Genome analysis

CRIS: Complete Reconstruction of Immunoglobulin V-D-J Sequences from RNA-seq data

Rashedul Islam¹,²,³, Misha Bilenky³, Andrew P. Weng⁴,⁵, Joseph M. Connors⁶ and Martin Hirst¹,²,³*¹

¹Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC V5Z 4S6, Canada
²Department of Microbiology and Immunology, Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z3, Canada
³Canada’s Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC V5Z 4S6, Canada
⁴Terry Fox Laboratory, BC Cancer, Vancouver, BC V5Z 1L3, Canada
⁵Department of Pathology & Laboratory Medicine, University of British Columbia, Vancouver, BC V6T 2B5, Canada
⁶Department of Medical Oncology, BC Cancer, Vancouver, BC V5Z 4E6, Canada

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX
Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: B-cells display remarkable diversity in producing B-cell receptors through recombination of immunoglobulin V-D-J genes. Somatic hypermutation of immunoglobulin heavy chain variable (IGHV) genes are used as a prognostic marker in B-cell malignancies. Clinically, IGHV mutation status is determined by targeted Sanger sequencing which is a resource intensive and low-throughput procedure. Here we describe a bioinformatic pipeline, CRIS (Complete Reconstruction of Immunoglobulin IGHV-D-J Sequences) that uses RNA sequencing (RNA-seq) datasets to reconstruct IGHV-D-J sequences and determine IGHV somatic hypermutation status.

Results: CRIS extracts RNA-seq reads aligned to immunoglobulin gene (Ig) loci, performs assembly of Ig-transcripts and aligns the resulting contigs to reference Ig sequences to enumerate and classify somatic hypermutations in the IGHV gene sequence. CRIS improves on existing tools that infer the B-cell receptor (BCR) repertoire from RNA-seq data using a portion IGHV gene segment by de novo assembly. We show that the somatic hypermutation status identified by CRIS using the entire IGHV gene segment is highly concordant with clinical classification in three independent chronic lymphocytic leukemia patient cohorts.

Availability: The CRIS pipeline is available under the MIT License from https://github.com/Rashedul/CRIS.

Contact: mhirst@bcgsc.ca

Supplementary information: Supplementary data are available at Bioinformatics Advances online.

Key words: Immunoglobulin, somatic hypermutation, V-D-J sequence, RNA-seq, chronic lymphocytic leukemia.
1 Introduction

During development in the bone marrow, B lymphocytes undergo rearrangement of immunoglobulin (Ig) heavy (V, D, and J) and light chain (V and J) gene segments through recombination (Fig. 1). Addition or deletion of nucleotides occurs at segment junctions during recombination.

In the germline centre, B-cells acquire additional somatic hypermutation (SHM) within the Ig variable regions as part of the adaptive immune response to generate a B-cell receptor (BCR) repertoire diversity estimated to be as much as ~10^18 (Janeway C. A., Travers P., Walport M., 2004; Briney et al., 2019). Following SHM, B-cells are positively selected for further differentiation into memory B-cells or antibody-secreting plasma cells (Akkaya et al., 2020).

Profiling of the B-cell Ig repertoire has become an essential component of immune research and is used clinically for malignant B-cell classification (Briney et al., 2019; Georgiou et al., 2014). B-cell malignancies arise at different stages of B-cell development and BCR diversification is used as both a prognostic and diagnostic marker (Georgiou et al., 2014; Monk et al., 2017). The presence of SHM and specific usage of IGHV genes are prognostic markers in different B-cell malignancies, including chronic lymphocytic leukemia (CLL), mantle cell lymphoma (MCL) and follicular lymphoma (Damle et al., 1999; Hamblin et al., 1999; Navarro et al., 2012; Berget et al., 2015). Malignant B-cells are classified into two major subtypes based on the SHM status, where cells with very low SHM are classified as “unmutated IGHV” subtype, while those cells with evidence of SHM are classified as “mutated IGHV” subtype. Unmutated IGHV subtypes of CLL and MCL show more aggressive disease compared to the mutated IGHV subtype (Damle et al., 1999; Hamblin et al., 1999; Navarro et al., 2012). IGHV gene usage is also used as a prognostic in follicular lymphoma (Berget et al., 2015).

