DEVELOPMENT OF BIOINFORMATICS TOOLS FOR THE CHARACTERIZATION OF B CELL NEOPLASMS

DESARROLLO DE HERRAMIENTAS BIOINFORMÁTICAS PARA LA

CARACTERIZACIÓN DE NEOPLASIAS LINFOIDES B

FACULTAD DE CIENCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOQUÍMICA Y BIOLOGÍA MOLECULAR
PROGRAMA DE DOCTORADO EN BIOMEDICINA Y BIOTECNOLOGÍA

AZAHARA MARÍA FUENTES TRILLO

SUPERVISORS:

Dr. Javier Chaves Martínez

Dra. M. José Terol Casterá

Dra. Blanca Navarro Cubells

TUTOR:

Dr. Miguel Ángel García Pérez

Valencia, October 2022

UNIVERSITAT DE VALÈNCIA

Programa de Doctorado en Biomedicina y Biotecnología

INCLIVA | VLC
Instituto de Investigación Sanitaria

INSTITUTO DE INVESTIGACIÓN SANITARIA-INCLIVA

Unidad de Genómica y Diabetes

DEVELOPMENT OF BIOINFORMATICS TOOLS FOR THE CHARACTERIZATION OF B CELL NEOPLASMS

Azahara María Fuentes Trillo

Supervisors:

Dr. Javier Chaves Martínez

Dra. M. José Terol Casterá

Dra. Blanca Navarro Cubells

Tutor:

Dr. Miguel Ángel García Pérez

Valencia, October 2022

The supervisors of the present doctoral thesis: Dr F. Javier Chaves, Dra. M. José Terol Casterá and Dra. Blanca Navarro Cubells and the tutor: Dr. Miguel A. García Pérez, CERTIFY that:

Azahara M. Fuentes Trillo, graduated in Biochemistry from University of Cordoba, has carried out the present Doctoral Thesis: "**Development of Bioinformatics tools for the characterization of B cell neoplasms**" and that, to their judgement, meets all the necessary requirements in order to qualify the candidate for a **PhD in Biomedicine and Biotechnology (3102**), for which purpose it will be presented at the University of Valencia. The work has been done under their direction, authorizing its presentation before the Qualifying Court. And for this purpose, this certificate is extended.

Valencia, October 2022.

Thesis Supervisors and tutor:

Dr. F. Javier Chaves | Dra. M. José Terol Casterá | Dra. Blanca Navarro Cubells

Dr. Miguel A. García Pérez

"There is no favorable wind for a man who does not know where to go."

Lucio Anneo Seneca (Córdoba 4 b.C.-Rome 65 a.C.)

Agradecimientos

El camino hacia la tesis doctoral ha sido una experiencia dura a la par que gratificante. Tengo que agradecer a todas las personas que han contribuido a que, a día de hoy, a pesar de todos los baches, pueda sentirme orgullosa del trabajo realizado:

En primer lugar, a mi director de tesis, Javier Chaves. Gracias por confiar en mí y elegirme para formar parte de tu laboratorio, allá por el año 2015. Sin ti como mentor y apoyo para todo, esto no hubiera sido posible. A los compañeros del laboratorio 23 que estaban por aquel entonces, gracias por tener paciencia conmigo y enseñarme tantas cosas: Cristina, Victoria, Griselda, Javi, Inma, Loles, Charo, Laura, Yovero, así como a las compañeras del laboratorio 2: Raquel, Olga y Ana. Gracias a Dani, Alba y Pablo en lo que concierne a la parte bioinformática en mis inicios en este grupo.

A mis codirectoras, Blanca Navarro y Maria José Terol. Gracias por dejarme colaborar con vosotras, por vuestros consejos y por depositar en mí la confianza para la parte bioinformática de muchos de vuestros proyectos, que me han llevado a aprender tanto. Gracias a Blanca Ferrer por su apoyo y confianza también desde el ámbito clínico.

Gracias a Pilar por estar dispuesta siempre a echar una mano, su inteligencia, personalidad todoterreno y por cuidar de nosotros. A Tani, por ser la voz de la sabiduría en el laboratorio y en otros ámbitos de la vida. Gracias a Kike, o mejor dicho MiKike, por la simbiosis con el MiSeq y por ser la alegría de la huerta y a la vez, una enciclopedia. Gracias por ~~Normal~~ Molar tanto. Roberto, por su naturalidad y buena disposición para todo, ya veremos si se cumple tu profecía sobre las tesis. En general a todos los precisos por los buenos ratos que todos ellos han brindado al laboratorio.

A Miguel y Alex de la Unidad de Bioinformática, que han sido el apoyo informático y también compañeros durante los primeros años en INCLIVA. Alex, aunque alguna vez nos sacabas de quicio con tus cosquillas, te queremos por tu humor y los dulces de la beata. A Miguel, gracias por la paciencia y por darme todos los recursos para el análisis en los servidores (además de alguna que otra visita a los montaditos). Gracias a vosotros los NFSs y VPNs ya no me parecen tan aburridos. A los chicos de bioestadística, con los que no pude compartir mucho tiempo pero que son la caña: Arce, Cristina, Jose Luis, Inma y Antonio.

A la nueva UGD, quiero agradeceros el haberme sentido arropada en una familia para los mejores y peores momentos (creedme que han sido más abundantes los mejores). Fran, el compi eterno de laboratorio y también de piso, gracias por tu paciencia y tu interés por sacar las cosas adelante, y por hacer la convivencia en el lab especial. Además, nos ayudas a ahorrar siempre, aunque luego siempre estás dispuesto a hacer planes. A Soraya, la anfitriona perfecta que no decepciona nunca, gracias por dejarnos formar parte de tu amplia vida social, por escuchar siempre los problemas y por inculcarnos tantas expresiones que son ahora el día a día del lab. Admiro tu naturalidad e intuición desde siempre. A Elena, la chica atenta que nos ha enseñado la cultura de México, gracias a ti sabemos lo que es "picoso" y que el castellano no suena allí como nosotros pensamos. A Rebeca, la chica "Rmarkdown", que cuenta las historias más emocionantes y que tiene valor para decir siempre lo que piensa, gracias por las risas. Airin, Irene, flautista de Hamelin, o como se diga según el continente en el que te encuentres, gracias por el tiempo que coincidimos en el laboratorio y espero que podamos coincidir en algún otro sitio. Evelyn, gracias por dejarnos conocerte mejor y apuntarte a los viajes. Cristina Rusu, la crack de los primers, puedes hacer lo que te propongas y eres una gran persona con la que se puede contar. Gracias también a Ana B por su paciencia con nosotros en las distintas etapas que ha vivido el laboratorio, los papeleos y los bizcochos.

A los que han llegado posteriormente y no he tenido el placer de conocer tanto pero sí coincidencias agradables: Malena, Celeste, Alba, Jonathan. Marina, María, y al resto de estudiantes de prácticas, TFG y TFM.

A Claudia, compi de aficiones violinísticas y conciertos en el Palau (aunque a veces no son en el Palau y nos llevamos una sorpresa). Gracias por ser tú, por ser tan fuerte y por escuchar siempre a los demás.

A los compis del máster de bioinformática más hackers: Cristian y Miguel. A Iris, porque tengo la suerte de tenerte como amiga y compañera de locuras desde que vivíamos en Burjassot hasta ahora y hemos compartido muchos momentos.

A Carmen, Loles y las compañeras de Seqplexing, gracias por contar conmigo para esta etapa post-tesis.

No puede faltar la dedicatoria a aquellas amistades que son y han sido un apoyo incondicional durante todo este proceso, en el trabajo y fuera de él. A la mejor bioinfo y data scientist: Carol. Gracias por tu inspiración y hacer un éxito de todo lo que tocas, y el apoyo durante la escritura. A Veronik, gracias por tu compañía en el laboratorio y en la salsa, tu apoyo, y sobre todo por tu amistad durante estos años que ya son unos cuantos. Lo que los exomas juntaron no se puede separar. A mi alma gemela de hemato, Alicia, por engancharme al mundo de las IGHs. Gracias por enseñarme tanto y tan bien, y por guiarme. Soy una afortunada de tenerte como amiga y como mentora al mismo tiempo. Espero que estés orgullosa de este trabajo ya que es tuyo también.

To the members of Kleinstein´s Lab at Yale. This internship was one of the most beautiful and inspiring experiences I ever had. Thanks to Nima, Mamie, Ken, Kevin, Edel, Julian, Toni, Roi, Hailong, David, and especially to Steve for having me. Susanna, thanks enormously for taking the time to do weekly meetings and for many good things that have happened since then. I owe you a plate of "bravas".

A la gente de mi Córdoba sultana, de la que me fui hace ya 7 años, pero que por supuesto siempre llevo en mi corazón. A todos mis amigos y familia, especialmente a Macarena, Rafa, Pedro y Laura, mis tíos y mis primos.

A mis padres Fernando y Mari Carmen, que han apostado por mi educación desde pequeña y me han inculcado tantos valores de los que ahora soy consciente más que nunca. Gracias a los dos por llevarnos de pequeños del conservatorio a casa y de casa al conservatorio y por ser los mejores profesores. A mi hermano pequeño Javi, por todos los momentos divertidos que hemos vivido.

A Dani, por aguantarme tanto estos meses difíciles y no dejarme tirar la toalla. Eres todo un ejemplo de duro trabajo que me ha enseñado a prestar atención a los pequeños detalles y a crecer como persona.

Resumen

El sistema inmune adaptativo está orquestado por una amplia batería de anticuerpos, secretados por linfocitos B, encargados del reconocimiento de moléculas antigénicas. Para expresar el receptor transmembrana tipo Inmunoglobulina en su superficie, las células B sufren un proceso de recombinación en las cadenas pesada y ligera del receptor. Esto ocurre durante su desarrollo, involucrando a los segmentos génicos V(D)J (uno de cada de los posibles genes V, D y J reordena, conformando el dominio variable del locus IGH). Al receptor se le confiere variabilidad adicional después de la exposición antigénica, con la introducción de mutaciones somáticas que aumentan la especificidad del reconocimiento antígeno-receptor. El grado de mutación en el locus IGH es especialmente determinante como factor pronóstico en la Leucemia Linfocítica Crónica (LLC), donde los pacientes clasificados como mutados tienden a llevar un curso de la enfermedad más indolente comparado con los pacientes con receptores de células B no mutados.

El objetivo principal de esta tesis es el desarrollo de herramientas bioinformáticas para la determinación clínica del locus IGH en pacientes diagnosticados de LLC. Para ello, se empleó un método simple y de bajo coste para la preparación de las librerías de secuenciación de la región VDJ del locus IGH, y se diseñó un *pipeline* bioinformático específico (BMyRepCLL) basado en análisis de clonas de células B orientado a la obtención de una secuencia consenso. Un método de corte específico se implementó en este flujo de trabajo para diferenciar los reordenamientos de la fracción clonal y subclonal de cada paciente. Un segundo *pipeline* llamado CLL Immcantation se adaptó del "Immcantation Framework"; una suite que reúne herramientas para el análisis de datos de repertorios inmunológicos. Ambos flujos de trabajo se han usado para analizar un set de datos de 314 pacientes con LLC diagnosticados con los protocolos y criterios estándar, y comparar la determinación del estado mutacional. Además, los resultados de BMyRepCLL han sido validados exhaustivamente con respecto al *gold-standard*. Debido a que los dos *pipelines* se basan en métodos diferentes, CLL Immcantation se usó para la comparativa de la caracterización de muestras con múltiples clonas por BMyRepCLL, así

como también para llevar a cabo análisis posteriores específicos para datos de repertorios inmunológicos.

Los resultados muestran acuerdo entre BMyRepCLL y SSeq en cuanto a la anotación de los genes *IGHV* e *IGHJ* y la extracción de las secuencias de CDR3. El estado mutacional fue caracterizado con éxito en el 99% de los reordenamientos. Las clonas fueron detectadas con una especificidad y sensibilidad del 97% y 100%, respectivamente. Con el desarrollo de estos métodos, contribuimos a la estandarización de los protocolos NGS para la determinación del locus IGH en LLC, el cuál será muy ventajoso por aumentar drásticamente el alcance de SSeq y permitir estudiar en profundidad la arquitectura clonal de la LLC.

Abstract

The adaptive immune system is orchestrated by a wide battery of antibodies, secreted by B lymphocytes, in charge of the recognition of antigenic molecules. In order to express the Immunoglobulin-type receptor transmembrane protein on their surface, B cells suffer recombination of the receptor Heavy and Light Chains genes. This occurs during their development and involving V(D)J gene segments (one of each of the possible V, D and J genes rearrange, conforming the IGH locus variable domain). Additional variability is conferred on the receptor after antigen exposure, when somatic mutations are introduced to augment the specificity of antigen-receptor recognition. The degree of mutation in the IGH locus is especially determinant as a prognostic factor in CLL, where patients classified as mutated tend to have more indolent disease course compared to patients with unmutated B cell receptors.

The main aim of this thesis work is the development of bioinformatics tools for clinical determinations of the IGH locus in CLL patients. For that purpose, a simple and cost-effective library preparation protocol was employed to sequence the VDJ region of the IGH locus, and a specific bioinformatics pipeline for analysis of B cell clones based on the construction of a consensus sequence was designed (BMyRepCLL). A specific cut-off method was implemented within this workflow to differentiate the clonal and subclonal rearrangements fraction among patients. A second pipeline (CLL Immcantation) was adapted from the Immcantation Framework; a suite which unites tools to analyze immune *repertoires* data. Both workflows have been used to analyze a dataset of 314 CLL patients previously diagnosed with the standard criteria and protocols, and to compare the assessment of mutational status. Moreover, the results from BMyRepCLL have been validated exhaustively against the gold-standard. Since both pipelines are based in different methodologies, CLL Immcantation was used for benchmarking the characterization of samples with multiple clones by BMyRepCLL, as well as to perform downstream analyses specific to immune *repertoire* data.

Results show agreement between BMyRepCLL and SSeq regarding *IGHV* and *IGHJ* genes annotation and CDR3 sequence extraction. The mutational status was characterized accordingly in 99% of the rearrangements. Clones were detected with a specificity and sensitivity of 97% and 100%, respectively. With the development of these methods we contribute to the standardization of NGS protocols for the determination of IGH locus in CLL, which will be highly advantageous for augmenting drastically the scope of SSeq and allowing to study in deep the clonal architecture in CLL.

Table of contents

List of figures

List of tables

Resumen extendido

Introducción

Las neoplasias linfoides B se caracterizan por una expansión de células B tras un proceso oncogénico cuyas causas son variables y desconocidas. En el caso de la LLC (Leucemia Linfocítica Crónica), es una de las neoplasias de linfocitos B más prevalentes en el mundo occidental, y se caracteriza por la acumulación de células B en sangre, médula ósea y órganos linfoides secundarios. De modo característico expresan en la superficie celular CD5, CD19, CD23 y a su vez baja expresión de CD20 y CD79b. A nivel molecular, poseen un perfil muy heterogéneo que se puede presentar tanto a nivel de aberraciones cromosómicas como mutaciones puntuales de pequeña escala, siendo algunas de ellas más deletéreas que otras. Las pruebas convencionales para completar el diagnóstico son: cariotipo, FISH, mutaciones de *TP53* y estado mutacional de la región variable del locus IGH. Las mutaciones en el gen supresor de tumores *TP53* que causan pérdida de función, tienen un valor pronóstico muy desfavorable (Chiorazzi et al. 2021). Actualmente, el desarrollo de fármacos para el tratamiento está siendo enfocado a la disrupción de la vía de señalización del receptor de las células B, como inhibidores de BTK, inhibidores de PI3K y bloqueantes de BCL2 (Fabbri et al. 2016).

Por otro lado, el estado mutacional de la región variable del gen de la cadena pesada de las inmunoglobulinas (*IGHV*), ha demostrado ser una variable de gran valor pronóstico, ya que se mantiene estable a lo largo de la enfermedad. Un 2% de mutación se ha descrito como el punto de corte para diferenciar entre pacientes mutados (identidad menor al 98% con respecto al alelo correspondiente de la línea germinal de *IGHV*) y no mutados (identidad mayor o igual al 98% con respecto al alelo de la línea germinal) (Damle et al. 1999, Hamblin et al. 1999, van Dongen et al. 2003).

Por tanto, el análisis de mutaciones en *TP53* y el estado mutacional de *IGHV* se determinan siempre antes de aplicar el algoritmo de decisión de tratamiento, como indican las guías iwCLL desde 2008 hasta la actualización de 2018 (Hallek et al. 2008, Hallek et al. 2018).

El repertorio de inmunoglobulinas en la LLC se caracteriza por un uso recurrente de ciertos genes de *IGHV*, *IGHD* e *IGHJ*. Este uso sesgado implica el reconocimiento de un número limitado de antígenos por parte del receptor (Duke et al. 2003, Fais et al. 1998, Donisi et al. 2006, Karan-Djurasevic et al. 2012). Además, un 30% de los pacientes poseen lo que se denomina como "reordenamientos estereotipados", de los que se está estudiando qué combinaciones concretas de los distintos genes VDJ y patrones especialmente en la secuencia de CDR3 (tanto en la cadena pesada como la ligera del receptor de la célula B), se relacionan con grupos pronósticos (Agathangelidis et al. 2012).

La técnica utilizada en la práctica clínica para determinar el estado mutacional de *IGHV* con el fin de estratificar clínicamente a los pacientes es, hasta el día de hoy, la secuenciación capilar clásica de Sanger, a pesar de las profundas limitaciones de la técnica para este propósito (Davi et al. 2020). A lo anterior, se suma la elevada variabilidad que existe derivado del proceso inherente de formación del reordenamiento a nivel genómico del locus IGH (Davi et al. 2020, Rosenquist et al. 2017, Stamatopoulos et al. 2017). Para la amplificación por PCR de la región, existen varios sets de oligonucleótidos, de los cuales se usan preferentemente el set de la región Leader o FR1 en su defecto, por ser los que cubren un mayor número de nucleótidos de *IGHV*.

El locus IGH, codifica para la cadena pesada de la proteína receptora de la superficie de las células B, con un dominio variable y otro constante. Su caracterización requiere el desarrollo de herramientas específicas ya que se expresa a raíz de un fenómeno de recombinación a nivel genómico, de tal forma que únicamente un segmento de cada una de las regiones (*IGHV*, *IGHD* e *IGHJ*) conforman el dominio variable del receptor (Ghia et al. 2007). Además, se produce adición y eliminación de nucleótidos en la zona de la unión de los tres segmentos, conocida como CDR3 o región determinante de la complementariedad (altamente variable), la cual forma el bucle de reconocimiento del antígeno y es por ello altamente específica. Posteriormente, el proceso conocido como hipermutación somática que tiene lugar en los centros germinales, es responsable de la maduración y supervivencia del linfocito B, si éste ha conseguido aumentar su afinidad ante un antígeno específico. Este proceso, junto con el cambio de isotipo de inmunoglobulina o CSR (del inglés "class switch recombination"), son altamente importantes para la diferenciación del receptor y el desarrollo de afinidad antígeno-específica, la cual es clave a su vez, para el reconocimiento de partículas patogénicas de origen exógeno (Ghia et al. 2007, Watson and Breden 2012, Moser and Leo 2010).

La LLC se ha definido como una proliferación de linfocitos B con origen en una única célula, ya que en la mayoría de los casos se describe un perfil monoclonal. Sin embargo, aproximadamente un 10-25% de los casos poseen varias clonas tumorales productivas (Davi et al. 2020, Stamatopoulos et al. 2017, Rosenquist et al. 2017). Del mismo modo, aunque siempre esté caracterizada por uno o varios clones predominantes, poseer la información de las subclonas presentes en menor proporción puede arrojar luz al entendimiento de la enfermedad.

Los avances experimentados en el ámbito de la secuenciación masiva durante los últimos 10 años, han permitido el estudio del repertorio de células B en un amplio abanico de áreas de investigación relacionadas con vacunas, virus, infecciones, enfermedades autoinmunes y también neoplasias linfoides. El hecho de poder obtener información de miles o millones de las secuencias presentes en el material genético extraído y seleccionado, en este caso a partir de clonas de células B tumorales, constituye una gran ventaja con respecto a las técnicas que se emplean en la práctica clínica actualmente. Sin embargo, debido a la cantidad y complejidad de datos que se obtienen, el procedimiento NGS no está estandarizado para la caracterización del grado de hipermutación somática de las inmunoglobulinas en un entorno clínico, aunque cada vez disponemos de más conocimiento y herramientas para sobrellevar estas limitaciones (Georgiou et al. 2014).

La determinación por NGS del locus IGH conlleva complicaciones derivadas de la elevada variabilidad y el tamaño de lectura de la secuenciación. Para la tecnología Illumina, la única plataforma que soporta secuenciación de 2x300 ciclos, y que por tanto permitiría la secuenciación completa de la región VDJ desde la región Leader hasta la zona consenso JH, es MiSeq. El mayor número de ciclos empleados para la secuenciación con este kit produce caídas en la calidad hacia el final de las lecturas de forma más drástica que kits de menos ciclos (como por ejemplo, 2x150 ciclos). Esta caída se debe al fenómeno de error por desfase en la incorporación de nucleótidos inherente a la tecnología Illumina.

También hay que sopesar las ventajas y los inconvenientes del uso de ADN o ARN como material genético de partida.

Para las determinaciones del estado mutacional en el locus IGH y el análisis de subclonas, hay que tener en cuenta que el análisis de estos biomarcadores genéticos es complejo, especialmente para la detección de mutaciones somáticas o clonas en proporciones muy bajas (menores al 5%) o en el estudio del gen IGH (por su complejidad, la presencia de hipermutación, etc), que son las que podrían ser más útiles para determinar resistencia o recaídas. Así, la metodología para NGS y el análisis bioinformático de los datos obtenidos no está totalmente definido. Por ello, es necesario desarrollar procedimientos específicos, especialmente para el estudio del gen IGH y los subclones a baja proporción (Greiff et al. 2015).

Objetivos

Los objetivos de este trabajo son en primer lugar, determinar el método más idóneo para la preparación de librerías de NGS a partir de la amplificación del locus IGH, en términos de coste y tiempo empleado. En segundo lugar, desarrollar un *pipeline* propio adaptado a las necesidades de las guías clínicas, que contemple la complejidad intrínseca a los resultados de NGS, y la resolución automática del ruido de fondo clonal y no clonal. Posteriormente, el desarrollo de un segundo *pipeline* empleando herramientas creadas por expertos en análisis de repertorio de células B para realizar comparaciones con el *pipeline* principal. Por último, con objeto de analizar la robustez de las herramientas diseñadas, se realizará una validación frente a la técnica de secuenciación Sanger, considerada el “gold-standard”.

Métodos

Para la prueba de métodos de preparación de las librerías de ADN, se extrajo ADN de 23 pacientes con LLC y se emplearon dos mixes diferentes con los sets de oligonucleótidos Leader y Framework estándar del consorcio BIOMED-2. El primer mix se empleó para amplificar el locus IGH desde la región Leader hasta la zona consenso *IGHJ*. El segundo mix se empleó para amplificar el locus IGH desde las 3 regiones Framework hasta la zona consenso *IGHJ*. Estas mezclas se utilizaron para 3 protocolos de preparación de librerías: El protocolo A consistió en el uso del mix Leader-JH, y secuenciación con el kit de Illumina MiSeq de 300pbx2. El protocolo B consistió en el uso del mix Leader-JH y posterior tagmentación de los fragmentos de ADN amplificados con el protocolo de Nextera (Illumina). La secuenciación se realizó con el kit de Illumina MiSeq de 150pbx2. El protocolo C, se constituyó con el mix de cebadores de las regiones Framework para su posterior secuenciación con el kit de Illumina MiSeq de 150pbx2 a partir de los fragmentos solapantes de los 3 amplicones. Este último mix se empleó también para poner a punto el mismo diseño para ADN copia, añadiendo el set de oligonucleótidos Leader a los anteriores con el fin de realizar la secuenciación con el kit de lecturas cortas de Illumina MiSeq (150pbx2).

En cuanto al análisis bioinformático, se diseñaron dos *pipelines* para el análisis de reordenamientos de inmunoglobulinas en LLC: BMyRepCLL y CLL Immcantation.

- BMyRepCLL: el *script* principal *pipeline.py*, integra el primer módulo del análisis, que se empleó para la caracterización y anotación de VDJ. El flujo de análisis se realizó partiendo de la harmonización de las muestras de partida en formato FASTQ con el módulo “fastq_merge”, preprocesado por calidad y longitud de las lecturas más eliminación de secuencias de los cebadores con los programas seqtk y bbduk. Posteriormente, se mapearon las lecturas contra los alelos de la base de datos de referencia IMGT (*IGHV* e *IGHJ*) empleando BWA *mem*. Para asignar la correspondencia IGHV-IGHJ, se emplearon módulos propios, para la extracción del conteo de lecturas mapeadas contra los alelos de IMGT de forma conjunta para *IGHV* e *IGHJ* y cálculo de los alelos mayoritarios.

Posteriormente, se aislaron las lecturas de cada reordenamiento para su mapeo contra una referencia simulada, construida con todas las posibles combinaciones de los alelos *IGHV* e *IGHJ.* Tras ello, le siguió un proceso de llamado de variantes con Freebayes, mediante el que se obtuvo una secuencia consenso por cada reordenamiento presente en una muestra, y las secuencias de *IGHD* y CDR3. El estado mutacional se obtuvo a partir del alineamiento de las secuencias consenso contra los alelos *IGHV* de la línea germinal. El segundo módulo, realizó eliminación de artefactos y priorización de los reordenamientos potencialmente clonales.

- CLL Immcantation: para este *pipeline* se empleó el programa pRESTO para el preprocesado de las secuencias (filtrado por calidad, eliminación de las secuencias de cebadores, marcaje de la secuencia con el cebador encontrado y ensamblaje). Posteriormente, cada secuencia se anotó con el programa IgBlast, dando lugar a un formato tabular con la información de cada lectura en una línea. Para determinar relaciones clonales entre las secuencias de un mismo paciente, se utilizó la función *defineClones* de Change-O. Finalmente, se calcularon las frecuencias mutacionales por secuencia con Shazam. Todas estas herramientas pertenecen a la suite del “Immcantation Framework”, desarrolladas por el laboratorio de Steve Kleinstein, del departamento de Patología de la Universidad de Yale.

Para establecer un umbral de distancia con el que agrupar las secuencias en clonas, se analizaron 238 muestras con CLL Immcantation. Previo al paso de definición de las clonas, se empleó la función *distToNearest*, que calcula la distancia de cada secuencia a su secuencia “vecina” más próxima (nearest-neighbor). El algoritmo empleado para establecer cómo de parecidas son las secuencias entre sí fue la distancia de haming (hamming distance). Tras ello, se eligió dicho umbral, y se incluyeron en los archivos de análisis de NGS las secuencias obtenidas mediante secuenciación Sanger en el laboratorio de Hematología del Hospital Clínico Universitario de Valencia, con el fin de establecer si la agrupación se realizaba correctamente en la clona predominante. Se añadió un paso de filtrado de variantes presentes a una baja proporción (<2%), para eliminar posibles artefactos de secuenciación.

Para la validación del método desarrollado con el *pipeline* BMyRepCLL, se emplearon 319 muestras de ADN de pacientes con LLC con más de 1000 lecturas en el reordenamiento mayoritario y 47 controles sanos con más de 1000 lecturas totales. El método de preparación de librerías empleado fue el protocolo C. Éstas se analizaron con BMyRepCLL y CLL Immcantation en un servidor ubicado en el CPD de INCLIVA (16 Intel ® Xeon ® CPU E5-2650 0 @ 2.00 GHz procesadores, 190 GB de RAM y 41 TB de espacio en disco. Para correr varias muestras en paralelo se ejecutó GNU Parallel. El control de calidad de los datos crudos se realizó con un repositorio propio (https://github.com/afuentri/QC).

Para poner a punto el cálculo de la fracción clonal por paciente, las muestras se dividieron aleatoriamente en un grupo de test (24 con una única clona, 10 con doble clona y 20 policlonales), y un grupo de validación (260 muestras con clona única, 20 con múltiples clonas y 27 policlonales), según el número de clonas detectadas previamente con Sanger en el laboratorio de Hematología del Hospital Clínico Universitario de Valencia. A las muestras del test se aplicó en primer lugar el cálculo de ratios del porcentaje de las clonas detectadas por orden de abundancia y el valor más alto se estableció como punto de corte entre la fracción clonal y subclonal. Los mismos métodos se replicaron con el grupo de la validación y se calculó la sensibilidad y especificidad en la detección de reordenamientos clonales. Las discordancias entre Sanger y este método se validaron repitiendo la secuenciación Sanger y mediante análisis de fragmentos. Por otro lado, la presencia y cuantificación relativa de las clonas secundarias se comprobó con CLL Immcantation y un tercer programa: MiXCR.

Resultados

En primer lugar, se evaluaron los métodos testados para la preparación de librerías. Los tres protocolos detectaron el reordenamiento de células B predominante previamente detectado con secuenciación Sanger en las muestras analizadas. Debido a un menor coste y tiempo de preparación de la librería, el protocolo C se eligió para realizar los experimentos de secuenciación.

Para el análisis de los datos, se diseñaron dos *pipelines* bioinformáticos con distintas aproximaciones. La descripción general de los desarrollos finales es la siguiente:

- BMyRepCLL es un flujo de análisis de diseño propio basada en una estrategia diseñada desde el punto de vista clonal, con la obtención de una secuencia consenso por reordenamiento (https://github.com/afuentri/B-MyRepCLL). Para llegar a ello, tras el preprocesado se mapean las lecturas independientemente contra los alelos de *IGHV* e *IGHJ* de la base de datos IMGT, y posteriormente, éstas se agrupan en reordenamientos empleando la correspondencia VJ de los mismos. Tras una serie de pasos ajustados específicamente a este tipo de análisis, se realiza un paso de llamado de variantes con el fin de generar una secuencia consenso por cada reordenamiento VJ. A partir de esta, se caracteriza el locus IGH y posteriormente, se aplican una serie de filtros para minimizar sesgos causados tras los pasos de mapeo contra la línea germinal de referencia y de los distintos amplicones secuenciados, ya que tienen longitudes diferentes. Además, se ha implementado el cálculo de la fracción clonal por paciente.
- CLL Immcantation: La suite de herramientas del "Immcantation Framework" (https://immcantation.readthedocs.io/en/stable/), es un conjunto de programas y *pipelines* bioinformáticos diseñados por el laboratorio Kleinstein, del departamento de patología de la escuela de Medicina de Yale. Tras la realización de una estancia en este laboratorio por parte de la doctoranda, se diseñó un flujo de análisis empleando estas herramientas ampliamente usadas entre la comunidad de investigadores de repertorios de inmunoglobulinas, adaptado a la LLC y a nuestro diseño específico de preparación de

librerías. Este método se usa con el fin de validar los resultados de BMyRepCLL y aplicar aproximaciones computacionales diseñadas por la comunidad de inmunoinformáticos para analizar la arquitectura clonal de los pacientes de LLC. El *pipeline,* se compone de 4 bloques independientes: preprocesado con pRESTO, anotación con IgBlast (Change-O), "clustering" clonal (Shazam), y cálculo de la carga mutacional (Shazam). A estos se añadieron *scripts* propios de filtrado de datos y representación gráfica de los mismos.

Tras la inspección de la distribución de distancias entre secuencias de una misma muestra, se comprobó que en ninguna de las muestras aparecía un patrón bimodal necesario para establecer el umbral clonal con los métodos empleados por CLL Immcantation. Esto es debido a que el patrón en la LLC es altamente clonal y, por tanto, no se pudo realizar una determinación del umbral clonal específica para el repertorio de inmunoglobulinas de cada paciente. Se eligió un umbral general de 0.1 para la definición de clonas.

Tras realizar la definición de las clonas en los pacientes con LLC, se observan artefactos en las frecuencias mutacionales calculadas con las lecturas del fragmento FR3. Por ello, se decide eliminar esas lecturas en el paso previo a la definición de las clonas.

Las secuencias Sanger agruparon correctamente en la clona mayoritaria de NGS en el 85% de las muestras. 11 secuencias Sanger agruparon en clonas menores a la clona de orden 2. Por otro lado, no hubo secuencias Sanger agrupadas en clonas independientes.

Se observa que, a pesar del filtrado de variantes a baja proporción, existen muestras con perfiles de frecuencias mutacionales ampliamente distribuidas, incluso en casos de pacientes con clona mayoritaria no mutada. Aun así, los valores medios de la distribución de las frecuencias mutacionales de todas las secuencias clasificadas en la clona mayoritaria se mantienen dentro de lo esperado para cada grupo de la clasificación por estado mutacional (mutado, no mutado y borderline).

314 muestras de LLC pasaron los filtros de calidad para la validación de las mismas con el "gold-standard". Tras el análisis, se puso a punto el cálculo de la fracción clonal. Las diferencias entre las

ratios máximas fueron significativas entre las comparaciones 1 clona vs policlonal y 2 clonas vs policlonal. Al aplicar el mismo método en las muestras de la validación, se obtuvo una sensibilidad del 100% y especificidad del 97%.

Los pasos de filtrado implementados en BMyRepCLL eliminaron reordenamientos artefactuales de forma eficiente, ya que el número de reordenamientos crudos detectados por muestra era del orden de 100, mientras que tras la aplicación de los filtros y el umbral clonal se obtuvieron 4 clonas como máximo. Del total del listado curado de clonas, 362 se clasificaron como clonales y 867 como subclonales, con un rango de porcentajes de 0.1 a 9.1. El porcentaje medio de amplitud de cobertura superior a 500X fue 85%, en clonas presentes desde el 2 al 100%.

Todos los reordenamientos múltiples encontrados previamente por secuenciación Sanger se detectaron con BMyRepCLL. Se detectaron 9 clonas adicionales por NGS que no se detectaban por secuenciación Sanger. Entre ellas, 8 eran dobles reordenamientos que compartían familia de *IGHV*, y que por tanto no se detectan fácilmente con la técnica estándar. 7 muestras que habían sido previamente clasificadas en única clona por Sanger se reclasificaron al grupo de 2 clonas ya que al repetir Sanger se encontraron las mismas clonas que reportaba el *pipeline* de NGS. 9 clonas se catalogaron como falsos positivos al no estar presentes tras la secuenciación Sanger ni el análisis de fragmentos. Los resultados de la validación con respecto al “gold-standard” fueron muy satisfactorios. Además, se evaluó la caracterización de los reordenamientos, con una alta tasa de acierto en la asignación de los genes *IGHV*, *IGHJ*, la secuencia de CDR3 y el estado mutacional ($r^2 = 0.862$). El estado mutacional determinado por BMyRepCLL y Sanger fue discordante en tres casos que se achacan al diseño empleado, ya que cada una de las muestras portaba una mutación que se encontraba aguas arriba del cebador FR1. En estos 3 casos, Sanger determinó que eran borderline mientras que BMyRepCLL los reportó como no mutados, con un 98% de identidad en todos los casos.

En algunos casos, los porcentajes clonales en las muestras con más de un reordenamiento clonal no coincidían entre BMyRepCLL y CLL Immcantation, empleando un umbral de 1.5 de fold-change para determinar las diferencias. Por ello, las muestras con múltiples reordenamientos, así como las que tenían clonas catalogadas como falsos positivos, se analizaron con el programa MiXCR. Solo hubo diferencias significativas entre los porcentajes de las clonas secundarias calculados por BMyRepCLL y CLL Immcantation. Sin embargo, 35 reordenamientos coincidían en porcentaje entre BMyRepCLL y MiXCR, mientras que entre CLL Immcantation y MiXCR coincidían 19. MiXCR obtuvo la mayor tasa de solapamiento con el resto de programas (84%).

La comparación del estado mutacional para las muestras con LLC entre CLL Immcantation y Sanger reveló que esta herramienta es más exacta que BMyRepCLL a la hora de determinar el porcentaje de identidad respecto a la línea germinal de *IGHV* (r^2 = 0.935). Sin embargo, hay más muestras discordantes entre el *pipeline* de Immcantation y Sanger (concordancia del 98.1%) que entre BMyRepCLL y Sanger (concordancia del 99%). Esto es debido a que CLL Immcantation es más sensible a la variabilidad intraclonal, al calcular la frecuencia mutacional independientemente en cada una de las lecturas. También se determinó que CLL Immcantation detecta un mayor número de casos con estado mutacional borderline (77.8% vs 66.7% en el caso de BMyRepCLL).

Por último, se detectaron 11 reordenamientos subclonales productivos con ambos métodos, ya que tanto CLL Immcantation como BMyRepCLL coincidían en el uso de *IGHV* e *IGHJ* del reordenamiento, secuencia de CDR3 (siendo ésta diferente a la de la clona mayoritaria) y estado mutacional. Entre éstas subclonas, hay varios casos en los que la clona predominante es mutada mientras que el reordenamiento subclonal es no mutado.

Conclusiones

1. El método propio de preparación de las librerías en multiplex ha sido probado como el método óptimo para la secuenciación NGS del locus IGH en pacientes de LLC, siendo fácilmente adaptable a la rutina clínica por implicar un menor coste y tiempo empleado en el procedimiento que los otros dos métodos testados. El método in-house emplea los 3 sets de cebadores de las regiones Framework (FR1, FR2 y FR3) del diseño del consorcio BIOMED-2, y el kit de 150pbx2 de Illumina. Éste kit es más corto que el comúnmente empleado (300bpx2), y compatible con plataformas de Illumina de más altas capacidades. Además, el mismo método partiendo de ADN copia se ha optimizado, incluyendo los cebadores Leader en el diseño para cubrir la región del locus IGH por completo.
2. Se ha diseñado un *pipeline* específico para reconstruir los genes VDJ a partir de lecturas que cubren el fragmento parcialmente, integrando *scripts* con programas bioinformáticos de uso libre. Además de detectar las clonas de LLC mayoritarias, el flujo de análisis proporciona una distinción directa de las fracciones clonal y subclonal tras un ajuste exhaustivo de los pasos de análisis implementados para la eliminación de artefactos y priorización de los reordenamientos.
3. Se ha desarrollado un segundo *pipeline* bioinformático, a partir de herramientas creadas por expertos en el análisis del repertorio inmune adaptativo. Dado que este *pipeline* emplea métodos computacionales específicos para estudiar en profundidad repertorios de células B, ha permitido observar variabilidad en las frecuencias mutacionales dentro de las clonas predominantes de algunos pacientes.
4. La validación con respecto a las técnicas "gold-standard", ha demostrado que los métodos desarrollados para la secuenciación y análisis bioinformático del locus IGH, son altamente robustos en la anotación de las características del reordenamiento VDJ y la información de clonas potencialmente expandidas, con una sensibilidad del 100% y especificidad del 97%.

Por último, el estado mutacional se ha caracterizado de forma idéntica a la secuenciación Sanger en el 99% de los pacientes estudiados.

