IG light chain enables the expression of  
11 a competent immunoglobulin M (IgM) and autoreactive B cell clones within the bone marrow  
12 microenvironment will be deleted, ensuring the production of functional non-autoreactive naïve  
13 B cells expressing IgM and IgD. IgM- and IgD-expressing naïve B cells exit the bone marrow  
14 and migrate to peripheral lymphoid organs (spleen, lymph nodes, mucosa-associated  
15 lymphoid tissues) where they will encounter antigen, and they will undergo class-switch  
16 recombination (CSR) and somatic hypermutation (SHM) in the presence of activation-induced  
17 cytidine-deaminase (AID) in a germinal center (GC) reaction at the centroblast (Cb) stage (dark  
18 zone). During SHM, Cb introduce point mutations in the IG variable region genes to mature  
19 affinity to antigen. Centrocytes (Cc) emerge in the light zone where their fate will depend on  
20 their BCR interactions with immune complexes on follicular dendritic cells (FDC) in the  
21 presence of T follicular helper ( $T_{FH}$ ) cells. Cc with the BCR of the right affinity to antigen receive  
22 survival signals and differentiate into memory B cells or plasma cells, while the others will  
23 undergo apoptosis.

24 B-cell tumors preserve the structural and functional features of the *IGHV-IGHD-IGHJ* and IG  
25 constant (IGHC) region of the cell of origin. Chronic Lymphocytic Leukemias with unmutated  
26 IG genes (U-CLL) arise from pre-GC B-cells and have an aggressive clinical course, while  
27 those with mutated IG genes (M-CLL) arise from post-GC B cells and display an indolent  
28 clinical course. In endemic Burkitt lymphoma (eBL), FL, and some DLBCL, there is intraclonal  
29 heterogeneity of the *IGV* gene sequences to indicate that the SHM process is ongoing, as in  
30 a GC B cell. Diffuse Large B-cell Lymphoma (DLBCL) can be classified into two major  
31 subtypes: GC B-cell-like (GCB) and activated B-cell-like (ABC). Asparagine-x-serine/threonine  
32 N-glycosylation motifs (where X is any amino acid except proline) are introduced by SHM,  
33 allowing occupation of the sites by oligomannose-type glycans in ~80% of FL and ~30% of  
34 GCB-DLBCL. Multiple myeloma (MM) is characterized by the clonal expansion of plasma cells,  
35 which carry mutated IG and secrete a monoclonal IG in the serum (paraprotein).

### 36 37 **Figure 1. Transcript Recovery by Sanger Sequencing or IgSeqR in samples with** 38 **different tumor B-cell purity**

39 We evaluated the performance of Sanger sequencing (left side, in red) or IgSeqR (right, in  
40 blue) in recovering IG transcripts from samples with known tumor contamination. Sanger  
41 sequencing was performed on a cohort of locally collected lymphoma samples (n=37) with a  
42 tumor B-cell purity (frequency of tumor cells in the sample) of  $\geq 10\%$ , while IgSeqR was applied  
43 to a publicly available RNA-seq dataset (n=489). Samples were grouped into every 10%  
44 percentile of tumor B-cell purity. Tumor B-cell purity was assessed by phenotype of the  
45 mononucleated cells for the Sanger cohort and CIBERSORTx-B cell estimation for the IgSeqR  
46 cohort. The number of samples with successfully (dark shade) and unsuccessfully (no  
47 transcripts meeting all criteria defined in Procedure step 12; light shade) recovered tumor IG  
48 transcripts is shown within each bar. A significantly higher probability to identify the tumor IG  
49 transcript was observed with IgSeqR (69%) compared to Sanger sequencing (30%) ( $X^2$ -test  
50 with Yates' correction,  $p < 0.0001$ ).