SHM analysis of the IGHV gene is commonly performed using multiplex PCR and Sanger sequencing following the best practice guidelines by ERIC (Ghia et al., 2007). However the PCR-Sanger method is resource intensive and technically challenging in both clinical and research applications and suffers from a 9-18% failure rate (Stamatopoulos et al., 2017). Massively parallel sequencing of targeted genomic DNA regions or RNA has emerged as an alternative method to reliably sequence SHM of the IGHV J rearrangement repertoire (Blachly et al., 2017). The gold standard for transcriptome analysis, applied in both clinical and research settings and has been used in limited cases to identify BCR rearrangement repertoire (Blachly et al., 2015; Mose et al.; Iglesia et al., 2014; Monk et al., 2017).

Several bioinformatic pipelines have been developed to infer BCR repertoire from RNA-seq data, including ABRA (Iglesia et al., 2014), TRUST (Hu et al., 2019), ImReP (Mandric et al., 2020), MiXCR (Bolotin et al., 2015), V’DJer (Mose et al.), IgId (Blachly et al., 2015). Among them ABRA (Iglesia et al., 2014) and IgId (Blachly et al., 2015) were not published with stand-alone code to allow for replication. The remaining IGHV-D-J reconstruction tools (e.g., TRUST, ImReP, MiXCR, V’DJer) were designed to reconstruct only the CDR3 region, representing only a portion of the IGHV gene, while the entire IGHV gene segment is required to determine the SHM status in B-cell malignancies. In addition, these tools have not been validated against gold standard PCR-Sanger datasets for SHM classification. To address these gaps in determining IGHV mutational status in B-cell malignancies, we developed a bioinformatic pipeline, CRIS, that extracts RNA-seq reads aligned to putative Ig loci, assembles the complete IGHV gene, identifies the most abundant Ig transcript, and enumerates somatic hypermutations by comparison with germline reference sequences. Classification of IGHV mutational subtypes by CRIS was validated against PCR-Sanger based clinical classification in three independent cohorts of CLL patients and shown to be comparable.

Fig. 1. IGHV-D-J rearrangement and somatic hypermutation during B-cell development. BCRs are generated by ordered assembly of the Ig heavy chain gene segments (V, D and J) during B-cell development. Addition and deletion of junctional nucleotides (N) contributes to the diversity of BCR repertoires. BCR sequences undergo affinity maturation upon antigen stimulation through somatic hypermutations in the variable domain (indicated in black arrows). Somatic hypermutations of Ig are enriched at the complementarity-determining regions (CDRs).

2 Methods

2.1 CLL samples

In the Centre for Epigenomic Technology (CEMT) cohort, peripheral blood samples were obtained from chronic lymphocytic leukemia patients undergoing treatment at BC Cancer (n=16) and used according to procedures approved by the Research Ethics Board (REB H12-01767) of the University of British Columbia (Table S1). RNA was purified from those peripheral blood samples and extraction was performed on CD19+ sorted cells with >90% purity as described (Pellacani et al., 2016).

2.2 RNA sequencing

The CEMT CLL RNA-seq datasets were generated as described (Pellacani et al., 2016). RNA extraction, library construction and sequencing were performed following the guidelines formulated by the International Human Epigenome Consortium (http://www.ihc-epigenomes.org). These guidelines as well as the standard operating procedures for RNA-seq library construction and sequencing are available at https://thisisepigenetics.ca/for-scientists/protocols-and-standards and by request. Additional CLL patient RNA-seq datasets with matching IGHV mutation status were collected from published datasets: GSE66228 (Blachly et al., 2015), EGAD00001004046 (Beekman et al., 2018) and phs000435.v3(Wang et al., 2011).

2.3 Identification of putative Ig loci

We identified five putative Ig loci enriched with reads that were used to reconstruct Ig containing contigs in the 16 CEMT samples (Table 1). The detailed procedure of identifying Ig loci is described in (Fig. S1a).

Table 1: Genomic coordinates of the putative Ig loci in the GRCh38 reference.

Need to insert Table 1 here.