List of abbreviations and acronyms

AID	Activation-Induced cytidine Deaminase
AIRR	Adaptive Immune Receptor Repertoire
BAM/SAM	Binary/Sequence Alignment Format
BcR	B cell receptors
BD	Borderline
BD-CLL	Borderline CLL patient
CDR	Complementary Determining Regions
CLL	Chronic Lymphocytic Leukemia
DNA	Deoxyribonucleic acid
cDNA	Complementary DNA
ERIC	European Research Initiative on CLL
FISH	Fluorescence in Situ Hybridization
FFPE	Formalin-fixed Paraffin Embedded
FWR	Framework Regions
GC	Germinal Centers
Ig	Immunoglobulin
IGK	Immunoglobulin Kappa Chain
IGL	Immunoglobulin Lambda Chain
IGH	Immunoglobulin Heavy Chain
IGHV	Variable region genes of the B cell Receptor Heavy Chain locus
IGHD	Diversity region genes of the B cell Receptor Heavy Chain locus
IGHJ	Joining region genes of the B cell Receptor Heavy Chain locus
IGV	Integrative Genomics Viewer
IMGT	International ImMunoGeneTics information system
MBL	Monoclonal B cell lymphocytosis
M-CLL	Mutated CLL patient
MM	Mutated (referred to B cell clones)
MRD	Minimum Residual Disease
NGS	Next Generation Sequencing

PCR	Polymerase Chain Reaction
RNA	Ribonucleic acid
RSS	Recombination signals
SHM	Somatic Hypermutation
sIgM	surface Ig M
SSeq	Sanger Sequencing
TcR	T cell receptors
TS	Targeted Sequencing
U-CLL	Unmutated CLL patient
UM	Unmutated (referred to B cell clones)
UMIs	Unique Molecular Identifiers
VAF	Variant Allele Frequency
VCF	Variant Calling Format
WGS	Whole Genome Sequencing
WES	Whole Exome Sequencing

1 Introduction

1.1 B cell receptors

The adaptive immune system is orchestrated by a specific battery of antibody molecules, designed for the recognition of a vast variety of antigens. These molecules are first expressed on the surface of B cells conforming transmembrane receptor proteins (BcRs) which are secreted after differentiation stages, by antibody-secreting cells. B cell immunoglobulin-type receptor is key for antigen recognition during secondary immune response. The protein is composed of two identical heavy chains associated with two identical light chains, whose loci are 14q32.33 (IGH; Immunoglobulin Heavy Chain), 2p11.2 (IGK; Immunoglobulin kappa chain) and 22q11.2 (IGL; Immunoglobulin light chain) (1,2). Structurally, each chain is composed of a variable domain (V), responsible for antigen binding, and a constant domain (C). The V domain is further divided into Framework (FR) and Complementary determining regions (CDR) regions. The VH domain is encoded by variable (V), diversity (D) and joining (J) genes, whereas VL chains contain V and J genes (2–5). D genes are very short and have an approximate length of 10 bases whereas V genes can be up to 290 bases long (6,7). A process of recombination of V(D)J gene segments occurs prior to the expression of the receptor in the B cell surface, in order to generate a virtually unlimited number of BcRs capable of recognizing a wide variety of external antigens (8–12).

The Heavy chain is rearranged firstly, at the pro B cell stage (13). The recombination process is mediated by endonucleases which produce DNA-breaks and the posterior ligation of a DH segment with a JH segment, and finally a VH segment to the DH-JH combination. Each segment possesses a signal of recombination (RSS) which is recognized by RAG1 and RAG2 proteins, forming a protein complex, and breaking the DNA double chain between the RSS and the flanking sequence, after forming a loop structure (14,15). The DNA breakpoints remaining after the breakage are asymmetric, with a hairpin loop structure in the coding side (7,16,17). As a consequence of the recombination process, the junction has suffered addition and elimination of random nucleotides in the junction region flanking the *IGHD* gene (CDR3), which is the third hypervariable variable segment, and encodes a highly specific contact antigen-binding loop (18–21). If the recombination process with one of the allele copies is capable of generating a functional and variable rearranged transcript, RAG1/2 complex is silenced to ensure the expression of a single BcR per cell (allelic exclusion). B cells with successful rearrangements of IGH transcripts are selected to fulfill the rearrangement of light chains. For that purpose, RAG1/2 complex is activated again to rearrange VL and JL segments (21–23) (Figure 1.1). After this accomplishment, these cells scale from pre B cell to naïve B cells (24).

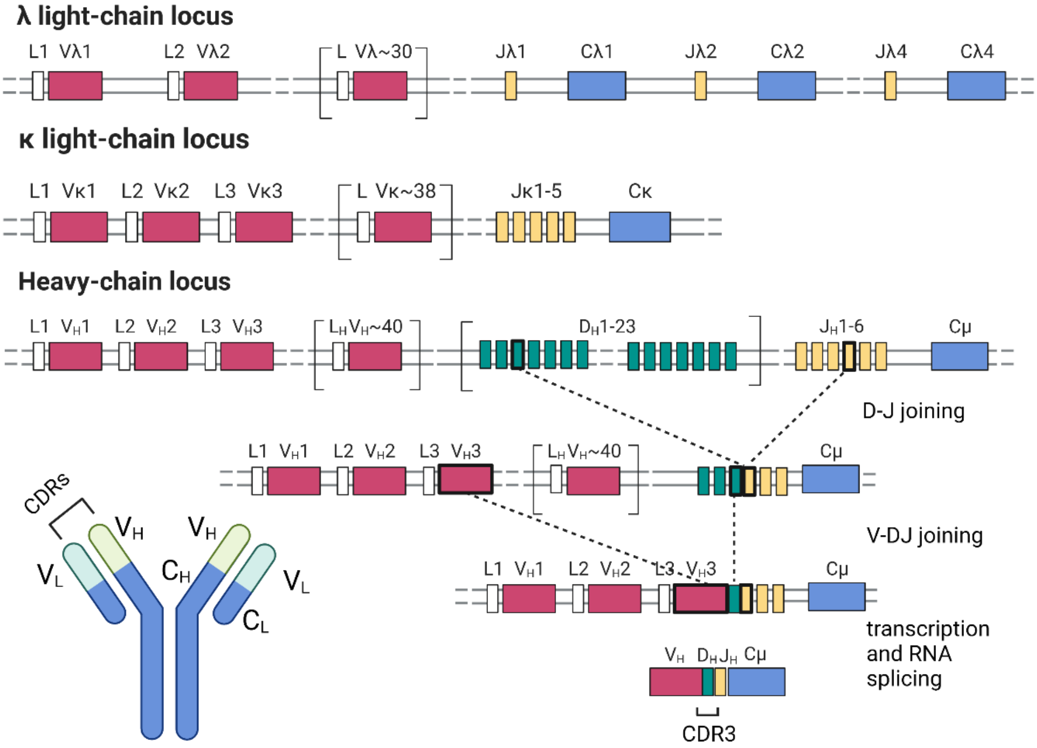


***Figure 1.1. IGL and IGH loci germline structure**. Recombination of IGH and IGL chains occur similarly, with the difference that D gene segments are not part of IGL chains, and the rearrangement involves V and J gene segments. For the IGH counterpart, the recombination process is schematized in the figure, as first a D and J segment are joined, and that complex is consecutively joined with a V segment, forming the rearranged DNA which is transcribed into mRNA. Adapted from BioRender.com template.*

Once the BcR is expressed, B cells are released as naïve lymphocytes and start affinity maturation in the germinal centers (GC), where various rounds of antigen-specific mutations along the *IGHV* region occur through a process termed Somatic Hypermutation (SHM), involving AID enzyme (activation-induced cytidine deaminase). B cells capable of recognizing an antigen with high specificity, are selected for survival and clonal expansion. In a posterior stage, B cells suffer class switch recombination to define the Ig isotype after changing their constant region (IgM to IgM + IgD to non-IgM) (1,25,26) (Figure 1.2). With combinatorial diversity of more than ~50 V, 6 J and ~30 D gene segments, junctional diversity (non-template nucleotides), and SHM processes conferring enormous variability to B cell receptors, the number of possible combinations has been estimated to be 10^{13} different specific receptors responding against a diverse amount of antigens (27,28).

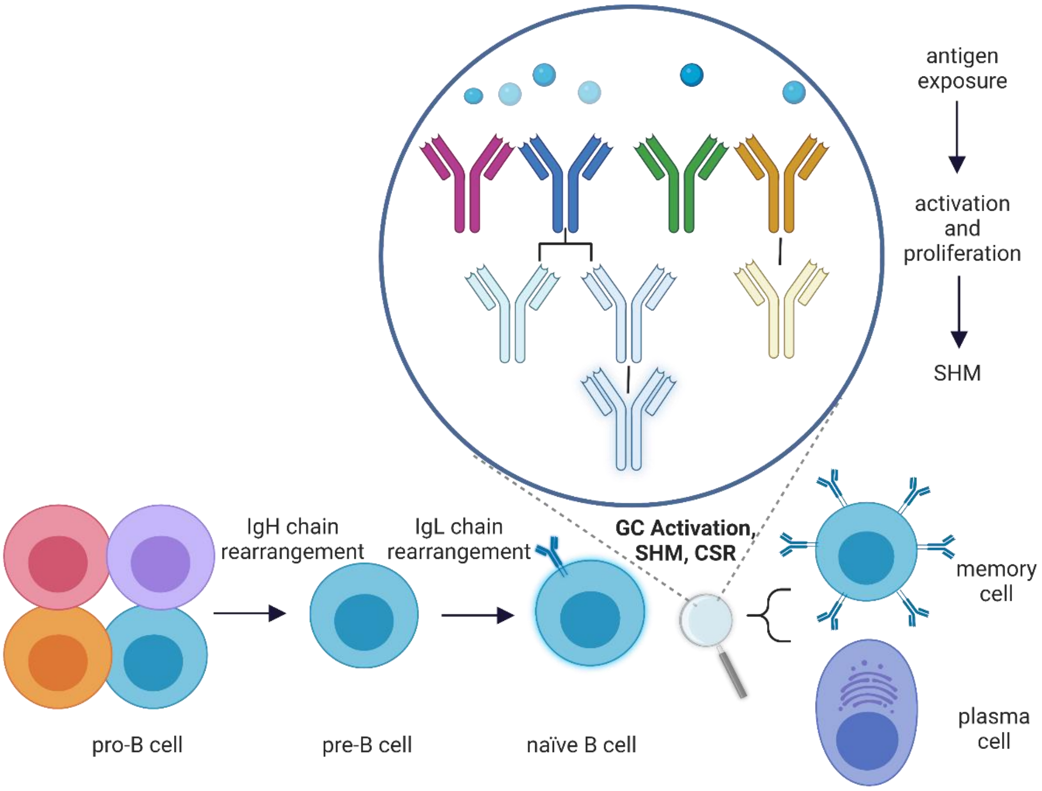


Figure 1.2. B cell maturation and differentiation. *Created with BioRender.com and from* (29).

After using cloning techniques, Matsuda and collaborators (30), reported the first complete VDJ region, of approximately 0.95Mb. Soon after, VDJ genes were determined (44 functional/open reading frame VH, 85 VH pseudogenes, 27 DH genes; of whom 23 were functional, and 9 JH genes (6 of them functional) (Figure 1.3). IGH genes nomenclature was established by the International ImMunoGeneTics information system (IMGT) (31), and approved by the HUGO Nomenclature Committee (HGNC), in 1999, and one year later, IMGT gene names were shared with the NCBI for their annotation in the human genome assembly (32). IMGT set rules and unique organization schemes for IGH/IGL/IGK/TRA/TRB/TRG/TRD loci to align sequences against the germline alleles and determine mutations, termed the 'IMGT Unique Numbering´ (33,34). IMGT numbering uses conserved amino acids positions to delimitate FWR and CDR regions: Cysteine 23 (1st cysteine), Tryptophan 41 (conserved TRP), Leucine 89, Cysteine 104 (2nd cysteine), as well as hydrophobic amino acids of the FWR regions (33) (Figure 1.4).

IGH locus genes sequencing has allowed the creation of highly extensive databases, harboring 55 ORF/functional *IGHV* genes (11 of them are not part of the reference human genome). In the genomic locations 15q11.2 and 16p11.2, and other *IGHV* genes catalogued as orphons (1) (Figure 1.3).

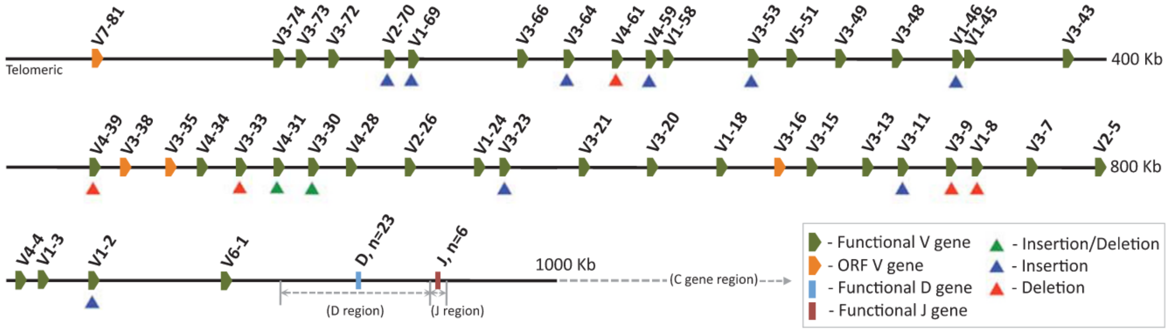


Figure 1.3. IGH locus scheme. *From* (1).

On the other hand, *IGHV* genes are divided into 7 families or subgroups (*IGHV1-IGHV7*) following phylogenetic classification methods (35). A substantial part of the mutations described for *IGHV* alleles are missense, mainly on CDR regions, which is where the contact with the antigen is established. However, this does not occur in all *IGHV* loci, meaning that they must be driven by different selection pressures (36). It is the case of *IGHV3-23*03*, whose specificity against Haemophilus i*nfluenzae* b-type polysaccharide is higher than the most frequent allele variant for the same gene (allele *01) (37).

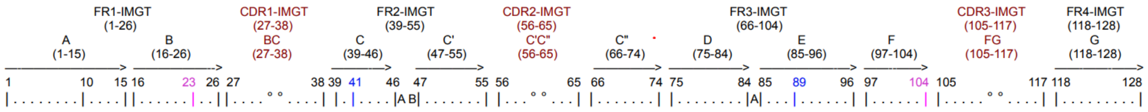


Figure 1.4. VDJ Framework and CDR3 regions with IMGT codon numbering. *From* (33).

1.2 Chronic Lymphocytic Leukemia

Chronic Lymphocytic Leukemia (CLL) is characterized by the proliferation of malignant B lymphocytes expressing CD5, CD19, CD23 and low CD20 and CD79b, which accumulate in blood, bone marrow and secondary lymphoid organs (38). It is the most prevalent leukemia type in Western countries (39). CLL presents high heterogeneous course and genomic changes varying from small-scale variants to chromosome abnormalities. This genetic heterogeneity is translated into different disease phenotypes, which can therefore, vary between highly indolent with no need of treatment at all to completely the opposite, so different prognostic biomarkers have been studied and described (40). Common chromosome aberrations present in these patients are deletion 13q14 (55% of cases), trisomy 12q (10-20% of cases), deletion 11q22-q23 (10% of cases) and deletion 17p13 (5-8% of cases). 13q and 12q are considered initiating aberrations (41,42), whereas the other two involving chromosome regions 11q and 17p, are found in more advanced disease stages, causing disruptions of the genes *ATM* and *TP53*, which are involved in responses to DNA damage (43–45). WGS (Whole Genome Sequencing) and WES (Whole Exome Sequencing) studies identified a wider list of additional target genes with recurrent somatic mutations in CLL: *NOTCH1, SF3B1, BIRC3, MYD88, NFKBIE, XPO1* among the most studied (46,47).

Thereafter, the most important prognostic factors considered until now are *TP53* somatic mutations and *IGHV* mutational status. They are determined always before the treatment decision algorithm in CLL as stated in the iwCLL guidelines from 2008 and the 2018 update (48,49).

1.2.1 TP53 mutations

TP53 gene is found mutated in over 50% of human cancers and encodes the protein p53, which is involved in the regulation of cell cycle and apoptosis. It acts after DNA damage in response to cellular stress signals and triggers arrest of cell cycle until the DNA repair has taken place, preventing the replication of harmful mutations. It also activates the expression of apoptotic genes when cell damage is irreversible, thus being cataloged as a tumor suppressor gene (50).

Genetic aberrations on *TP53* gene that cause loss of one or both copies of *TP53* can be originated by loss of function mutations or deletions in chr17p.13. They lead to decreased survival and impaired response to chemoimmunotherapy, and thus, it is used as one of the most important predictive markers for clinical decisions in CLL. Approximately 80% of cases with del(17p), have mutations in both allele copies, causing complete disruption of the *TP53* signaling pathway (45).

Recently, new therapies have become available for patients with *TP53* aberrations of any kind, such as ibrutinib (BTK inhibitor), idelalisib (PI3K inhibitor), and venetoclax (BCL2 inhibitor). While big scale deletions in 17p are tested using FISH (Flourescence In Situ Hybridization), small scale genome variants have to be detected using Sanger Sequencing (SSeq) or NGS (Next Generation Sequencing) (51). "TP53 network" is a certification program initiated by the European research initiative on CLL (ERIC), for clinical laboratories performing these studies, as a measure of standardizing these determinations. NGS is the preferred technique for detecting somatic mutations in this gene due to its augmented resolution compared to SSeq, but standardization is needed to decide the interpretation of variants with lower variant allele frequency (VAF). Those practically undetectable using Sanger Sequencing (<10-12%), have to be studied and reported carefully until the guidelines include further stratification (52).

1.2.2 IGHV mutational status

Over all, the presence and load of SHM within the rearranged IG heavy variable (*IGHV*) genes of the clonotypic B cell receptor, remains stable over time and dichotomizes CLL into two broad categories: unmutated (U-CLL) which includes cases with no or limited SHM and mutated CLL (M-CLL) with cases with significant SHM load (53,54). Both are markedly different, not only in the clinical course of the diseases, but also in terms of biological features: U-CLL is associated with adverse prognostic genomic aberrations, increased BcR signaling capacity, shorter time to progression and an overall inferior outcome compared to M-CLL. The current gold standard to determine *IGHV* status in CLL is SSeq using multiplex primers (preferentially Leader and if not possible, FR1, because they cover *IGHV* region almost entirely) (49,55,56).

Hematopoietic stem cells have been proposed to be the origin in CLL (57). Gene expression studies have elucidated that both U-CLL and M-CLL contain expression patterns similar to memory B cells rather than naïve which was the initial theory for U-CLL cells, as apparently they had not undergone SHM events (58,59). Besides, they have a surface phenotype typical of antigen-experienced B cells irrespective of the prognostic stratification (60). The biological mechanisms that differentiate so evidently the degree of mutations of U-CLL and M-CLL is therefore, not known, even though one of the most accepted theories is that U-CLL derive from GC-independent cells and M-CLL have a post-GC origin (58,59,61) (Figure 1.5). On the other hand, the existence of biased repertoire in both U-CLL and M-CLL, which implies the recognition of a limited set of antigens, and the more recent, further stratification into *stereotyped* receptors, is explained by the presence of antigenic selection. Usually, *IGHV1* family is more common in U-CLL whereas *IGHV3* and *IGHV4* are found with higher mutational loads (M-CLL). *IGHV1-69* is a very recurrent gene in CLL whose usage is mostly given on U-CLL rearrangements. Others like *IGHV3-23*, *IGHV4-34*, *IGHV3-7* and *IGHV3-48* are frequently used in M-CLL rearrangements (62–67). Regarding *IGHJ* genes usage, *IGHJ6*, which contains the largest sequences, is predominant in U-CLL, whilst *IGHJ4* is common in M-CLL (63,66,68).

However, gene expression differences also exist between the U-CLL and M-CLL groups, and between stereotyped subsets (69). Extensive BcR signaling is present in U-CLL rather than M-CLL where sIgM expression is scarce (70). Multireactivity in CLL B cells is also related to stereotyped receptors responding against a restricted set of antigens, proved also on a structural modelling basis (71,72).

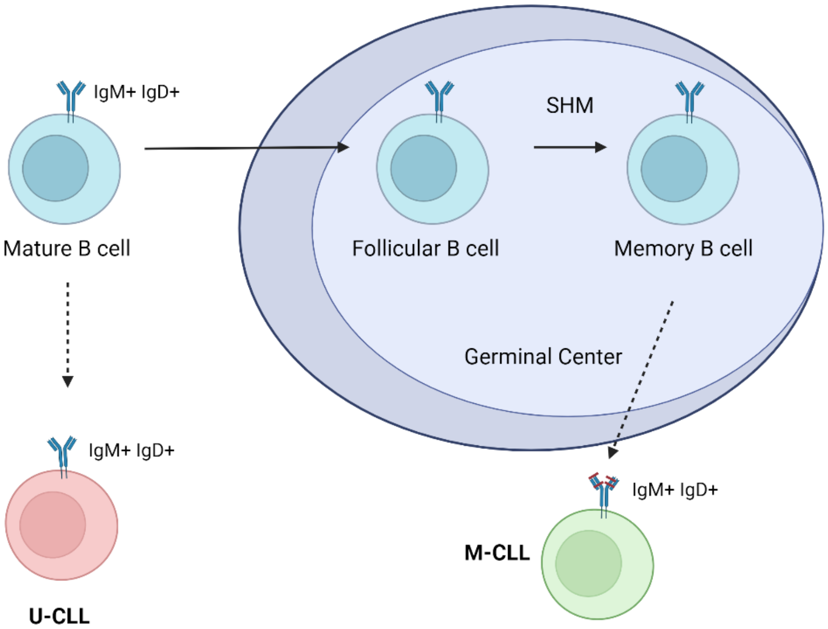


Figure 1.5. Origin of U-CLL and M-CLL cells. One of the accepted theories is that both U-CLL and M-CLL derive from mature activated B cells, but independently of GC antigenic reactions in the case of U-CLL. Adapted from (73), with BioRender.com.

1.2.3 Stereotyped subsets

Antigen-driven selection of CLL clones has been connected to the recognition of bacterial and auto-antigens (74,75). Additionally, tumor microenvironment plays an important role, through the involvement of T cells and other immune effectors, which can activate survival and cell proliferation (76,77). Stereotyped subsets reported by different groups, are a recent evidence of antigen role in the development of CLL malignant clones (78–81). Approximately 30% of patients carry stereotyped subset BcRs (82), and even two groups have been differentiated, so called "stereotyped" and "heterogeneous" (81).

It was discovered that certain heavy and light Ig chain genes were associated to concrete patterns in the CDR3 sequence, suggesting similar functional features and are also used to define prognostic groups (82–84). For instance, CLL-2 subset represented by *IGHV3-21/IGLV3-21* usage, is considered a poor prognosis marker regardless of the mutational status (85,86), whereas subset CLL-4 (*IGHV4-34/IGKV2-30*) has a remarkably indolent course (87,88). The current guidelines consider subsets CLL-2 and CLL-8 in the newest update for prognostic decisions (89).

Stereotypy has also been found in other B cell malignancies, such as mantle cell lymphoma (MCL), but not as recurrent as in CLL, and the subsets found were not coincident between malignancies (71,82).

1.2.4 B cell receptor signaling pathway

On the other hand, stratification groups from antigen-independent pathogenesis, regarding mutations in genes in the B cell receptor signaling pathway, are being studied for playing an important role in disease progression (90,91). U-CLL harbor more active signaling whereas M-CLL have been described to be more “anergic”, responding mildly to antigenic contact. This in part, explains how U-CLL cells proliferate faster and aggressively, being multireactive and probably being continuously driven by ongoing stimuli in the B cell receptors. Individualized treatment approaches are advancing towards the disruption of BcR signaling pathway, and it is reasonable to study the antigenic origin driving these responses (71).

Many signaling pathways have been related to the development or progression of CLL (genes with recurrent driver mutations or progressed after treatment or relapse). The most relevant biological mechanisms implicated are DNA damage responses, NOTCH1 signaling, RNA splicing, cell cycle, BcR signaling, chromatin modification and Toll-like receptors inflammatory pathway (61,91) (Figure 1.6). Novel treatment methods targeting the BcR signaling pathway are BTK inhibitors (ibrutinib, acabrutinib), PI3K inhibitors (idelalisib) and BCL2 blockers (venetoclax) (92). BTK inhibitors have been proven to be beneficial in cases with poor outcome such as *TP53* mutations, but resistances appear, with mutations in *BTK* and *PLCG2* (93).

About recurrent mutations, a study conducted by the ERIC consortium screened nearly 3500 CLL patients and determined that mutations in *NOTCH1, SF3B1 and TP53* were clinically aggressive, with shorter TFTT (time-to-first-treatment), and were present in both U-CLL and M-CLL cases. Special focus was added to *SF3B1* and *TP53* mutations, for having an adverse prognosis (94). Genetic lesions in *NOTCH1*, *SF3B1* and *BIRC3*, coupled with common chromosome aberrations in CLL were employed for stratification of patients in groups from low to high risk (95). Profound revisions regarding CLL treatment have been performed from different studies in other works focused on autonomous BcR signaling pathogenesis and therefore, they are not going to be further developed in this work (61).

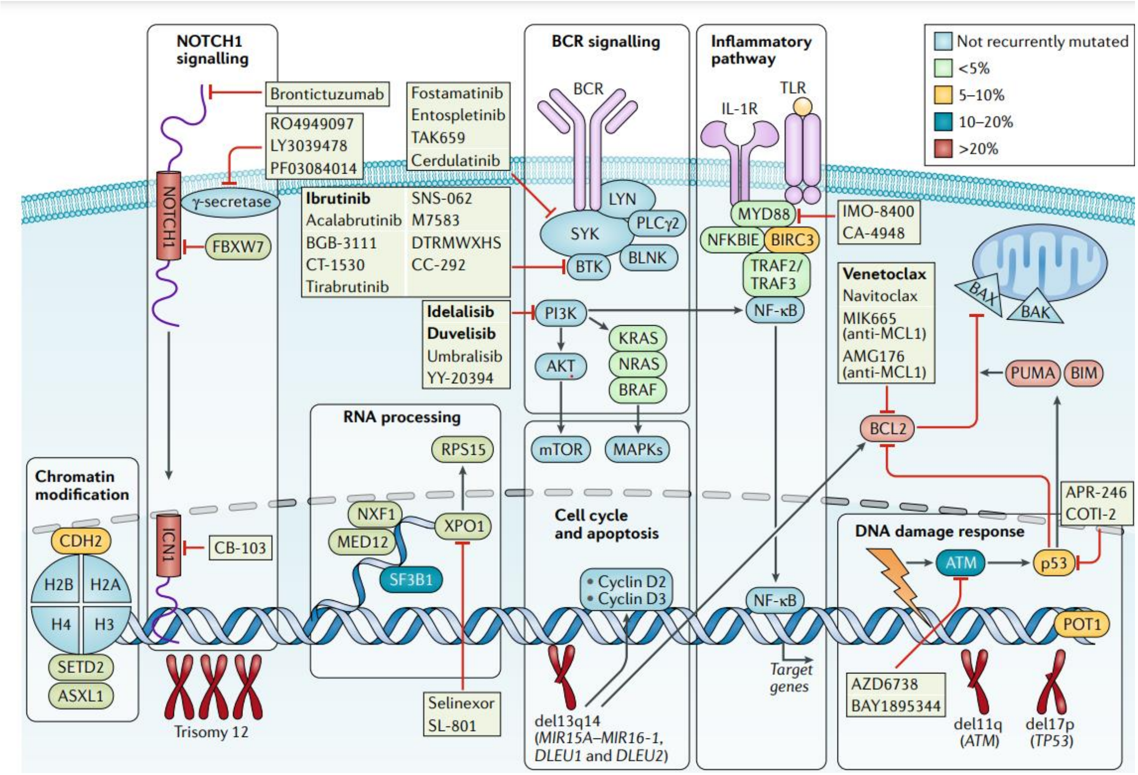


Figure 1.6. Genetic landscape of CLL revealed by FISH and NGS. Signaling pathways from recurrently mutated genes, and drug targets. From (61).

1.2.5 CLL repertoire architecture

Figure 1.7 illustrates a simplification of the types of clonal scenarios in CLL. The most common is the monoclonal state with a single productive rearrangement, whereas monoclonal with biallelic (productive + not productive) rearrangements can also be seen when allelic exclusion phenomena occurs. If by contrary, a single clonal population exists but allelic exclusion was not successful, two productive receptors can be expressed in the same B cell, detecting various clones with the conventional sequencing techniques. On the other hand, when various clones arise from different B cell populations, the scenario is biclonal or oligoclonal, depending on the number of clones, being multiple productive rearrangements (8).

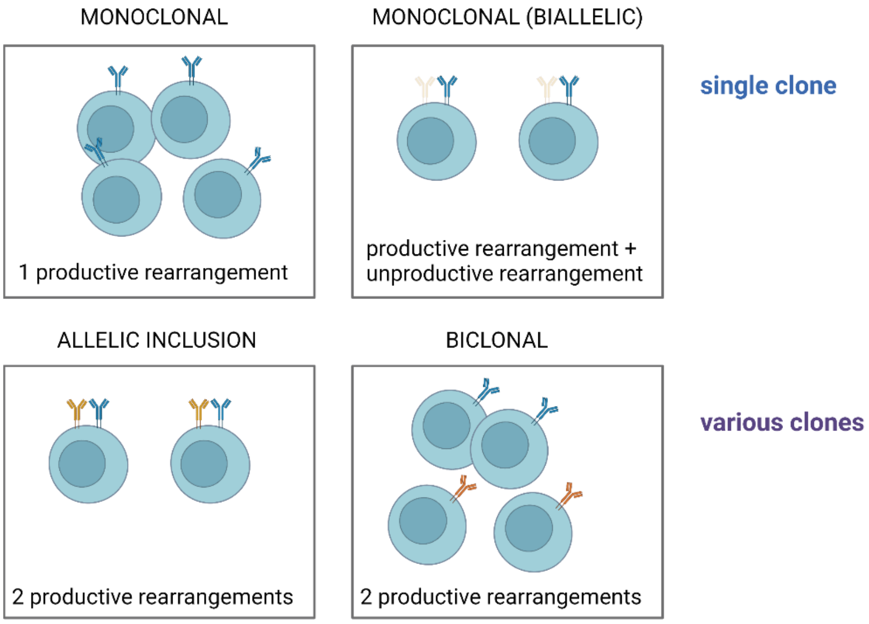

MONOCLONAL
1 productive rearrangement
MONOCLONAL (BIALLELIC)
productive rearrangement +
unproductive rearrangement
single clone
ALLELIC INCLUSION
2 productive rearrangements
BICLONAL
2 productive rearrangements
various clones

Figure 1.7. Types of clonal scenarios in CLL B cell clones. *"Single clone" and "various clones" refer to how these cases would be detected by the conventional molecular biology techniques. Adapted from* (8) *using BioRender.com.*

1.2.5.1 Double rearrangements

Scenarios where B cells express more than a single productive surface Ig receptor due to lack of allelic exclusion (termed allelic inclusion), have been described in autoreactive B cells probably intending to conform a second, functional receptor (biallelic populations) (56,96). A study by Plevova and collaborators reported that most of the multiple productive clone patients, derived from independent B cells populations (cases of biclonal/oligoclonal populations) (97), and the first study of these characteristics employing single-cell, proved that 7 cases harboring multiple productive rearrangements, accomplished allelic exclusion (98).

1.2.5.2 *Intraclonal diversification*

Several works have described cases of ongoing antigenic stimulation suffered by B cells, and causing intraclonal diversity from the predominant clone(s) (99). Sutton and collaborators (100), suggested antigen-driven ongoing SHM in rearrangements from subset number 4 with *IGHV4-34*. Other works where pyrosequencing was used to analyze the B cell *repertoire* of CLL patients, revealed the presence of so-called “satellite clones”, being lower proportion clones with minimal variation with respect to the most dominant clone (101,102) and also suggesting intraclonal diversification. Ongoing SHM has also been described in the MBL (Monoclonal B cell lymphocytosis) state, prior to CLL development (103). However, these cases have to be studied in depth to ensure the clear separation line between sequencing artifacts and minority variants (104,105). Clinical implications are not known.

1.2.6 Complicated cases

In 2011, Anton Langerak and collaborators (56), described a list of cases with challenging interpretations after their experience in the clinical determination of the mutational status of *IGHV* domain in CLL, and set some recommendations to treat these cases accordingly. These are: double rearrangements with discordant mutational status, single unproductive rearrangements and lack of the IMGT anchoring junction region amino acids.

1.2.6.1 *Double rearrangements with discordant mutational status*

In 10% of CLL cases, double Ig rearrangements are detected. Mainly, these imply the presence of 1 productive + 1 not productive rearrangement. However, 19% of the former present various productive Ig rearrangements and in different works using different methodologies, these percentages are variable (8). In general, it was described that 5% double rearrangement cases consist of two productive rearrangements, maybe due to lack of allelic exclusion or to unrelated clones (106). If the mutational status is discordant between 2 or more productive rearrangements, the interpretation is challenging. Currently, the guidelines recommend prioritizing the U-CLL rearrangement, especially if dominant, and encourages the use of NGS to assess these cases in depth (89). Stamatopoulos and collaborators (107), detected 25% of cases with multiple productive rearrangements and performed a new stratification based on the mutational status present in multiple clones (equal, discordant, etc), and unraveled a complex biological background in CLL that goes beyond the monoclonality described until then. Even though more studies proving these results are needed, cases with two B cell clones have shown earlier need for treatment than cases with a single CLL clonal rearrangement (108).

1.2.6.2 Single unproductive rearrangement.

Only one rearrangement is detected and the functionality is not clear. The interpretation will remain inconclusive until a productive VDJ rearrangement is detected, using different sample types, or genetic material (cDNA, gDNA), and using different analysis methods. The updated ERIC guidelines also recommend unraveling the clonal architecture using NGS if the other approaches fail (89).

1.2.6.3 Undetected junction anchor amino acids

IMGT amino acid anchors Cys 104 and Trp 118 from the junction region (crucial for the integrity of antigen-binding loop) are not identified and therefore the junction sequence region remains undetermined. These cases are also considered inconclusive. Exemptions are made when the final G-X-G (Gly, any amino acid, and Gly) motif preceded by an amino acid different to Trp or the expression is proven positive by other methodology, being considered productive (56).

1.2.7 Borderline mutational status

Borderline mutational status (BD-CLL) is not included officially among the complicated cases but special attention is needed in this scenario. Ig CLL rearrangements whose percentage identity against the closest germline allele is between 97-97.9%, are considered borderline (BD). The guidelines recommend caution when interpreting these cases with mutational status in the marginal zone of the arbitrary SHM cut-off between U-CLL and M-CLL (98%), even though they fall into the M-CLL group in the classification (109). Similar survival curves were described in BL-CLL and M-CLL, both showing better prognosis than U-CLL patients, in 759 patients studied (110).

1.2.8 Subclonal fraction

Current studies have depicted a more complex reality in the clinicobiological features of CLL, pointing for possible prognostic repercussion of minor rearrangements (97,98,107). Subclonal rearrangements can be stable over time or present clonal drift, changing the relative frequencies between clones over time, when these clones are selected after acquiring genetic characteristics which increase their fitness (111). The presence and accumulation of driver alterations displays worse prognosis and subclones with certain driver mutations can be unfavorable for the patient even at low proportions (112). Stable equilibrium is more common for untreated patients whereas branched evolution is more favorable after treatment, with the acquisition of resistant subclones (113) (Figure 1.8).

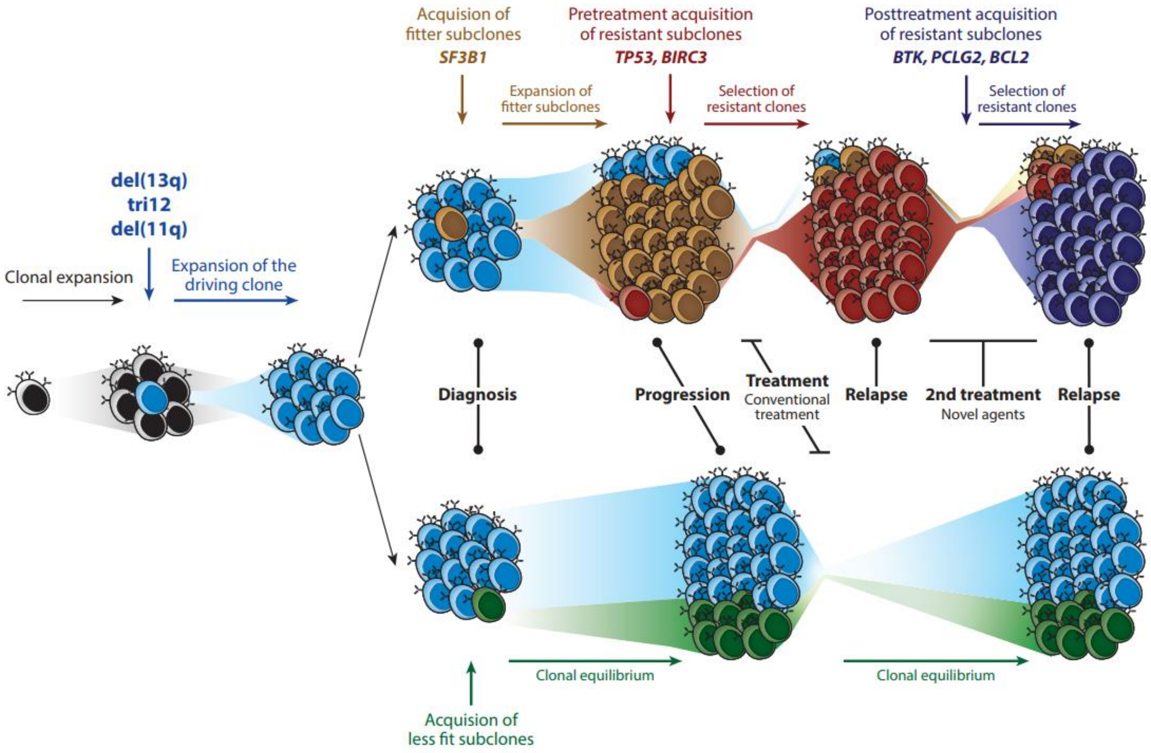


Figure 1.8. Clonal evolution. *Stable evolution maintains relative proportions of clones whereas branched evolution occurs when strongly fit subclones start to proliferate and become dominant after treatment resistances. From* (113).

1.3 Next Generation Sequencing

After two decades, the culmination of the Human Genome Project in 2003 (114,115), opened the door to novel studies in the fields of human genetics and understanding diseases. In the early 2000s, NGS technologies were introduced, revolutionizing many molecular biology research areas. Gradually, technologies and platforms for massively parallel sequencing started to evolve, being rapidly applied to *de novo* assembly of microbial genomes (metagenomics), RNA sequencing to measure gene expression (transcriptomics), or DNA sequencing for the detection of variants along the genome (genomics), among many others omics sciences that continue developing unceasingly. Thanks to the rise of these methodologies, nowadays it is possible to sequence a complete human genome with reasonable costs and increased throughput (116,117).