51  
52 **Figure 2. Comparison between the transcripts recovered by IgSeqR, MiXCR, and**  
53 **TRUST4.** Shown is a direct comparison of *IGHV-IGHD-IGHJ-IGHC* transcripts recovered from  
54 bulk whole transcriptome RNA sequencing data from 18 chronic lymphocytic leukemia

1 samples with high tumor purity (80-97%). IgSeqR, MiXCR (v 4.3.2), and TRUST4 (v1.0.12)  
2 were run using the Iridis5 high-performance computing cluster at the University of  
3 Southampton, utilizing 8 x 2.0 GHz CPU cores and 32 GB RAM to simulate a typical desktop  
4 workstation. The resulting transcripts were assessed for recovery of a full-length, productive  
5 *IGHV-IGHD-IGHJ* (V-Region) transcript and concordance with matched Sanger sequencing in  
6 the V-region. MiXCR recovered *IGHV-IGHD-IGHJ* transcripts for all 18 samples, with 17 (94%)  
7 having productive and full V-Region coverage, however only 17 matched the IGHV of Sanger,  
8 and 14 (78%) had 100% identity with Sanger. TRUST4 generated *IGHV-IGHD-IGHJ*  
9 transcripts from 17 (94%) of the samples, all of which had productive and full V-Region  
10 coverage and full concordance with Sanger. IgSeqR demonstrated productive and full V-  
11 Region coverage and full concordance with Sanger in all 18 (100%) samples. IgSeqR also  
12 produced the longest tumor transcripts, averaging a length of 2036 nucleotides, compared to  
13 589 and 769 nucleotides by MiXCR and TRUST4 respectively. Notably, the majority (78%) of  
14 the IgSeqR transcripts were long enough to cover the full *IGHM* transcript from leader to the  
15 membrane domains (M1 and M2) of the constant region (C-Region), a feature not detectable  
16 in the shorter transcripts generated by MiXCR or TRUST4.

17

18 **Figure 3. Schematic representation of the IgSeqR pipeline.** The procedure for running  
19 IgSeqR is divided into four key stages: **(a)** data pre-processing – RNA sequencing data  
20 (RNA-seq) can be supplied in either BAM or FASTQ format. If starting from a BAM file,  
21 reads are first extracted in FASTQ format using the Samtools fastq command. The  
22 data are re-aligned to a reference transcriptome by HISAT2, producing a BAM file  
23 which is filtered to retain reads mapping to *IG* gene coordinates, and reads unable to  
24 be mapped to the reference; **(b)** *de novo* transcriptome assembly - Trinity is used to  
25 assemble transcripts *de novo* from the filtered BAM file; **(c)** *IG* transcript selection and  
26 quantification – the assembled transcripts are run through a BLAST query to identify  
27 transcripts overlapping *IG* reference sequences. The abundance of the *IG*-derived  
28 transcripts is then estimated using Kallisto pseudoalignment; **(d)** *IG* transcript  
29 annotation and interpretation – the five most abundant transcripts by TPM are then run  
30 through IMG-T/V-QEUEST for *IG* alignment and annotation which is used to recover  
31 the dominant *IG* transcript originating from the putative tumor *IG* gene using a 5-step  
32 hierarchical selection process.