2.4 CRIS pipeline

Step 1: Read extraction prior to assembly of Ig-transcripts: bg38-bam file was created by aligning the reads to the GRCh38 reference genome using BWA mem (v0.7.6a) (Li and Durbin, 2009). Using sambamba (v0.7.0) (Tarasov et al., 2015), we extracted reads that were aligned to the putative IGHV loci (Table 1) and saved them in fastq format using picard SamToFastq (v2.20.3) (Broad Institute, 2009) (Fig. 2). These resultant paired-end reads originated from the putative Ig loci were used as input

https://mc.manuscriptcentral.com/bioadv
CRIS: Complete Reconstruction of Immunoglobulin V-D-J Sequence

for Trinity (v2.1.1) (Grabherr et al., 2011) for de novo transcriptome assembly.

Step 2: Identification of Ig-transcripts and their abundances: Trinity assembly performed in the previous step produced around 250 transcripts per sample. To filter the transcripts that have similarity (expectation value <= 20) with the germline IGy sequences, we used blastn (v2.9.0) (Altschul et al., 1990) with default parameters with a custom database of IGy sequences downloaded from IMGT (Ronique Giudicelli et al.). The resultant Ig-transcripts were used in Salmon (v0.8.1) (Patro et al., 2017) to quantify their abundances with a k-mer of 31bp. Transcript with the highest TPM (Transcripts Per Million) value was marked as the dominant clone.

Step 3: SHM and clonotype analysis: The Ig-transcript sequences identified in step 2 were queried in IgBLAST (v1.14.0) (Ye et al.) against the germline V, D and J gene database of IMGT. IgBLAST returned the percent identity of the IGy segment of Ig-transcripts compared to the germline alleles and clustered the similar Ig-transcripts into clonotypes. Productive Ig-transcript with highest TPM value was used to determine IGy mutation status of CLL sample and further compared with available clinical PCR-Sanger data. Transcripts having TPM varying within one log10 of the highest expressed transcript were also considered while comparing with the PCR-Sanger data according to (Blachly et al., 2015).

2.5 Analysis of SHM status using V’D’Jer, TRUST and MIXCR: V’D’Jer, TRUST (v3.0.3) and MIXCR (v3.0.3) were run on the RNA-seq bam file generated by STAR (v2.7.5a) aligner (Dobin et al., 2013) using GRC38 genome as reference. During STAR alignment ‘--outSAMunmapped Within’ was used to include the unmapped reads in the bam file. All three tools were run with default parameters to generate VDJ contigs of IGy. VDJ contigs were analyzed by IgBLAST to generate the percent identity of IGy sequences compared to the germline database.

Need to insert Fig. 2 here.

Fig. 2. CRIS workflow. CRIS extract reads from the putative Ig loci prior to assembly of Ig-transcripts and quantify transcript abundances. The percent of IGy mutations of Ig-transcripts are calculated by comparing to the germline sequences.

3 Results

3.1 De Novo assembly-based Ig detection from RNA-seq

De novo assembly using Trinity (Grabherr et al., 2011) for 16 deeply sequenced (~300M read pairs) CLL RNA-seq libraries generated an average of ~450,000 contigs per sample with 6-29 contigs demonstrating IGy sequence homology. However, de novo assembly of the complete RNA-seq read sets required significant computational resources (Hölzer and Marz, 2019) and thus we sought to identify the fraction of reads in the RNA-seq libraries corresponding to the Ig loci. Using the resulting assemblies, we found that on average 99.85% of the sequence reads used to reconstruct IGy containing contigs originated from five putative Ig loci in the GRC38 reference (Table S2). These putative Ig loci consist of human Ig locus, Ig pseudogene loci and unlocalized contigs at chromosome 14, 15 and 16 (Table 1, fig. S2). This suggests that sequence reads used to reconstruct Ig sequence not only map to the reference Ig locus but also to pseudogene regions both within the current assembly and in unlocalized contigs. We hypothesized that this novel set of loci could be used as a highly specific filter to reconstruct IGy-D-J sequence.

3.2 CRIS pipeline development

Given the time required to complete a full assembly from an RNA-seq library we sought to extract Ig sequences prior to performing assembly. For this we leveraged the putative Ig loci identified in our pilot set of libraries, retrieved sequences aligned by BWA-mem (Li and Durbin, 2009) within these co-ordinates (~1% of all reads) and subjected these to de novo assembly using Trinity (Grabherr et al., 2011). Enriching for Ig sequences from the bulk RNA-seq sequence set reduced the run time for assembly by two orders of magnitude while not significantly impacting the subsequent SHM analysis of Ig transcripts (Table S3). We confirmed that our approach also successfully assembled the IGy-D-J and N-junctional segments (Fig. 3a, Table 2 and S4).