There are three main approaches for sequencing the human genome, depending on the questions to address. WGS, is the most expensive as it covers the sequencing of all coding and noncoding regions (~3x10^{9} bp) and coverage depth is very limited (30-60X). A common alternative for studying human diseases is WES, which represents 1-2% of the whole genome. Consisting of the protein coding regions, where the majority of variants causing disease are based, the amount of bases sequenced is drastically reduced (118). However, the cheapest and most practical option is targeted sequencing (TS), which consists of direct sequencing using capture hybridization or PCR amplification methods to enrich a few regions of interest among the genome (119,120). The depth of coverage obtained depends on the size of the panel but compared to WES, ultra-high depths can be obtained, as much smaller fractions of the genome are targeted (Figure 1.9). The decision on whether to use one or another depends on the purpose and costs. For example, detecting somatic mutations requires reaching high levels of sensitivity, where exome sequencing is not recommendable as it compromises sequencing depth, assuming very high costs to reach the desired support for calling variants. In contrast, with targeted sequencing, the gene panel is designed to amplify the sequences of interest and coverage depth can be adapted to the limit of detection (118).

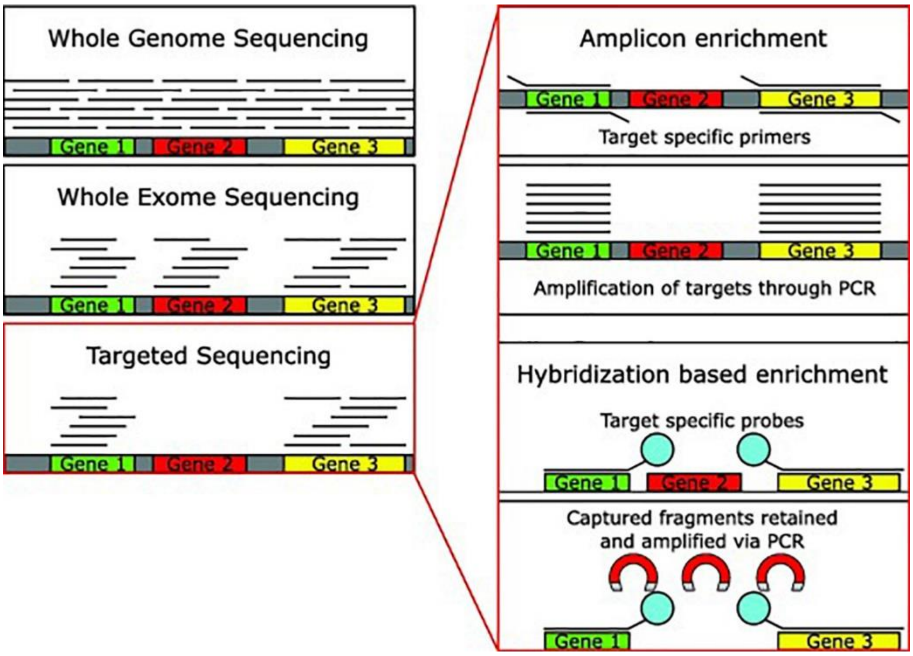


Figure 1.9. DNA-sequencing strategies used in NGS. *From* (118).

Among NGS platforms, the difference can be drawn regarding sequencing read lengths, being named short-reads and long-reads platforms, respectively. The most reliable short read platforms are Illumina (California, EEUU), based on solid-phase bridge amplification, and Ion Torrent (TermoFisher; Massachusetts, EEUU), based on emulsion PCR. Third generation sequencing platforms have the advantage of single molecule resolution, being able to skip PCR steps and the consequent errors (116,117,121). On the other hand, Pacific Biosciences (PacBio; California, EEUU), and Oxford Nanopore (Oxford, UK) can reach long read lengths. In the recent completeness of the human genome sequencing, they used Oxford Nanopore ultralong > 100kbp reads to span highly repetitive regions and centromeric areas of some chromosomes (122). Currently, Illumina platforms are the most used among the existing NGS technologies. Due to the lower error rates and costs compared to other sequencing modalities, they are especially valuable for clinical purposes regarding gene panels for variant detection. In the last decade, Illumina technology has adjusted to the growing demands and designed machines of increasingly high capacities (HiSeq and NovaSeq) (123).

1.3.1 Comparison of SSeq and Illumina sequencing by synthesis

Illumina sequencing is based on "sequencing by synthesis" technology, like its predecessor, capillary SSeq. The term refers to the reaction of DNA elongation carried out by the polymerase enzyme. In capillary SSeq, each DNA molecule allows the addition of a single nucleotide as the reaction is blocked completely afterwards by the ddNTP inserted (common nucleotide molecules lacking the 3'-OH necessary to elongate DNA chains) (Figure 1.10a). The main difference is that in Illumina, the terminator molecule blocking the addition of each nucleotide is reversible, and thus cleaved after each elongation cycle, allowing massively parallel sequencing of thousands/millions of molecules at a time. First of all, DNA molecules are attached to the surface termed "flow-cell", by complementary sequences of the sequencing adapter, hybridized or ligated to DNA inserts during library preparation. Afterwards, amplification of this molecules is performed, to form clusters of each DNA sequence. The step receives the name of bridge amplification due to the shape that DNA molecules acquire when they are amplified from one end and then from the other. The cluster of molecules allows sequencing in parallel of many copies of the same fragment and amplifies the fluorescent signals during sequencing (Figure 1.10b). The main advantage of reversible terminators is that homopolymer errors (common in Roche 454 and Ion Torrent machines) are minimized because nucleotides are not added simultaneously (121,124).

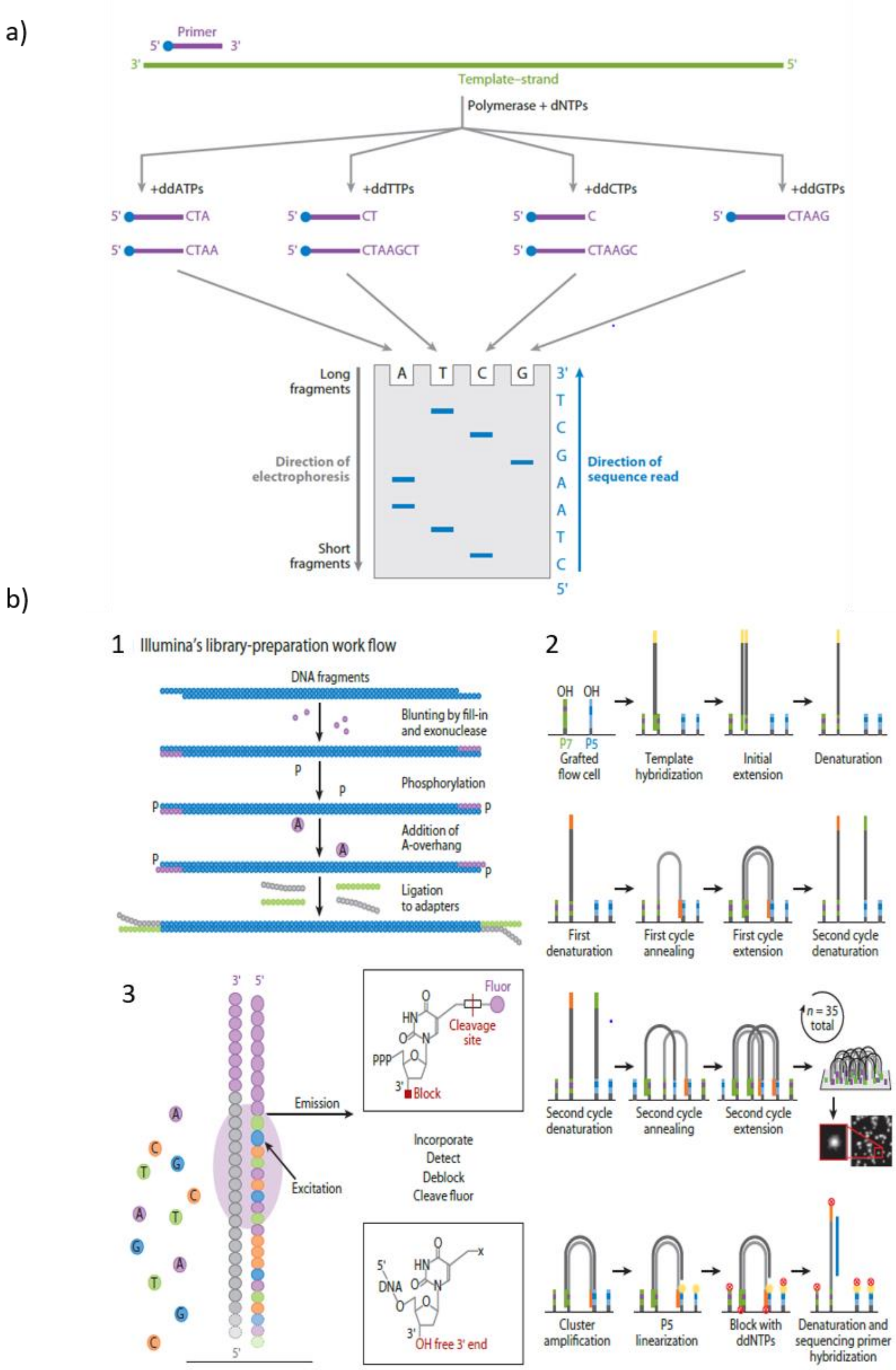


Figure 1.10. SSeq vs Illumina sequencing technologies. *a) SSeq scheme and b) Illumina sequencing technology scheme: 1) Library preparation, 2) generation of clusters by bridge amplification and 3) sequencing by synthesis process. Adapted from (124).*

1.4 Deep sequencing B cell receptors

The first NGS studies focused on the adaptive immune repertoire were released in 2009 (125–127). Since then, the use of NGS for TcR/BcR (T cell and B cell receptors) sequencing has opened a wide range of possibilities and applications in the fields of vaccines, infection, autoimmune, and hematological malignancies, among others (128).

Specific to CLL, it has uncovered a more complex reality in the biological ontogeny of this disease. Now it is clear, that the use of NGS for characterizing lymphoid malignancies offers many advantages, but the methods need standardization. The 2022 updated ERIC guidelines, had taken into consideration the use of NGS under certain conditions, and they will continue to grow in upcoming versions of these recommendations (89). Some aspects like characterizing the subclonal fraction, intraclonal diversity and minimal residual disease (MRD) are gaining importance in CLL and NGS can present an advantage for all of them as they can be assessed as a whole, with the suitable methodologies. The works describing this heterogeneity in CLL clonal architecture, suggest and create the necessity to study if these cases can have clinical implications. However, the intrinsic variability and characteristics of the IGH locus poses challenges to address the determination using NGS (104). Some consortiums like the NGS Euroclonality in Europe are working in validating methods using multicentric studies to set standards for this purposes (129,130).

There are mainly two methods used for library preparation in BcR sequencing (also applicable to TcRs). The first one is the multiplexed PCR amplification from RNA or gDNA. Consensus primers designed for the Leader, Framework, and JH or constant region reverse primers are used. These primers were designed per *IGHV* families with a degenerate basis to hybridize with most of the allele segments (55). The second is template switching 5´RACE, applied to RNA. The main disadvantage of the multiplexed PCR method is the introduction of primer amplification biases. On the other hand, with 5´RACE, full-length sequences are not always retrieved (131).

The use of one or another depends on the application and the source of the genetic material. 5´RACE is used with RNA and the use of UMIs (Unique Molecular Identifiers) can be integrated. On the other hand, multiplexed PCR is one of the most used methods, with the disadvantage of amplification biases. When targeting whole B cell repertoires with many rare clones, these biases are magnified. In approaches where the aim is targeting clonally expanded B cells, sequencing errors and primer biases are less significant (132). In a protocol designed by Cole et al, they use an in-house tagmentation method to strategically sequence the region entirely (133).

1.4.1 Limitations

The imitations for NGS sequencing the IGH locus start with the variability and length of the rearranged gene segments. Adjusting to the region length with SSeq is not a challenge, but most reliable NGS sequencing technologies used nowadays, employ short reads. Long-read sequencing NGS platforms, on the other hand, could solve the issue but their error rates are still high and can cause many difficulties when assigning VDJ gene calls. PacBio has an error rate of 13%, which is strikingly high compared to short-read sequencing platforms in Table 1.1. Due to error rates and costs ratios, Illumina and Ion Torrent are the most used platforms for this purpose, and due to the read-length limitation, many approaches adopted Illumina MiSeq 300bpx2 sequencing kit, for being the longest sequencing kit in Illumina technology, as HiSeq offers higher sequencing depths but at a cost of read length (134). Therefore, HiSeq is commonly used mostly for targeting the junction region (from the end of *IGHV* to *IGHJ*) (135). Illumina sequencing quality decreases with the number of cycles introduced, and in practice, the 600 bidirectional-cycles kit yield decreases to 450 bp sequenced at acceptable quality score values (133). Ion Torrent has also been described as acceptable even though the error rate is 10 times higher than Illumina (104) (Table 1.1).

Regarding the use of RNA or gDNA as starting material, using RNA as genetic material has the main advantage of making sequencing more straightforward in means of length for lacking intronic sequences and thus, the possibility of reaching the constant region. On the other hand, gDNA reflects in a more unbiased manner the proportion of cells, whereas with RNA clonal architecture can be distorted as some types of cells have intrinsically higher surface Ig expression (136). Also gDNA, is more stable to work with and does not require reverse transcription, but also requires higher concentrations. As advantages and disadvantages exist with both sources, it depends on the main goal of the study. For instance, whether expression or relative quantification of the clones is needed for characterization, or the presence of both productive and unproductive rearrangements (137,138).

For clinical determinations of *IGHV* mutational status, the recommendation is to obtain full-length *IGHV* sequences to be able to calculate unbiased mutation frequencies. The commercial LymphoTrack assay (Invivoscribe Inc., San Diego, CA, United States), details in their protocols the use of 300bpx2 or 250bpx2 Illumina kits. Using the Leader primers always require 300bpx2 sequencing whereas 250bpx2 can be used with FR1 primers.

Platforms	**Roche's 454 GS FLX**	**Illumina MiSeq**	**Illumina HiSeq**	**PacBio**	**Ion torrent**
Read length	700 bp	300 bp × 2	250 bp × 2	860–1,100 bp	>100 bp
Run time	18–20 h	26 h	8 days	0.5–2 h	2 h
Reads/run	1M	3.5M	2B	0.01M	60–80M
Error rate (%)	1	~0.1	~0.1	~13	~1
Type of errors	Indel	Substitution	Substitution	Indel	Indel
Cost/mbp ($)	12.40	0.74	0.10	11–180	<7.5
Region of antibody covered	FWR1-CR	FWR1-CR	FWR1-CR	Amplification of linked H and L chains	FWR3 to CR

Table 1.1. Common platforms used for immunoglobulin repertoire sequencing. *From* (134).

1.5 Analysis methods for immune repertoires

Frequently, bioinformatic analysis methods for immune repertoires include several preprocessing steps to tackle PCR and sequencing artifacts, which are normally difficult to distinguish from SHM events. Those include quality correction, pair read merging, consensus sequence building, and similarity grouping, based on sequence distance clustering methods or identity thresholds (139–143).

After preprocessing, reads are aligned against the reference IMGT alleles as there is no conventional reference genome. Some commonly annotation bioinformatics tools employed for the analysis of immune repertoires are MiXCR (pipeline that performs VDJ assignments at gene level) (144), IgBlast (uses BLAST algorithm to annotate VDJ genes at allele level) (6), and IMGT/V-QUEST, which has a version for larger sequence sets (IMGT/HighV-QUEST) (145). Most of these methods rely on the standards of the IMGT reference alleles and unique numbering for CDR3 delineation (146).

Afterwards, several downstream analyses can be performed in order to explore repertoire characteristics such as gene usage, repertoire overlap analyses (Morisita-Horn index, Jaccard index), repertoire diversity based on ecology parameters (Hill, Shannon), etc. The Immcantation Framework is a suite of tools that integrates the use of different pipelines including bulk and single-cell repertoire analysis, with the integration of the IgBlast tool for allele annotation and many functionalities for downstream analyses including construction of phylogenetic trees with SHM models, mutation analysis, clustering methods to infer clonal relationships and CDR3 amino acid properties, among others (147).

2 Hypothesis and Aims

2.1 Hypothesis

Until the current ERIC guidelines for the characterization of the IGH locus in CLL, released in May 2022, the use of NGS was not considered (148). Besides, the application of the same is recommended only under unclear scenarios (89). However, simplification and automation of the method with NGS protocols is advisable, as 3-4% of CLL cases remain unclear after SSeq.

The use of NGS for the characterization of IGH locus in B cell neoplasms for clinical procedures can be standardized in the near future using simple library preparation methods and automated bioinformatics tools. The use of NGS can unify mutational status, clonality and MRD analysis with more straightforward determinations in methods validated against the gold standard.

The general hypothesis of this work is that by using short-read massive sequencing of Ig rearrangements and a reliable bioinformatics approach, we can detect and differentiate the rearrangements present in CLL samples and their mutational status in an automated way.

2.2 Aims

The main aim of this work is the development and automation of bioinformatic tools that can be applied in the clinical routines for IGH locus characterization in B cell neoplasms, based on the analysis of NGS data.

To accomplish this, the specific aims are:

- To find suitable methods for the construction of IGH locus NGS DNA libraries, in terms of coverage, turnaround time and costs.
- To develop an in-house bioinformatic pipeline for the characterization of the IGH locus, filling the needs of clinical procedures applied in CLL, while overcoming the intrinsic

complexity of the results obtained after NGS with automatic resolution of clonal/non-clonal background.

- To develop a second pipeline from tools created by experts in B cell repertoire sequencing, and use it for benchmarking of the primary pipeline.
- To evaluate the reliability of the aforementioned methods by comparing their performance against the gold-standard techniques used for clinical purposes.

3 Methods

3.1 Patients, sample collection and preparation, and DNA/RNA extraction

Peripheral blood (PB) and bone marrow (BM) aspirate samples were obtained at diagnosis from patients with CLL, and PB samples were obtained from healthy donor patients. Genomic DNA was isolated from PB by the Maxwell® 16 Blood DNA Purification Kit (Promega). Total RNA was extracted from PB Maxwell® 16 Total RNA Purification Kit (Promega; Wisconsin, EEUU). Complementary DNA (cDNA) was then generated by retrotranscription using the Quantitect Reverse Transcription Kit (Qiagen, Germany). Cases where the number of absolute lymphocytes was below 5000 (cells/µL), purification of tumoral CD19+ cells was performed. This study was approved by the local Institute Ethics Committee and CLL patients and control samples were used after written informed consent. Samples were obtained according to the National Cancer Institute Working Group guidelines in our institution between 1986 and 2019. Special mention to medical doctors Dr. MJ. Terol, Dr. B. Navarro and Dr. B. Ferrer.

3.2 Classical PCR Sanger Sequencing method

Genomic DNA (gDNA; 50–100 ng) and/or cDNA (50–100 ng), were used for *IGHV* analysis. gDNA/cDNA, was amplified using IGH locus-specific primer sets to allow the amplification of all known alleles of the germline IGH sequence. Leader primers (forward primers) and consensus *IGHJ* (reverse primer) allow the whole sequence of the *IGHV* region to be obtained and thereby the precise definition of the percentage of identity to the closest germline gene (55). In occasional cases, other primer sets (including FR1, FR2 and FR3 primers) were employed (149). After performing PCR reactions, the presence of rearranged bands was checked by capillary electrophoresis by means of the QIAxcel Advanced system (Qiagen, Germany). Direct sequencing of the PCR reaction products (SSeq) with forward and reverse primers was performed with BigDye Primer Sequencing Kit (Thermo Fisher, MA, USA). The consensus sequences obtained by means of SSeq were analyzed using IMGT/VQUEST tool (https://www.imgt.org/IMGT_vquest/analysis), which provides automatically the calculation of the percentage of *IGHV* identity to germline, the number and description of mutations per FR-IMGT and CDR-IMGT, and the identification and localization of the hot spots in the germline (150).

The two aforementioned steps (3.1,3.2) were performed in the Hematology Department of the Clinical University Hospital of Valencia (HCUV), especially by A. Serrano.

3.3 Preparation of primers pools

Leader and BIOMED-2 consortium (55) *IGHV* gene family consensus primers were pooled into Leader, FR1, FR2 and FR3 primer pools: 7 *IGHV* families + *IGHJ* in the case of Leader, 6 *IGHV* families + *IGHJ* for FR1, 7 *IGHV* families + *IGHJ* for FR2, and 7 *IGHV* families + *IGHJ* for FR3 (Appendix 8.1). Primers were mixed into a final concentration of 0.2 µM. Different combinations of these pools were used to generate the three library approaches (detailed in Methods section 3.4).

3.4 NGS sequencing libraries testing

For the amplification and sequencing of the IGH locus with NGS Illumina technology, three different strategies with different primer sets combinations were tested, as well as different library preparation protocols, and finally, different Illumina sequencing kits with different read-length yield (Figure 3.1).

We evaluated which method was the most reliable in terms of costs and performance characterizing the predominant CLL rearrangements. Two of the methods were based on the amplification of the whole region, sequenced with 300bpx2 Illumina MiSeq kit in one case, and tagmentation of the amplificated region using Nextera XT commercial Kit (Illumina), followed by 150bpx2 reads Illumina MiSeq sequencing kit in the second. The third, consisted of an in-house method including various standard primers sets for amplification of the VDJ region from different starting points. The three approaches were analyzed with a preliminary in-house IGH pipeline.

Samples from 23 CLL patients were used after gDNA extraction as described in 3.1, and subjected to sequencing using these methods, explained next.

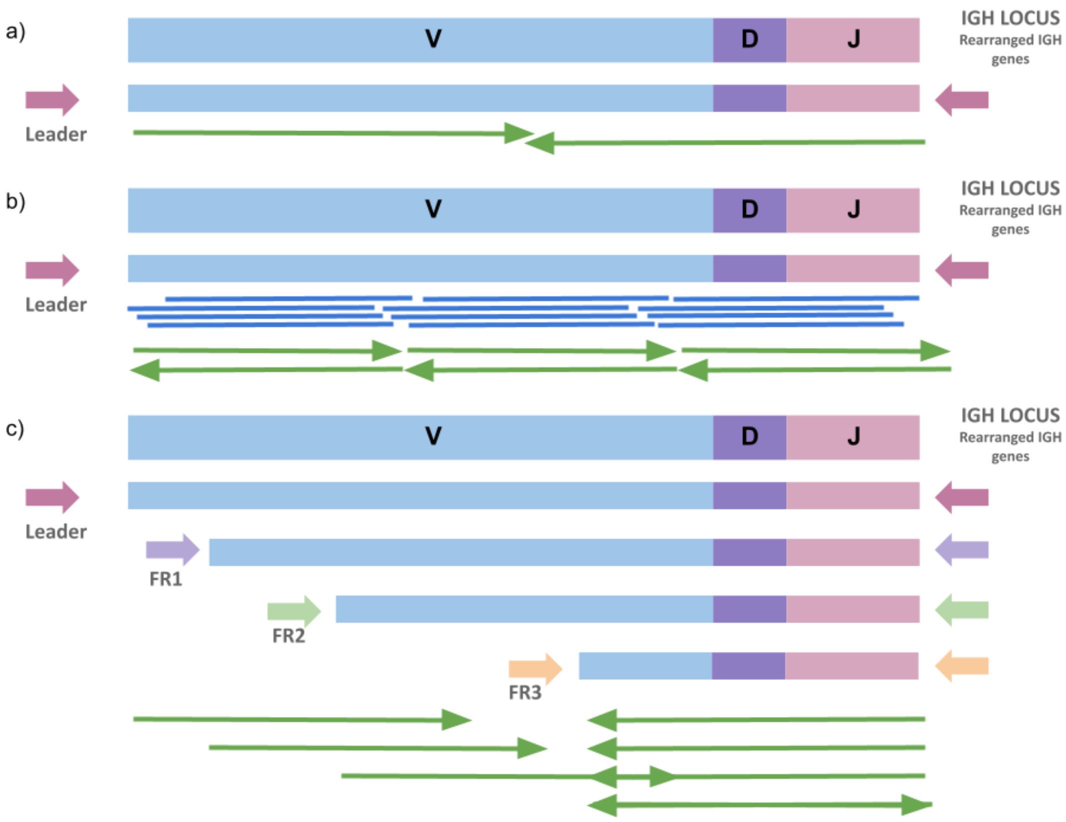


*Figure 3.1. **Library preparation methods.** a) Leader IGHV and IGHJ consensus primers used to cover the whole IGH rearranged locus. Sequencing of the whole fragment is performed with paired 300 read-length in the MiSeq platform. b) Leader IGHV and IGHJ consensus primers are used to amplify the whole IGH rearranged locus and tagmented using Nextera XT Illumina protocol so as to be sequenced using paired 150 read-length. c) Use of the 3 Framework primer sets (BIOMED-2) consortium for multiplex amplification and obtention of partial reads with paired 150-length Illumina sequencing.*

3.4.1 Library approach A: Leader primers with MiSeq V3 (300bpx2)

This strategy (Figure 3.1a) consisted of the amplification of the IGH locus variable region from gDNA using Leader *IGHV* families set of forward primers and the reverse consensus *IGHJ* primer. In order to perform sequencing of the entire amplified region, 300bpx2 cycles were employed with the v3 Illumina MiSeq sequencing kit (Illumina; California, EEUU). 1 µL of DNA (50ng/µL) was amplified using 2x QIAGEN PCR Master Mix (Qiagen; Germany) from reference Leader primers with Nextera adapters complementary ends (Illumina; California, EEUU) for all *IGHV* family subtypes in multiplex (55,151), combined as in 3.3. Afterwards, a second PCR step was performed using the same enzyme with Nextera custom adapter sequences. PCR products of the different samples were pooled, purified using 0.6X Magsi-NGS Prep magnetic beads (Magnamedics Diagnostics; United Kingdom), and quantified with Quantifluor dsDNA system (Promega; Wisconsin, EEUU). DNA libraries were normalized (10nM) before introducing them into the Illumina MiSeq platform for sequencing with MiSeq Reagent kit V3 300bpx2 (Illumina; California, EEUU) following commercial specifications.

3.4.2 Library approach B: Illumina Nextera XT

The second strategy (Figure 3.1b), consisted of the amplification of the region Leader-JH with the different *IGHV* family leader primers set and *IGHJ* consensus. In this case, reads were tagmented using Nextera XT library preparation kit (following kit instructions), to obtain shorter DNA fragments and sequence reads 150 bp long. 1 µL of DNA (50ng/µL) was amplified using reference Leader primers (55,151) with 2x QIAGEN PCR Master Mix (Qiagen; Germany) as in 3.4.1. The product of PCR1 was purified using Magsi-NGS Prep magnetic beads (Magnamedics Diagnostics; United Kingdom), and quantified with Quantifluor dsDNA system (Promega; Wisconsin, EEUU), and tagmented. Samples were pooled, and DNA libraries were normalized (10nM). Illumina MiSeq V2 150bpx2 sequencing protocol (Illumina; California, EEUU) was used to load the library in the MiSeq platform for sequencing using the commercial specifications.

3.4.3 Library approach C: In-house multiplexed primer fragments method

The last strategy (Figure 3.1c), consisted of an in-house method combining 3 Framework BIOMED-2 consortium primer sets (Framework regions 1, 2 and 3), and JH consensus. Primers previously pooled into separate FR1, FR2 and FR3 mixes as described in Methods section 3.3 were pooled into a FR1-FR2-FR3 unique pool. The aim of this strategy was to use shorter read sequencing (150bp), obtaining partial reads with the support of the 3 amplicons whose overlap will give enough information to characterize CLL B cell clones.

1 µL of DNA (50ng/µL) was amplified using a mix of primer sets in multiplex to obtain nested fragments in a single reaction (55,151). Using all FR primer sets (FR1, FR2, FR3) with Nextera adapters complementary ends (Illumina; California, EEUU). PCR was performed using 2x QIAGEN PCR Master Mix (Qiagen; Germany). A second amplification step was performed with the same master mix using Nextera custom adapter sequences. Samples were pooled, purified using 0.6X Magsi-NGS Prep magnetic beads (Magnamedics Diagnostics; United Kingdom), quantified with Quantifluor dsDNA system (Promega; Wisconsin, EEUU), and then normalized (10nM). Illumina v2 150bpx2 cycles sequencing protocol (Illumina; California, EEUU) was used to load the library in the MiSeq platform for sequencing, following the commercial specifications.

3.5 Setting of the multiplex primer fragments in-house method

3.5.1 gDNA FR regions

Library preparation method C (3.4.3) was tested with different proportions of FR primer sets in 6 CLL samples and positive polyclonal (100µg/ml) and clonal (200µg/ml) controls samples (Vitro; Master Diagnostica, Spain), and sequenced in the Illumina MiSeq platform. Proportions of FR1:FR2:FR3 primers used: 1:3:6 in MIX1, 1:4:8 in MIX2 and 1:8:12 in MIX3. Efficiency was evaluated taking into account the number of samples correctly characterized and the percentage of reads mapped against *IGHV* alleles.

3.5.2 cDNA Leader + FR regions

Apart from the primer mixes employing FR region primers sets, another mix of oligonucleotides was prepared adding Leader primers set to the mix described in 3.5.1. Leader primers ensure complete coverage of the IGH locus, but intron 1 present downstream the Leader region does not allow to sequence the rearranged complex from gDNA using short 150 reads as it does not reach *IGHV* exon 2 (Figure 3.2). For that reason, we included the set of Leader primers from all *IGHV* families for the cDNA approach. 6 CLL samples were amplified with the primer mix MIXLFA (Leader-Framework A), prepared as described in 3.3 (proportions 1:8:12:16). Polyclonal and clonal controls were included (Vitro; Master Diagnostica; Spain). Library preparation and sequencing were performed as described for library approach C (3.4.3). Accuracy of rearrangement determinations was assessed after bioinformatics analysis.

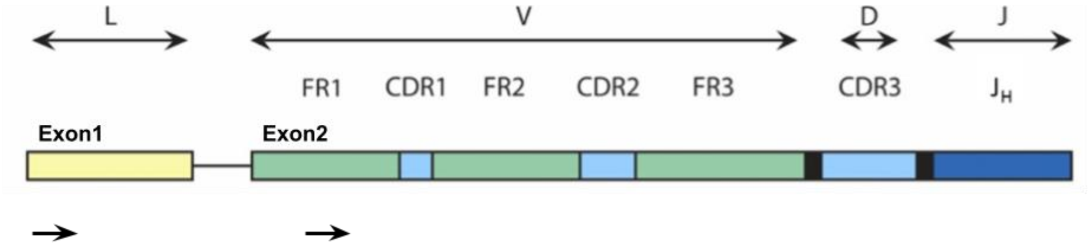


***Figure 3.2. Schematic structure of IGH locus**. Arrows represent Leader and FR1 primers location. Adapted from https://catalog.invivoscribe.com/product/igh-somatic-hypermutation-assay-megakit-v2-0-gel-detection/.*

3.6 Bioinformatic analysis

3.6.1 In-house pipeline for the characterization of CLL BcR clones: BMyRepCLL

The steps in this workflow can be divided into three different categories: conventional, determinant and specific. Conventional steps are processes common to variant calling NGS pipelines (they are required whenever working with DNA-seq data). Determinant steps are those where the process itself is frequently used in NGS pipelines but the parameters have been tightly adapted to BcR heavy chain’s analysis and the library design employed herein. Finally, specific processes are those applied exclusively to BcR clone sequencing, most of them constructed with in-house scripts.

For a comprehensive description of the procedure, analysis steps detailed in this section are illustrated with examples.

3.6.1.1 VDJ region clone characterization

The script *pipeline.py* integrates this block of the analysis, importing modular scripts present in the same repository. The final output is a tabular format CSV file called “homology_table.csv”, with the information of the rearrangements detected at *IGHV* allele level, for all the samples included in the analysis.

The code is encapsulated in different modules and functions that allow the preparation of each step and command construction. For time optimization, each command is written in a CMD file which is afterwards executed by the shell using GNU parallel (152), which manages the execution of processes simultaneously. These commands are used to call different open-source bioinformatics programs and custom scripts and modules. CMD and LOG files are named with the execution date and time for documentation of the experiment.

FASTQ MERGE (conventional)

The script *fastq_unifier.py* was designed for FASTQ files unification in conventional DNA-seq pipelines (original repository NGStools https://github.com/afuentri/NGS-tools.git). The main function of the script is the generation of a Python dictionary storing the features of each FASTQ file, with automatic detection, along with the file names of downstream analysis files. The module *fastq_merge* was adapted to be included within the B-MyRepCLL repository. The script checks if FASTQ files are replicated or unique (concatenates FASTQ files if they are split by lanes or replicated), and then it will merge the required files, checking the extensions to manage both compressed and uncompressed FASTQ files and seeking in descending folders to cope with FASTQs from different samples named identically. Both single and paired FASTQ files are supported and detected automatically.

A CSV format file with the concatenations information is created (input FASTQ files and output merged FASTQ files). The process has to be done carefully and in the same order for paired FASTQ files (R1 and R2). A control step is performed after the merging process, ensuring that each FASTQ file has the expected number of reads. FASTQ files which are not in need of concatenation are copied into the output folder named “merged” together with the merged files as they are the input for the next step.

TRIMMING (conventional)

Trimming low quality bases is a determinant preprocessing step in NGS pipelines, and specifically for B cell receptors, the use of high quality data is decisive for subsequent VDJ allele/gene assignment, avoiding sequencing errors.

- seqtk: default option. Trimming reads by quality using Phred score with the default parameters (trimming up to 30 bp from each side following a 0.05 error rate threshold). Applicable for cases where only trimming by quality is needed (e.g. Nextera tagmentation workflow described in 3.4.2). Seqtk version 1.2-r101-dirty.
- bbduk (argument --primers): if this argument is specified the primer sequences given in FASTA format (*IGHV* families forward primers) will be removed from 5’ end and quality trimming below Q30 Phred score on both read ends will take place. Bbduk (v38.26), is a program included in the bbtools suite (https://sourceforge.net/projects/bbmap/) and finds and removes (or masks) all the coincidences of the sequences provided in FASTA format using k-mer search. The maximum number of nucleotides trimmed at the ends is set to 30 to avoid trimming primer matches in the middle of the reads. Reads below 50 bp are removed. By default, only left primers will be trimmed with this option whereas adding argument –*bothsides*, primer sequences will be trimmed from both read-ends.

For experiments performed with in-house MIX3 library preparation method (3.4.3,3.5) the second option with primers trimming was employed.

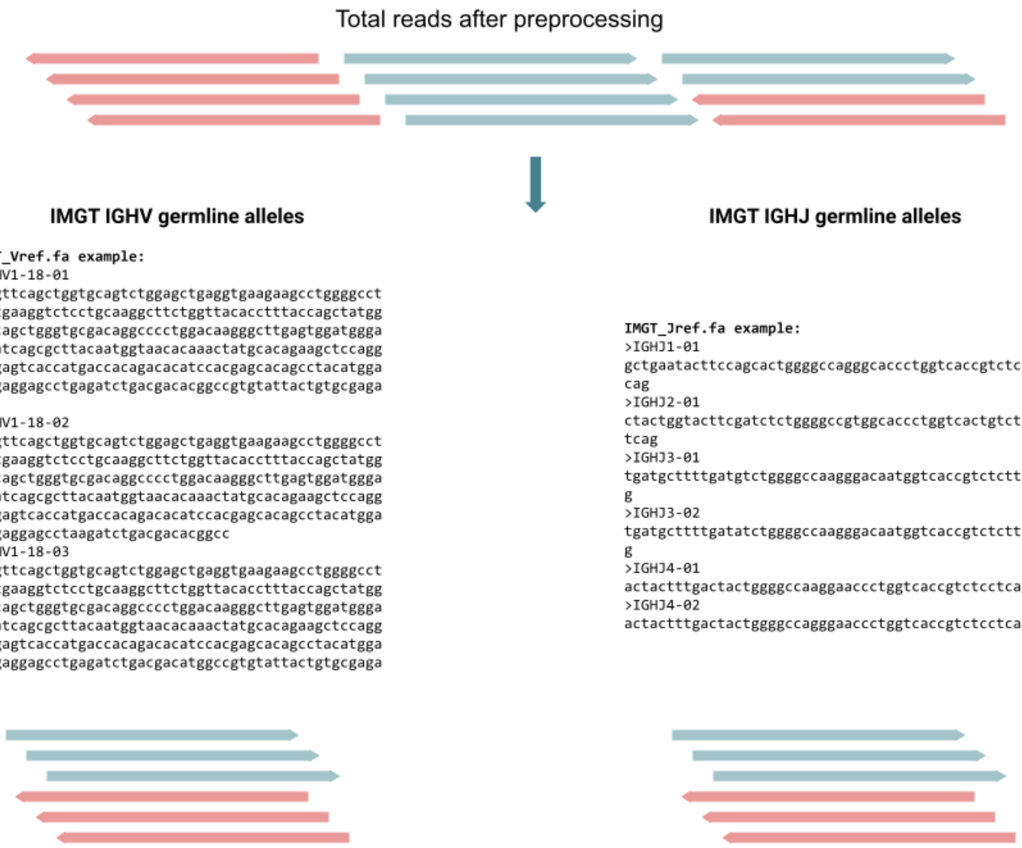
ALIGNMENT AGAINST IMGT (determinant)

Aligning reads against a reference genome assembly is a common step in bioinformatic workflows to map each read to its genomic coordinates. However, the rearranged VDJ region is not mappable against the species whole reference genome due to the recombination process that takes place in the B cell receptor. Instead, most bioinformatic workflows for TcR and BcR characterization use the IMGT database (31). The reference files used suffer some modifications after they are downloaded from the IMGT database. Gaps are removed and the allele number separator "*" is replaced by the character "-". Two different reference multi-FASTA files are used, against whom reads are mapped simultaneously (Figure 3.3):

- IMGT V alleles
- IMGT J alleles
- IMGT D alleles are not used for mapping as the assignment is complicated by mapping with partial reads. The region is too short and contains the junction variability.

Each FASTQ or pair of FASTQs is mapped against the two references FASTA files separately using BWA *mem* 0.7.15-r1140 (153). BWA is an efficient mapper based on the Burrows Wheeler Transform, widely used for mapping reads against the human genome. The algorithm *mem*, is fast and accurate for Illumina and other technologies, as it sets an alignment seed (subset of a read) with maximum exact matches (MEM), and then extends the seed with the Smith-Waterman algorithm. BAM files are afterwards sorted and indexed with samtools v1.7 (154,155) (samtools *sort* and samtools *index*, respectively).