33

34

Box 1

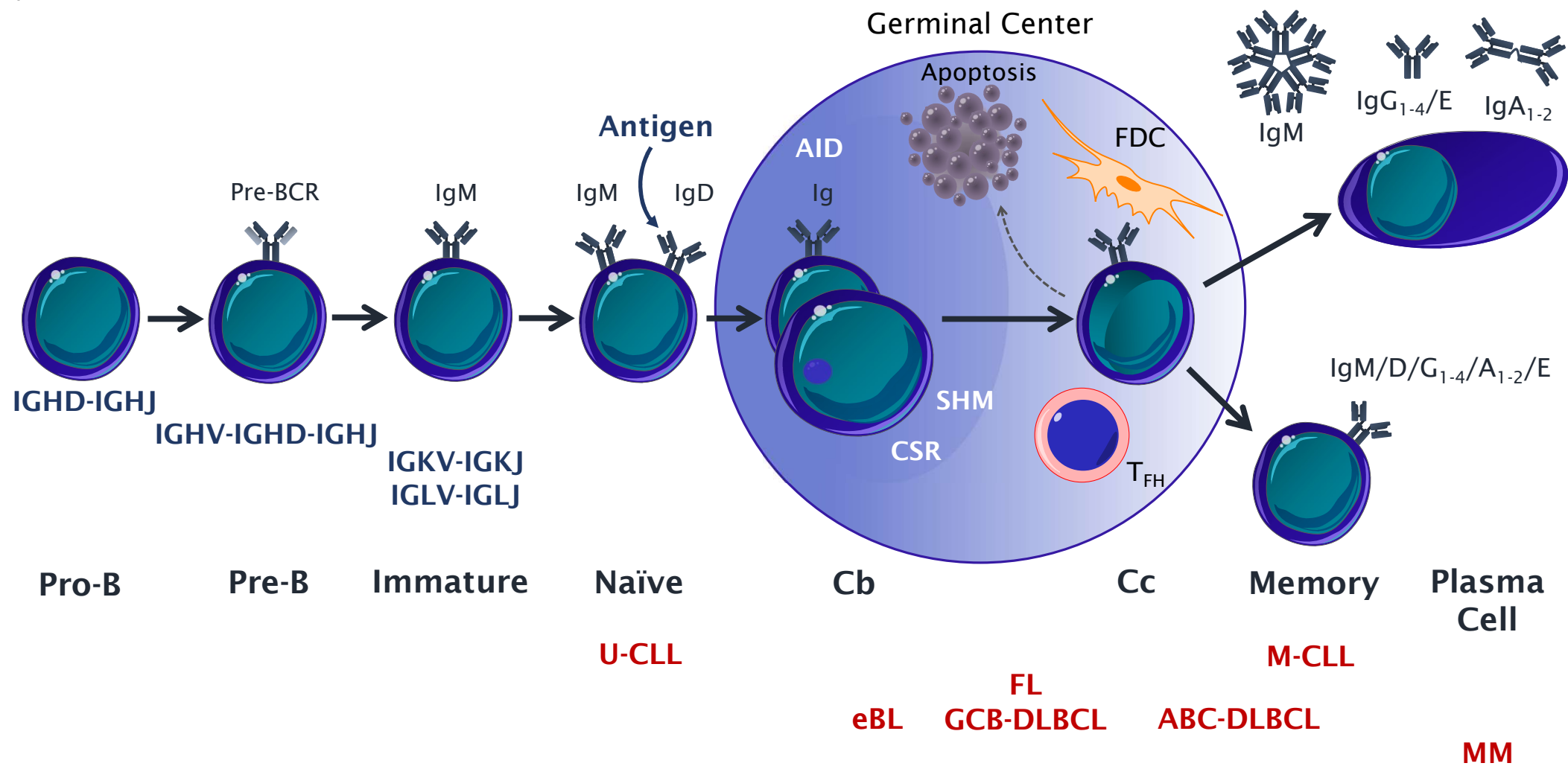


Figure 1

	Sanger (n = 37)	IgSeqR (n = 489)	P-value
IG Transcript Recovered	11 (29.7%)	339 (69.3%)	<0.0001
IG Transcript Not Recovered	26 (70.3%)	150 (30.7%)	