Having established that enrichment of sequences using our Ig feature set significantly reduced compute resources without a reduction in sensitivity we next examined the impact of RNA-seq sequencing depth. For this we leveraged a set 16 CLL RNA-seq libraries with an average of ~28 million paired reads from GSE66228 (Blachly et al., 2015) and compared these to the results obtained from our deeply sequenced CEMT libraries (~300 million paired reads). We found no significant difference in the fraction of the IGy gene assembled between deep and shallow RNA-seq libraries (Fig. 3b). This appears to be in part due to the high expression level of the dominating clone in the GSE66228 dataset (Blachly et al., 2015) driving sufficient sequence read coverage (at least 10^4 reads) for Trinity to assemble the Ig transcript. However, as expected, the overall number of Ig-transcripts identified correlated with the sequencing depth. Based on this analysis we developed a pipeline called Complete Reconstruction of Immunoglobulin V-D-J Sequences (CRIS) and benchmarked its ability to call IGy mutation status. In the CRIS pipeline, we automate the process of read extraction from our novel putative Ig coordinates, perform quality trimming of selected sequence reads, assemble transcripts, enumerate transcript abundances and identify somatic mutations using reference germline sequences for SHM classification.

Need to insert Fig. 3 here.

Fig. 3. Evaluation of CRIS to reconstruct IGy-D-J sequences. a) The most abundant Ig-transcript from US-142278 sample was aligned to the germline database using IgBLAST where top hit germline genes are shown. In the alignment mismatches are represented as nucleotide bases and matches as dots. The alignment length, number of matches and mismatches are 296, 280 and 16 respectively. Total number of matched nucleotides between query and germline IGy sequence is used to calculate percent identity e.g., 100*(280/296) = 96.4%. N-junctional sequences are highlighted in gray boxes. b) Fraction of the IGy gene assembled in two CLL RNA-seq datasets with different sequence depths and lengths as indicated. An unpaired two-tailed t-test demonstrated no significant (p > 0.15) difference between the two distributions (NS) c) Scatter plot comparing the percent of mutation of IGy as predicted by CRIS and clinical PCR-Sanger based analysis for 16 CLL patient samples obtained from GSE66228.

3.3 CRIS is concordant with clinical IGy mutation status

Having established that CRIS could efficiently assemble IGy transcripts we explored its ability to call SHM mutation status in CLL. Unmutated CLL (uCLL) is clinically defined by IGy sequence alignments of >98% identity to the reference sequence (Georgiou et al., 2014; Monk et al., 2017). To benchmark CRIS against gold standard Sanger-based clinical classification we analyzed a series of published RNA-seq libraries from...
CLL patients with matched Sanger sequencing classifications. CRIS reported SHM on clonally amplified Ig-transcripts and its classification of mutated/unmutated CLL (mCLL/uCLL) showed perfect concordance with Sanger based clinical calls in the GSE66228 dataset (Blachly et al., 2015) (Table 2). The percent mutations reported by CRIS and the clinical test was also highly correlated (Pearson’s r=0.95, 95% CI 0.86-0.98). Fig. 3c. The reported IGHV mutational frequency was identical in 8/16 cases with the remaining cases showing small deviations (mean deviation 0.22%) that did not change the SHM classification. In 7 of the 8 divergent cases, the percent IGHV identity reported by the Sanger based test was higher compared to CRIS (Table 2). Closer inspection of the alignments revealed that this likely an artifact in the Sanger calls due to incomplete IGHV coverage by the PCR product used as denominator to calculate percent identity (Blachly et al., 2015). In addition to calling mCLL/uCLL status, CRIS also reported 2-8 dominant clonotypes in the GSE66228 dataset, a feature not detected by clinical Sanger-based classifiers.