After mapping, soft clipped reads (term used for reads that map partially with the reference) mapped before coordinate 200 are filtered out from *IGHV* BAM files.



Total reads after preprocessing
IMGT IGHV germline alleles
IMGT IGHJ germline alleles

```
IMGT_Vref.fa example:
>IGHV1-18-01
caggttcagctggtgcagtctggagctgaggtgaagaagcctggggcct
cagtgaaggtctcctgcaaggcttctggttacacctttaccagctatgg
tatcagctgggtgcgacaggcccctggacaagggcttgagtggatggga
tggatcagcgcttacaatggtaacacaaactatgcacagaagctccagg
gcagagtcaccatgaccacagacacatccacgagcacagcctacatgga
gctgaggagcctgagatctgacgacacggccgtgtattactgtgcgaga
ga
>IGHV1-18-02
caggttcagctggtgcagtctggagctgaggtgaagaagcctggggcct
cagtgaaggtctcctgcaaggcttctggttacacctttaccagctatgg
tatcagctgggtgcgacaggcccctggacaagggcttgagtggatggga
tggatcagcgcttacaatggtaacacaaactatgcacagaagctccagg
gcagagtcaccatgaccacagacacatccacgagcacagcctacatgga
gctgaggagcctaagatctgacgacacggcc
>IGHV1-18-03
caggttcagctggtgcagtctggagctgaggtgaagaagcctggggcct
cagtgaaggtctcctgcaaggcttctggttacacctttaccagctatgg
tatcagctgggtgcgacaggcccctggacaagggcttgagtggatggga
tggatcagcgcttacaatggtaacacaaactatgcacagaagctccagg
gcagagtcaccatgaccacagacacatccacgagcacagcctacatgga
gctgaggagcctgagatctgacgacatggccgtgtattactgtgcgaga
ga
```

```
IMGT_Jref.fa example:
>IGHJ1-01
gctgaatacttccagcactggggccagggcaccctggtcaccgtctcct
cag
>IGHJ2-01
ctactggtacttcgatctctggggccgtggcaccctggtcactgtctcc
tcag
>IGHJ3-01
tgatgcttttgatgtctggggccaagggacaatggtcaccgtctcttca
g
>IGHJ3-02
tgatgcttttgatatctggggccaagggacaatggtcaccgtctcttca
g
>IGHJ4-01
actactttgactactggggccaaggaaccctggtcaccgtctcctcag
>IGHJ4-02
actactttgactactggggccagggaaccctggtcaccgtctcctcag
```

Figure 3.3. Mapping against IGHV and IGHJ IMGT alleles*. FASTQ files are mapped simultaneously against IGHV and IGHJ allele references. Reads can either map against IGHV alleles, IGHJ alleles, or both.*

IMGT ALIGNMENT STATISTICS (conventional)

On each sample, the samtools command *idxstats*, reports the number of reads mapped against each entry in the reference (IMGT *IGHV* or *IGHJ* alleles in this case). *IGHV* and *IGHJ* BAM files are taken as input.

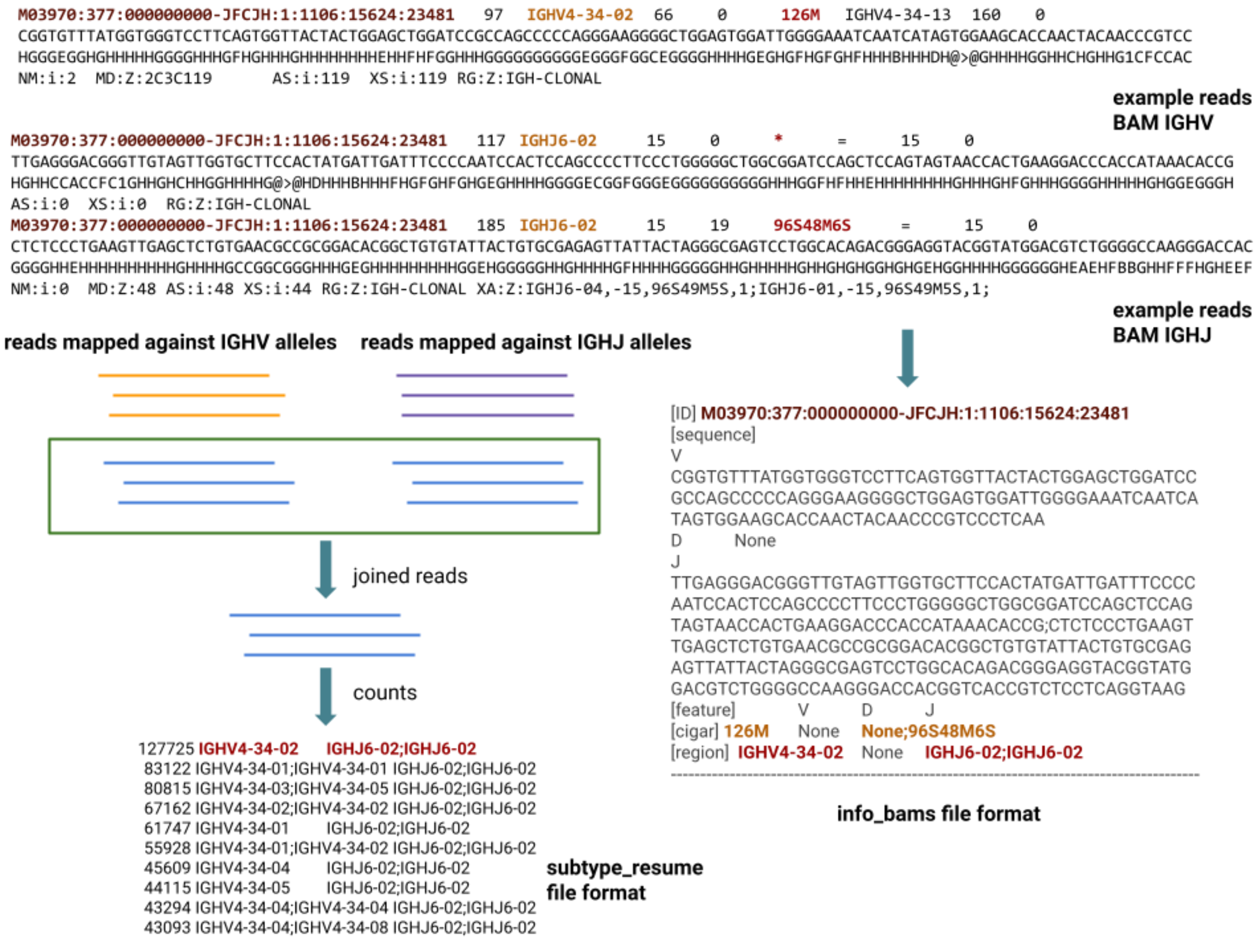
PROBABLE REGIONS EXTRACTION (specific)

The script *probable_regions.py* parses the alignment stats files previously generated with samtools *idxstats*, annotating entries for *IGHV* and *IGHJ* alleles with at least 1 read mapped in probable_vregions.csv and probable_jregions.csv tables, respectively.

Whereas the probable_vregions.csv and probable_jregions.csv tables provide information of the reads mapped against each of the IMGT alleles provided in the reference allele database, the major 10 are stored in the resume_vregions.csv and resume_jregions.csv tables.

IGH BAM PARSING (specific)

Until this step, reads are independently assigned to *IGHV* and/or *IGHJ* alleles. Since the library design does not cover the entire VDJ region at once, overlapping V-J reads are needed to map information of VJ correspondence. A custom script called *IGHBamsParser.py* is used to parse *IGHV* and *IGHJ* BAM files using the *pysam* Python module. A file of files (FOF) listing *IGHV* BAMs is passed as input to the script and for each sample, reads in *IGHV* BAM files are searched for their corresponding reads in the *IGHJ* BAM files by sequence ID. pysam.AlignmentFile() function reads the binary format BAM file, being the fields accessible as attributes. The fields sequence, cigar and reference name are gathered in a new file format (results/bamparse/bamparsing_out/info_bams_SAMPLE.txt) (info_bams format in Figure 3.4), and its summary (results/bamparse/subtypes_resume/ subtypes_resume_SAMPLE.txt) (subtype_resume format in Figure 3.4) with counts of the appearances of IGHV-IGHJ alleles combined, sorted from most abundant to less abundant. In Figure 3.4, the 10 first coincidences correspond to allele *IGHJ6*02* paired with *IGHV4-34* gene in many of its allele variants. The most dominant appearance matches the most represented *IGHV* allele (*IGHV4-34*02*).



```
M03970:377:000000000-JFCJH:1:1106:15624:23481    97   IGHV4-34-02   66     0       126M   IGHV4-34-13   160    0
CGGTGTTTATGGTGGGTCCTTCAGTGGTTACTACTGGAGCTGGATCCGCCAGCCCCCAGGGAAGGGGCTGGAGTGGATTGGGGAAATCAATCATAGTGGAAGCACCAACTACAACCCGTCC
HGGGEGGHGHHHHHGGGGHHHGFHGHHHGHHHHHHHHEHHFHFGGHHHGGGGGGGGGGEGGGFGGCEGGGGHHHHGEGHGFHGFGHFHHHBHHHDH@>@GHHHHGGHHCHGHHG1CFCCAC
NM:i:2  MD:Z:2C3C119       AS:i:119  XS:i:119 RG:Z:IGH-CLONAL
```

```
M03970:377:000000000-JFCJH:1:1106:15624:23481   117  IGHJ6-02      15     0       *       =      15     0
TTGAGGGACGGGTTGTAGTTGGTGCTTCCACTATGATTGATTTCCCCAATCCACTCCAGCCCCTTCCCTGGGGGCTGGCGGATCCAGCTCCAGTAGTAACCACTGAAGGACCCACCATAAACACCG
HGHHCCACCFC1GHHGHCHHGGHHHHG@>@HDHHHBHHHFHGFGHFGHGEGHHHHGGGGECGGFGGGEGGGGGGGGGHHHGGFHFHHEHHHHHHHHGHHHGHFGHHHGGGGHHHHHGHGGEGGGH
AS:i:0  XS:i:0  RG:Z:IGH-CLONAL
M03970:377:000000000-JFCJH:1:1106:15624:23481   185  IGHJ6-02      15     19      96S48M6S      =      15     0
CTCTCCCTGAAGTTGAGCTCTGTGAACGCCGCGGACACGGCTGTGTATTACTGTGCGAGAGTTATTACTAGGGCGAGTCCTGGCACAGACGGGAGGTACGGTATGGACGTCTGGGGCCAAGGGACCAC
GGGGHHEHHHHHHHHHHGHHHHGCCGGCGGGHHHGEGHHHHHHHHHGGEHGGGGHHGHHHHGFHHHHGGGGHHGHHHHHGHHGHGHGGHGHGEHGGHHHHGGGGGGHEAEHFBBGHHFFFHGHEEF
NM:i:0  MD:Z:48 AS:i:48 XS:i:44 RG:Z:IGH-CLONAL XA:Z:IGHJ6-04,-15,96S49M5S,1;IGHJ6-01,-15,96S49M5S,1;
```

```
[ID] M03970:377:000000000-JFCJH:1:1106:15624:23481
[sequence]
V
CGGTGTTTATGGTGGGTCCTTCAGTGGTTACTACTGGAGCTGGATCC
GCCAGCCCCCAGGGAAGGGGCTGGAGTGGATTGGGGAAATCAATCA
TAGTGGAAGCACCAACTACAACCCGTCCCTCAA
D        None
J
TTGAGGGACGGGTTGTAGTTGGTGCTTCCACTATGATTGATTTCCCC
AATCCACTCCAGCCCCTTCCCTGGGGGCTGGCGGATCCAGCTCCAG
TAGTAACCACTGAAGGACCCACCATAAACACCG;CTCTCCCTGAAGT
TGAGCTCTGTGAACGCCGCGGACACGGCTGTGTATTACTGTGCGAG
AGTTATTACTAGGGCGAGTCCTGGCACAGACGGGAGGTACGGTATG
GACGTCTGGGGCCAAGGGACCACGGTCACCGTCTCCTCAGGTAAG
[feature]        V        D        J
[cigar] 126M     None     None;96S48M6S
[region]  IGHV4-34-02     None     IGHJ6-02;IGHJ6-02
--------------------------------------------------------------------------------
```

```
127725 IGHV4-34-02       IGHJ6-02;IGHJ6-02
 83122 IGHV4-34-01;IGHV4-34-01 IGHJ6-02;IGHJ6-02
 80815 IGHV4-34-03;IGHV4-34-05 IGHJ6-02;IGHJ6-02
 67162 IGHV4-34-02;IGHV4-34-02 IGHJ6-02;IGHJ6-02
 61747 IGHV4-34-01       IGHJ6-02;IGHJ6-02
 55928 IGHV4-34-01;IGHV4-34-02 IGHJ6-02;IGHJ6-02
 45609 IGHV4-34-04       IGHJ6-02;IGHJ6-02
 44115 IGHV4-34-05       IGHJ6-02;IGHJ6-02
 43294 IGHV4-34-04;IGHV4-34-04 IGHJ6-02;IGHJ6-02
 43093 IGHV4-34-04;IGHV4-34-08 IGHJ6-02;IGHJ6-02
```

***Figure 3.4. Process of obtaining IGHV-IGHJ reads correspondence**. Reads mapping IGHV alleles are searched in IGHJ BAM files and read ID, IGHV germline allele, IGHJ germline allele, cigar and sequence are annotated in info_bams*.txt file format. Subsequently, the information is summarized into counts of paired IGHV-IGHJ mappings, in the subtype_resume file format. The first lines of the subtype_resume file in the example indicates that there is overlap supported by thousands of reads between the allele IGHJ6*02 and different alleles of IGHV4-34 gene, being the allele variant *02, the most abundant.*

IGHV REGION BASAL FILTER (specific)

This step is optional since by default, no filters are applied regarding clonal percentage in this module. If option *–basal* is included, *IGHV* alleles represented below 3.6% will be discarded. Alternatively, a polyclonal (healthy) *repertoire* sample can be used to perform this filter along the sequencing experiment to calculate a specific threshold. Using table probable_vregions.csv, the function looks for a sample containing the word 'polyclonal' in its name. If it exists, the filter will be applied calculating the proportion of the most represented *IGHV* gene in that sample and will define the basal filter. If several polyclonal samples are found, the greatest proportion among them will be taken as filter. In the same way, a filter is applied for *IGHJ* regions, but always using 3.6% threshold.

After applying this filter, *IGHV* regions within a sample with a proportion above the filter will be written in the filtered table (homology_table-filters.csv). Output table without filters (homology_table.csv) will always be included in the results folder. Additionally, pie plots with the represented *IGHV* genes per sample are created. If the basal filter is performed, pie plots will be generated for both unfiltered and filtered data.

IGHJ ANNOTATION (specific)

With the subtype_resume files from BAM parsing step, *IGHJ* region annotation is performed (output column J_assigned in homology_table.csv) for each *IGHV* allele in the list, using the information of the most represented *IGHJ* alleles in that sample and joined information (number of reads overlapping both *IGHV* and *IGHJ*). Combinations of *IGHV* and *IGHJ* alleles in a proportion less than 0.8% of *IGHV* allele reads are not taken into account, and the combination with the greatest count is given as *IGHJ* assignment. If there is no joined information available or it does not pass the filters mentioned, the major *IGHJ* allele is assigned. If the greatest Jregion and Jassigned match, another column called J coincidence is set to 'yes'. Example of J_region annotation decision algorithm in Figure 3.5.

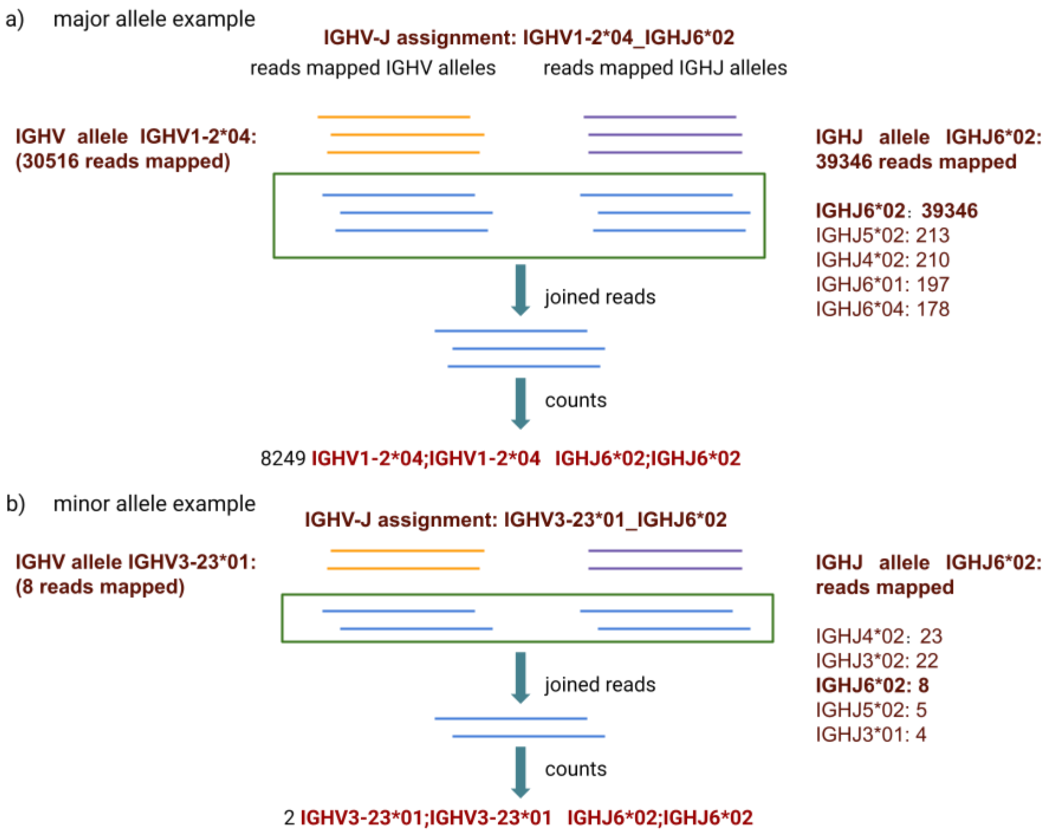

a) major allele example
IGHV-J assignment: IGHV1-2*04_IGHJ6*02
reads mapped IGHV alleles
reads mapped IGHJ alleles
IGHV allele IGHV1-2*04: (30516 reads mapped)
IGHJ allele IGHJ6*02: 39346 reads mapped
IGHJ6*02: 39346
IGHJ5*02: 213
IGHJ4*02: 210
IGHJ6*01: 197
IGHJ6*04: 178
joined reads
counts
8249 IGHV1-2*04;IGHV1-2*04 IGHJ6*02;IGHJ6*02
b) minor allele example
IGHV-J assignment: IGHV3-23*01_IGHJ6*02
IGHV allele IGHV3-23*01: (8 reads mapped)
IGHJ allele IGHJ6*02: reads mapped
IGHJ4*02: 23
IGHJ3*02: 22
IGHJ6*02: 8
IGHJ5*02: 5
IGHJ3*01: 4
joined reads
counts
2 IGHV3-23*01;IGHV3-23*01 IGHJ6*02;IGHJ6*02

Figure 3.5. IGHJ annotation. *a) Case of a major rearrangement: allele IGHV1-2*04 is the major IGHV allele in a sample (30516 reads mapped). The maximum number of reads joined information is found in the correspondence of this allele with IGHJ6*02, which is also the major IGHJ allele (39346 reads mapped). Since IGHV-IGHJ joined information is supported by 8249 reads and that value surpasses 0.8% of reads mapped against IGHV1-2*04 (30516x0.008=244), it is added to IGHV-IGHJ alleles rearrangement assignment. b) Case of an IGHV allele supported by 8 reads (minor). The IGHJ allele found with the highest number of supporting reads after joined information of allele IGHV3-23*01 is IGHJ6*02. In this case, IGHJ6*02 is not the major IGHJ allele, but since it complies the joined information filter, the assignment can be fulfilled with the information of overlapping IGHV-IGHJ reads. This allows not only the correct assignment of the predominant rearrangements but also those represented by a smaller fraction of reads.*

REARRANGEMENT-SPECIFIC READ MAPPING (specific)

In prior steps, Ig rearrangements were defined by pairs of *IGHV-IGHJ* alleles. For the purpose of characterizing these rearrangements as a whole, in the current block composed of two steps, reads are isolated and mapped against a whole-rearrangement sequence. The reference sequence is created after all combinations of *IGHV* and *IGHJ* alleles (IMGT-IGHV-J.fa FASTA format file). This step is decisive for the extraction of *IGHD* sequence in posterior steps (Figure 3.6).

- Subset and merging of *IGHV-IGHJ* reads

Reads corresponding to each IGHV-IGHJ combination are extracted from the separate *IGHV* and *IGHJ* BAM files, respectively, and then merged into a single BAM file. In order to map these reads against the simulated rearrangement reference, the BAM file is converted to FASTQ format. Duplicated reads in the FASTQ file are filtered before mapping.

- Mapping against combined IGHV-IGHJ alleles reference

FASTQ files are aligned against a multiFASTA file with the combination of *IGHV* and *IGHJ* alleles, which is constructed at this point. Mapping is performed using software BWA *mem* with the parameter -L set to 50. This value corresponds to clipping penalty, and it was tuned (increased) to avoid definition of reads mapping partially against the reference as clipped, in order to take into account *IGHD* segment when performing variant calling, as it does not map the simulated reference.

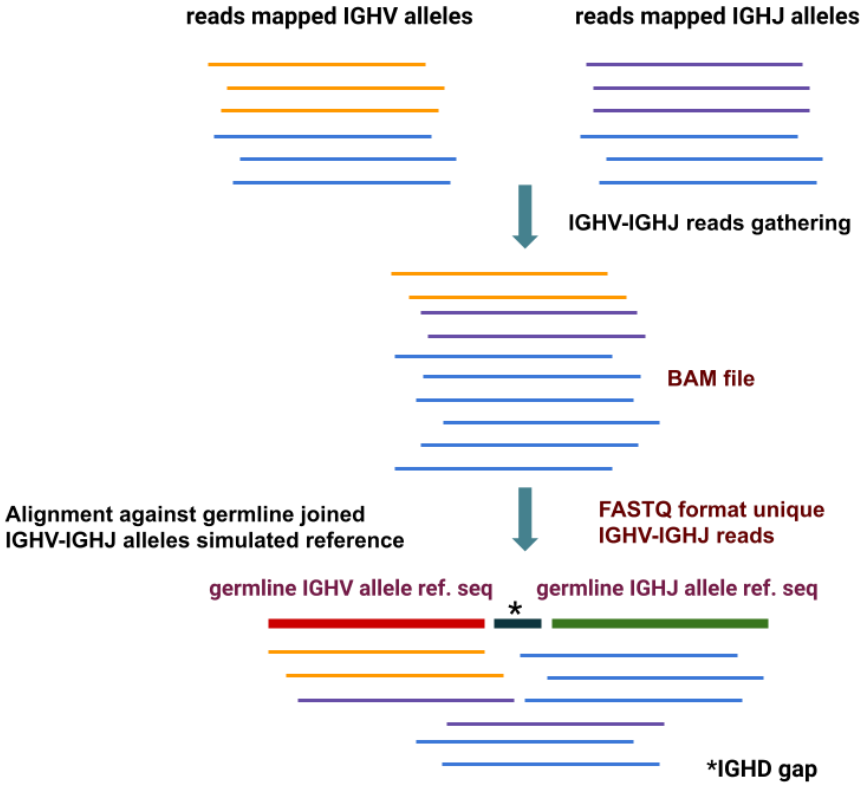

reads mapped IGHV alleles
reads mapped IGHJ alleles
IGHV-IGHJ reads gathering
BAM file
Alignment against germline joined
IGHV-IGHJ alleles simulated reference
FASTQ format unique
IGHV-IGHJ reads
germline IGHV allele ref. seq
*
germline IGHJ allele ref. seq
*IGHD gap

***Figure 3.6. Specific rearrangement mapping**. All reads belonging to a specific IGHV-IGHJ rearrangement are reunited in a BAM file, converted to FASTQ format with removed duplicated read entries and aligned against the simulated IGHV-IGHJ allele reference. * represents the region where the IGHD segment should be, gapped area that will be detected as an insertion between IGHV and IGHJ genes.*

COMPLETE CONSENSUS SEQUENCE (specific)

A consensus sequence for each rearrangement defined by *IGHV-IGHJ* combinations is created performing a variant calling step with Freebayes v1.1.0 (156). Freebayes is a Bayesian haplotype variant caller, which means that equal variant sequences can yield different alignments against a reference sequence and these variant calling algorithms are aware of that fact. Moreover, it simplifies the variant calling process by integrating and avoiding steps required for the GATK best practices reference variant calling (157) (indel realignment, base quality score recalibration, etc), obtaining high fidelity results in a single step. After obtaining a VCF file (Variant Calling Format), bcftools *consensus* is employed to obtain the consensus sequence (Figure 3.7). A minimum sequencing depth of 50 reads, frequency 0.5 and minimum 2 alternative reads are used to report variants.

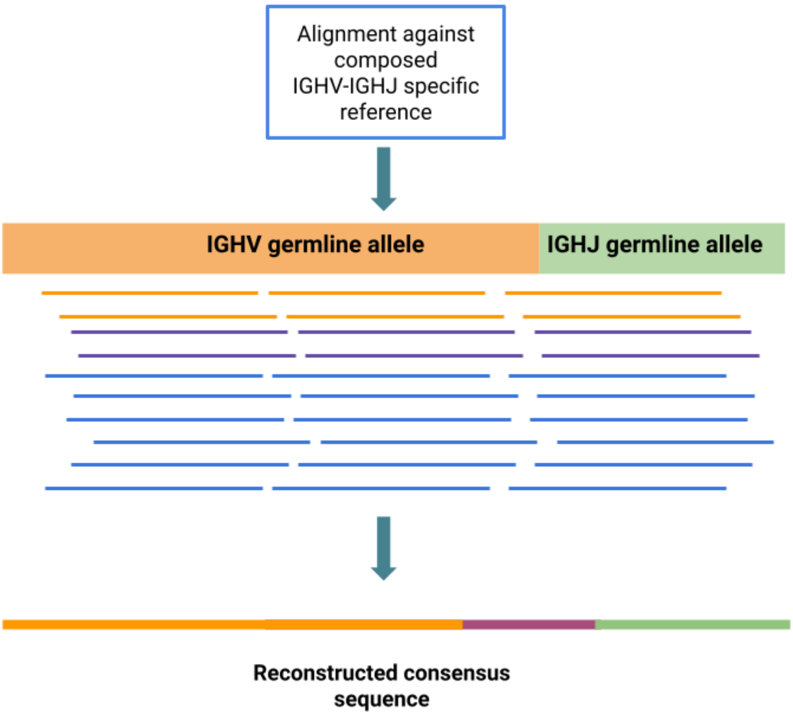

Alignment against composed IGHV-IGHJ specific reference
IGHV germline allele
IGHJ germline allele
Reconstructed consensus sequence

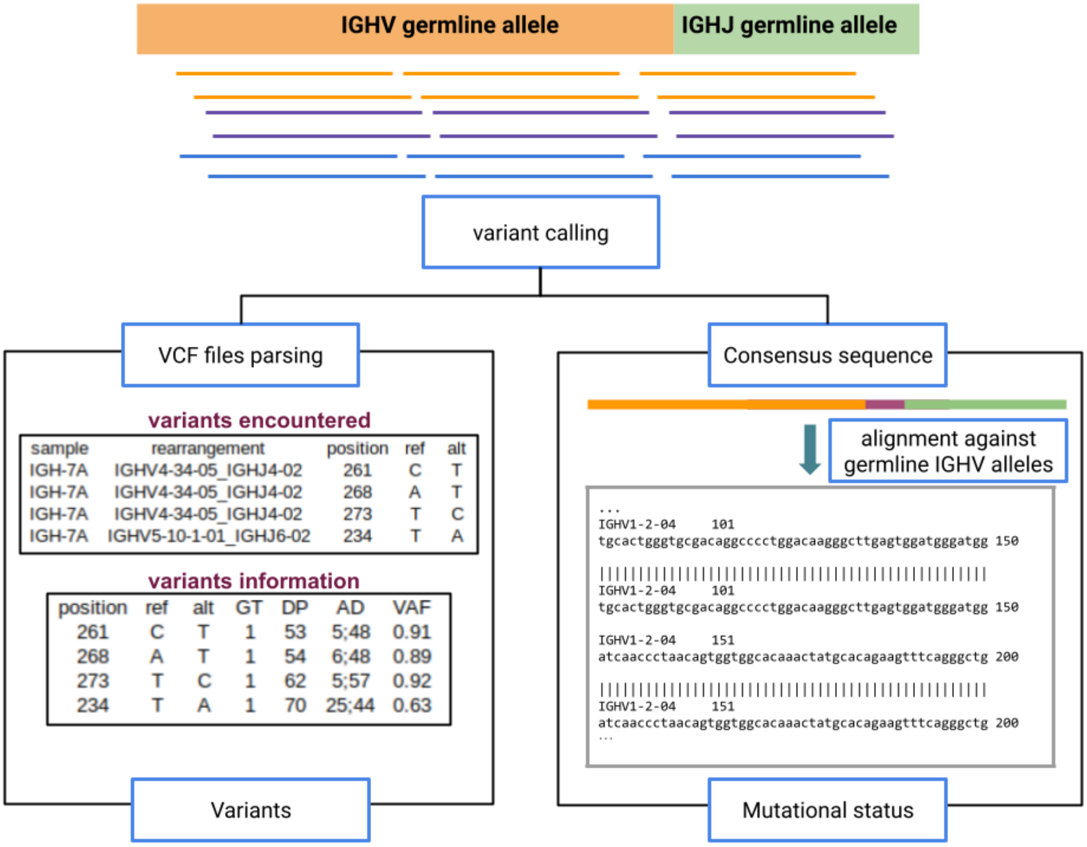
Figure 3.7. Consensus sequence extraction. *A consensus sequence is extracted per rearrangement found, being characterized by an IGHV and IGHJ allele. The consensus sequence obtained with the parameters used in the variant calling step intends to represent the variants present in the clonal Ig rearrangement, reflecting the level of SHM in the clone, and to reconstruct IGHD and CDR3 sequence (purple region in the sequence), that were unknown until this point and therefore, not included in the recombined reference sequence.*

MUTATIONAL STATUS CALCULATION (specific)

Variations in the IGH locus are inspected at two levels: conventional variant calling method, and the assessment of the mutational status. For the last purpose we decided to align the consensus sequence with traditional local alignment methods (Figure 3.8).

Consensus sequences are aligned against the closest IMGT *IGHV* alleles with command line EMBOSS *water* v6.6.0 (pairwise nucleotide local alignment) (158). Alignment output is parsed afterwards to obtain the identity percentage and alignment length. If the identity percentage is below 85, results in mutational status will be tagged as ‘not valid’. Otherwise, mutational status is annotated for UM (unmutated; % ≥ 98) and MM (mutated; % < 98). Number of mismatches is corrected when gaps are present in the alignment to consider only one variation per gap.

From the previous step consensus sequence generation procedure, VCF files are kept and are parsed to obtain CSV tabular format files containing all variants called per sample and *IGHV* rearrangement (Figure 3.8).



variants encountered

sample	rearrangement	position	ref	alt
IGH-7A	IGHV4-34-05_IGHJ4-02	261	C	T
IGH-7A	IGHV4-34-05_IGHJ4-02	268	A	T
IGH-7A	IGHV4-34-05_IGHJ4-02	273	T	C
IGH-7A	IGHV5-10-1-01_IGHJ6-02	234	T	A

variants information

position	ref	alt	GT	DP	AD	VAF
261	C	T	1	53	5;48	0.91
268	A	T	1	54	6;48	0.89
273	T	C	1	62	5;57	0.92
234	T	A	1	70	25;44	0.63

```
...
IGHV1-2-04     101
tgcactgggtgcgacaggcccctggacaagggcttgagtggatgggatgg 150

||||||||||||||||||||||||||||||||||||||||||||||||||
IGHV1-2-04     101
tgcactgggtgcgacaggcccctggacaagggcttgagtggatgggatgg 150

IGHV1-2-04     151
atcaaccctaacagtggtggcacaaactatgcacagaagtttcagggctg 200

||||||||||||||||||||||||||||||||||||||||||||||||||
IGHV1-2-04     151
atcaaccctaacagtggtggcacaaactatgcacagaagtttcagggctg 200
...
```

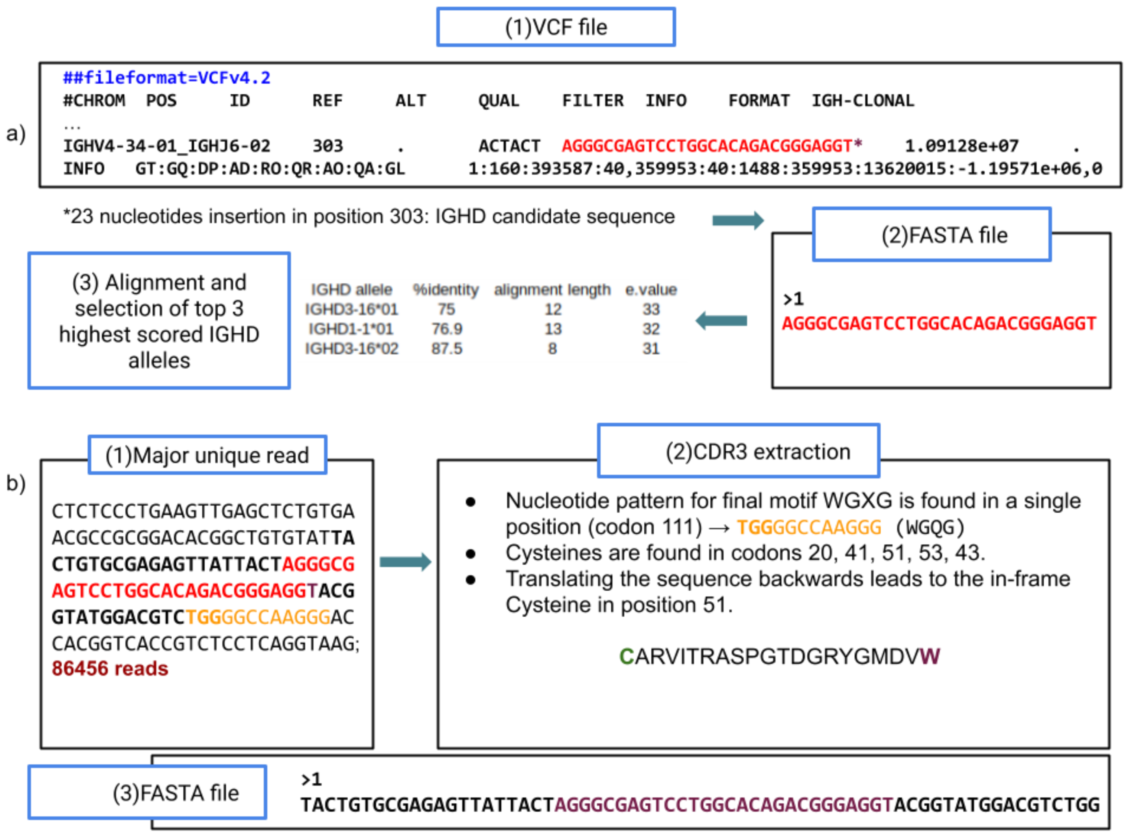
Figure 3.8. Variant calling and alignment against germline IMGT IGHV alleles*. The variant calling steps is used to parse VCF files, keeping the variants found within patient in tabular CSV format (left square), and to generate a consensus sequence per rearrangement (defined until this stage by pairs of IGHV-IGHJ alleles, that will be used to determine the percentage of identity against the closest germline IGHV allele (right square).*

PRODUCTIVITY AND CDR3 (specific)

In previous versions of the pipeline, to infer CDR3 sequence and *IGHD* alleles, the consensus sequences were translated into amino acids sequences and checked in the 6 reading frames. The current version extracts CDR3 from the major productive junction sequence in the BAM specific-rearrangement files. A Python module called *consensus2CDR3*, seek firstly all possible nucleotide patterns for the final Tryptophan amino acid motif WGXG (Trp-Gly-any-Gly). From those patterns, the sequence is translated backwards, codon by codon, until a Cysteine which is in-frame with the final motif is found. Since the final motif is less frequent than a single Cys, there are less positions to begin the search with. If there are various possibilities, a decision is made based on productivity and length of the sequences. The process is repeated starting from the most abundant unique sequence to the least, until a productive CDR3 is found (with a minimum number of reads threshold), and ensuring the sequence is paired with the *IGHV* allele determined for that rearrangement (Figure 3.9a). Otherwise, CDR3 sequence is set as “None”. This decision algorithm was implemented with the help of A. Serrano (Unit of Hematology HCUV). With argument *--cdr3simp*, rearrangements represented in less than 100 reads are not annotated CDR3 sequence to obtain faster results (e.g., projects with many samples). *IGHD* sequence is extracted parsing the VCF files and looking for an insertion with length greater than 6 nucleotides.

CDR3 ALIGNMENT (IGHD DETERMINATION) (specific)

If an insertion longer than 6 nucleotides is found in the VCF files, FASTA files containing *IGHD* region sequences, are aligned against *IGHD* IMGT alleles using EMBOSS *water*. Output alignment files are parsed and the 3 alleles with the highest alignment scores (e-value score) are annotated in *IGHD* calls column in the final table. In the example with the clonal control sample in Figure 3.9b, *IGHD* sequence can also be found embedded in the junction sequence found using the major unique sequence. However, the delimitation of this sequence is more exact with the insertion method as it is anchored by the *IGHV* and *IGHJ* sequences.


(1)VCF file
a)
*23 nucleotides insertion in position 303: IGHD candidate sequence
(2)FASTA file
(3) Alignment and selection of top 3 highest scored IGHD alleles
b)
(1)Major unique read
(2)CDR3 extraction
(3)FASTA file

```
##fileformat=VCFv4.2
#CHROM  POS      ID       REF      ALT      QUAL     FILTER  INFO     FORMAT  IGH-CLONAL
...
IGHV4-34-01_IGHJ6-02     303      .       ACTACT   AGGGCGAGTCCTGGCACAGACGGGAGGT*     1.09128e+07     .
INFO    GT:GQ:DP:AD:RO:QR:AO:QA:GL       1:160:393587:40,359953:40:1488:359953:13620015:-1.19571e+06,0
```

```
>1
AGGGCGAGTCCTGGCACAGACGGGAGGT
```

IGHD allele	%identity	alignment length	e.value
IGHD3-16*01	75	12	33
IGHD1-1*01	76.9	13	32
IGHD3-16*02	87.5	8	31

```
CTCTCCCTGAAGTTGAGCTCTGTGA
ACGCCGCGGACACGGCTGTGTATTA
CTGTGCGAGAGTTATTACTAGGGCG
AGTCCTGGCACAGACGGGAGGTACG
GTATGGACGTCTGGGGCCAAGGGAC
CACGGTCACCGTCTCCTCAGGTAAG;
86456 reads
```

- Nucleotide pattern for final motif WGXG is found in a single position (codon 111) → TGGGGCCAAGGG (WGQG)
- Cysteines are found in codons 20, 41, 51, 53, 43.
- Translating the sequence backwards leads to the in-frame Cysteine in position 51.