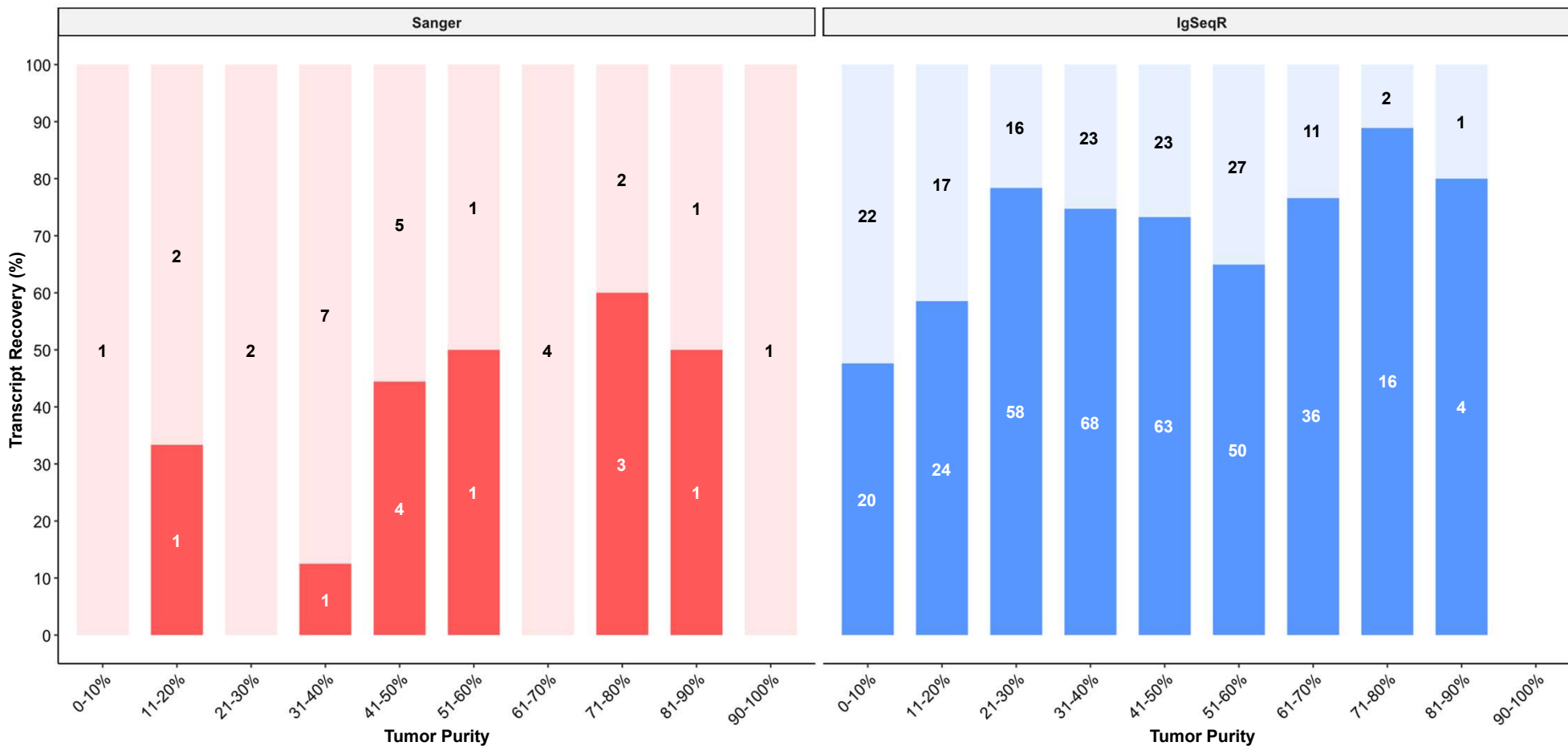


Figure 2