We further benchmarked CRIS using two independent CLL RNA-seq datasets with matched IGHV mutation status determined by Sanger sequencing. In the phs000435.v3 dataset (Wang et al., 2011), CRIS calls were identical to the Sanger based calls in 50/51 cases with 98.3% accuracy, 100% sensitivity and 97.3% specificity (Fig. 4a). A single sample (DFCI-5121) was reported as mCLL (Wang et al., 2011) however, CRIS determined it as uCLL. In the third independent dataset, EGAD00001004046 (Beekman et al., 2018), CRIS agreed with clinical classification in all cases and determined the identical IGHV gene as the dominant clone (Fig. 4b, fig. S1b).

### Table 2: Concordance of IGHV gene prediction and percent mutation between PCR-Sanger based analysis and CRIS. CRIS reconstructed V-D-J segments of Ig-transcripts and identified multiple transcripts per sample that belong to different clonotypes. NA is used in cases where IGHV genes were absent.

| Sample ID   | IGHV Gene Call | CRIS Prediction | PCR-Sanger Call |
|-------------|----------------|----------------|-----------------|
| DFCI-5121   | IGHV1-69       | IGHV3-74*93    | IGHV3-74*93    |
| DFCI-5122   | IGHV1-69       | IGHV3-74*93    | IGHV3-74*93    |
| DFCI-5123   | IGHV1-69       | IGHV3-74*93    | IGHV3-74*93    |

### Need to insert Table 2 here.

#### 3.4 Comparison of CRIS against existing tools

We next compared CRIS with previously published tools: V’DJer (Mose et al.), TRUST (Hu et al., 2019) and MiXCR (Bolotin et al., 2015) that reconstruct BCR repertoires from short read RNA-seq data. In 16 CLL RNA-seq samples obtained from GSE66228 (Blachly et al., 2015), V’DJer did not produce full length IGHV gene as it is designed to generate contigs of fixed length (360bp) spanning the CDR3 region. Thus, on average, V’DJer assembled 75.44% of the IGHV gene whereas CRIS reconstructed 99.74% (Fig. 4e). Partial reconstruction of the IGHV gene could lead to mis-classification of IGHV mutation status especially for samples with IGHV sequence identity near the established 98% cutoff. For example, CRIS reconstructed 295bp out of 296bp of the IGHV3-74*93 sequence whereas V’DJer assembled 226bp in US-1422368 (Fig. S3a,b). The additional 69bp reported by CRIS contained two mutations that resulted in a 1.2% difference in reported percent identity between CRIS (91.2%) and V’DJer (89.4%). TRUST assembled only 41.5% of the IGHV gene on average using the GSE66228 dataset (Fig. 4e). Furthermore, V’DJer and TRUST did not produce a contig for US-142282 that contained IGHV1-69 gene whereas CRIS generated IGHV1-69 containing contig in agreement with the clinical call.

To compare the computational performance between CRIS and V’DJer, both of the pipelines were configured to use up to 16 threads. In the shallow libraries from GSE66228 dataset, CRIS had ~14% faster total run time (average 3.07 wall-clock minutes) compared to V’DJer (average 3.50 wall-clock minutes). Using deeper RNA-seq datasets (~300 million reads) V’DJer took 5 times more time to run than CRIS (87 vs. 16 wall-clock minutes on average). Using 16 threads, TRUST took 36 wall-clock minutes on average using GSE66228, an order of magnitude longer than CRIS. MiXCR (Bolotin et al., 2015) generated partial CDR3 sequence contigs with <10% of IGHV gene sequence in the GSE66228 dataset of 75bp read length. MiXCR recommends >=100bp read length to extract CDR3 repertoires from RNA-Seq data. Thus, our comparisons suggest that existing BCR reconstruction tools developed to extract just CDR3 regions perform poorly compared to CRIS in the determination of SHM status because they are designed to generate and analyze partial IGHV sequences. Overall, CRIS showed increased sensitivity and specificity and reduced run time over existing RNA-seq based BCR reconstruction tools.

### Need to insert Fig. 4 here.

**Fig. 4. Comparison of CRIS with clinical data and existing tools.** a,b) Confusion matrix represents the classification accuracy of CRIS compared to Sanger-PCR data in two independent CLL cohorts. The P-value was calculated by one sided binomial test. c) Comparison of CRIS, V’DJer and TRUST to reconstruct the proportion of IGHV sequences in GSE66228 (Blachly et al., 2015) dataset. Average fraction of IGHV gene length for each tool is represented by dashed horizontal lines.