CARVITRASPGTDGRYGMDVW

```
>1
TACTGTGCGAGAGTTATTACTAGGGCGAGTCCTGGCACAGACGGGAGGTACGGTATGGACGTCTGG
```

Figure 3.9. Scheme of IGHD and CDR3 extraction. *Clonal sample extraction of a) IGHD: Insertions over 6 nucleotides are selected on the rearrangement VCF files as IGHD candidate sequences, with are subsequently stored in FASTA files and aligned against IGHD IMGT alleles. The top 3 highest score matches are kept. b) CDR3: The major productive sequence contains a productive CDR3 and it is represented by 86456 reads (1). The region is extracted (region in bold) with the procedure in step (2) and stored in a FASTA file. The internal region in bold highlighted in red corresponds to the IGHD sequence detected in a). Orange nucleotides represent WGXG motif (WGQG in this example).*

3.6.1.2 Artifact filtering and rearrangement prioritization

OUTPUT AND GENE GROUPING

The final rearrangements output table of module 1 is the file homology_table.csv. In this table, rearrangement information is reported at *IGHV* allele level (example in Figure 3.10 part 1, some columns are not shown). A series of steps are used over this table to generate final results:

1) Identity Join: *filter_equalrearrangements.py* performs pairwise sequence alignment (EMBOSS *water*) between the consensus sequences of rearrangements belonging to the same *IGHV* gene but differing in allele. Those pairs of rearrangements whose consensus sequences share ≥95% identity are joined into the predominant Ig rearrangement (number of mapped reads is

added to the majority allele). The information of the alleles that have been combined is kept in a new column called "joined alleles". Example step 1 to step 2 in example Figure 3.10, different rearrangements with alleles from *IGHV* gene *IGHV3-11* have been combined in the same row.

2) Identity Join: *filter_equalrearrangements2.py* performs a consecutive step of *IGHV* simplification as in the aforementioned step, joining in this case rearrangements from the same *IGHV* family and different genes whose sequences share 95% identity. In step 2 to step 3 in Figure 3.10, rearrangement with allele *IGHV3-21*04* has been added to the row from the main clone (*IGHV3-11*01*). *IGHV3-23*01* will be later removed or added to other rearrangement, as it is only supported by FR3 reads.

3) Final table generation. *onlyclonality.py* carries out the following actions:

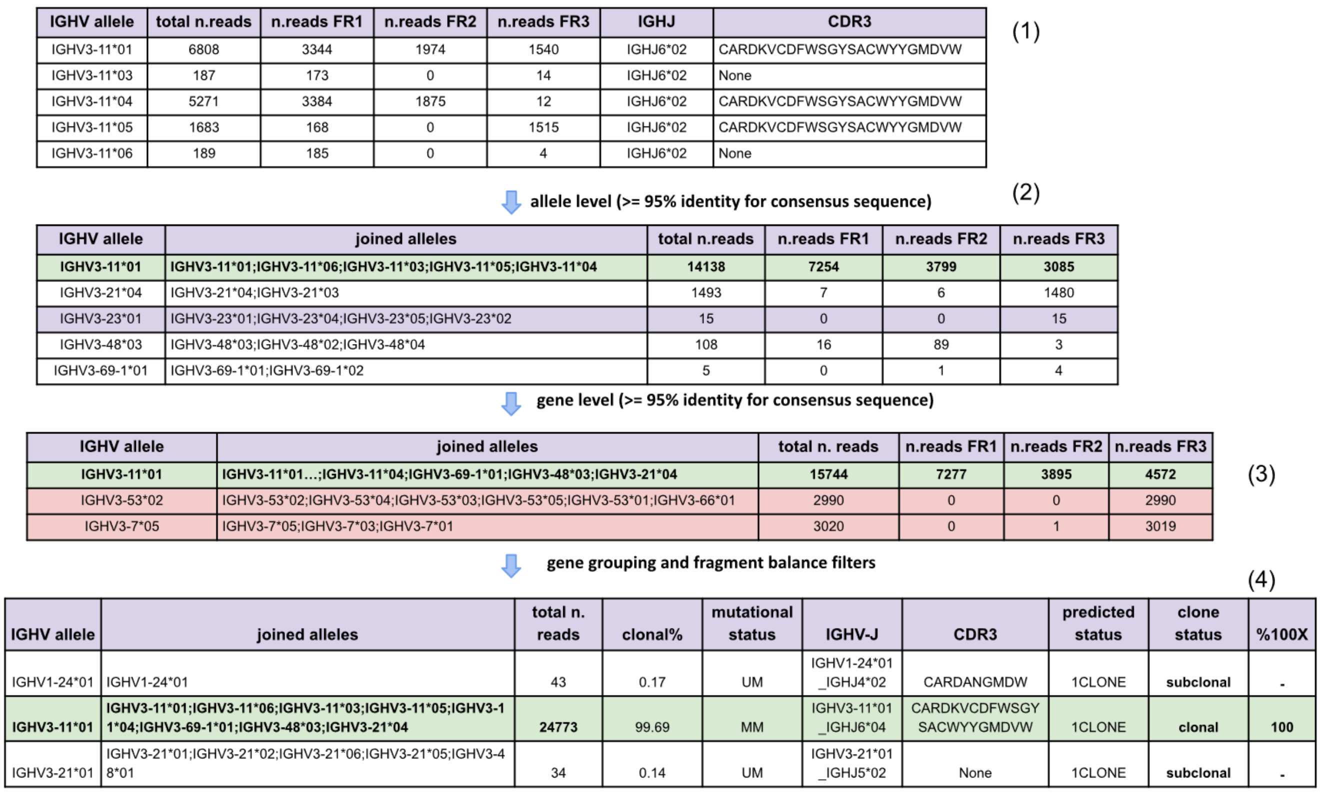
   - Fragment-wise filters: In order to remove artifacts from unbalanced alleles, rearrangements in which any of the fragments supports the totality of mapped reads, are dropped from the final table and read counts are added to the major rearrangement with the same *IGHJ* gene and CDR3, if there is any. CDR3 field needs to have a defined sequence. Likewise, but more strictly for the FR3 amplicon, when read counts account for at least 92% of total mapped reads, the same conditions are applied. There is an exception for this when the rearrangement is the major in that sample (> 80%). In Figure 3.10, in steps 3 to 4, the two rearrangements colored in red have been reallocated for the final results table, as they are only supported by FR3 fragment.
   - Summary of rearrangements at gene level: Choosing the major productive *IGHV* allele per gene. The major rearrangement will be highlighted in the excel table (both XLSX and CSV file are saved).

- Predicted status: Determination of the clonal threshold per sample. Clones with a percentage < 0.1% are filtered out prior to applying this cut-off. The maximum difference read ratio between consecutive clones, termed MAX_DIFF, is used to adjust the cut-off of the B cell clonal fraction per sample, calculated with the formula: **%reads_mapped(N)/%reads_mapped(N+1)**, where N is the current clone and N+1 the consecutive clone in abundance order. MAX_DIFF is the maximum value among these ratios, and it is used for two important predictions. The first is determining whether the sample has a predicted clonal profile, when MAX_DIFF >=5 (otherwise, it will be tagged as polyclonal). The second, if the sample is predicted as clonal, the clone with the MAX_DIFF value in that sample is determined as the cutoff for the clonal counterpart, considering the clones below it as subclonal. Samples are thus tagged as "NCLONE" (being N the number of predicted CLL clones). Clones are tagged with their clone status as well (clonal or subclonal).

- Coverage breadth percentage calculation of the rearrangements tagged as clonal with *coverage_IGHs* Python module. Coverage breadth of the rearrangements determined clonal over a given threshold (e.g., 100X), are annotated in the output table.

- Gene usage plots. *IGHV* alleles/genes represented per sample and for all the samples introduced in the analysis, and stacked bar plots representing *IGHV* and *IGHJ* genes usage.

homology_resume*.xlsx is the output corresponding to this second module. The final table with this information example is in Figure 3.10 steps 3 to 4 (showing only some of the columns). The major and top3 rearrangements per sample files are found in homology_resume*_principal-rearrangement.csv and homology_resume*_top3-rearrangement.csv, respectively.



(1)

IGHV allele	total n.reads	n.reads FR1	n.reads FR2	n.reads FR3	IGHJ	CDR3
IGHV3-11*01	6808	3344	1974	1540	IGHJ6*02	CARDKVCDFWSGYSACWYYGMDVW
IGHV3-11*03	187	173	0	14	IGHJ6*02	None
IGHV3-11*04	5271	3384	1875	12	IGHJ6*02	CARDKVCDFWSGYSACWYYGMDVW
IGHV3-11*05	1683	168	0	1515	IGHJ6*02	CARDKVCDFWSGYSACWYYGMDVW
IGHV3-11*06	189	185	0	4	IGHJ6*02	None

allele level (>= 95% identity for consensus sequence)

(2)

IGHV allele	joined alleles	total n.reads	n.reads FR1	n.reads FR2	n.reads FR3
IGHV3-11*01	IGHV3-11*01;IGHV3-11*06;IGHV3-11*03;IGHV3-11*05;IGHV3-11*04	14138	7254	3799	3085
IGHV3-21*04	IGHV3-21*04;IGHV3-21*03	1493	7	6	1480
IGHV3-23*01	IGHV3-23*01;IGHV3-23*04;IGHV3-23*05;IGHV3-23*02	15	0	0	15
IGHV3-48*03	IGHV3-48*03;IGHV3-48*02;IGHV3-48*04	108	16	89	3
IGHV3-69-1*01	IGHV3-69-1*01;IGHV3-69-1*02	5	0	1	4

gene level (>= 95% identity for consensus sequence)

(3)

IGHV allele	joined alleles	total n. reads	n.reads FR1	n.reads FR2	n.reads FR3
IGHV3-11*01	IGHV3-11*01…;IGHV3-11*04;IGHV3-69-1*01;IGHV3-48*03;IGHV3-21*04	15744	7277	3895	4572
IGHV3-53*02	IGHV3-53*02;IGHV3-53*04;IGHV3-53*03;IGHV3-53*05;IGHV3-53*01;IGHV3-66*01	2990	0	0	2990
IGHV3-7*05	IGHV3-7*05;IGHV3-7*03;IGHV3-7*01	3020	0	1	3019

gene grouping and fragment balance filters

(4)

IGHV allele	joined alleles	total n. reads	clonal%	mutational status	IGHV-J	CDR3	predicted status	clone status	%100X
IGHV1-24*01	IGHV1-24*01	43	0.17	UM	IGHV1-24*01 _IGHJ4*02	CARDANGMDW	1CLONE	subclonal	-
IGHV3-11*01	IGHV3-11*01;IGHV3-11*06;IGHV3-11*03;IGHV3-11*05;IGHV3-11*04;IGHV3-69-1*01;IGHV3-48*03;IGHV3-21*04	24773	99.69	MM	IGHV3-11*01 _IGHJ6*04	CARDKVCDFWSGYSACWYYGMDVW	1CLONE	clonal	100
IGHV3-21*01	IGHV3-21*01;IGHV3-21*02;IGHV3-21*06;IGHV3-21*05;IGHV3-48*01	34	0.14	UM	IGHV3-21*01 _IGHJ5*02	None	1CLONE	subclonal	-

Figure 3.10. Module 2 filtering steps of the in-house pipeline.

3.6.1.3 Pipeline execution.

Bioinformatic analysis of the sequencing experiments and tests described along prior Methods sections were performed with the in-house IGH pipeline (BMyRepCLL) in the corresponding development stages. BCL files were demultiplexed by MiSeq Reporter Software v.2.6 and the FASTQ files generated were used as input for the pipeline.

```
DEVELOPMENT VERSIONS COMMAND:
python pipeline_fragments.py --pipeline -f $fastqs_folder -o $output_folder -v –
p $nproc –basal
```

$nproc = number of processes

The newest versions include steps for minimization of artifacts and summary of results (3.6.1.2).

```
## IGH pipeline
python3.5   B-MyRepCLL/src/pipeline.py   --pipeline   -f   $fastqs_folder   -o
$output_folder -v –p $nproc --basal --primers primers_5-noleader.fa

## QC
QC/main-parallel.sh -p $folder_results -b $folder_primerfiles -t $nproc

## mapping stats
/QC/flagstat.sh -b $folder_results/bamsV

## artifact filtering (allele level)
python3 B-MyRepCLL/src/filter_equalrearrangements.py
$folder_results/results/homology_table.csv
$folder_results/results/consensus_complete

## artifact filtering (gene level)
python3
B-MyRepCLL/src/filter_equalrearrangements2.py
$folder_results/results/homology_tablesimpalleles.csv
$folder_results/results/consensus_complete

## coverage analysis and summary tables
python3 B-MyRepCLL/src/onlyclonality.py
$output_folder/results/homology_tablesimpallelessimpalleles.csv $name
$output_folder $output_folder/QC/fastq_stats.xlsx
$output_folder/bamsV/flagstat/resume.csv $mincov
```

$folder_results = IGH pipeline $output_folder (first step)
$name = project name

The whole set of commands can be launched with the following script:

```
time python3.5 B-MyRepCLL/launch-default.py $projectname $mincov $output_folder
$primers_fasta
```

$mincov = coverage threshold chosen by the user.
$primers_fasta = file in FASTA format containing the primer sequences to trim.

The code of the pipeline is available on GitHub (https://github.com/afuentri/B-MyRepCLL). The instructions for setting the conda environment with the required software and Python modules/packages are detailed in the GitHub repository Readme.

Analyses were performed on a local server (16 Intel ® Xeon ® CPU E5-2650 0 @ 2.00 GHz processors, 190 GB of RAM and 41 TB disk space) using up to 16 CPU threads, administered by M.

Herreros, the head of the IT department at INCLIVA Health Research Institute. Use of computational resources was coordinated using GNU Parallel (152).

On the course of pipeline development, Leader fragment was taken into consideration in the updates to make it compatible with MIX3 and MIXLFA primer amplification mixes. The analysis of cDNA MIXLFA samples described in 3.5.2 was repeated with the updated 2022 pipeline to ensure that compatibility and show the performance in the Results section.

Upon finalization of the IGH pipeline, quality control analyses were performed using the main script (BASH) of QC repository (https://github.com/afuentri/QC), which executes and parses the output files of the program FastQC (v0.11.5, Babraham Bioinformatics) to generate summary tables and plots per sequencing experiment. The quality on the FASTQ files was evaluated before and after preprocessing and quality trimming steps, and the presence of primer sequences was checked by means of the Python module *primer_QC*. Mapping against IMGT *IGHV* alleles was afterwards evaluated with the BASH script *flagstat* which uses the program *flagstat* from samtools v1.7 (155). To use further parallelization with GNU parallel, main-parallel instead of main BASH pipeline was used in cases with high sample loads.

3.6.2 Immcantation pipeline

This pipeline was developed during a predoctoral stay at the Kleinstein Lab (Yale School of Medicine). The Immcantation suite was employed to adapt a workflow specific for B cell neoplasms, especially with the help of Dr. Steven Kleinstein and Susanna Márquez. The workflow is composed by a preprocessing module, annotation of VDJ genes, clonal clustering and downstream analyses.

3.6.2.1 Preprocessing (pRESTO)

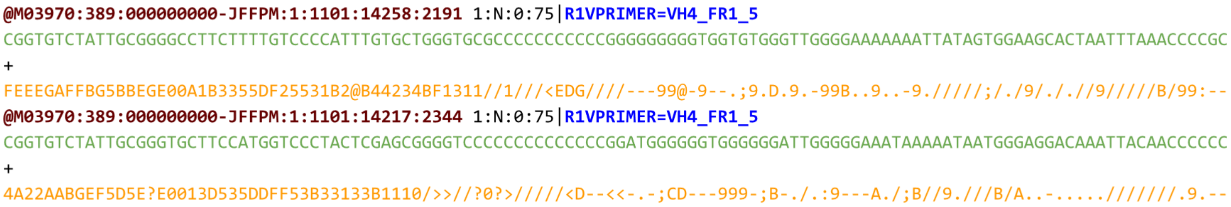
pRESTO is an Immcantation toolkit for preprocessing steps performance in high-throughput Immunoglobulins repertoire data. It covers raw data processing prior to germline gene segments assignment (159).

Unzipped FASTQ files are used as input. The following command example for one sample includes the options employed in the customized script. Specific steps added to the pipeline for our analysis are going to be specified and detailed. The Immcantation team has released their software in different platforms such as docker and singularity, to provide the user with an environment to run their pipelines, and docker is employed in this case for that purpose:

```
docker run -v $path:/data:z immcantation/suite:azahara bash
/data/filter_sepassembly.sh -1 A-1.fastq -2 A-2.fastq -v
primers_R2LPresto.fasta -j primers_R1LPresto.fasta -o IGH-A -p15
```

First of all, low quality reads are filtered out from forward and reverse reads separately using the Immcantation script *FilterSeq.py (read threshold)*. The commands and arguments used inside this script to call the different modules and the following, are detailed in the code files (docker image immcantation/suite:azahara; bash script *filter_sepassembly.sh*). To trim primer sequences from both ends of the reads, Immcantation pipelines use the script *MaskPrimers.py*. We performed a specific combination of steps involving this script to adapt it to our library preparation method (3.4.3, 3.5.1):

- Primers are trimmed on the left side separately from forward and reverse reads. The primer sequence found within the read is annotated in R1 FASTQ header ("r1vprimer") (Figure 3.11). In Figure 3.12, these primers are represented with black crosses in the fragments sequenced (on the left side, *IGHV* primers are trimmed on R1 FASTQ files and *IGHJ* consensus primer is trimmed on R2 FASTQ files).



```
@M03970:389:000000000-JFFPM:1:1101:14258:2191 1:N:0:75|R1VPRIMER=VH4_FR1_5
CGGTGTCTATTGCGGGGCCTTCTTTTGTCCCCATTTGTGCTGGGTGCGCCCCCCCCCCCGGGGGGGGGTGGTGTGGGTTGGGGAAAAAAATTATAGTGGAAGCACTAATTTAAACCCCGC
+
FEEEGAFFBG5BBEGE00A1B3355DF25531B2@B44234BF1311//1///<EDG////---99@-9--.;9.D.9.-99B..9..-9./////;/./9/././/9/////B/99:--
@M03970:389:000000000-JFFPM:1:1101:14217:2344 1:N:0:75|R1VPRIMER=VH4_FR1_5
CGGTGTCTATTGCGGGTGCTTCCATGGTCCCTACTCGAGCGGGGTCCCCCCCCCCCCCCGGATGGGGGGTGGGGGGATTGGGGGAAATAAAAATAATGGGAGGACAAATTACAACCCCCC
+
4A22AABGEF5D5E?E0013D535DDFF53B33133B1110/>>//?0?>/////<D--<<-.-;CD---999-;B-./.:9---A./;B//9.///B/A..-.....///////.9.--
```

Figure 3.11. FASTQ format with pRESTO primers annotation. *Trimmed primers on R1 are added with the tag "R1VPRIMER" (blue). The rest of the element are the standard for the FASTQ format (green: sequence, yellow: qualities per base, red: read ID).*

- *PairSeq.py* is used to transfer *IGHV* primer annotation from R1 to R2 reads FASTQ files, so that the *IGHV* family and the FR fragment that originated the read are known.
- This step and the following are performed with the purpose of trimming primer sequences from the right side of the reads only on FR3 fragment reads. For the rest of fragments, reads do not reach the primer location on the right end. For both forward and reverse reads, FR3 reads are split into separate FASTQ files, leaving FR1 and FR2 reads in other output FASTQ file (script *SplitSeq.py*).
- *MaskPrimers.py* is used once more to perform trimming of primer sequences but this time on the right end of reads, for both R1 and R2 in the FR3 fragment FASTQ files. These are represented with red crosses in Figure 3.12 (*IGHV* primers on R2 and *IGHJ* consensus primer on R1).

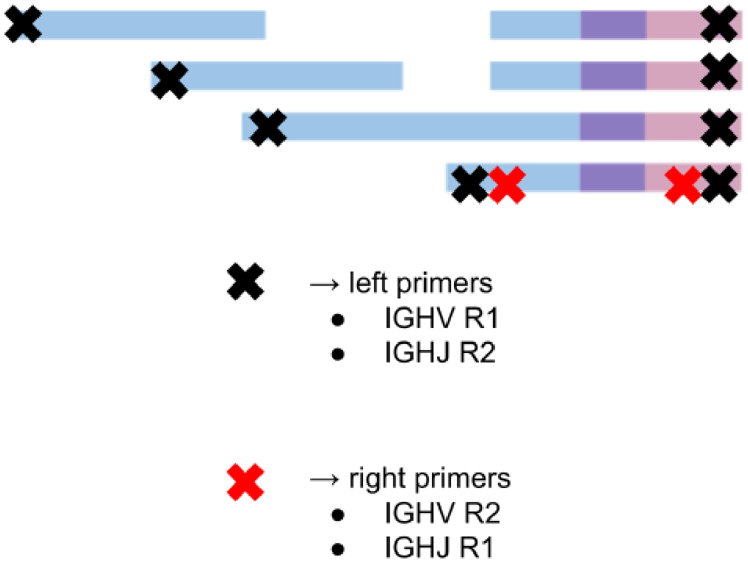


Figure 3.12. Primers trimming scheme. *Right primers are shown with read crosses and left primers with black crosses. Depending on the sequencing read (R1 or R2), IGHV and IGHJ primers are trimmed on right or left read-ends. Leader primers were not included but it is included in the scheme for when cDNA samples are used.*

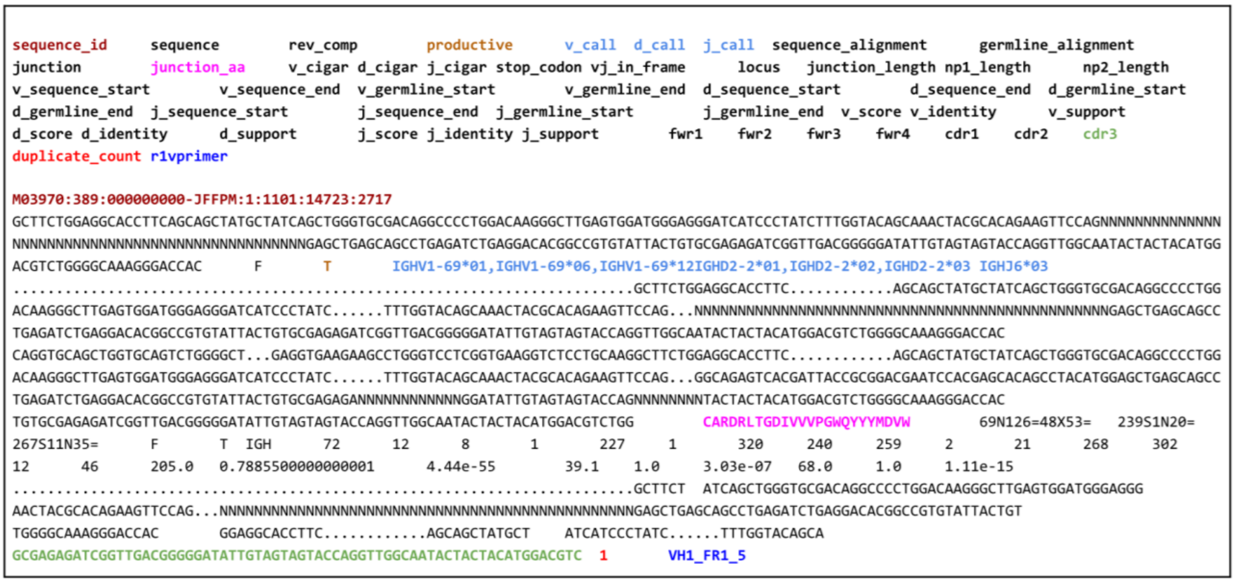
- To assemble paired reads, FASTQ files are split once more into FR fragments combinations (*SplitSeq.py*). Since FR1 reads do not overlap, FR1 and FR2 reads are split and FR2 reads are joined with the previously right-trimmed FR3 reads, leaving FR2FR3 reads FASTQs and FR1 reads FASTQs.

- *PairSeq.py* is used to sort and match sequence records with matching coordinates across R1 and R2 files (performed separately for FR1 and FR2FR3 FASTQ files).
- Paired reads are assembled with two different strategies, distinguishing fragments with and without overlap:
    - *AssemblePairs*.py mode *align* assembles overlapping R1 and R2 reads from FR2FR3 fragments.
    - *AssemblePairs.py* mode *reference* assembles R1 and R2 reads from FR1 fragment using guide reference sequences.
- *CollapseSeq.py* removes sequence redundancy from the final FASTQ files. Finally, log files are parsed and pRESTO reports are generated, for quality parameters. The abundance of each unique read is specified in the column “duplicate_count”.

3.6.2.2 IgBlast annotation (Change-O)

Change-O is a suite of utilities to perform specific analyses focused on B cell repertoires (160). A script was used for the annotation of IGH gene calls and junction sequence assignment using IgBlast (6). Arguments -g, -t, and -f allow to choose the species, BcR/TcR type data, and the output format standards, respectively. Separate FR1 and FR2-FR3 FASTQ files from the pRESTO block are concatenated in a single FASTQ file containing preprocessed and collapsed reads and given as input to the script (*changeo.sh*; bash). The output table per sample, contains each read in a different line, with the corresponding annotation (Figure 3.13).

```
docker run -v $path:/data:z immcantation/suite:azahara bash /data/changeo.sh -s
IGH-A-finalFR1FR2FR3_collapse-unique.fastq -g human -t ig -f airr -n IGH-A -o IGH-
A -p5 -k
```



```
sequence_id     sequence        rev_comp        productive      v_call  d_call  j_call  sequence_alignment      germline_alignment
junction        junction_aa     v_cigar d_cigar j_cigar stop_codon vj_in_frame      locus   junction_length np1_length      np2_length
v_sequence_start        v_sequence_end  v_germline_start        v_germline_end  d_sequence_start        d_sequence_end  d_germline_start
d_germline_end  j_sequence_start        j_sequence_end  j_germline_start        j_germline_end  v_score v_identity      v_support
d_score d_identity      d_support       j_score j_identity j_support        fwr1    fwr2    fwr3    fwr4    cdr1    cdr2    cdr3
duplicate_count r1vprimer

M03970:389:000000000-JFFPM:1:1101:14723:2717
GCTTCTGGAGGCACCTTCAGCAGCTATGCTATCAGCTGGGTGCGACAGGCCCCTGGACAAGGGCTTGAGTGGATGGGAGGGATCATCCCTATCTTTGGTACAGCAAACTACGCACAGAAGTTCCAGNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGAGCTGAGCAGCCTGAGATCTGAGGACACGGCCGTGTATTACTGTGCGAGAGATCGGTTGACGGGGGATATTGTAGTAGTACCAGGTTGGCAATACTACTACATGG
ACGTCTGGGGCAAAGGGACCAC       F        T        IGHV1-69*01,IGHV1-69*06,IGHV1-69*12IGHD2-2*01,IGHD2-2*02,IGHD2-2*03 IGHJ6*03
...................................................................................GCTTCTGGAGGCACCTTC...........AGCAGCTATGCTATCAGCTGGGTGCGACAGGCCCCTGG
ACAAGGGCTTGAGTGGATGGGAGGGATCATCCCTATC......TTTGGTACAGCAAACTACGCACAGAAGTTCCAG...NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGAGCTGAGCAGCC
TGAGATCTGAGGACACGGCCGTGTATTACTGTGCGAGAGATCGGTTGACGGGGGATATTGTAGTAGTACCAGGTTGGCAATACTACTACATGGACGTCTGGGGCAAAGGGACCAC
CAGGTGCAGCTGGTGCAGTCTGGGGCT...GAGGTGAAGAAGCCTGGGTCCTCGGTGAAGGTCTCCTGCAAGGCTTCTGGAGGCACCTTC...........AGCAGCTATGCTATCAGCTGGGTGCGACAGGCCCCTGG
ACAAGGGCTTGAGTGGATGGGAGGGATCATCCCTATC......TTTGGTACAGCAAACTACGCACAGAAGTTCCAG...GGCAGAGTCACGATTACCGCGGACGAATCCACGAGCACAGCCTACATGGAGCTGAGCAGCC
TGAGATCTGAGGACACGGCCGTGTATTACTGTGCGAGAGANNNNNNNNNNNNGGATATTGTAGTAGTACCAGNNNNNNNNNTACTACTACATGGACGTCTGGGGCAAAGGGACCAC
TGTGCGAGAGATCGGTTGACGGGGGATATTGTAGTAGTACCAGGTTGGCAATACTACTACATGGACGTCTGG        CARDRLTGDIVVVPGWQYYYMDVW        69N126=48X53=    239S1N20=
267S11N35=      F        T  IGH      72       12       8        1        227      1        320      240      259      2        21       268      302
12      46      205.0   0.7885500000000001      4.44e-55        39.1    1.0     3.03e-07   68.0    1.0     1.11e-15
...................................................................................GCTTCT  ATCAGCTGGGTGCGACAGGCCCCTGGACAAGGGCTTGAGTGGATGGGAGGG
AACTACGCACAGAAGTTCCAG...NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGAGCTGAGCAGCCTGAGATCTGAGGACACGGCCGTGTATTACTGT
TGGGGCAAAGGGACCAC       GGAGGCACCTTC...........AGCAGCTATGCT     ATCATCCCTATC......TTTGGTACAGCA
GCGAGAGATCGGTTGACGGGGGATATTGTAGTAGTACCAGGTTGGCAATACTACTACATGGACGTC  1       VH1_FR1_5
```

Figure 3.13. IgBlast format annotation. *Important fields highlighted with colors, matching column name(s) and information.*

3.6.2.3 Clonal threshold tuning (Shazam)

After assigning VDJ alleles and complete IgBlast annotation fields to each of the sequences, clonal relationships between these sequences need to be assessed to define CLL B cell clones. Hierarchical clustering methods were tested initially, to define the threshold to be used to infer those clonal relationships. Note that this is not part of the modular CLL immcantation pipeline as it has to be tuned previously, to decide the parameters to use in 3.6.2.4.

Shazam R package includes the *distToNearest* function, by which nearest neighbor distances are calculated for each sequence in the IgBlast table. The mode hamming distance is the default model employed, which counts single-nucleotide differences, measuring the minimum number of changes that could have transformed one string into another (same length). The *distToNearest* function was applied to each sample and a polyclonal control sample used for crossvalidation, with the commands below:

```
## cross query sample with polyclonal sample (52 V genes represented)
dist_ham <- distToNearest(table, sequenceColumn="junction",
        vCallColumn="v_call", jCallColumn="j_call",
        model="ham", normalize="len", cross="sample", nproc=1)
```

```
## within
dist_ham2 <- distToNearest(samp_norm, sequenceColumn="junction",
       vCallColumn="v_call", jCallColumn="j_call",
       model="ham", normalize="len", nproc=1)
```

The documentation of this package specifies that B cell repertoires sequence distances plotted in a histogram follow a bimodal distribution, and the threshold corresponds to the intersection (Figure 3.14; shown in red discontinuous line). This can be detected after histogram manual inspection or automatically, using the *findThreshold* function from the same package.

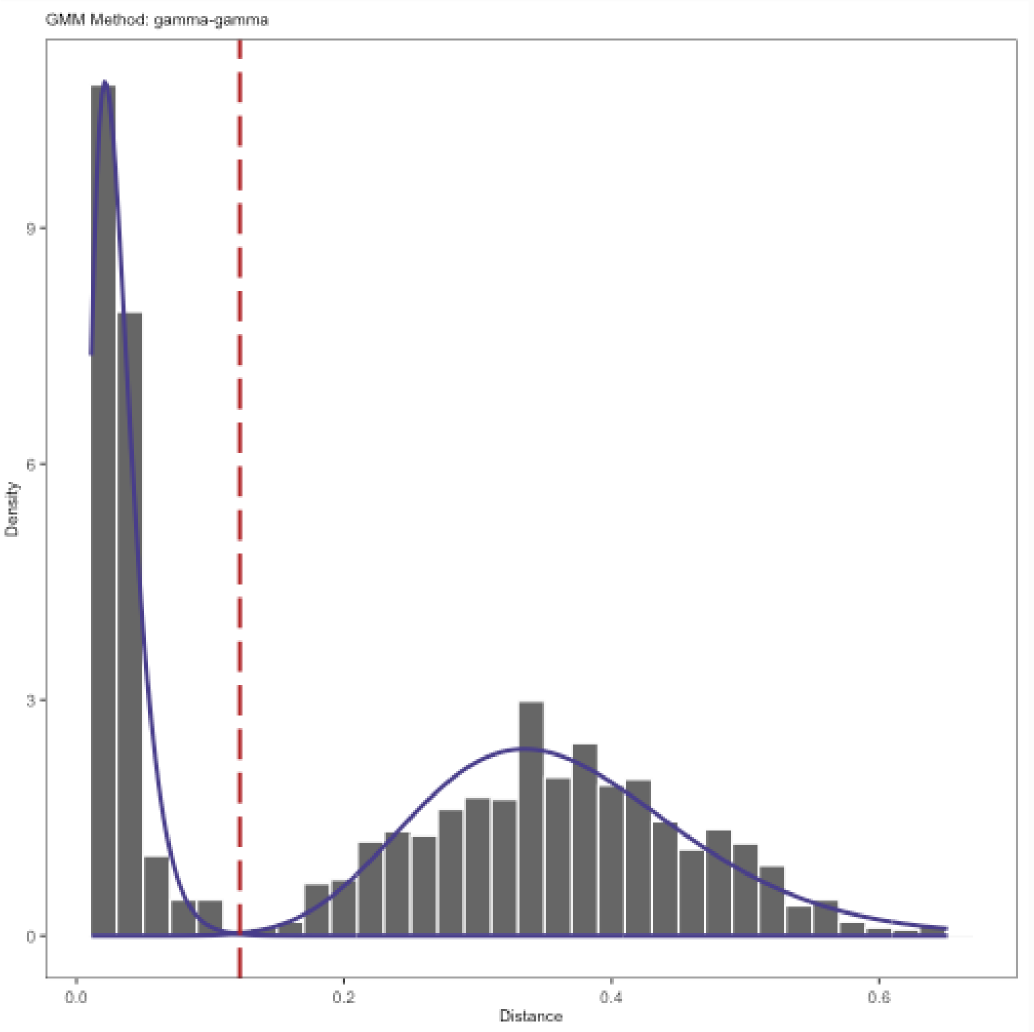


Figure 3.14. distToNearest histogram. *From the Immcantation ReadtheDocs documentation (https://shazam.readthedocs.io /en/stable/vignettes/DistToNearest-Vignette/). distToNearest output adds "dist_nearest" column to the table used as input. The plot results after calculating a histogram with such sequence distances. The threshold can be determined manually or automatically by finding the minimum density value, which conforms a valley between the two modes. The assumption is that the smaller distances peak represents intraclonal distances and the larger distances peak represents interclonal distances.*

Plotting was done following the recommendations detailed in the Immcantation documentation (https://shazam.readthedocs.io/en/stable/vignettes/DistToNearest-Vignette/), for each sample individually, and then with all the samples against a polyclonal background.

3.6.2.4 Clonal clustering (Change-O)

Clustering threshold is used to infer clonal relationships and group individual sequences using distance to the nearest neighbor with the Change-O *defineClones* function. A standard script from the Immcantation team for cloning and germline reconstruction was customized to add different steps required for the analysis of our data (*define_clones_ori.sh*; bash). IgBlast output tables per sample are used as input for this script. Argument -x indicates the threshold, previously tuned (3.6.2.3), employed to define clonally related sequences and -m specifies the model that will be used to calculate such distances ("ham" for hamming in this case). -a specifies to clone the full dataset and not only productive (functional) sequences.

```
docker run -v $path:/data:z immcantation/suite:azahara bash
/data/define_clones_ori.sh -d IGH-A_db-pass.tsv -a -x 0.1 -m ham -n IGH-A_db-pass
-o $output_folder -f airr -p10
```

- *DefineClones.py* is used to group sequences sharing hamming distance below the threshold chosen. Reads are first grouped regarding *IGHV* and *IGHJ* alleles, and junction region length, and then further subdivided into groups with the clonally related sequences following the distance metric and the threshold chosen.

  Inner parameters:

  - *model* (distance metric): ham (hamming distance)
  - *norm* (method for normalizing distances): normalize by length
  - *dist* (distance threshold): 0.1
  - *mode* (use allele or gene level for initial grouping): gene
  - *act* (use only the first of ambiguous VDJ calls, or all of them as a whole): set

In the output table, each sequence (line) is tagged with the clone number assigned, which is given arbitrarily ("clone_id" column).

- *CreateGermlines.py* reconstructs the germline alleles for each sequence using the alignment information. In this case we use the flag *–cloned* to generate a specific germline sequence per clone.

The output table contains annotation of the germline alignment in the type chosen (column "germline_alignment_d_mask", and the columns "germline_v_call", "germline_d_call" and "germline_j_call").

The following steps also included in *define_clones_ori.sh* were implemented specifically for downstream analyses:

- Filter of low frequency mutations:

The aim of the steps in this section is to remove possible sequencing artifacts, and for that purpose, mutations present in a fraction below 2% of the reads are removed. In order to do that, variants are identified on each sequence with their alignment positions. The frequencies on each position are calculated within the sample.