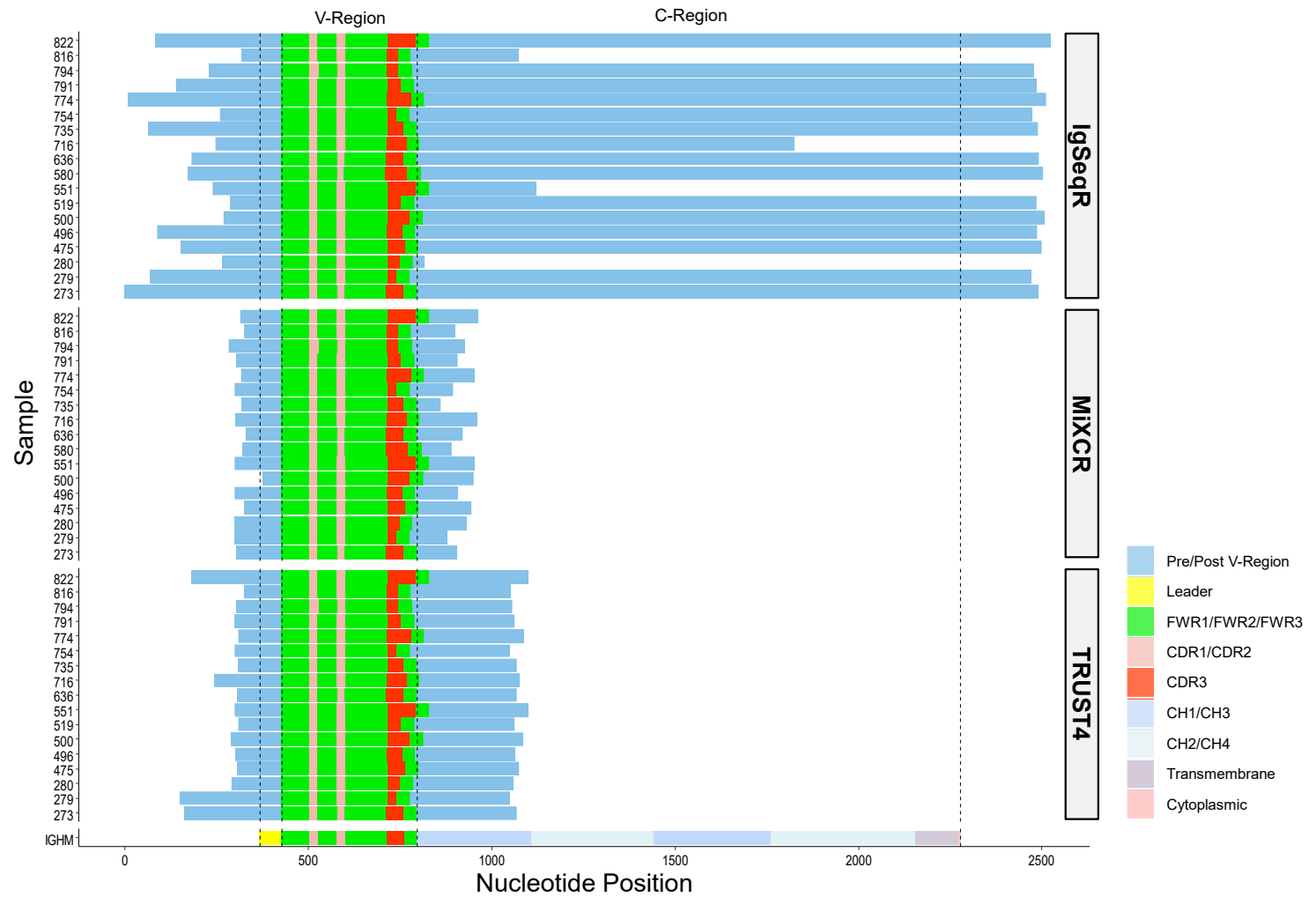
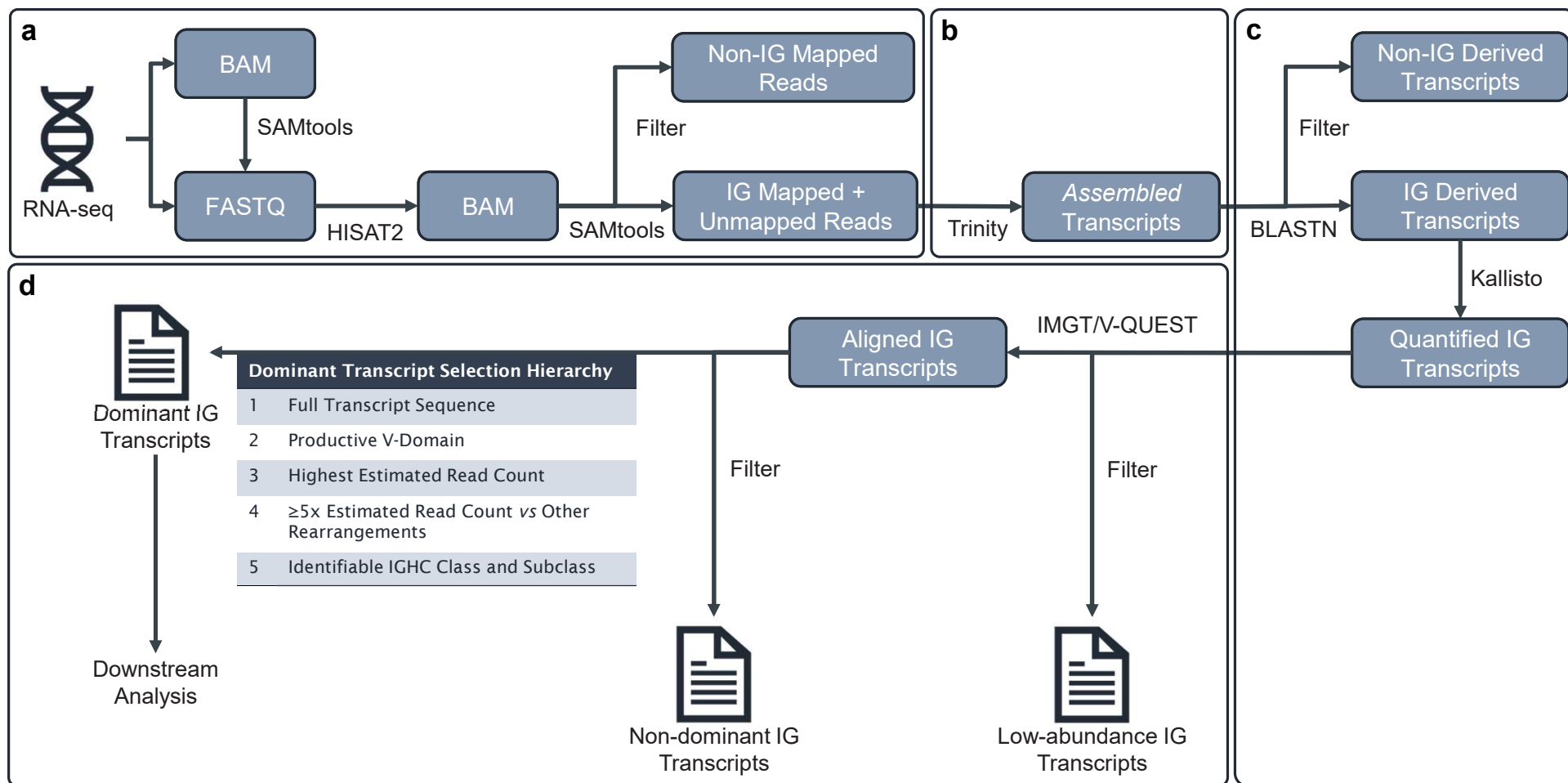