#### 4 Discussion

PCR-Sanger based Ig SHM classification is resource intensive, subject to PCR bias, and suffers from an ~9-18% failure rate (Ghia et al., 2007; Stamatopoulos et al., 2017). In contrast, RNA-seq is now routinely applied in the clinical setting, eliminates the need for targeted amplification of Ig locus and can be used to identify BCR rearrangement repertoire (Blachly et al., 2015). Here we showed that CRIS can rapidly analyze RNA-seq to detect IGHV mutation status in CLL at a sensitivity and specificity equivalent to current Sanger-based clinical tests. Furthermore, CRIS was able to reconstruct the entire IGHV sequence thus increasing the accuracy of SHM classification. This is in contrast to a majority of existing pipelines designed to infer only CDR3 derived sequences (Hu et al., 2019; Bolotin et al., 2015; Mose et al.).

A registry of ~1,500 CLL patients showed that 90% of patients were not screened for IGHV mutations (Mato et al., 2016). In the public domain there are thousands of RNA-seq data available for different B-cell malignancies but their SHM status of IGHV genes is either not reported or partially reported. Furthermore, for a majority of publicly available RNA-seq datasets where SHM status is reported, detailed IGHV mutation reports with gene name, percent identity and clonal frequency are not available restricting the ability assess mutational values. To meet this need we developed CRIS and demonstrated its ability to rapidly classify IGHV mutational status with clinical accuracy. We anticipate that CRIS will prove to be useful in the mining of available B-cell RNA-seq datasets and that it will provide a framework to incorporate RNA-seq as a diagnostic tool to examine the BCR clonal rearrangement and SHM status.

### Acknowledgements

The authors wish to acknowledge Canada’s Michael Smith Genome Sciences Centre, Vancouver, BC Canada for computational resources and support. The results published here are in part based upon data generated by The Canadian Epigenetics, Epigenomics, Environment and Health Research Consortium (CEHRC) initiative.

---

https://mc.manuscriptcentral.com/bioadv
funded by the Canadian Institutes of Health Research (CIHR), Genome BC, and Genome Quebec. Information about CEEHRC and the participating investigators and institutions can be found at http://www.ciehrc-rccie.gc.ca/e/43734.html. This study makes use of data (EGAD00010004046) generated by the Blueprint Consortium. A full list of the investigators who contributed to the generation of the data is available from www.blueprint-epigenome.eu. Funding for the project was provided by the European Union’s Seventh Framework Programme [FP7/2007-2013] under grant agreement no 282510 – BLUEPRINT. This study makes use of phs000435.v3 from NIH dbGaP, which was supported by the National Human Genome Research Institute [S54HG003067], the Blavatnik Family Foundation, NCI (5R21CA115043-2), the Howard Hughes Medical Institute and the Damon-Runyon Cancer Research Foundation [CI-38-07].

Funding
This work was supported by the Canadian Institutes of Health Research (EPI-120589 & CEE-151619) and Genome Canada (C41EMT & R32EMT) under the Canadian Epigenetics, Environment and Health Research Consortium (to M.H.) and by the Terry Fox Research Institute Program Projects [TF-122869, TFR1-1074] awarded to M.H. R.I. was supported by CIHR Bioinformatics Training Program for Health Research, University of British Columbia, Vancouver, BC Canada.

Conflict of Interest: none declared.