  - *mutations.R*: *observedMutations* function from the Shazam R package is used to annotate the mutation frequencies on *IGHV* region per sequence. Such frequency is added in a separate column named "mu_freq", added to the *defineClones* and germline annotated IgBlast format table.
  - *mutation_freqs.R*: *calcObservedMutations* function from the Shazam R package is used to retrieve the positions with mutations on each sequence and stored in tabular format (Table 3.1).
  - *igblast_filter0-2freq_perfragment_numpyfaster.py*: The table constructed in the previous step and the *defineClones* table are used as input for this script. Per read mutation information is used to calculate the frequencies of each

mutation among all sequences. Those considered low frequency variants are masked with the reference sequence nucleotide, and a new *defineClones* format table is saved. Python package Numpy was employed to store sequences and their corresponding information in array-type data instead of conventional lists and dictionaries to manipulate and perform operations faster. Reads per fragment and variant frequency plots are generated (*-reads.png, *-variantsfragnew.png).

seq_id	positions
M03970:327:000000000-J364L:1:2110:24622:21105	188;193;275;286;360
M03970:327:000000000-J364L:1:1101:14271:27662	168;194;253;257;295;298;303;344
M03970:327:000000000-J364L:1:1101:27523:12840	288;291;294;297;299;315;338;346;368
M03970:327:000000000-J364L:1:2106:9637:27165	90;133;135;168;172;198;208;213;258;…;376
M03970:327:000000000-J364L:1:2106:12315:21174	77;108;116;138;172;255;262;265;266;272;…;376
M03970:327:000000000-J364L:1:2106:2468:11526	74;83;105;121;124;140;145;…;332
M03970:327:000000000-J364L:1:2105:12588:25735	106;110;118;129;144;148;157;159;166;…;340
M03970:327:000000000-J364L:1:2105:4197:20415	83;108;116;118;141;157;165;169;195;215;…;362
M03970:327:000000000-J364L:1:2105:3979:19195	74;108;161;164;172;197;205;262;…;316

Table 3.1. Variant frequencies per sequence example table. *Sequence ID and the positions with variations respecting to germline alleles encountered.*

The downstream final steps in *define_clones_ori.py* include clone visualizations and final results reports, such as:

- Top10 clone plots with mutation frequencies: *mutations_dot_colourfragment.R* generates a plot with the mutation distributions per sample in the 10 most represented clones and a histogram of the mutation frequencies.
- *IGHV* and *IGHJ* genes usage barplots.
- Clone plots and summary files.

3.6.2.5 Validation using SSeq sequences

With the purpose of comparing SSeq results with CLL Immcantation, available sanger sequences were included along with NGS reads for the *defineClones* steps (converted to FASTQ file format, annotated using Change-O IgBlast and concatenated with each sample´s NGS sequences IgBlast file), to be clustered into clones (Figure 3.15). Checkings were performed to validate whether the Sanger Sequences were clustered into the NGS predominant clones or not. A 238 samples dataset was used in the first place to validate the methods employed and adjustments performed, excluding FR3 reads from clustering steps and low frequency variants filtering. Finally, the average of mutation frequencies within a clone was used to determine the mutational status in a patient.

CLL samples correctly classified regarding their mutational status, agreeing with SSeq, and with a single predominant clone, were used to plot mutation frequencies and compare mutation distributions and CDR3 lengths among the M-CLL, U-CLL and BD-CLL groups.

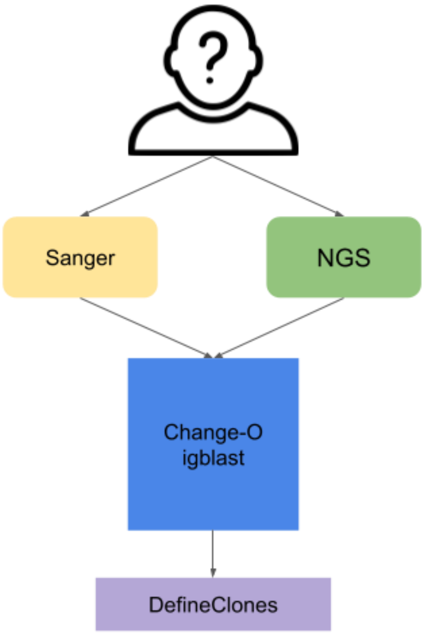


Figure 3.15. Integration of SSeq sequences into IgBlast and defineClones steps.

3.6.2.6 Pipeline execution

To integrate the in-house scripts with functionalities added for the analysis of CLL samples together with the Immcantation framework scripts/modules used, especially within the *defineClones* script (define_clones_ori.sh), a custom docker image was built from the original Immcantation suite v4.1. For that purpose, a configuration file has to be encapsulated along with the code, indicating the existing docker image that has to be imported:

```
FROM immcantation/suite:4.1.0
LABEL maintainer="Azahara Fuentes [afuentri@alumni.uv.es]" \
       description="Immcantation + local changes CLL pipeline"
ADD $script /usr/local/bin/$script
RUN $packageInstallation
```

Afterwards, the custom image for CLL immcantation was built with the following command:

```
docker build . --tag immcantation/suite:CLL
```

3.7 Evaluation of the methods developed and validation

319 CLL samples (314 PB and 5 BM) and 47 healthy donor samples were processed as described in 3.1, and sequenced in various sequencing experiments with Library approach C (3.4.3) with the MIX3 oligo proportions (3.5.1), and analyzed using BMyRepCLL and CLL Immcantation latest versions (3.6.1,3.6.2). After analysis, 319 samples with >1000 total reads assigned to the major Ig rearrangement were selected for validation against the gold standard method. 5 samples were not available at the stage of validation with fragment capillary analysis and were removed from the study. Healthy donor samples were selected with a minimum of 1000 total reads after the trimming stage (47 samples). Ig rearrangements detected were compared to those obtained by the gold-standard method (SSeq), that was used following clinical guidelines, as described (3.2).

Samples were split randomly into test and validation groups, and considering the clonal profiles determined previously by SSeq (Table 3.2). The clonal cutoff was trained on the test dataset and replicated afterwards with the validation samples, to evaluate the sensitivity and specificity of the method.

	sample group	
	test	validation
polyclonal	20	27

		sample group	
		test	validation
clonal	1CLONE	24	260
	>1CLONE	10	20
		34	280

***Table 3.2. Test and validation samples division**. Sample groups regarding the clonal profile previously determined by SSeq and the arbitrary inclusion of them into the test and validation groups.*

3.7.1 Test reliable cutoff

After detecting and characterizing accordingly predominant pathological clones, reaching a reliable cut-off for NGS minor clones was necessary to report only CLL clonal Ig rearrangements and determining the rest as subclonal background. For that purpose, we performed a test with 20 polyclonal-profile samples (healthy donor samples), and used 34 out of 314 validation samples for clonality testing (24 samples with 1 clone and 10 with double clonal profile determined previously by SSeq). The maximum difference read ratio between consecutive clones (MAX_DIFF parameter) described in 3.6.1.2 was used to define the clonal threshold.

Mann-Whitney U test was used to determine significant differences between the maximum difference ratios obtained on pairwise comparisons between the 3 groups, and a clonal/polyclonal background was chosen. Afterwards, the same analysis was reproduced for a validation dataset with 27 polyclonal samples and 280 clonal CLL samples (260 samples with a single SSeq-determined clone and 20 samples with multiple clones determined previously by SSeq). After this classification, samples are tagged as “polyclonal” or “NCLONE” (being N the number of potential pathological CLL clones). Inconsistencies with SSeq were assessed and additional Ig rearrangements detected by NGS were validated using fragment capillary sequencing as described in 3.7.2. Fragments confirmed both by NGS and SSeq were subjected to comparison regarding *IGHV*, *IGHJ* genes, mutational status and CDR3 sequence. The correlation of identity percentages against germline *IGHV* alleles was assessed and compared between both analysis pipelines.

3.7.2 Confirmation of additional rearrangements

Among the NGS results, we obtained B cell rearrangements that were not previously detected using the standard SSeq protocol. In cases where the additional and the previously-detected rearrangements did not belong to the same *IGHV* family (7 cases), SSeq was repeated by A. Serrano (Unit of Hematology HCUV), as described in the methods section 3.2. Afterwards, in the cases where there were still incongruences, along with cases of coexistence of rearrangements from the same *IGHV* family, amplification with leader, FR1, FR2 or FR3 consensus primers and fragment length analysis (GeneScan) was performed on ABI3730 capillary DNA analyzer (coordinated and performed mainly by A. Serrano; Unit of Hematology HCUV). Multiple rearrangements detected were compared with the rearrangements detected with CLL Immcantation.

Samples with multiple rearrangements were also analyzed with MixCR (144), to assess discrepancies in clonal percentages between BMyRepCLL and CLL Immcantation. Example command:

```
mixcr analyze amplicon --species hs --starting-material dna --receptor-type IGH
--5-end v-primers --3-end j-primers --adapters no-adapters A-1.fastq A-2.fastq
```

3.8 Downstream analyses

3.8.1 Repertoire diversity and abundance

Hill diversity numbers are mathematical equations commonly used in ecology to study the diversity of populations (161). Nowadays, it is also commonly applied to microbiome sequencing and immune repertoires. Richness refers to the number of species and evenness, to the relative population of each species (how abundances are distributed). Alpha diversity explains diversity within a sample, and different indexes comprise different ecology measures from 0 = species richness, 1 = Shannon-Weiner entropy Index, 2 = inverse Simpson Index. The higher the alpha, the more the account of high abundant clones on the diversity of the sample population (139,162,163).

estimateAbundance (estimates clonal relative abundance distribution) and *alphaDiversity* (Hill diversity) functions from the Immcantation R package Alakazam were employed to calculate repertoire diversity with default parameters and a 95% confidence interval among the CLL (U-CLL and M-CLL) and healthy donors group.

3.8.2 Statistical significance

Pairwise comparisons of numeric variables (continuous), were tested for normality among the groups using normaltest from scipy.stats Python package. If the distributions were normal, paired t-test (two-sided), was employed to obtain P values for statistical significance (scipy.stats). On the other hand, for no normal distributions, non-parametric Mann Whitney U test was performed (two-sided) using scipy.stats. P values were corrected with the Bonferroni method, with the Python package statannotations.

Pearson correlation statistical test was performed for extraction of the r-squared value for lineal data correlations and P value extraction.

For the comparisons of SSeq-NGS results, automatic scripts were constructed. Logarithmic scales were calculated for diversely distributed data.

4 Results

4.1 Comparison of library preparation methods

Initially, we tested 3 methods for the preparation of DNA libraries from the IGH locus: protocol A using 300bpx2 sequencing kit (Leader-JH primers), and protocols B and C, both using 150bpx2 sequencing kit (using amplification from Leader-JH primers and DNA tagmentation in B, whereas the last combined Framework primer sets with reverse JH).

Table 4.1 shows the comparison between the predominant rearrangements encountered with the different DNA libraries tested and SSeq. *IGHV* gene and *IGHJ* alleles were used to characterize the major clone. SSeq rearrangement was identified as the predominant in 18/23 (78.26%) samples in the case of 300bpx2 cycles Leader sequencing (Library Protocol A), and in 19/23 (82.60%) in the other two protocols, both using 150bpx2 cycles sequencing (Library Protocols B and C). Figure 4.1 shows the example coverage profiles obtained from the three library preparation methods tested. Having successfully generated sequencing libraries using the 3 methods, library preparation C was chosen for performing sequencing experiments, due to the lower costs (not depending on commercial kits like method B), and improvements in turnaround time and sequencing quality.

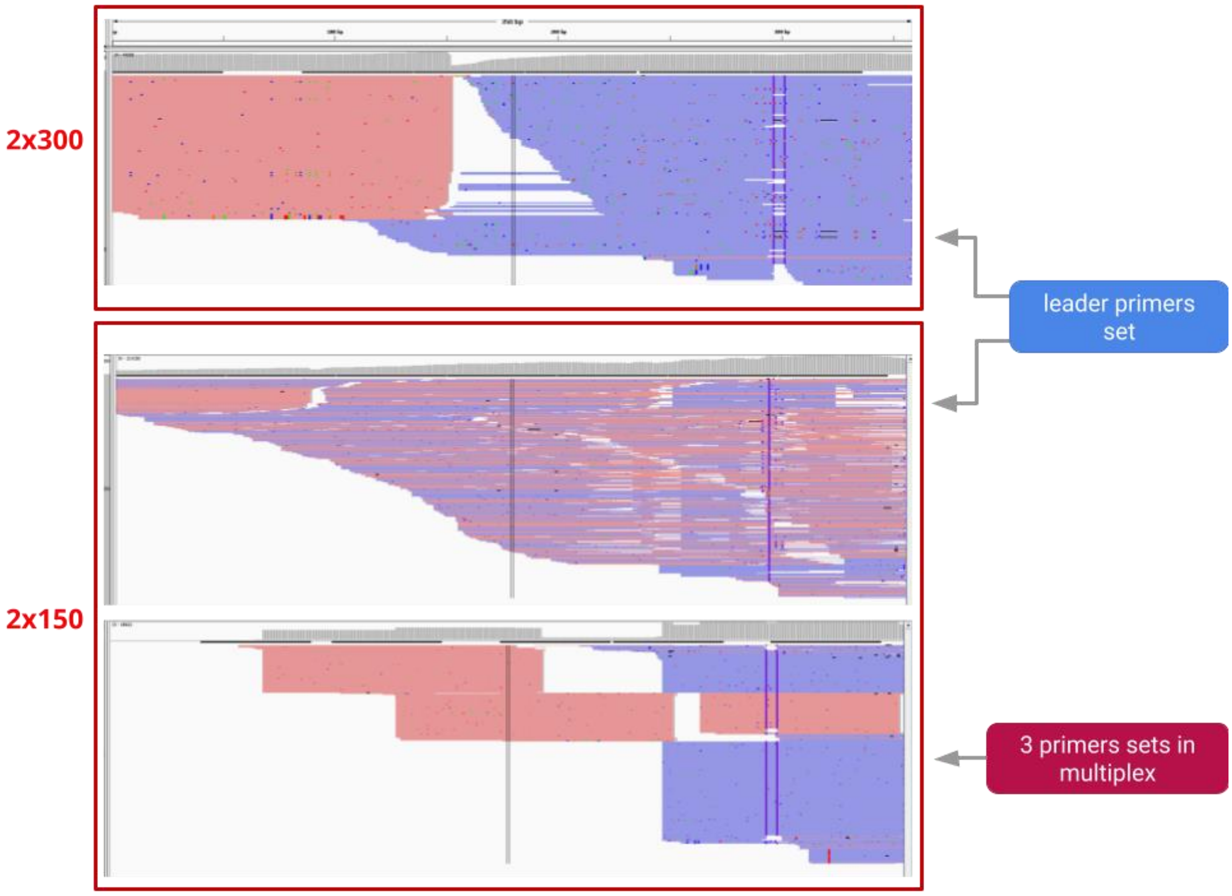


Figure 4.1. **IGV visualization example BAM files from the 3 library preparation methods tested.**

	Sanger sequencing		HTS 300bpx2		HTS Nextera XT (150bpx2)		HTS in-house (150bpx2)	
IGH-1	IGHV4-34_IGHJ6*02	CARGYGDTGVIRRYYYYGMDVW	IGHV4-34_IGHJ6*02	CARGYGDTGVIRRYYYYGMDVW	IGHV4-34_IGHJ6*02	CARGYGDTGVIRRYYYYGMDVW	IGHV4-34_IGHJ6*02	CARGYGDTGVIRRYYYYGMDVW
IGH-2	IGHV1-69_IGHJ4*02	CAREPRPIPAIPYYDFWSGYSPYFDYW	IGHV1-69_IGHJ4*02	NOT DETERMINED	IGHV1-69_IGHJ4*02	NOT DETERMINED	IGHV1-69_IGHJ4*02	NOT DETERMINED
IGH-3	IGHV3-30_IGHJ4*02	CASPLRRGFFDWAVAGTFGLDYW	IGHV3-30_IGHJ4*02	NOT DETERMINED	IGHV3-30_IGHJ4*02	CASPLRRGFFDWAVAGTFGLDYW	IGHV3-30_IGHJ4*02	NOT DETERMINED
IGH-4	IGHV4-39_IGHJ6*02	CANRPGYCSGGSCYDYYYYGMDVW	IGHV3-64D_IGHJ6*02	NOT DETERMINED	IGHV3-64D_IGHJ6*02	NOT DETERMINED	IGHV3-64D_IGHJ6*02	NOT DETERMINED
IGH-5	IGHV1-69_IGHJ2*01	NOT DETERMINED	IGHV1-69_IGHJ6*04	NOT DETERMINED	IGHV1-69_IGHJ6*04	CARGTDNYDFWSGYSNGYYYYYGMDVW	IGHV1-8_IGHJ6*04	CARGTDNYDFWSGYSNGYYYYYGMDVW
IGH-6	IGHV3-23_IGHJ4*02	CAKDGGVYDFWSGYYPPYYFDYW	IGHV3-23D_IGHJ4*02	NOT DETERMINED	IGHV3-23_IGHJ4*02	CAKDGGVYDFWSGYYPPYYFDYW	IGHV3-23_IGHJ4*02	CAKDGGVYDFWSGYYPPYYFDYW
IGH-7	IGHV1-8_IGHJ6*02	CARGDLLRFLEWLSNYYYGMDVW	IGHV1-8_IGHJ6*02	CARGDLLRFLEWLSNYYYGMDVW	IGHV1-8_IGHJ6*02	CARGDLLRFLEWLSNYYYGMDVW	IGHV1-8_IGHJ6*02	CARGDLLRFLEWLSNYYYGMDVW
IGH-8	IGHV1-69_IGHJ6*02	CARETIFGVVNYNYYYYYGMDVW	IGHV1-69D_IGHJ6*02	NOT DETERMINED	IGHV1-69_IGHJ6*02	CARETIFGVVNYNYYYYYGMDVW	IGHV1-69_IGHJ6*02	CARETIFGVVNYNYYYYYGMDVW
IGH-9	IGHV2-5_IGHJ4*02	CGHRRGLWFGFYW	IGHV2-5_IGHJ4*02	CGHRRGLWFGFYW	IGHV2-5_IGHJ4*02	CGHRRGLWFGFYW	IGHV2-5_IGHJ4*02	CGHRRGLWFGFYW
IGH-10	IGHV4-59_IGHJ6*02	CARGRGDYYDSSGYLYYYYGMDVW	IGHV1-8_IGHJ6*02	NOT DETERMINED	IGHV1-8_IGHJ6*02	NOT DETERMINED	IGHV1-8_IGHJ6*02	CARGRGDYYDSSGYLYYYYGMDVW
IGH-11	IGHV3-7_IGHJ6*01	CAGGWADMEYYYYYYGMDVW	IGHV3-7_IGHJ6*02	CAGGWADMEYYYYYYGMDVW	IGHV3-7_IGHJ6*02	CAGGWADMEYYYYYYGMDVW	IGHV3-7_IGHJ6*02	CAGGWADMEYYYYYYGMDVW
IGH-12	IGHV4-59_IGHJ4*02	CARGGSNLRLDYFDYW	IGHV4-59_IGHJ4*02	NOT DETERMINED	IGHV4-59_IGHJ4*02	CARGGSNLRLDYFDYW	IGHV4-59_IGHJ4*02	CARGGSNLRLDYFDYW
IGH-13	IGHV3-9_IGHJ4*02	CAKDREYYDFWSGYRKAYSFDYW	IGHV3-9_IGHJ4*02	CAKDNYFDYW	IGHV3-9_IGHJ4*02	CAKDREYYDFWSGYRKAYSFDYW	IGHV3-9_IGHJ4*02	CAKDREYYDFWSGYRKAYSFDYW
IGH-14	IGHV3-7_IGHJ4*03	NOT DETERMINED	IGHV3-7_IGHJ4*02	CARDMGWSQFDSW	IGHV3-7_IGHJ4*02	CARDMGWSQFDSW	IGHV3-7_IGHJ4*02	CARDMGWSQFDSW
IGH-15	IGHV4-34_IGHJ4*02	CARGRTGWYPPGSW	IGHV4-34_IGHJ5*02	CARGRTGWYPPGS	IGHV4-34_IGHJ5*02	CARGRTGWYPPGS	IGHV4-34_IGHJ5*02	CARGRTGWYPPGS
IGH-16	IGHV1-69_IGHJ3*02	NOT DETERMINED	IGHV1-69_IGHJ3*02	CARDDAFDIW	IGHV1-69_IGHJ3*02	CARGGDYDSPYLPNDAFDIW	IGHV1-69_IGHJ3*02	CARGGDYDSPYLPNDAFDIW

*continues in next page

IGH-17	IGHV3-21_IGHJ6*02	CVW	IGHV3-21_IGHJ6*02	CVGDRNGMDVW	IGHV3-21_IGHJ6*02	CVGDRNGMDVW	IGHV3-21_IGHJ6*02	CVGDRNGMDVW
IGH-18	IGHV3-30_IGHJ4*02	NOT DETERMINED	IGHV3-30_IGHJ4*02	NOT DETERMINED	IGHV3-30_IGHJ4*02	CASDRKWLPHYTQFDYW	IGHV3-30_IGHJ4*02	NOT DETERMINED
IGH-19	IGHV3-30_IGHJ2*01	CAGDGHCRGFGCYFTVFSYYFDLW	IGHV3-30_IGHJ2*01	CAGDGHCRGFGCYFTVFSYYFDLW	IGHV3-30_IGHJ2*01	CARVSYYFDLW	IGHV3-30_IGHJ2*01	CARVSYYFDLW
IGH-20	IGHV1-69_IGHJ6*02	CARAHPGHDDFWSGYPYQYLYYYYYYGMDVW	IGHV1-69_IGHJ6*02	NOT DETERMINED	IGHV1-69_IGHJ6*02	CARAHPGHDDFWSGYPYQYLYYYYYYGMDVW	IGHV1-69_IGHJ6*02	CARAHPGHDDFWSGYPYQYLYYYYYYGMDVW
IGH-21	IGHV3-7_IGHJ4*02	CASRAVPRDSWYYLDYW	IGHV3-21_IGHJ6*02	NOT DETERMINED	IGHV3-7_IGHJ4*02	CASRAVPRDSWYYLDYW	IGHV3-7_IGHJ4*02	CASRAVPRDSWYYLDYW
IGH-22	IGHV1-46_IGHJ3*02	CARVYYYDSSGYYYKGVHDAFDIW	IGHV1-46_IGHJ3*02	CARDDAFDIW	IGHV1-46_IGHJ3*02	CARHDAFDIW	IGHV1-46_IGHJ3*02	CAADDAFDIW
IGH-23	IGHV1-46_IGHJ4*02	CARMPHPYSSSWYPFDYW	IGHV1-46_IGHJ4*02	NOT DETERMINED	IGHV1-46_IGHJ4*02	CARMPHPYSSSWYPFDYW	IGHV1-46_IGHJ4*02	CARMPHPYSSSWYPFDYW

Table 4.1. Results comparison between the 3 library preparation methods. *Rearrangements obtained with the three library preparation methods tested and comparison with SSeq (predominant rearrangement and CDR3).*

4.2 Adjustment of Framework primers proportions for the in-house method

Library preparation Method C (3.4.3) was tested with different proportions of primer sets FR1, FR2 and FR3 (3.5.1), in 6 different CLL samples and a polyclonal control in order to ensure the correct amplification of the region of interest and the proportion of reads mapped against it. Results obtained after visualizing by means of capillary electrophoresis the DNA fragments amplified with the 3 mixes, show more homogeneous DNA sizes in MIX3, with less unspecific amplification. In the same way, results after sequencing show less condensed mismatch areas in the visualization of BAM files with the program IGV (BAM files employed are from the major rearrangement encountered) (Figure 4.2).

There were no significant differences in the percentage of reads mapped against IGHV alleles between MIX1, MIX2 or MIX3 (p.values Mann-Whitney-Wilcoxon test two-sided: MIX1 vs MIX2 = 7.012e-01, MIX1 vs MIX3 = 1.00, MIX2 vs MIX3 = 7.981e-01) (Figure 4.3). Further, MIX2 and MIX3 had equal efficiency percentage on characterizing the predominant CLL clones (Table 4.2). Thus, MIX3 was chosen for future sequencing experiments due to its more specific amplification.

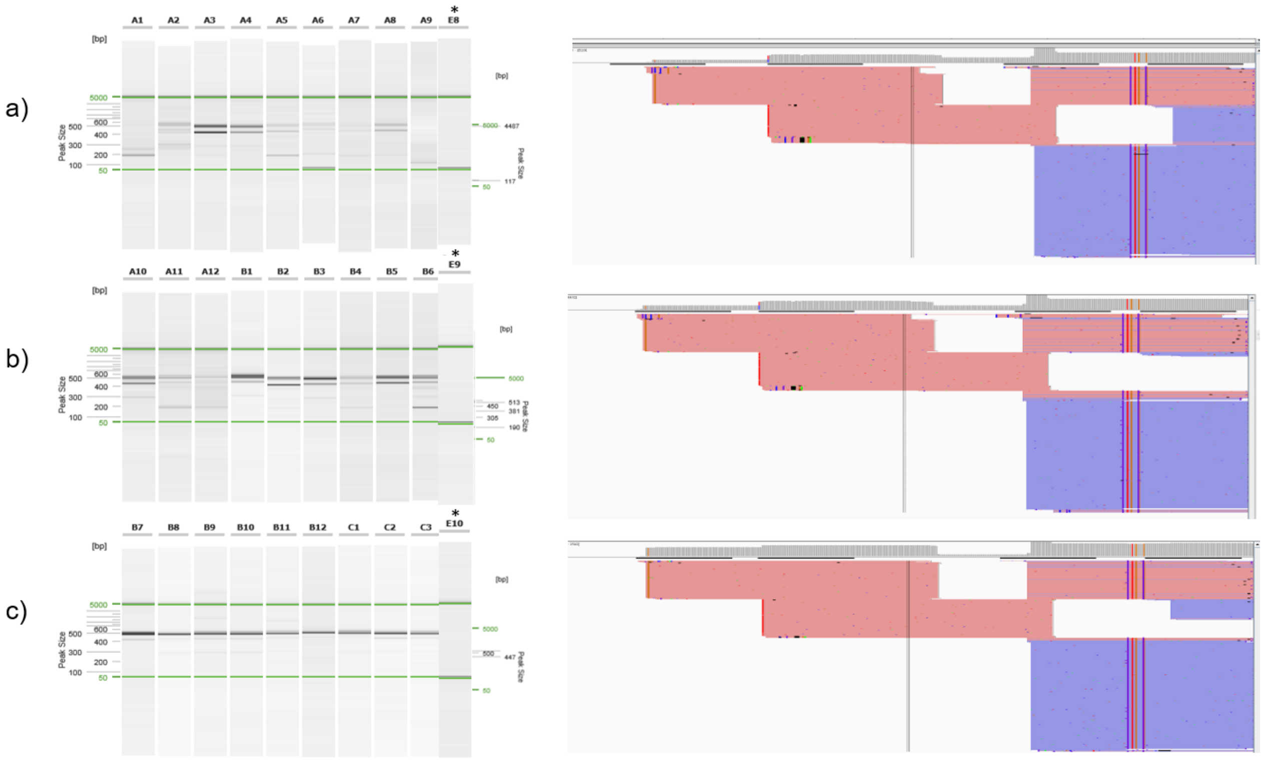


Figure 4.2. Adjustment of Framework primers proportions. *Qiaxcel DNA bands (left) and IGV visualization of rearrangement BAM files (right) of a) MIX1, b) MIX2 and c) MIX3. *Negative control.*

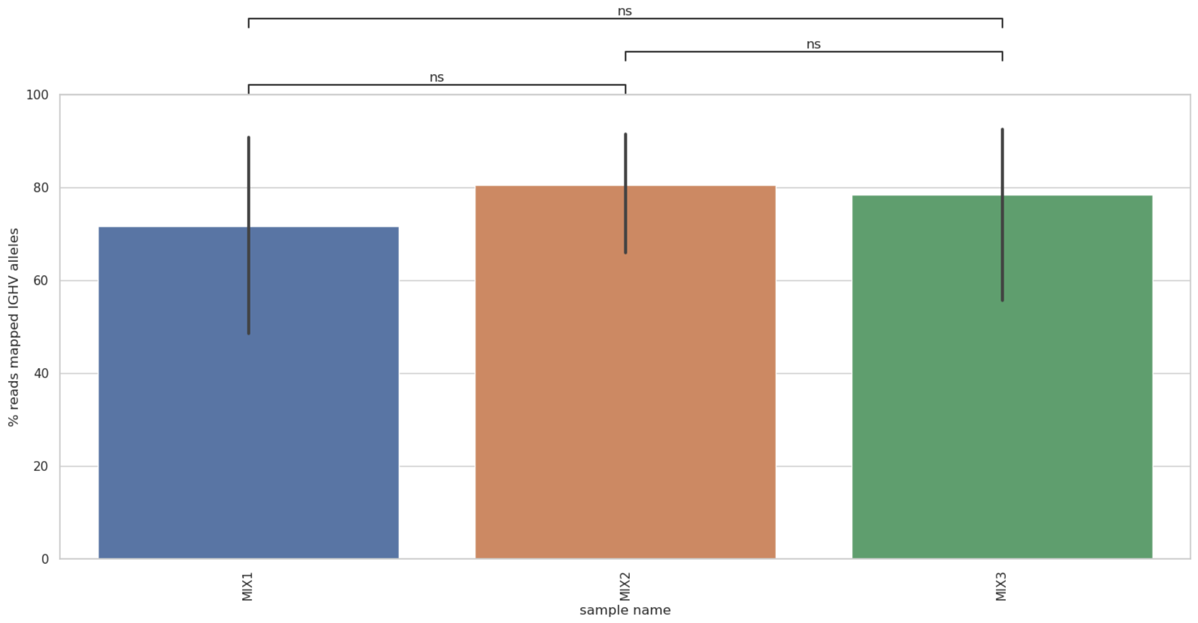


***Figure 4.3. Percentage of reads mapped against IGHV alleles with MIX1, MIX2 and MIX3**. MIX1 vs. MIX2: Mann-Whitney-Wilcoxon test two-sided, P_val:7.012e-01. MIX2 vs. MIX3: Mann-Whitney-Wilcoxon test two-sided, P_val:7.981e-01. MIX1 vs. MIX3: Mann-Whitney-Wilcoxon test two-sided, P_val:1.000e+00. (ns: p <= 1.00e+00; *: 1.00e-02 < p <= 5.00e-02; **: 1.00e-03 < p <= 1.00e-02; ***: 1.00e-04 < p <= 1.00e-03; ****: p <= 1.00e-04).*

	average total reads	average % of reads mapped IGHV alleles	% correctly characterized samples
MIX1	28783.57 ± 19449.34	72.07 ± 32.04	60
MIX2	24834.57 ± 12992.45	81.12 ± 19.48	80
MIX3	22258.00 ± 13495.36	78.81 ± 28.29	80

Table 4.2. Performance results of the different primer mixes *(MIX1, MIX2 and MIX3).*

4.2.1 cDNA leader + framework

6 CLL samples with a predominant CLL clone determined by SSeq were compared with the results obtained after amplification with primers mix MIXLFA (3.5.2), which includes the Leader *IGHV* family primer sequences added to the Framework regions primer sets. The performance was evaluated after sequencing, and agreement with SSeq was proven for all samples.

Table 4.3 shows the major rearrangement detected, where there is a single difference in allele in CDNA case n.5, ensuring that the analysis is compatible with both primer mixes and allow the inclusion of Leader primers using cDNA, representing an alternative to the MIX3 method for covering the entire IGH variable (VDJ) region. As expected, Leader primers combined with the FR primer sets are successful in covering the whole IMGT combined *IGHV-IGHJ* allele references used with short 150 reads (Figure 4.4), with an example of an UM (a) and MM (b) case.

Sample	Ig Rearrangement	Clonal %	Mutational Status	CDR3	N. clones
CDNA1	IGHV1-2*04_IGHJ6*02	98.8	UM	CARDGYDILTGYPQDYYYYYGMDVW	1CLONE
CDNA2	IGHV1-69*09_IGHJ4*02	98.5	UM	CARAYYDFWSGYSEFDYW	1CLONE
CDNA3	IGHV5-10-1*03_IGHJ4*02	100	MM	CARHWGRAWNYRPDYW	1CLONE
CDNA4	IGHV1-69D*01_IGHJ6*02	98.8	UM	CARSPYCSSTSCYLVDYYYGMDVW	1CLONE
CDNA5	IGHV3-7*04_IGHJ6*02^	95.9	MM	CARALSEGYCPSCGMDVW	1CLONE
CDNA6	IGHV3-11*06_IGHJ5*02	99.9	UM	CAREKLIYYGSGSYYNWFDPW	1CLONE

Table 4.3. Major rearrangement reported after bioinformatics analyses on the cDNA experiment with primer mix MIXLFA. *Equal results to SSeq were obtained for all samples regarding the predominant rearrangement detected, CDR3 sequence and mutational status.* ***^SSeq different IGHV allele: IGHV3-7*02****.*

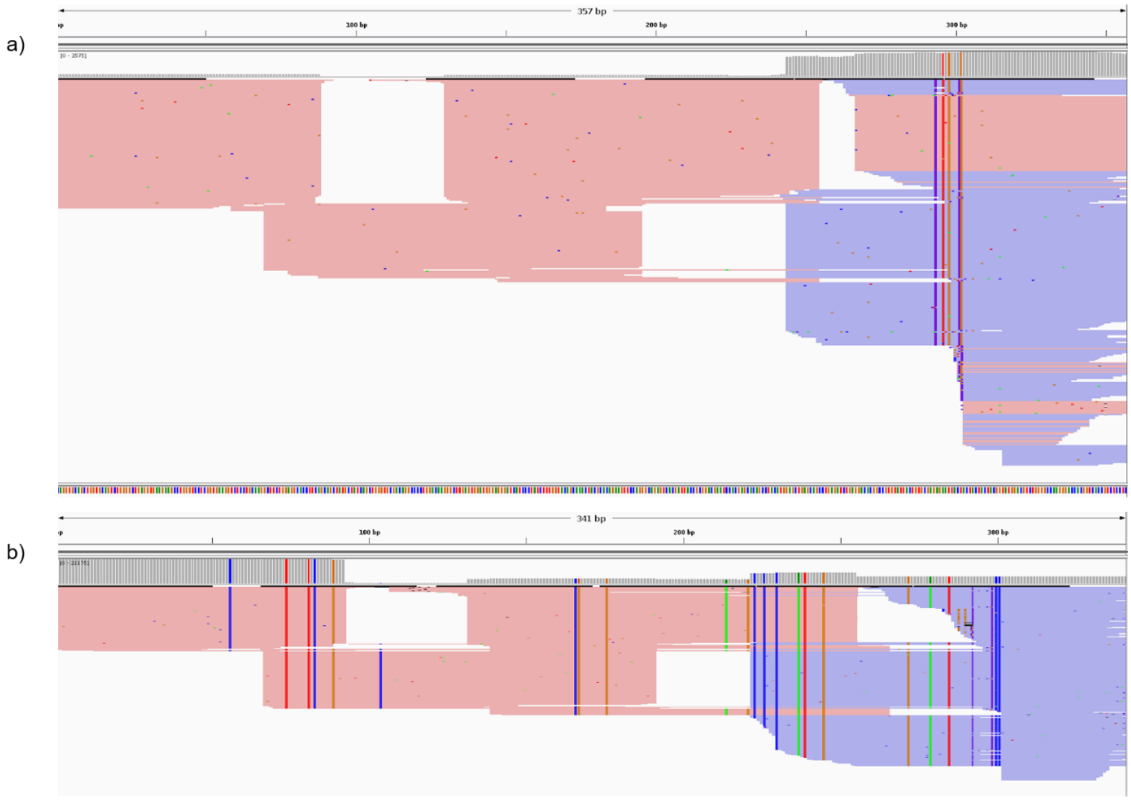


Figure 4.4. Example of IGV VDJ rearrangement BAM file visualization. *a) UM case (sample CDNA4 from Table 4.3), IGHV1-69D*01_IGHJ6*02. b) MM case example (sample CDNA3 from Table 4.3), IGHV5-10-1*03_IGHJ4*02.*

4.3 Bioinformatic analysis: pipelines development

The focus of this thesis work has been the development of tools for clinical determination of Ig rearrangements, and their subsequent validation. Following the library preparation in-house method chosen (3.4.3), due to the lower costs and turnaround time for the use in clinical procedures, two pipelines have been strategically developed:

1. BMyRepCLL: in-house pipeline, whose automation and parallelization is performed with a main script written in Python programming language. The strategy consists of mapping reads separately against the different IMGT gene segments references, following a clone-centered determination which is achieved with the obtaining of a consensus sequence. B cell rearrangements are defined after IGHV-IGHJ alleles correspondence determination and a

specific procedure has been designed to cope with unspecific mapping and gene-call fragment biases, and for the calculation of the clonal fraction per patient.

2. CLL Immcantation: The Immcantation Framework (https://immcantation.readthedocs.io/en/stable/) is a suite of tools and pipelines written in Bash, Python and R, developed by the Kleinstein lab (Yale School of Medicine; Pathology department). A trimester internship took place in this laboratory with the purpose of creating a workflow using software developed by a community of experts in computational immunology. Different tools of the Immcantation Framework were used, and combined with in-house scripts for specific functions regarding the library preparation method employed to generate the data and the purpose of detecting and characterizing CLL B cell clones accordingly. Since the algorithmic basis in this pipeline varies from BMyRepCLL and the troubleshooting of the Immcantation had already been fulfilled by the developers of Kleinstein lab, both methods were compared to obtain a double-check in the results. The main difference between both pipelines is that Immcantation VDJ gene assignment and the rest of features are annotated per read, using IgBlast tool, and clone grouping of the sequences is performed after annotating this information, using clonal clustering methods. R packages such as Shazam and Alakazam, are designed for downstream analyses such as physicochemical property analysis, *repertoire* diversity, clonal lineage reconstruction, mutation profiles, etc. The Immcantation group participates in the AIRR community (Adaptive Immune Receptor *Repertoire* Community of the Antibody society), which is setting standards for the analysis of BcR/TcR *repertoires*, and their software is up-to-date with AIRR recommended formats. Special mention to the collaboration of Dr. Steven Kleinstein and Susanna Marquez.

4.3.1 Development of BMyRepCLL

The development of this pipeline started in the year 2018, and the steps have been fine-tuned over time until the current 2022 version, available on GitHub (https://github.com/afuentri/B-MyRepCLL).

The workflow is restricted to B cell clone detection, whose main strategy is generating a consensus sequence per rearrangement, defined by a combination of IGHV-IGHJ unique alleles. The steps of the pipeline are detailed in the Methods section (3.6.1).

The analysis process has been divided into two modules: “VDJ region clone characterization”, in which raw reads are preprocessed and aligned against the reference germline IMGT alleles in the succession of several steps to characterize the clones present on each sample and extract information of VDJ calls, mutational status, CDR3 amino acid sequence, and productivity (Figure 4.5). The second module, so called “Artifact filtering and rearrangement prioritization”, consists of using the output of the first module to filter artifact rearrangements arising from unspecific mapping in the allele assignment steps due to *IGHV* gene/allele similarities and length differences between the fragments sequenced. After obtaining a list of high confidence rearrangements per sample, a clonal threshold regarding the clonal profile of each patient is calculated to assort B cell clones into clonal and subclonal rearrangements.

The whole process together with the update and generation of IMGT alleles database, can be automated with a simple script (example in the GitHub repository parent directory “launch-default.py”), along with the quality control steps. Different files from the first analysis module and the quality control are needed to generate a final summary with a report using a final script called “onlyclonality.py” (second module).