Figure 3



Box 1

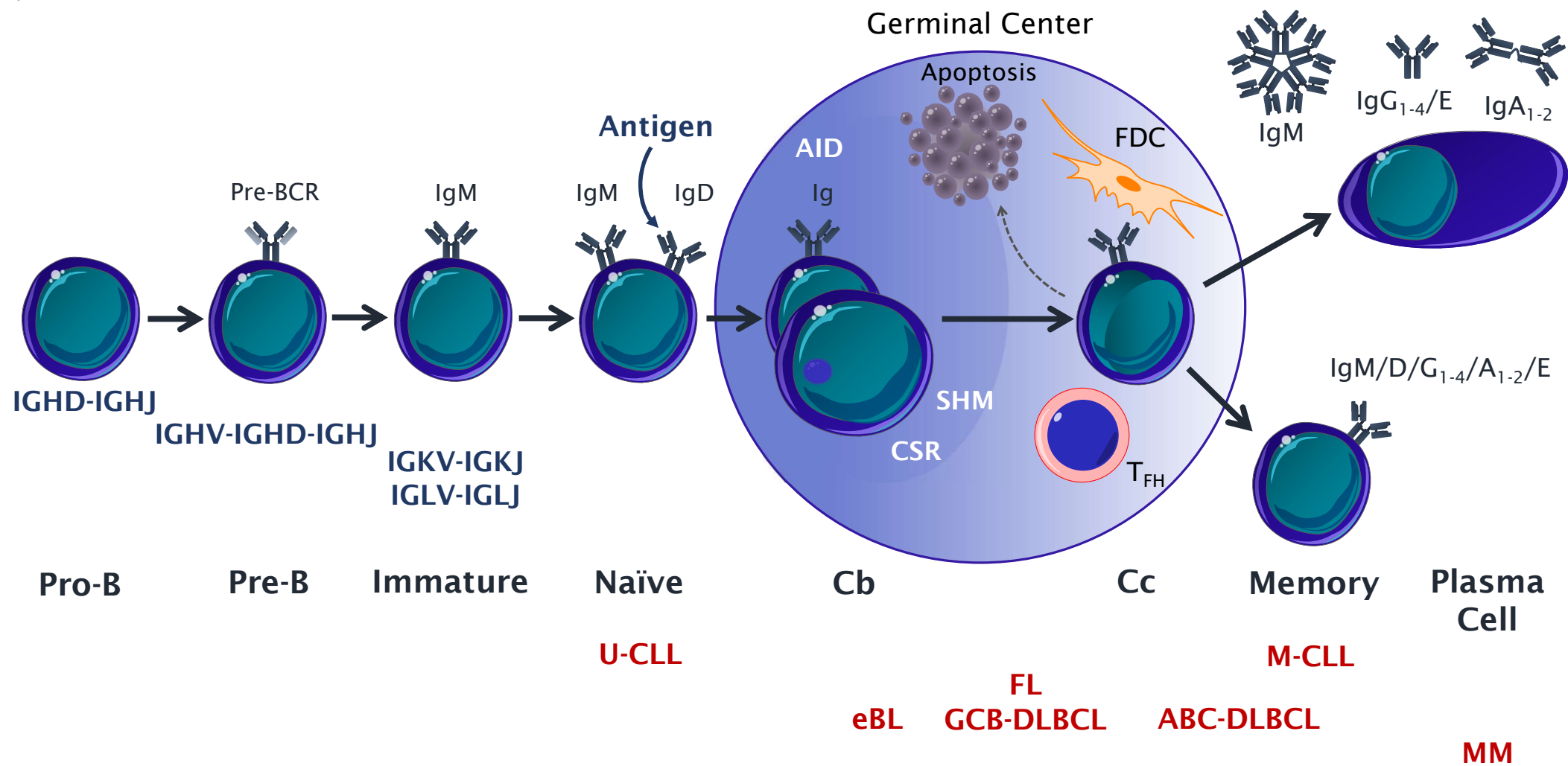


Figure 1

	Sanger (n = 37)	IgSeqR (n = 489)	P-value
IG Transcript Recovered	11 (29.7%)	339 (69.3%)	<0.0001
IG Transcript Not Recovered	26 (70.3%)	150 (30.7%)	