References
Akkaya,M. et al. (2020) B cell memory: building two walls of protection against pathogens. Nat. Rev. Immunol., 20, 229–238.
Altschul,S.F. et al. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403–410.
Beekman,R. et al. (2018) The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. Nat. Med., 24, 868–880.
Berget,E. et al. (2015) IGHV gene usage and mutational status in follicular lymphoma: Correlations with prognosis and patient age. Leuk. Res., 39, 702–708.
Blachly,J.S. et al. (2015) Immunoglobulin transcript sequence and somatic hypermutation computation from unselected RNA-seq reads in chronic lymphocytic leukemia. Proc. Natl. Acad. Sci. U.S.A., 112, 4322–4327.
Boletin,D.A. et al. (2015) MXCR: Software for comprehensive adaptive immunity profiling. Nat. Methods, 12, 380–381.
Boyd,S.D. and Joshua,S.A. (2014) High-Throughput DNA Sequencing Analysis of Antibody Repertoires. Microbiol. Spectr., 2.
Briney,B. et al. (2019) Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature, 566, 383–397.
Broad Institute (2009) Picard Tools - By Broad Institute. Github.
Damle,R.N. et al. (1999) Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. Blood, 94, 1840–7.
Dobin,A. et al. (2013) STAR: Ultrafast universal RNA-seq aligner. Bioinformatics, 29, 15–21.
Georgiou,G. et al. (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat. Biotechnol., 32, 158–168.
Ghia,P. et al. (2007) ERIC recommendations on IGHV gene mutational status analysis in chronic lymphocytic leukemia. Leukemia, 21, 1–3.
Grabherr,M.G. et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol., 29, 644–652.
Hamblin,T.J. et al. (1999) Unmutated Ig VH genes are associated with a more aggressive form of chronic lymphocytic leukemia. Blood, 94, 1984–54.
Hölzer,M. and Marz,M. (2019) De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience, 8, 1–16.
Hu,X. et al. (2019) Landscape of B cell immunity and related immune evasion in human cancers. Nat. Genet., 51, 560–567.
Iglesia,M.D. et al. (2014) Prognostic B-cell signatures using mRNA-seq in patients with subtype-specific breast and ovarian cancer. Clin. Cancer Res., 20, 3818–3829.
Janeway C. A., Travers P., Walport M.,S.M.J. (2004) Immunobiology Garland Science.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754–60.
Mandric,I. et al. (2020) Profiling immunoglobulin repertoires across multiple human tissues using RNA sequencing. Nat. Commun., 11, 1–14.
Mato,A. et al. (2016) Real-world clinical experience in the Connect ® chronic lymphocytic leukemia registry: a prospective cohort study of 1494 patients across 199 US centres. Br. J. Haematol., 175, 892–903.
Menzel,U. et al. (2014) Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. PLoS One, 9.
Monk,J.M. et al. (2017) Antigen receptor repertoire profiling from RNA-seq data. Nat. Biotechnol., 35, 908–911.
Mose,L.E. et al. Assembly-based inference of B-cell receptor repertoires from short read RNA sequencing data with VDJer.
Navarro,A. et al. (2012) Molecular subsets of mantle cell lymphoma defined by the IGHV mutational status and SOX11 expression have distinct biologic and clinical features. Cancer Res., 72, 5307–5316.
Patro,R. et al. (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods, 14, 417–419.
Pellacani,D. et al. (2016) Analysis of Normal Human Mammary Epigenomes Reveals Cell-Specific Active Enhancer States and Associated Transcription Factor Networks. Cell Rep., 17, 2060–2074.
Ronique Giudicelli,V. et al. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes.
Stamatopoulos,B. et al. (2017) Targeted deep sequencing reveals clinically relevant subclonal IgHV rearrangements in chronic lymphocytic leukemia. Leukemia, 31, 837–845.
Tarasov,A. et al. (2015) Sambamba: fast processing of NGS alignment formats. Bioinformatics, 31, 2032.
Wang,L. et al. (2011) SF3BI and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia. N. Engl. J. Med., 365, 2497–2506.
Yaari,G. and Kleinstein,S.H. (2015) Practical guidelines for B-cell receptor repertoire sequencing analysis. Genome Med., 7, 1–14.
Ye,J. et al. IgBLAST: an immunoglobulin variable domain sequence analysis tool. academic.oup.com.
Table 1: Genomic coordinates of the putative Ig loci in the GRCh38 reference.

| Chromosome/Contig       | Start     | End       | Length (bp) |
|-------------------------|-----------|-----------|-------------|
| Chr14                   | 105,550,001 | 106,880,000 | 1,329,999   |
| Chr15                   | 21,710,000  | 22,190,000  | 480,000     |
| Chr16                   | 31,950,001  | 33,970,000  | 2,019,999   |
| chr14_KI270726v1_random | 1         | 43,739     | 43,739      |
| chr16_KI270728v1_random | 1         | 1,872,759   | 1,872,759   |
Table 2: Concordance of \textit{IGHV} gene prediction and percent mutation between PCR-Sanger based analysis and CRIS. CRIS reconstructed \textit{V-D-J} segments of Ig-transcripts and identified multiple transcripts per sample that belong to different clonotypes. NA is used in cases where \textit{IGHD} genes were absent.