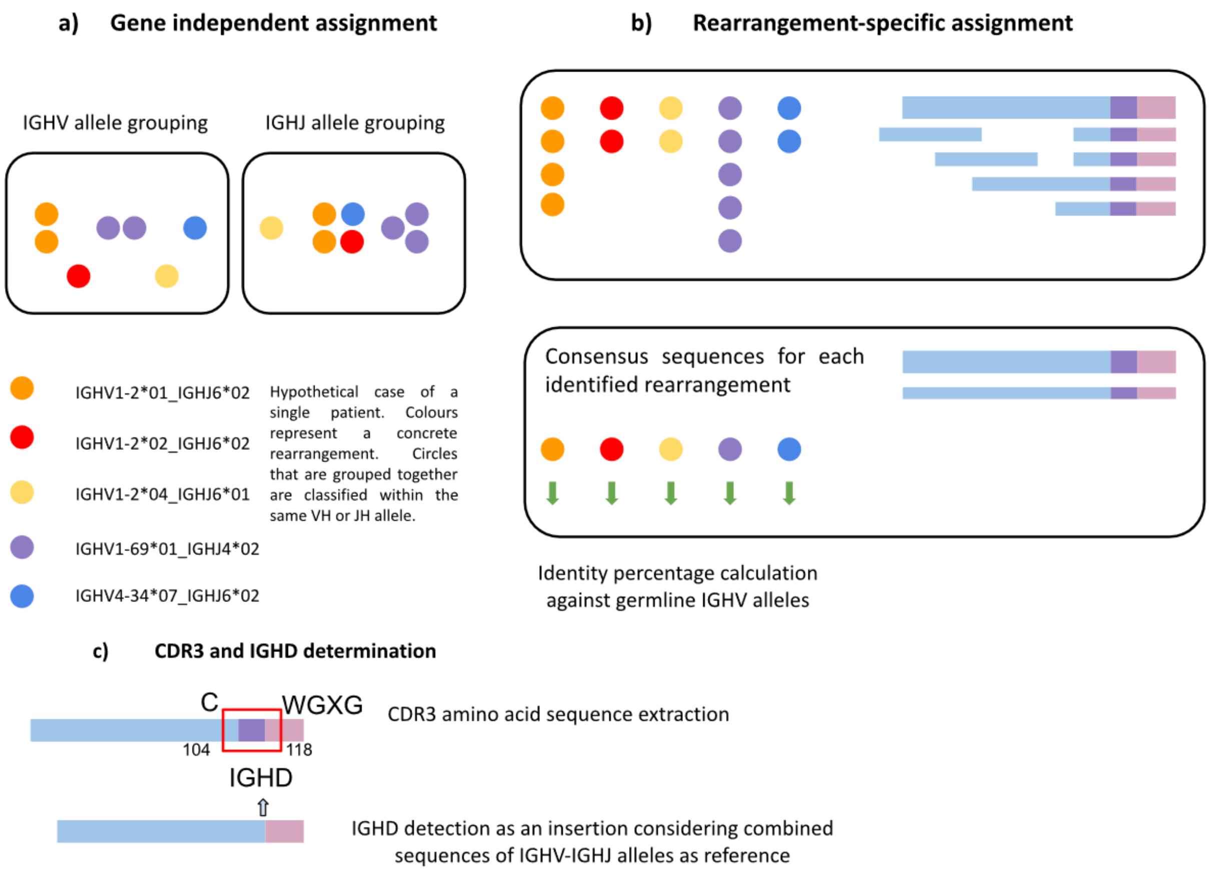

a) Gene independent assignment
b) Rearrangement-specific assignment
IGHV allele grouping
IGHJ allele grouping
IGHV1-2*01_IGHJ6*02
IGHV1-2*02_IGHJ6*02
IGHV1-2*04_IGHJ6*01
IGHV1-69*01_IGHJ4*02
IGHV4-34*07_IGHJ6*02
Hypothetical case of a single patient. Colours represent a concrete rearrangement. Circles that are grouped together are classified within the same VH or JH allele.
Consensus sequences for each identified rearrangement
Identity percentage calculation against germline IGHV alleles
c) CDR3 and IGHD determination
C
WGXG
104
118
CDR3 amino acid sequence extraction
IGHD
⇧
IGHD detection as an insertion considering combined sequences of IGHV-IGHJ alleles as reference

Figure 4.5. BMyRepCLL first module pipeline scheme.

4.3.2 Development of CLL Immcantation

The pipeline developed employing the Immcantation suite, consists of 4 independent blocks: preprocessing (pRESTO), IgBlast annotation (Change-O), clonal clustering (Shazam), and mutational load calculation (Shazam). Different in-house methods for filtering and plotting have been integrated within these analysis modules (Figure 4.6). The steps within each aforementioned block are detailed in the Methods section (3.6.2).

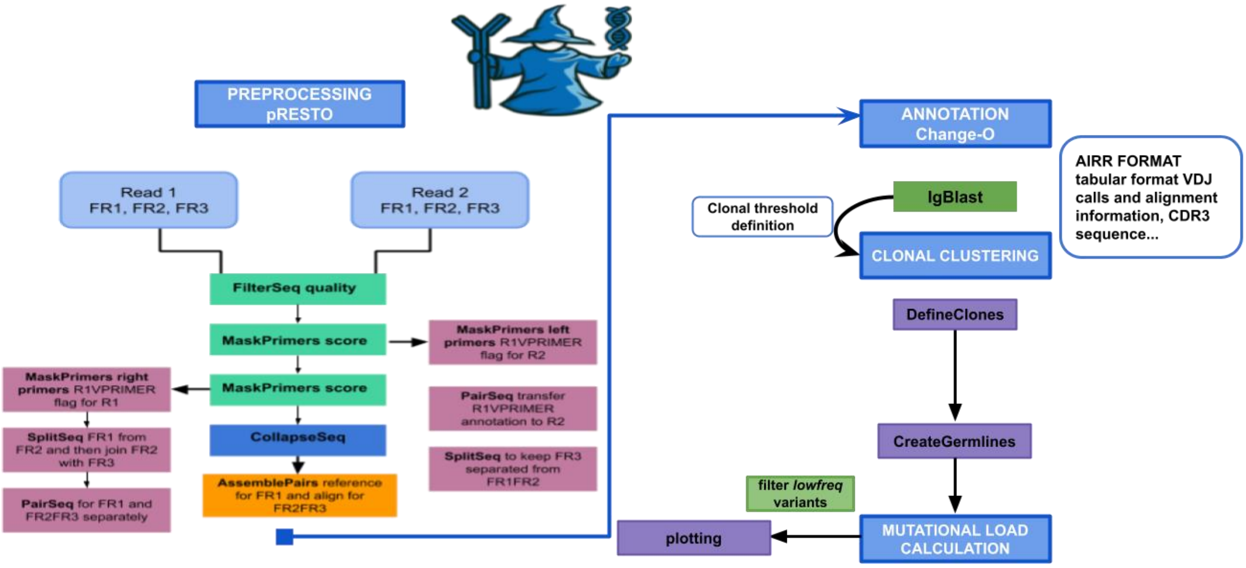

PREPROCESSING
pRESTO
Read 1
FR1, FR2, FR3
Read 2
FR1, FR2, FR3
FilterSeq quality
MaskPrimers score
MaskPrimers left primers R1VPRIMER flag for R2
MaskPrimers score
MaskPrimers right primers R1VPRIMER flag for R1
PairSeq transfer R1VPRIMER annotation to R2
CollapseSeq
SplitSeq FR1 from FR2 and then join FR2 with FR3
SplitSeq to keep FR3 separated from FR1FR2
AssemblePairs reference for FR1 and align for FR2FR3
PairSeq for FR1 and FR2FR3 separately
ANNOTATION
Change-O
AIRR FORMAT tabular format VDJ calls and alignment information, CDR3 sequence...
Clonal threshold definition
IgBlast
CLONAL CLUSTERING
DefineClones
CreateGermlines
filter lowfreq variants
plotting
MUTATIONAL LOAD CALCULATION

Figure 4.6. General overview of the CLL Immcantation pipeline.

- Preprocessing with pRESTO: pRESTO is a tool from the Immcantation Framework for the performance of preprocessing steps in high-throughput Immunoglobulins repertoire data. The preprocessing pipeline used on our data was created using a predefined Immcantation pipeline as a model and modifying specific steps. The original pRESTO AbSeq pipeline (https://github.com/czbiohub/bcell_pipeline/blob/master/src/presto-abseq.sh) and the custom developed here (CLL Immcantation) pRESTO block are compared in Figure 4.7. The steps include trimming by quality, primer sequences annotation, primer sequences masking and assembly of reads. Fragment-wise steps are based on the detachment of reads from different fragments in two occasions, one for primer sequences trimming and the other for assembly (Figure 4.7).
- IgBlast Annotation with Change-O: after preprocessing, reads are annotated employing IgBlast, and a tabular file is created with the information of VDJ alleles, alignment information, junction, productivity, etc. Each row represents the information for a single sequence.
- Clonal clustering with Shazam: sequence grouping employing hierarchical clustering with the Hamming distance method. Sequences sharing *IGHV* and *IGHJ* alleles and junction

length are measured in distance to their nearest neighbor and those with distances below a threshold chosen by the user, are grouped into clones.

- Mutational Load Calculation with Shazam: Shazam R package calculates sequence mutation frequencies. Afterwards, this information is managed with in-house scripts to perform filtering of variants at low proportion (possible sequencing artifacts).

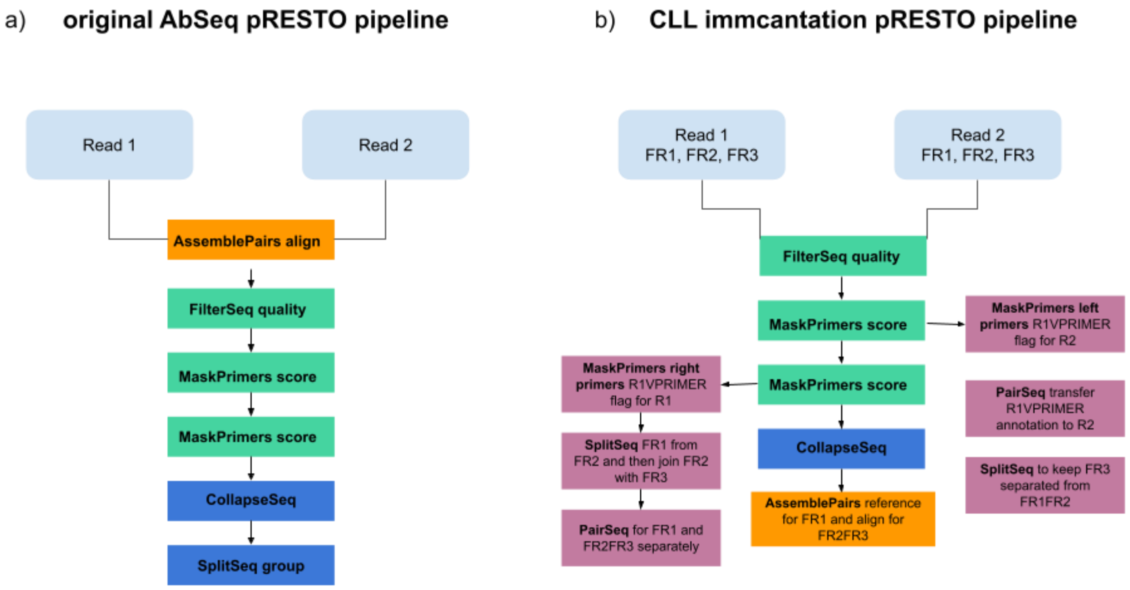

a) original AbSeq pRESTO pipeline
Read 1
Read 2
AssemblePairs align
FilterSeq quality
MaskPrimers score
MaskPrimers score
CollapseSeq
SplitSeq group
b) CLL immcantation pRESTO pipeline
Read 1 FR1, FR2, FR3
Read 2 FR1, FR2, FR3
FilterSeq quality
MaskPrimers score
MaskPrimers left primers R1VPRIMER flag for R2
MaskPrimers right primers R1VPRIMER flag for R1
MaskPrimers score
PairSeq transfer R1VPRIMER annotation to R2
SplitSeq FR1 from FR2 and then join FR2 with FR3
CollapseSeq
SplitSeq to keep FR3 separated from FR1FR2
PairSeq for FR1 and FR2FR3 separately
AssemblePairs reference for FR1 and align for FR2FR3

Figure 4.7. Comparison between the diagrams of the original AbSeq pRESTO pipeline used as a template (a) and customized preprocessing steps in CLL Immcantation workflow (b).

4.4 Performance of BMyRepCLL analysis pipeline

For the evaluation of the in-house pipeline performance, we will focus on the use of MIX3 primers pool with 150bpx2 Illumina sequencing kit, as it is the method used for the validation (section 3.7). For benchmarking and a better understanding of the workflow, the performance is shown after the analysis of two control commercial samples with clonal and polyclonal B cell clone profiles, respectively.

4.4.1 Quality control

High quality reads above Q30 were obtained on average after trimming steps performed by BMyRepCLL (Figure 4.8). 43 and 34% of reads were removed during preprocessing steps in the polyclonal and clonal samples, respectively. The number of effective reads remaining after trimming was 553409 for the polyclonal and 683648 for the clonal sample. Percentage of reads mapped against *IGHV* alleles was 85.25% in the polyclonal and 79.68% in the clonal sample.

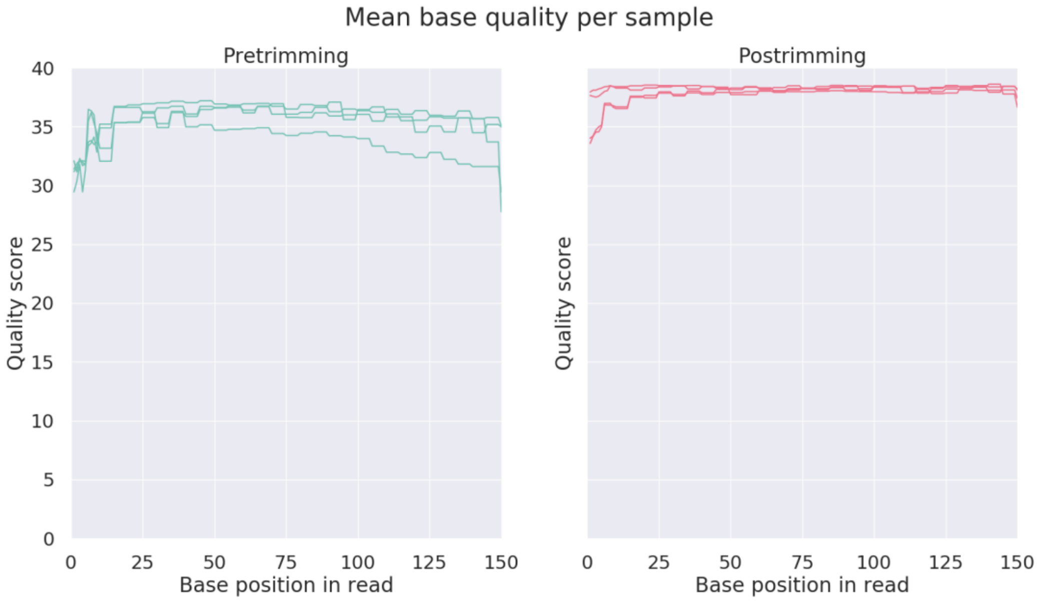


Figure 4.8. Raw data quality control. *Base quality encoded in the Phred score scale for Illumina sequencing, per read base position in the pretrimming steps and after the preprocessing steps performed throughout the analysis pipeline. Each line represents the quality score on a FASTQ file (R1 and R2 are included, thus having 2 FASTQ files per sample). Quality scores are parsed from the output files of FastQC program.*

4.4.1.1 Primers trimming

Left primers corresponding to *IGHV* FR1, FR2 and FR3 regions were removed during preprocessing. Differences in the alignment of reads with or without removing primers is shown in Figure 4.9. Mismatch bases at the beginning of forward reads are removed successfully when trimming primers.

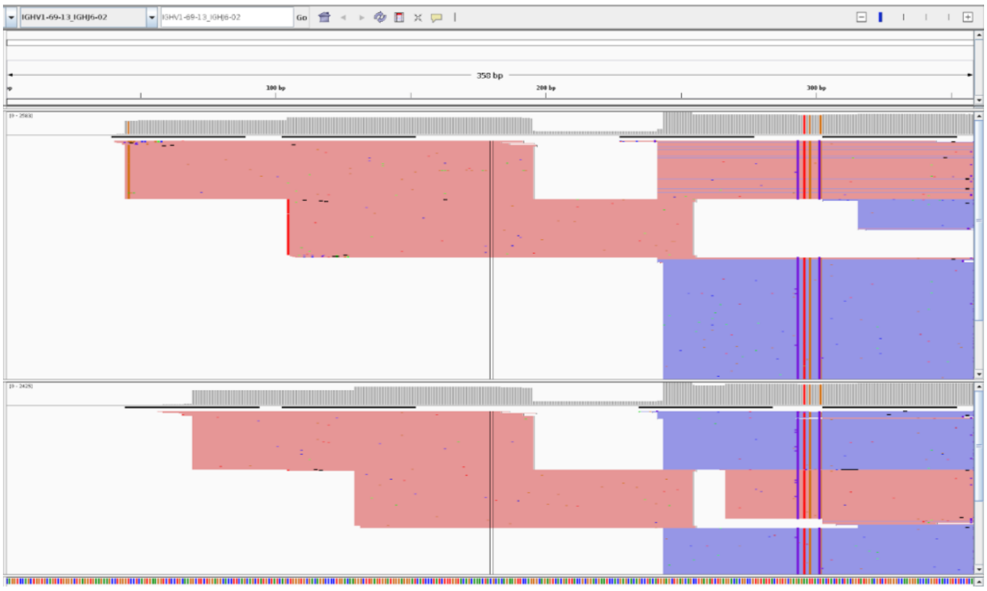


Figure 4.9. Clonal sample prior (up) and after (down) primer sequences removal by trimming in the preprocessing steps. *Visualization shown at the BAM file step, where reads have been assigned to a rearrangement corresponding to IGHV-IGHJ alleles pairing.*

The in-house QC program calculates primer content in different orientations in the steps previous and posterior to trimming primer sequences. Before trimming, 1039576 (899090 R1 + 140486 R2) and 1136184 (952731 R1 + 183453 R2) total primer sequences are found within reads in the polyclonal and clonal samples, respectively. In FASTQs R1, forward *IGHV* primer sequences are found exclusively on the left side of the read, whereas for FASTQs R2, forward primers are found on the right side in reverse complement. After removing primer sequences from reads using bbduk, 51815 and 63335 primer sequences remain in both reads FASTQ files for the polyclonal and clonal samples, respectively (20 and 18-fold lower than the raw files) (Figure 4.10). We can observe that the diversity in the primer sequences encountered is higher in the case of the polyclonal.

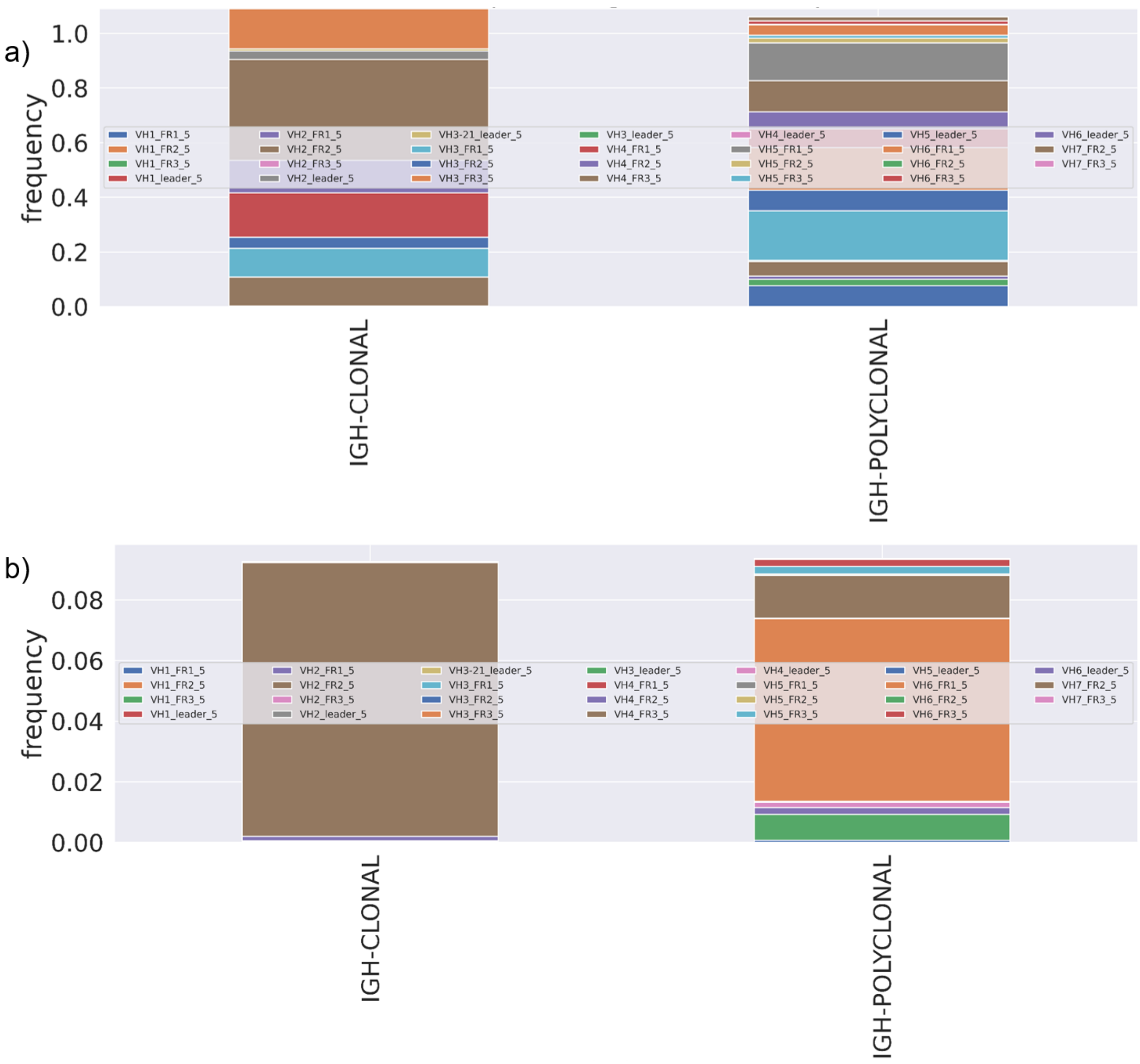


***Figure 4.10. Primers content frequency**. Example clonal and polyclonal samples before trimming in a), and after removing primers sequences in b). The stacked bar plots show the abundance of each IGHV primer grouped by family (IGHV1-7) and fragment (FR1-3).*

4.4.2 Artifact filtering

After alignment against IMGT alleles database, soft clipped reads mapping before coordinate 200 are filtered out from *IGHV* BAM files with BMyRepCLL. These reads are avoided on the coordinate interval where FR3 forward reads map against *IGHV* region, minimizing allele miscalls. Most times these reads mapped to a different *IGHV* allele than FR1/FR2 fragments, producing mapping artifacts that are solved employing this filter (Figure 4.11).

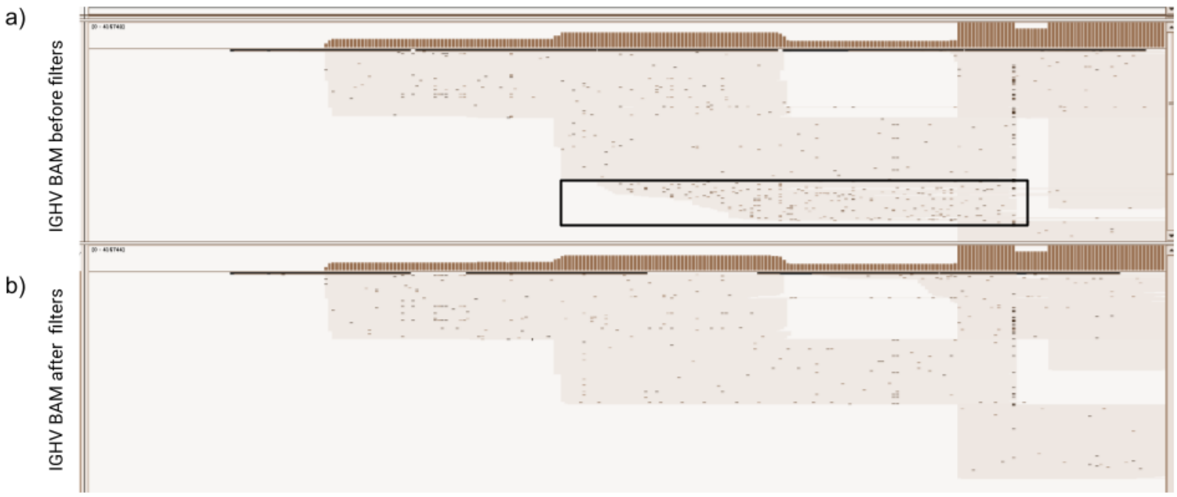


Figure 4.11. FR3 reads artifact filter in BAM files. *Reads in a) show higher mutational noise. This corresponds to FR3 R1 reads belonging to a different IGHV allele. After filtering by the cigar BAM field in b, noise is removed, minimizing IGHV incorrect assignments.*

4.4.3 Gene usage

4.4.3.1 Probable regions

Mapping reads counts for each allele are annotated in intermediate files as in the example in Table 4.4. At this step reads are counted by fragment using mapping coordinates from the reference alleles used (example shown in Figure 4.12).

sample name	region	reads mapped	region length	reads leader	reads FR1	reads FR2	reads FR3
IGH-CLONAL	IGHV1-18*02	12	276	0	2	9	1
IGH-CLONAL	IGHV1-18*03	6	296	0	2	3	1
IGH-CLONAL	IGHV1-18*04	173	296	0	127	5	41
IGH-CLONAL	IGHV1-2*01	9	296	0	7	0	2
IGH-CLONAL	IGHV1-2*02	18	296	0	13	0	5
IGH-CLONAL	IGHV1-2*03	16	296	0	0	0	16
IGH-CLONAL	IGHV1-2*04	40	296	0	29	3	8

Table 4.4. Example of probable_Vregions.csv (first 7 rows), showing read count per allele and fragment*. Since the standard library preparation method used does not include Leader primers, 0 reads appear in that column. The same table for IGHJ alleles, does not contain this distinction, only the total number of read counts.*

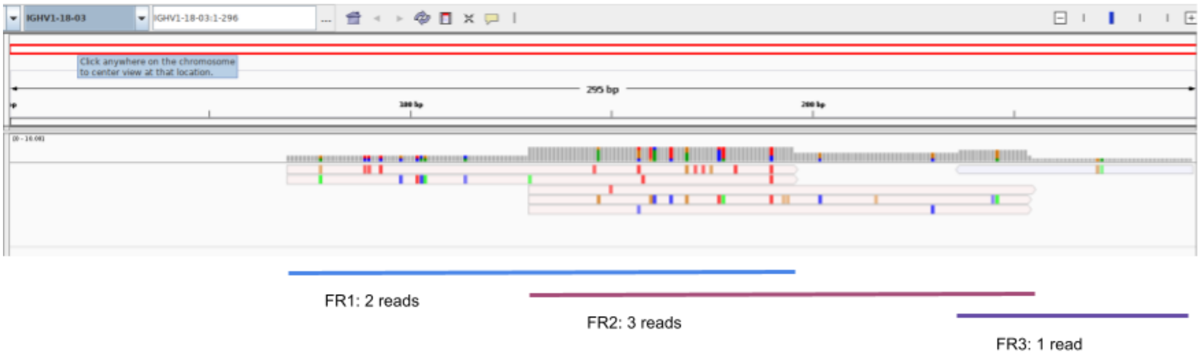


*Figure 4.12. Example with a minority allele to illustrate the number of sequencing reads grouped by fragment mapped against IGHV alleles: IGHV1-18*03 allele with number of reads represented in Table 4.4, with 2, 3 and 1 reads in FR1, FR2 and FR3 amplicons, respectively.*

Visualization with software IGV (Integrative Genomics Viewer) of the most represented *IGHV* and *IGHJ* alleles in BAM files can be used for inspection using the pipeline IMGT alleles FASTA files as reference (Figure 4.13).

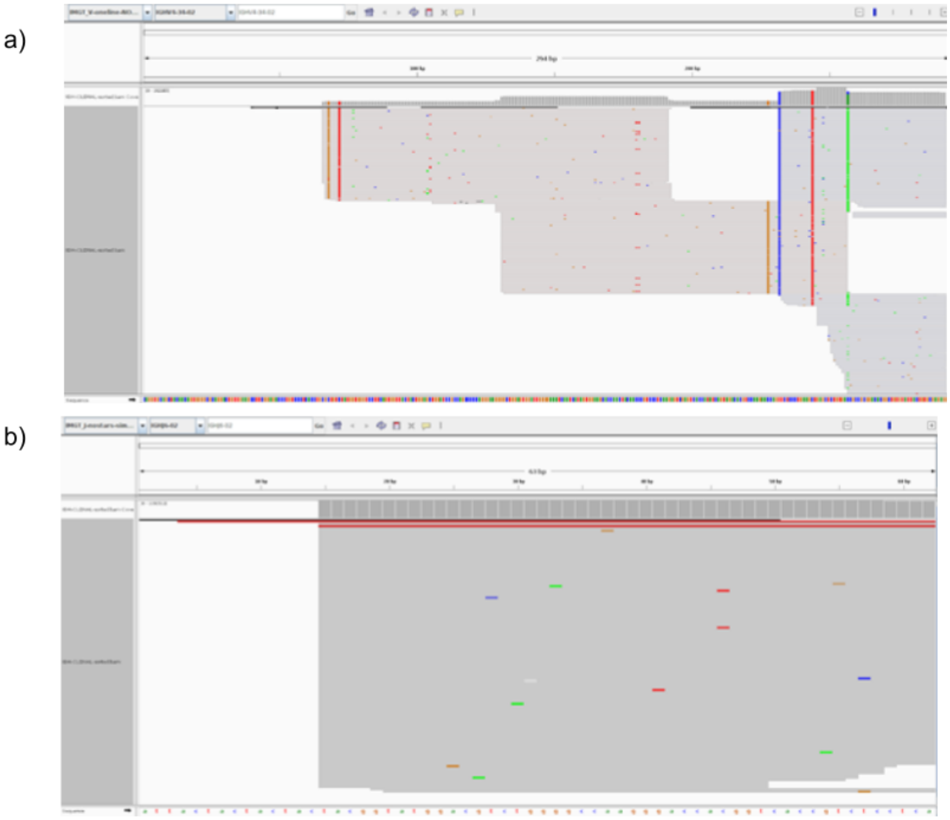


*Figure 4.13. IGV inspection of the most represented alleles in the commercial clonal sample a) IGHV4-34*02 b) IGHJ6*02.*

4.4.3.2 Final report

In Table 4.5, final rearrangements reported by BMyRepCLL for the clonal control are shown. 99.12% reads are assigned to *IGHV4-34* gene (Figure 4.14a.; allele *IGHV4-34*01*), coming from the combined alleles *IGHV4-34*01, IGHV4-34*03, IGHV4-34*04, IGHV4-34*06, IGHV4-34*07, IGHV4-34*08, IGHV4-34*09, IGHV4-34*10, IGHV4-34*11* and *IGHV4-34*12*. There are 673572 reads assigned to the major *IGHJ* allele: *IGHJ6*02*, and 143992 of them (21.4%) overlap with *IGHV4-34*01* allele. The same *IGHJ* allele is assigned by joined reads to the rest of IGHV4-34 gene variant alleles. 0.88% of *IGHV* reads are assigned to noise genes, and the most predominant among them is represented by 0.32% of rearrangement-assigned reads. Additionally, 2/3 of the rearrangements tagged as subclonal in Table 4.5, share the same CDR3 with the major rearrangement, showing evidence of unspecific *IGHV* mapping in clones reported at very low proportions. 100X coverage breadth is 100%, as illustrated in Table 4.5 (coverage information is annotated for clonal rearrangements only). *IGHV* gene usage in the polyclonal sample counterpart shows a completely different clonal profile (Figure 4.14).

The final consensus sequence from the major rearrangement reported in Table 4.5 has been inserted in IMGT/V-QUEST software tool to prove fidelity in the results given *IGHV-IGHJ* assignments, CDR3 sequence and the mutational status, which is reported as 97.9% by both programs (97.89% in the case of IMGT/V-QUEST) (Figure 4.15).

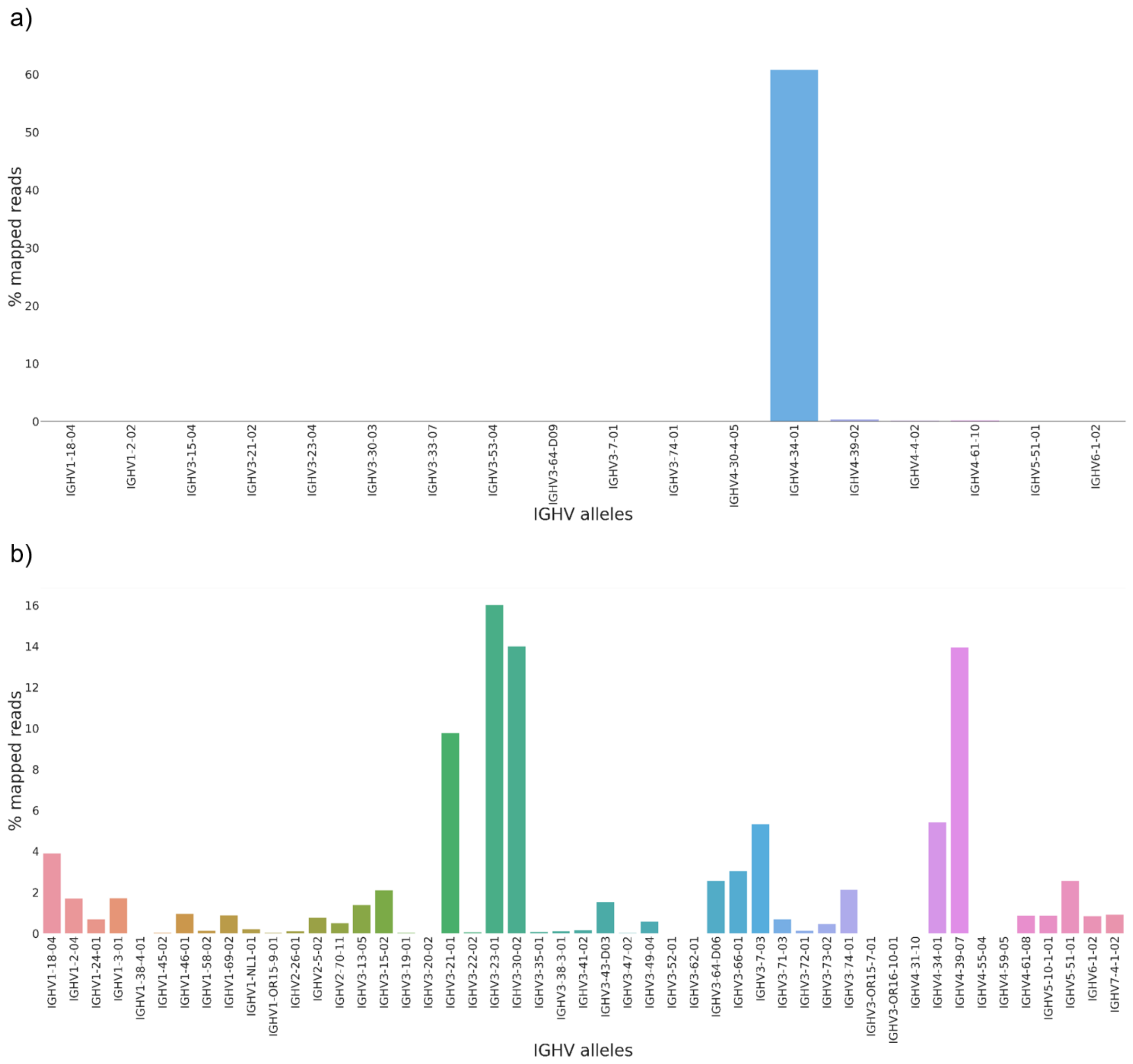
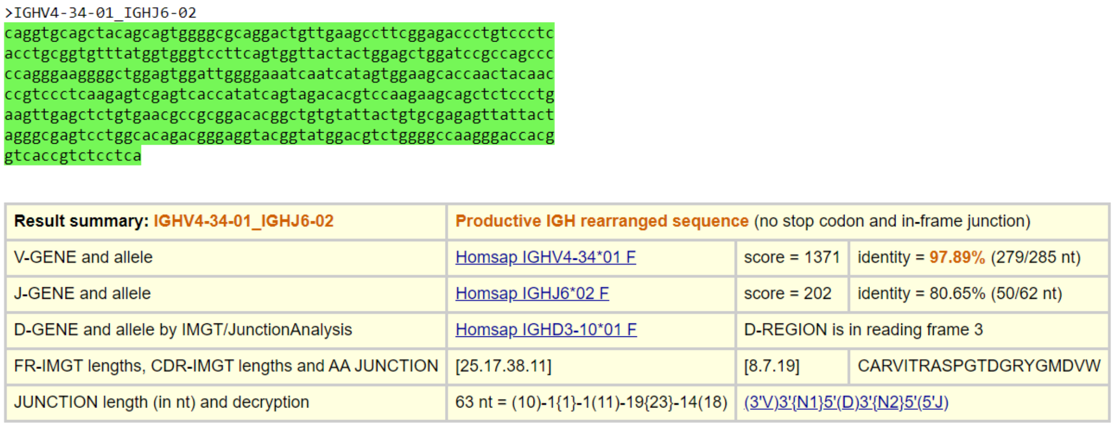


Figure 4.14. Gene usage shown after merging IGHV genes and alleles by identity %. *Clonal profile in a) and polyclonal profile in b).*



```
>IGHV4-34-01_IGHJ6-02
caggtgcagctacagcagtggggcgcaggactgttgaagccttcggagaccctgtccctc
acctgcggtgtttatggtgggtccttcagtggttactactggagctggatccgccagccc
ccagggaaggggctggagtggattggggaaatcaatcatagtggaagcaccaactacaac
ccgtccctcaagagtcgagtcaccatatcagtagacacgtccaagaagcagctctccctg
aagttgagctctgtgaacgccgcggacacggctgtgtattactgtgcgagagttattact
agggcgagtcctggcacagacgggaggtacggtatggacgtctggggccaagggaccacg
gtcaccgtctcctca
```

Result summary: IGHV4-34-01_IGHJ6-02	Productive IGH rearranged sequence (no stop codon and in-frame junction)		
V-GENE and allele	Homsap IGHV4-34*01 F	score = 1371	identity = 97.89% (279/285 nt)
J-GENE and allele	Homsap IGHJ6*02 F	score = 202	identity = 80.65% (50/62 nt)
D-GENE and allele by IMGT/JunctionAnalysis	Homsap IGHD3-10*01 F	D-REGION is in reading frame 3	
FR-IMGT lengths, CDR-IMGT lengths and AA JUNCTION	[25.17.38.11]	[8.7.19]	CARVITRASPGTDGRYGMDVW
JUNCTION length (in nt) and decryption	63 nt = (10)-1{1}-1(11)-19{23}-14(18)	(3'V)3'{N1}5'(D)3'{N2}5'(5'J)	

Figure 4.15. Confirmation with IMGT/V-QUEST*. To double check rearrangement information, consensus sequences obtained after analyses with BMyRepCLL pipeline can be inserted in the reference software: IMGT/V-QUEST web page. Here, results display is shown for the major rearrangement consensus sequence in the clonal control sample.*

Sample name	IGHV allele	joined alleles	reads mapped	clonal%	alignment IGHV	mutational status	IGHV-J	CDR3	IGHD	predicted status	clone status	Fraction covered 100X
IGH-CLONAL	IGHV4-34*01	IGHV4-34*01;IGHV4-34*03;IGHV4-34*04;IGHV4-34*06;IGHV4-34*07;IGHV4-34*08;IGHV4-34*09;IGHV4-34*10;IGHV4-34*11;IGHV4-34*12	539977	99.12	Length: 292; Identity: 286/292 (97.9%)(valid)	MM	IGHV4-34*01_IGHJ6*02	CARVITRASPGTDGRYGMDVW	IGHD3-16*01 75.0 12 33.0 IGHD1-1*01 76.9 13 32.0 IGHD3-16*02 87.5 8 31.0	1CLONE	CLONAL	1
IGH-CLONAL	IGHV4-39*02	IGHV4-39*02;IGHV4-39*01;IGHV4-39*03;IGHV4-39*06;IGHV4-39*07;IGHV4-39*08;IGHV4-39*09;IGHV4-31*03;IGHV4-38-2*02;IGHV4-59*12;IGHV4-61*08	2847	0.32	Length: 298; Identity: 291/298 (97.7%); gaphomology: 98.66(valid)	UM	IGHV4-39*02_IGHJ6*02	CARVITRASPGTDGRYGMDVW	IGHD3-22*01 69.2 26 62.5 IGHD3-10*01 50.0 44 60.0 IGHD3-10*02 72.7 22 57.0	1CLONE	SUBCLONAL	-
IGH-CLONAL	IGHV4-4*02	IGHV4-4*02;IGHV4-4*04;IGHV4-4*05;IGHV4-4*10;IGHV4-28*06;IGHV4-38-2*02;IGHV4-55*08;IGHV4-OR15-8*02	984	0.20	Length: 295; Identity: 289/295 (98.0%); gaphomology: 98.64(valid)	UM	IGHV4-4*02_IGHJ6*02	CARVITRASPGTDGRYGMDVW	IGHD3-22*01 69.2 26 62.5 IGHD3-10*01 50.0 44 60.0 IGHD3-10*02 72.7 22 57.0	1CLONE	SUBCLONAL	-
IGH-CLONAL	IGHV4-61*10	IGHV4-61*10;IGHV4-61*01;IGHV4-61*02;IGHV4-61*03;IGHV4-61*06;IGHV4-61*07;IGHV4-61*11;IGHV4-28*03;IGHV4-30-2*05;IGHV4-31*11;IGHV4-38-2*01;IGHV4-39*05;IGHV4-4*09;IGHV4-59*02	1417	0.16	Length: 298; Identity: 291/298 (97.7%); gaphomology: 98.32(valid)	UM	IGHV4-61*10_IGHJ6*02	not calculated		1CLONE	SUBCLONAL	-

Table 4.5. Final report for the clonal commercial sample. *B cell rearrangements (clonal and subclonal), reported after BMyRepCLL analysis of the clonal commercial sample, reporting a single high-proportion clone with great specificity. Noise B cell background accounts for 0.88% of total reads assigned to IGHV rearrangements.*

4.5 CLL Immcantation adjustment

4.5.1 Clonal threshold tuning

For the definition of a threshold to infer clonal relationships, a first round of 238 samples was analyzed. The *distToNearest* function, which performs calculation of the distance between the sequences in a dataset and their nearest-neighbor, was employed with the Hamming distance method. The histogram plotted after the *distToNearest* output dataframe, following the Immcantation documentation, showed similar distributions among the CLL patients tested. No bimodal profile as seen in the examples provided in the documentation of their methods was observed (Figure 3.14), and therefore automatic threshold definition could not be performed (Figure 4.16). Clonal relationships of B cell expansions were very narrow and for that reason, a threshold of 0.1 was chosen, being an intermediate value between clonal relationships found within samples and among different samples.