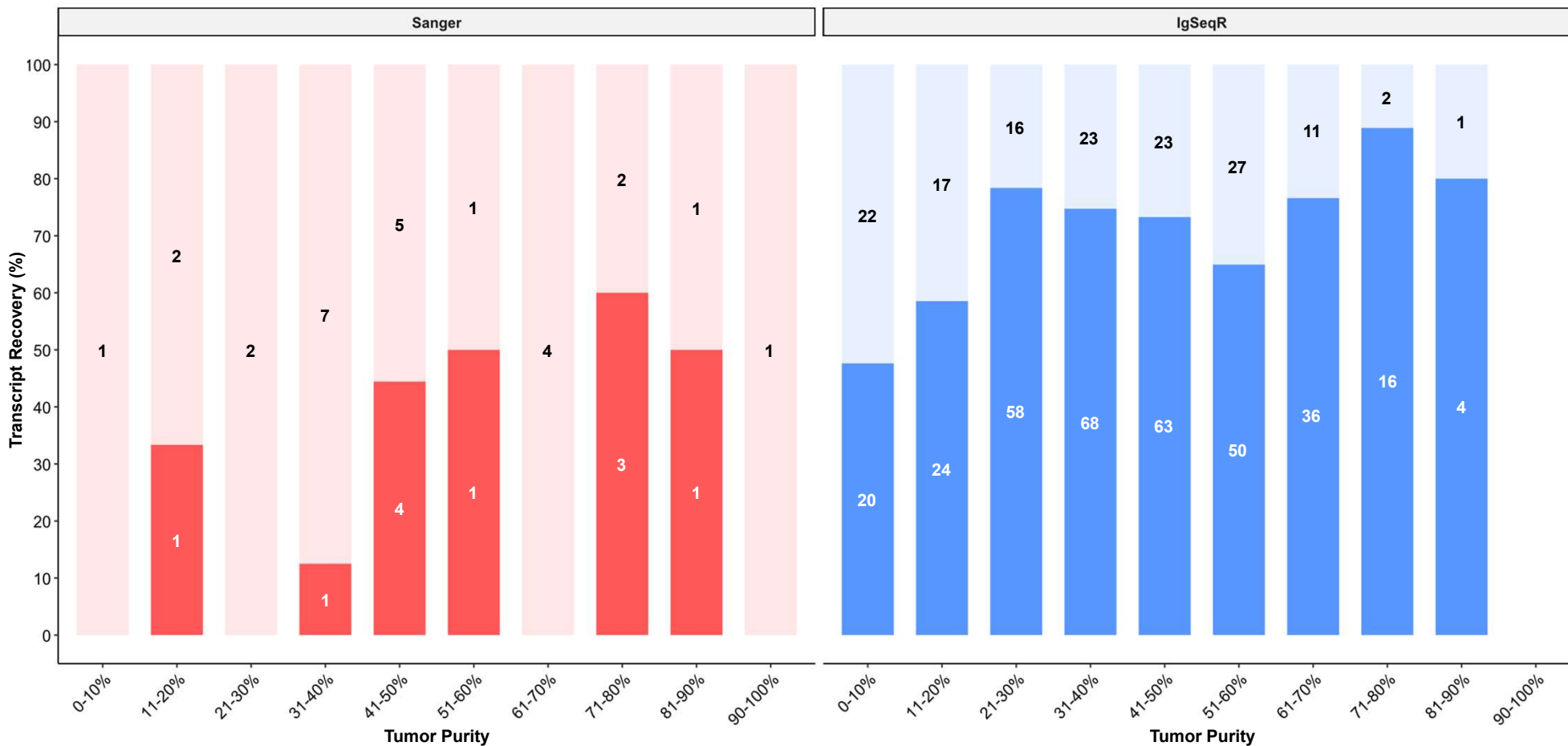


Figure 2