| Sample ID   | IGHV | % mutation | IGHV     | % \textit{IGHV} mutation | IGHD     | IGHJ     | # Ig-transcript | # clonotype |
|-------------|------|------------|----------|--------------------------|----------|----------|----------------|------------|
| US-1422282  | V1-69| 0.4        | IGHV1-69*04| 0.3         | IGHD6-19*01| IGHJ4*02| 7              | 4          |
| US-1422366  | V1-18| 0.34       | IGHV1-18*04| 0           | IGHD3-3*01| IGHJ6*02| 21             | 5          |
| US-1422311  | V3-11| 2          | IGHV3-11*01| 2           | IGHD4-17*01| IGHJ4*02| 5              | 4          |
| US-1422278  | V3-74| 5.4        | IGHV3-74*01| 5.4         | IGHD5-18*01| IGHJ6*02| 5              | 3          |
| US-1422335  | V4-59| 10.2       | IGHV4-59*02| 8.5         | IGHD3-10*01| IGHJ4*02| 3              | 2          |
| US-1422321  | V3-66| 0.7        | IGHV3-66*02| 0.7         | NA        | IGHJ4*02| 9              | 4          |
| US-1422333  | V4-34| 0          | IGHV4-34*01| 0           | IGHD3-3*01| IGHJ6*02| 6              | 3          |
| US-1422356  | V2-70| 0.8        | IGHV2-70*01| 0.3         | IGHD3-16*01| IGHJ3*02| 15             | 8          |
| US-1422368  | V3-74| 6.1        | IGHV3-74*03| 8.8         | IGHD1-1*01| IGHJ5*02| 2              | 2          |
| US-1422309  | V3-53| 8.8        | IGHV3-53*01| 6.1         | IGHD3-10*01| IGHJ6*03| 4              | 3          |
| US-1422302  | V2-70| 0.3        | IGHV2-70*01| 0.3         | IGHD2-15*01| IGHJ4*02| 20             | 4          |
| US-1422351  | V1-46| 0          | IGHV1-46*01| 0           | IGHD3-10*01| IGHJ4*02| 6              | 3          |
| US-1422314  | V1-3 | 0.7        | IGHV1-3*01| 0           | IGHD6-19*01| IGHJ4*02| 5              | 3          |
| US-1422342  | V3-21| 0          | IGHV3-21*01| 0           | IGHD3-16*01| IGHJ4*02| 4              | 2          |
| US-1422350  | V3-48| 2.8        | IGHV3-48*03| 2.4         | IGHD3-22*01| IGHJ4*02| 3              | 2          |
| US-1422352  | V1-46| 0          | IGHV1-46*01| 0           | IGHD3-22*01| IGHJ6*02| 17             | 4          |
Figure 1

Germline genome of Ig heavy (IGH) chain

V-D-J recombination

Somatic hypermutations

Immunoglobulin

C = constant
D = diversity
J = joining
V = variable
L = light
H = heavy
FW = framework
CDR = complementarity-determining regions
N = junctional diversity sequence

81x108mm (300 x 300 DPI)
CRIS pipeline

RNA-seq hg38 aligned bam file

Germline IGHV sequences

- Extract reads aligned to Ig loci
  - Sambamba
  - bam file of putative Ig loci
- Convert bam to fastq
  - Picard
  - Paired-end fastq files
- Quality trimming of reads
  - Trimmomatic
  - Trimmed fastq files
- Assembly of transcripts
  - Trinity (de novo)
  - Transcripts fasta

- Selecting Ig-transcripts
  - BLASTn; IGHV-db
  - Ig-transcripts
- Ig-transcript expression
  - Salmon
  - Rank transcripts by expression
- IGHV mutation and clonotype
  - IgBLAST
  - Percent of IGHV mutation, clonotypes

Figure 2

76x57mm (300 x 300 DPI)
Figure 3

94x111mm (300 x 300 DPI)
Figure 4

97x104mm (300 x 300 DPI)