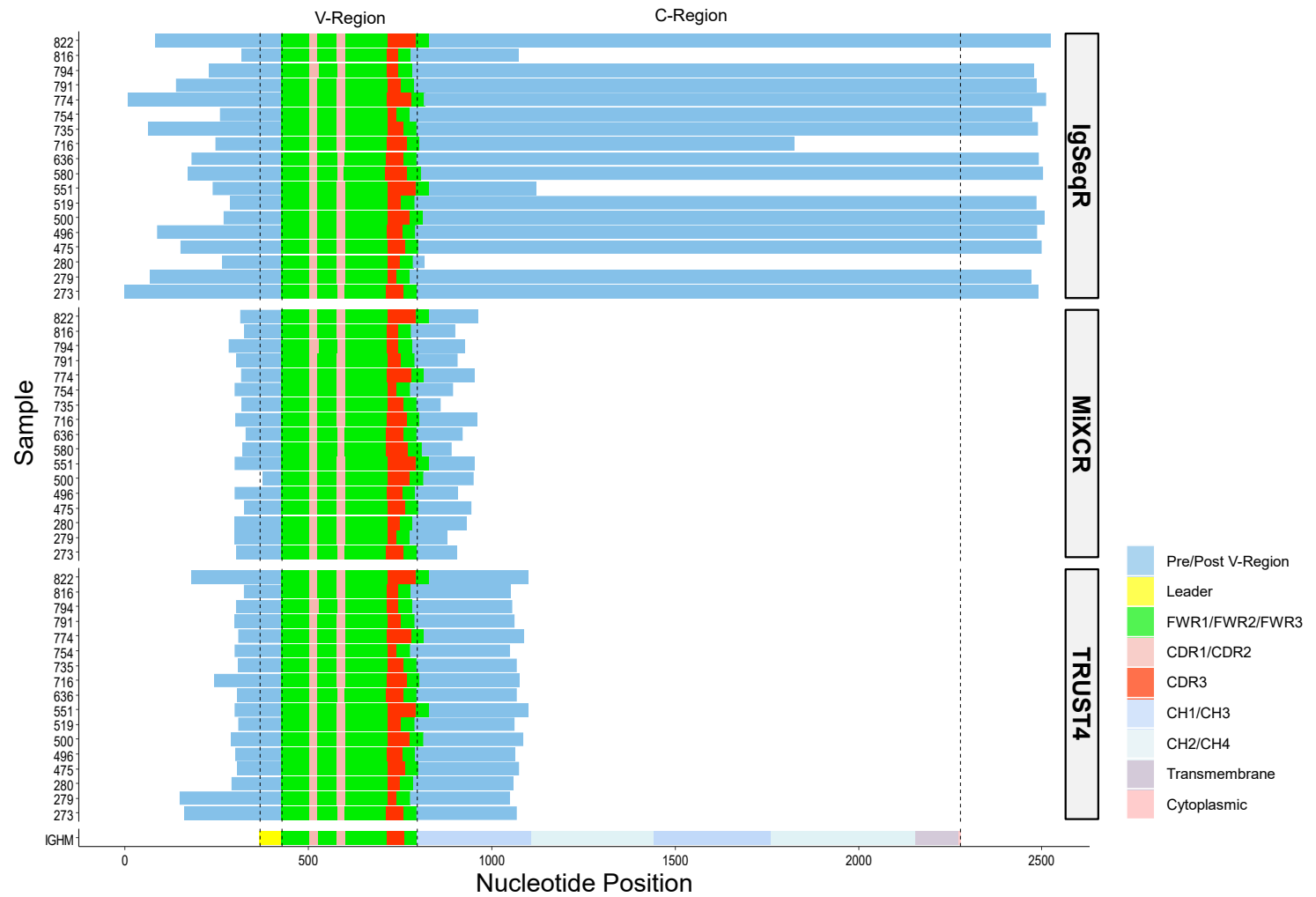


Figure 3