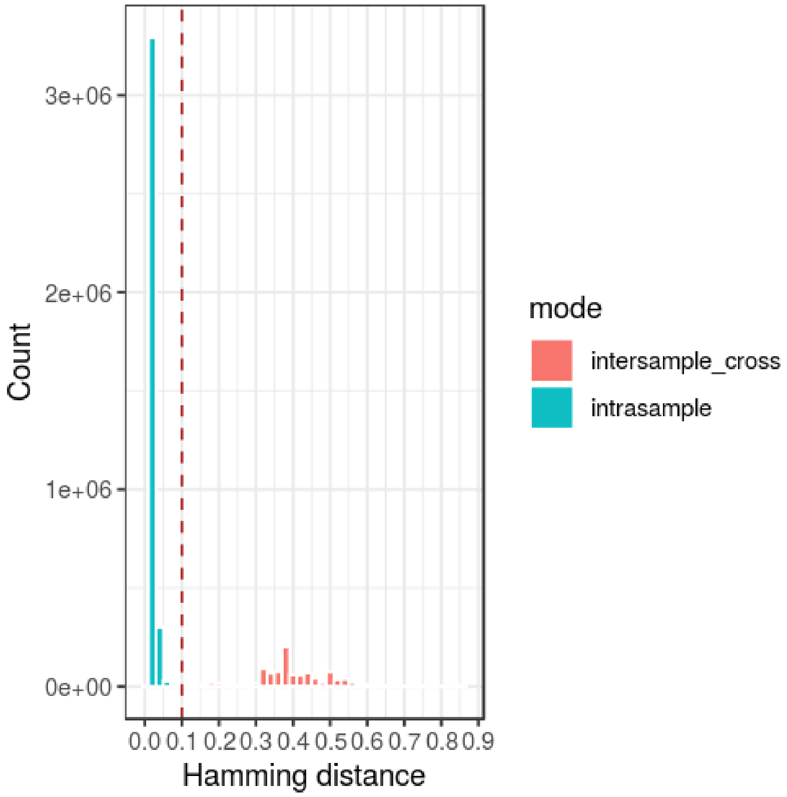


Figure 4.16. distToNearest dataframe histogram from 238 CLL patients plotted together. *Individual profiles were similar. Here, distances in blue represent hamming distances between sequences in the same sample, and distances in red are the cross validation between samples (distances between sequences in the CLL sample against a polyclonal). The majority of sequences within a sample correspond to a clonally expanded B cell and thus, the distances are very small, which leads to a wide gap between clonal and non-clonal relationships. The threshold was chosen in between the intersample and intrasample profiles (vertical red discontinuous line).*

4.5.2 FR3 sequences noise

Clone definition was performed with a 0.1 distance threshold for the *defineClones* function. After inspecting clone plots, it was noticeable that mutation frequencies were variable in some cases in FR3 sequences, due to sequence length differences (Figure 4.17 shows much more unmutated sequences in M-CLL examples 1 and 2, and differentially mutated in the U-CLL example 3). Other cases like example 2a in Figure 4.17, also harbored whole-FR3 clones, in which FR3 sequences clustered independently into clones due to *IGHV* gene calls that were more ambiguous in this fragment for covering a small fraction of *IGHV* gene segments. Clone average mutation frequencies are found distorted in FR3 fragment (Figure 4.18) and thereafter, it was decided to filter out these reads before the *defineClones* step to avoid biased calculation of mutation frequencies.

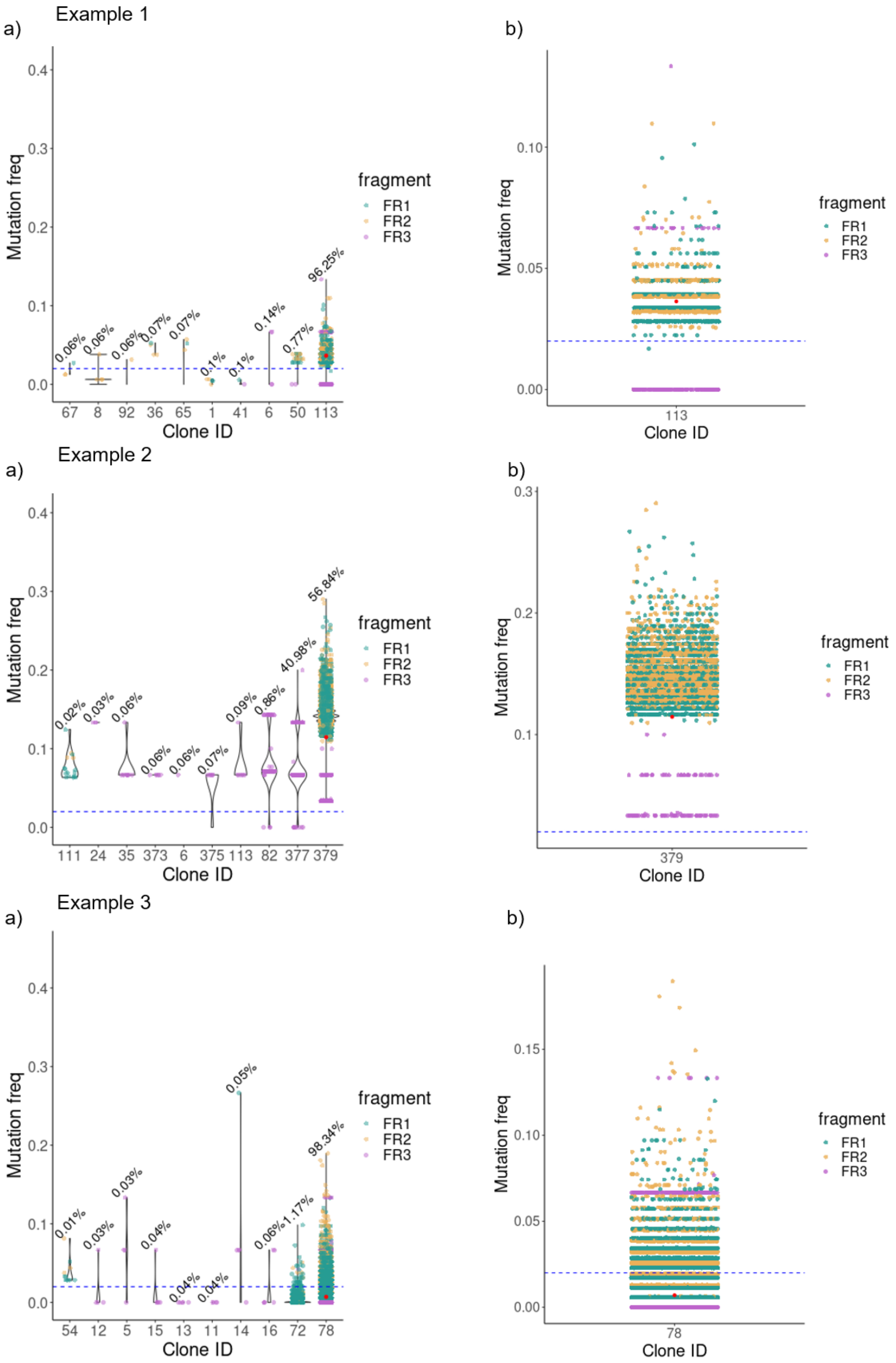


__Figure 4.17. Three different examples of CLL samples whose major clones have notable differences in mutational frequencies distribution caused by FR3 fragment reads.__ a) Represents the top 10 clones in these samples with mutation frequencies distributions. SSeq is represented with a red dot, clustered in the predominant clones in the 3 examples. b) Shows mutation frequencies distribution on the predominant clone, differentiated by FR regions reads. Discontinuous y-axis blue line represents mutational status frequency threshold (0.02).

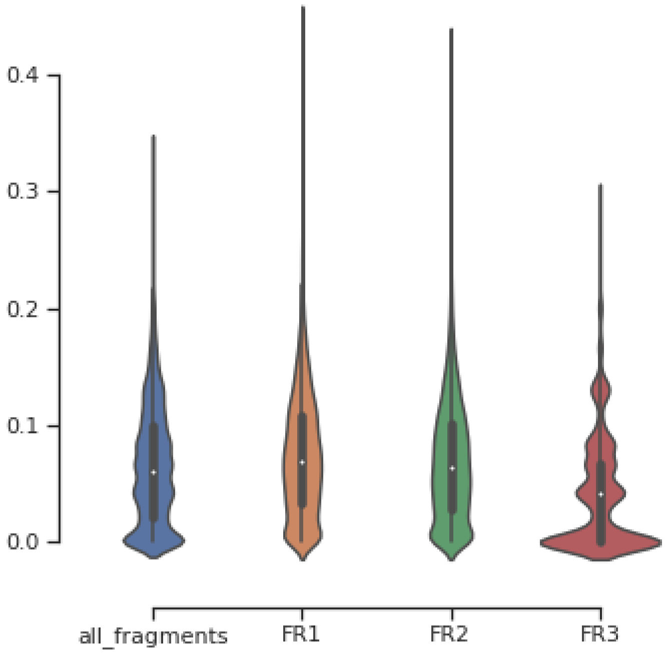


Figure 4.18 Average mutation frequencies in the major clone by fragment.

4.5.3 Sequencing artifact filtering: elimination of sequencing noise

The last steps of the CLL Immcantation pipeline included minimizing sequencing artifacts, by ignoring variants below 0.02 read proportion (a variant must be represented at least in 2% of the reads to be considered real). Figure 4.19a shows the presence of a single clone in the clonal commercial sample, which confirms that clustering methods used worked accordingly. By performing low frequency variants filtering after cloning steps, the range in mutation frequencies was narrower (Figure 4.19b).

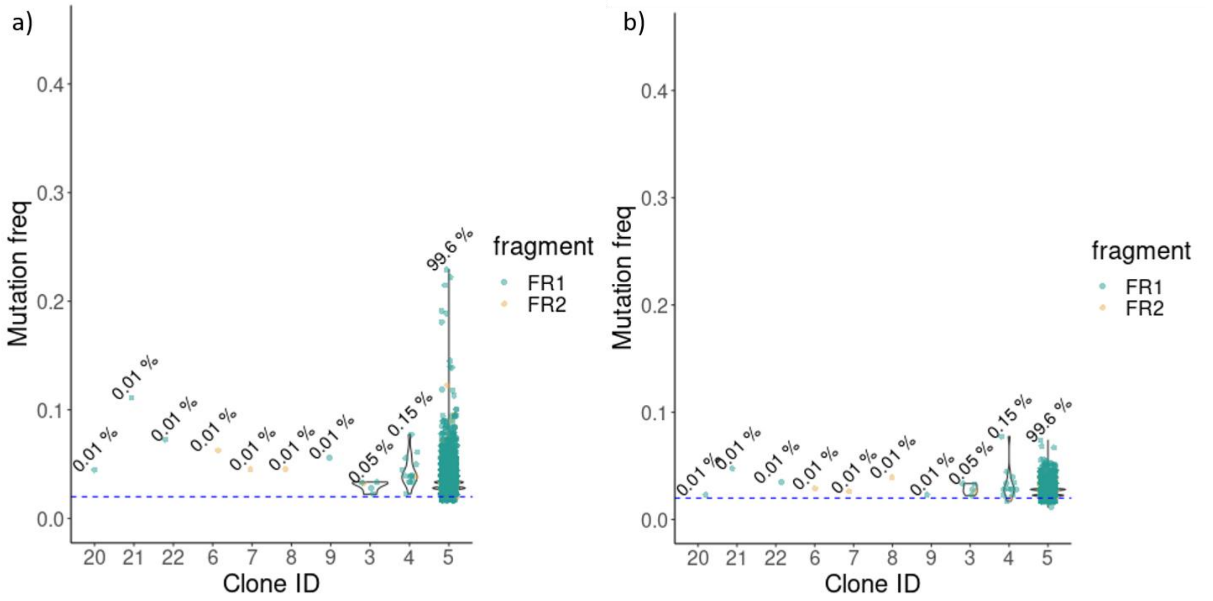


Figure 4.19. Top 10 clone plots with mutation distributions in the clonal commercial sample. a) without performing low frequency variants filters. b) after performing such filters.

4.5.4 Validation using SSeq sequences

To evaluate the clustering methods employed with CLL Immcantation pipeline, SSeq sequences from 238 CLL patients were included for clone definition using FR1 and FR2 reads. In 85% of the samples, SSeq sequences clustered together with NGS clones in rank 1 (205/238). There are no Sanger sequences grouped in separate clones. 22/29 SSeq sequences grouped in secondary rearrangements are grouped in rank 2 (Table 4.6).

Count	Clone rank
205	1
22	2
3	4
4	5
1	6
1	9
1	14
1	26

Table 4.6. Counts of NGS clone rank SSeq sequence clustering.

4.5.5 Performance with different patient profiles

After performing analyses with the Immcantation pipeline, characterizing patient profiles is straightforward by inspecting the figures that are generated in the last steps, specifically implemented for rapidly determining the characteristics of CLL clones. In Figure 4.20, example of UM, MM, BD and double clonal profiles as a comparison.

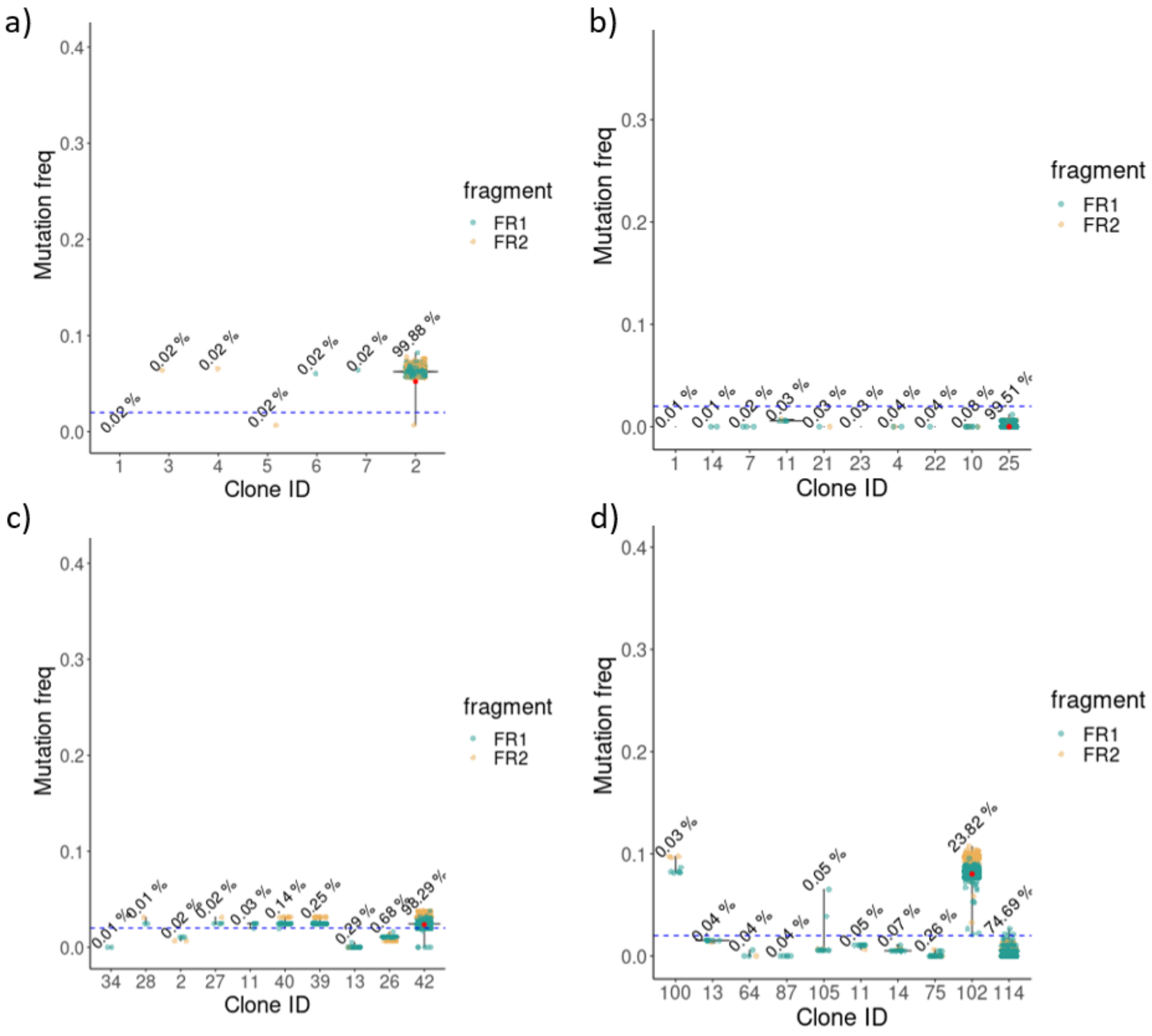


Figure 4.20. Examples of top 10 clone plots from CLL samples analyzed with CLL Immcantation*. If the SSeq sequence is included within the defineClones step, it will be represented with a red dot overlaid to the NGS sequences. a) Single clone; MM. b) Single clone; UM. c) Single clone; BD. d) Double productive clone; UM and MM.*

4.5.6 Mutation frequencies distributions in the different mutational statuses

After clone plot inspection, we found wide mutation frequencies distributions in some cases, in spite of the previous filtering of low proportion variants. Among these, we found cases with highly mutated MM clones, even in secondary clones (Figure 4.21) and more interestingly, cases of predominant UM clones with sequences spanning to the mutated frequencies threshold (> 0.02) (Figure 4.22). Even though different groups of sequences in that clones held mutation frequencies up to 0.1 approximately, the average identity percentage against germline *IGHV* alleles was coincident with SSeq. Identity percentages determined by CLL Immcantation were 100% in sample example 1 (Figure 4.22ab) and 99.65% in sample example 2 (Figure 4.22cd).

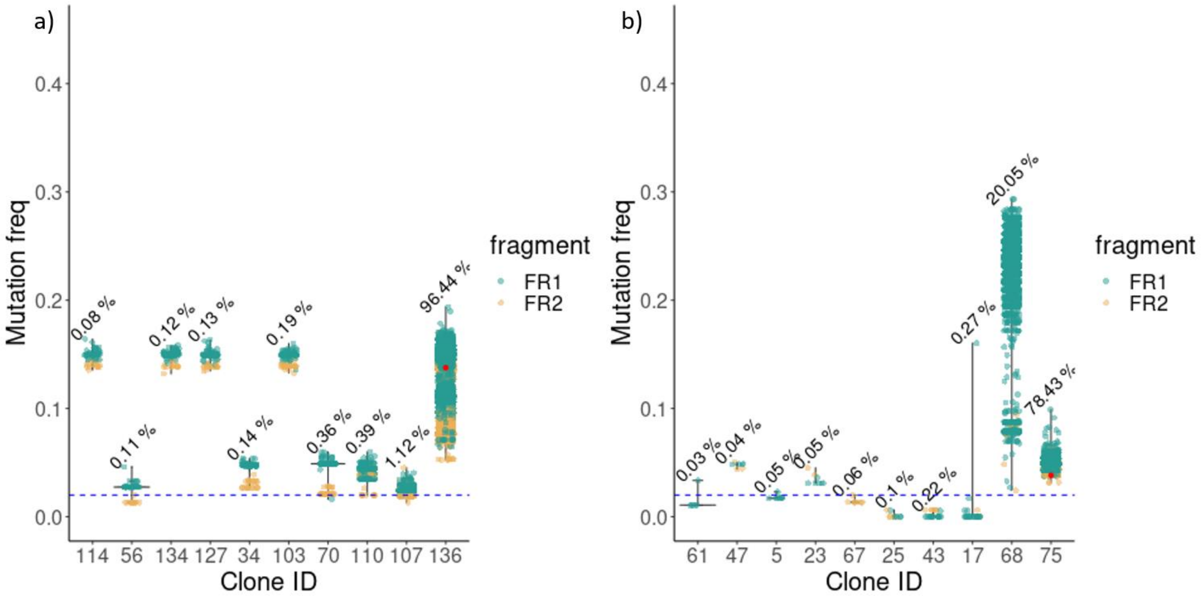


Figure 4.21. Cases of MM-CLL samples with wide mutation distributions*. a) CLL sample top 10 clone plot with a single predominant clone represented by 96.44% of reads and classified as MM, with mutation frequencies surpassing 0.15 in a group of sequences. IGHV gene is IGHV1-2. b) CLL sample top 10 clone plot, with double MM rearrangements. The second clone is highly mutated, with groups of sequences between 0.2 and 0.3 mutation frequencies. IGHV genes are IGHV3-30 and IGHV1-2.*

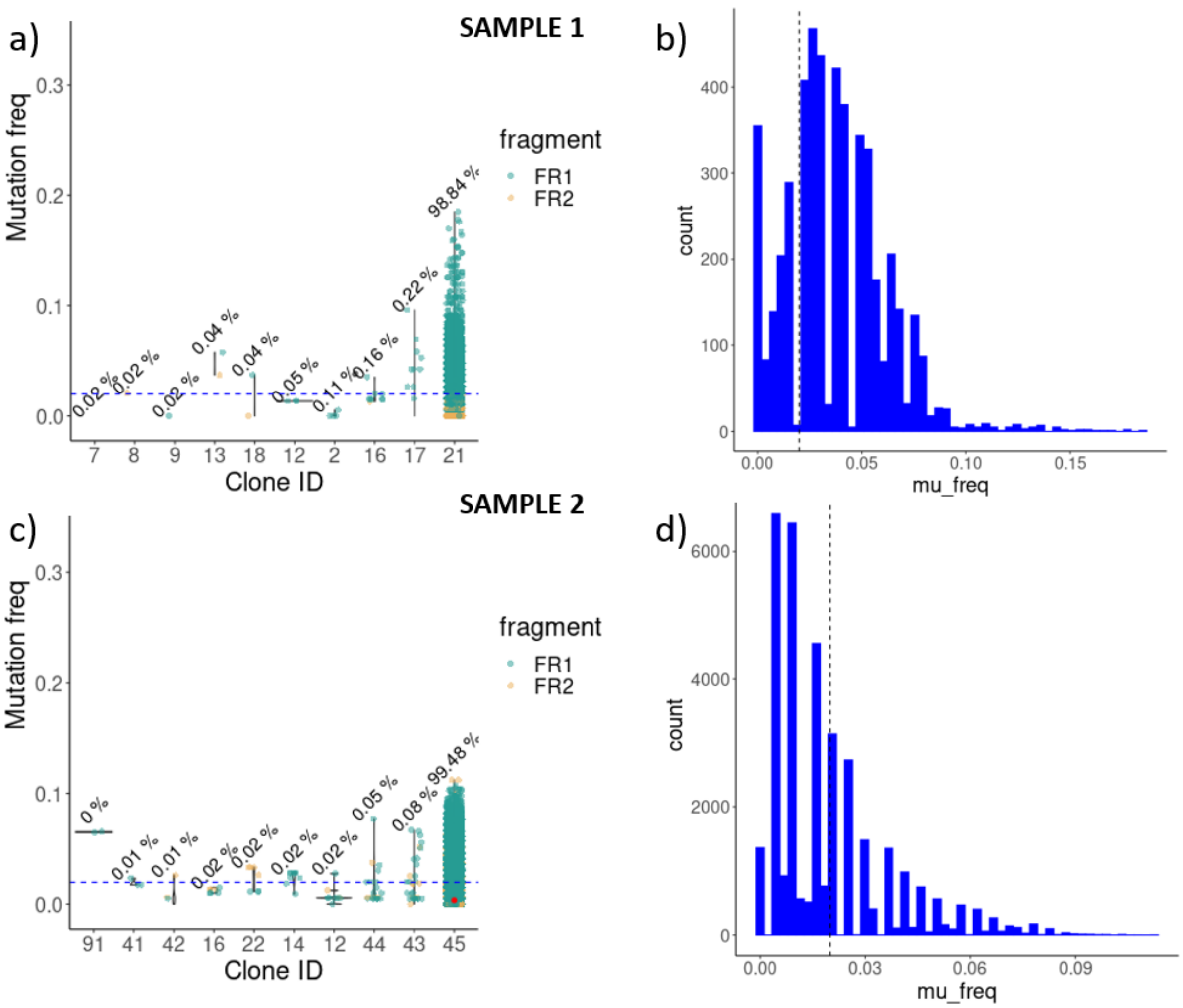


***Figure 4.22. Cases of UM-CLL samples with wide mutation distributions**. a) Top ten clones plot for sample 1 example with UM rearrangement with wide mutation frequencies distribution. Single clone represented by 98.84% of reads with IGHV4-39 gene. b) Mutation frequencies histogram for sample 1 example. Discontinuous black line represents 0.02 frequency threshold. c) Top ten clones plot for sample 2 example with UM rearrangement with wide mutation frequencies distribution. Single clone represented by 99.48% reads and IGHV3-30 gene. d) Mutation frequencies histogram for sample 2 example. Discontinuous black line represents 0.02 frequency threshold.*

To prove the level of expansion of mutation frequencies considering the whole set of sequences in the predominant clones, samples with a single predominant clone whose mutational status was accordingly characterized regarding SSeq results, were plotted to compare mutation frequencies distributions among the UM, MM and BD patient groups. 5 BD patients stuck to this criteria, whereas UM and MM groups contained 122 and 65 samples, respectively. The distributions observed in Figure 4.23a, showed that mutation profiles in sequences of BD predominant clones are intermediate between MM and UM CLL clones (average mutation frequencies 0.026 ± 0.004 for BD, 0.073 ± 0.029 for MM and 0.003 ± 0.0065 for UM clones; p.values Mann-Whitney-Wilcoxon with Bonferroni correction: <0.0001 UM vs MM, <0.0001 UM vs BD, <0.0001 MM vs BD). The majority of

BD clones had mutation frequencies in the borderline margins, with mutation frequencies between 0.02 and 0.03. Therefore, the tendency showed significant differences among the mutational frequencies distributions following the classification determined by the clone average mutations (log2 fold change for MM vs UM=4.6; log2 fold change for MM vs BD=1.45).

CDR3 lengths distributions were also significantly different between groups considering the whole set of sequences among predominant clones. Average values shown are 41.69 ± 4.68, 49.06 ± 9.62 and 61.14 ± 9.73 respectively for BD, MM and UM groups (Figure 4.23b) (p.values Mann-Whitney-Wilcoxon with Bonferroni correction: <0.0001 UM vs MM, <0.0001 UM vs BD, <0.0001 MM vs BD) (log2 fold change for UM vs MM=0.32; log2 fold change for MM vs BD=0.23, log2 fold change for UM vs BD=0.55).

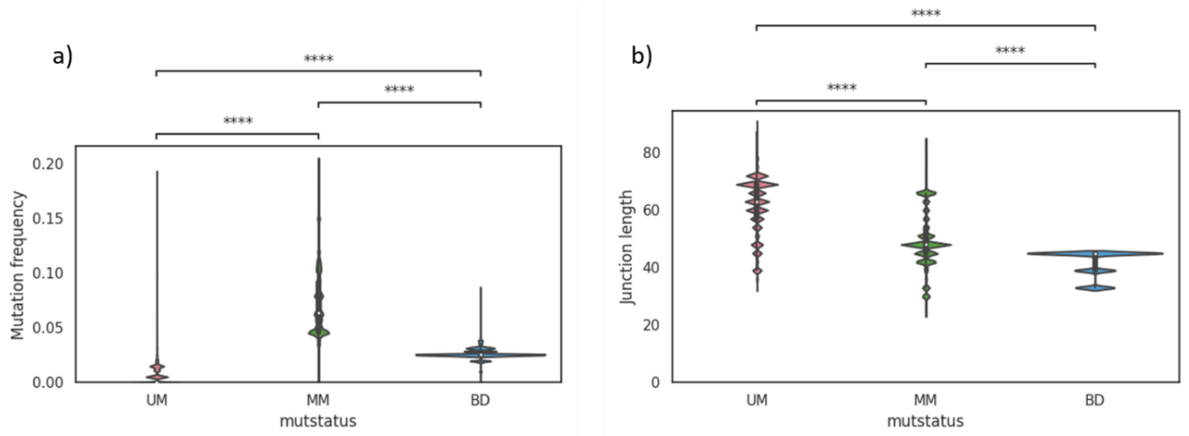


*Figure 4.23. Mutation frequencies and CDR3 length pairwise comparisons among mutational status classification groups. a) Mutation frequencies distributions in all sequences of predominant CLL clones from UM, MM and BD groups and b) CDR3 length distributions among the predominant CLL clones sequences for UM, MM and BD stratification. (ns: p <= 1.00e+00; *: 1.00e-02 < p <= 5.00e-02; **: 1.00e-03 < p <= 1.00e-02; ***: 1.00e-04 < p <= 1.00e-03; ****: p <= 1.00e-04).*

4.5.7 Clonal quality control in healthy donors vs CLL: repertoire diversity

CLL B cell repertoires are expected to be highly clonal and in fact, most patients present a monoclonal profile, whereas a healthy repertoire is highly diverse in the number of clones represented. The average number of clones is 449.18 ± 336.86 for healthy donors and 162.77 ± 271.85 in CLL samples (p.value for Mann-Whitney-Wilcoxon test with Bonferroni correction: 4.175e-13) (Figure 4.24a). Such significant differences can be observed also in the number of unique *IGHV* genes represented in both groups (average number of *IGHV* genes represented is 37.89 ± 10.39 and 16.74 ± 8.98 for healthy and CLL groups, respectively. p.value: 1.150e-21) (Figure 4.24c), and distribution with tendency to higher clonal percentages in the case of CLL patients and the opposite in healthy donor repertoires (average maximum clonal percentages in healthy donors is 16.64 ± 26.00%, whereas for CLL patients ins 85.45 ± 21.83%; p.value: 9.010e-23) (Figure 4.24b).

Diversity methods are commonly used to dissect the architecture of adaptive immune repertoires, employing classical ecology measures; from richness of different clones (q = 0), to evenness of the clonal population distribution (q = 4). Alpha diversity and abundance calculations show higher abundance (Figure 4.25 and Figure 4.26a), whilst lower diversity (Figure 4.26b) in both CLL groups (MM and UM), than healthy donor groups.

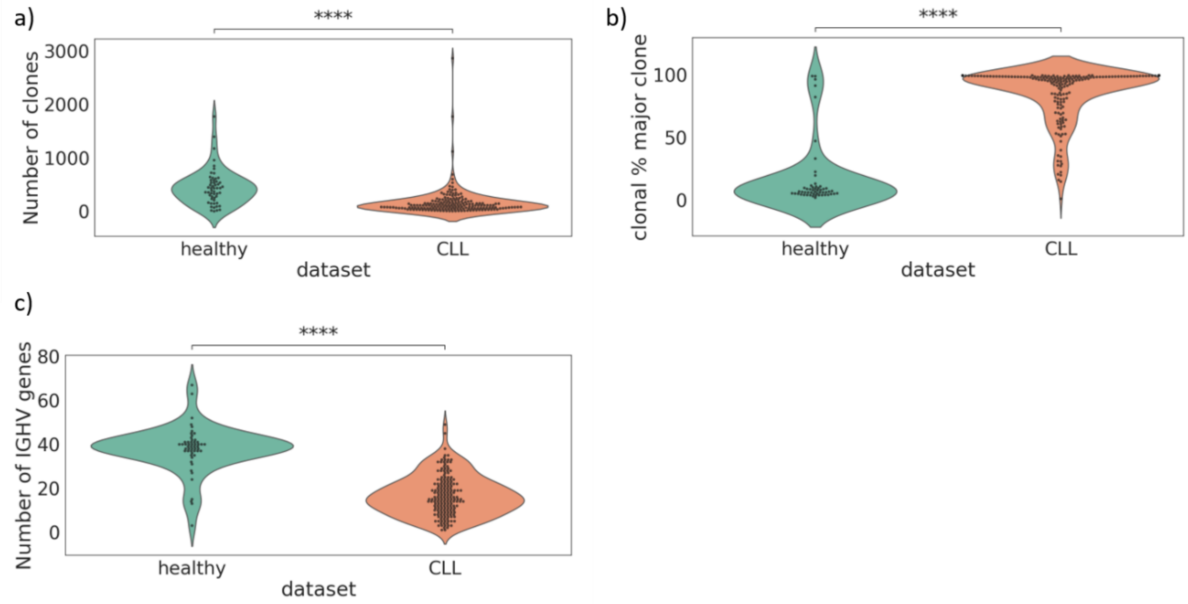


***Figure 4.24. Comparison of repertoire architecture from healthy donor and CLL samples**. a) Number of B cell clones defined after clustering, b) percentages at which the predominant clones per sample are represented and c) number of unique IGHV genes represented. (ns: p <= 1.00e+00; *: 1.00e-02 < p <= 5.00e-02; **: 1.00e-03 < p <= 1.00e-02; ***: 1.00e-04 < p <= 1.00e-03; ****: p <= 1.00e-04).*

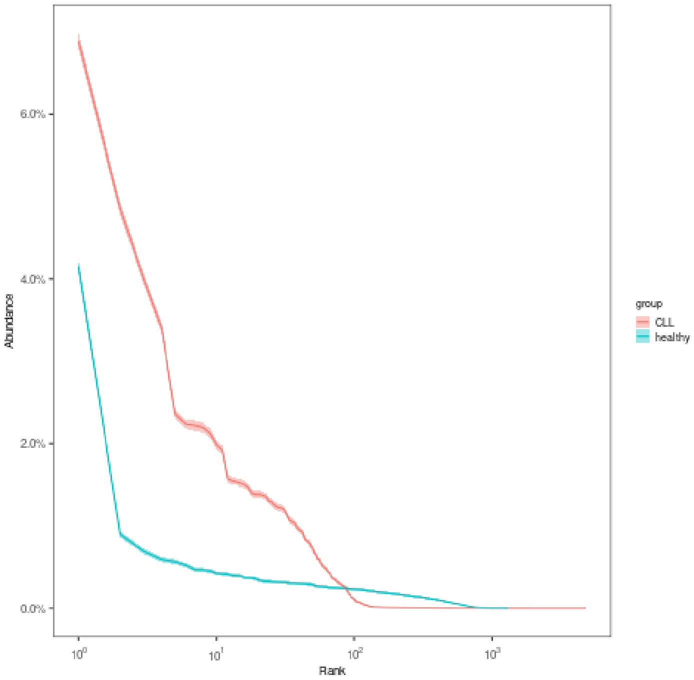


Figure 4.25. calcAbundance Alakazam package output plots. *We can observe notable differences in the abundance of clone ranks, being clone abundance higher in CLL repertoire samples.*

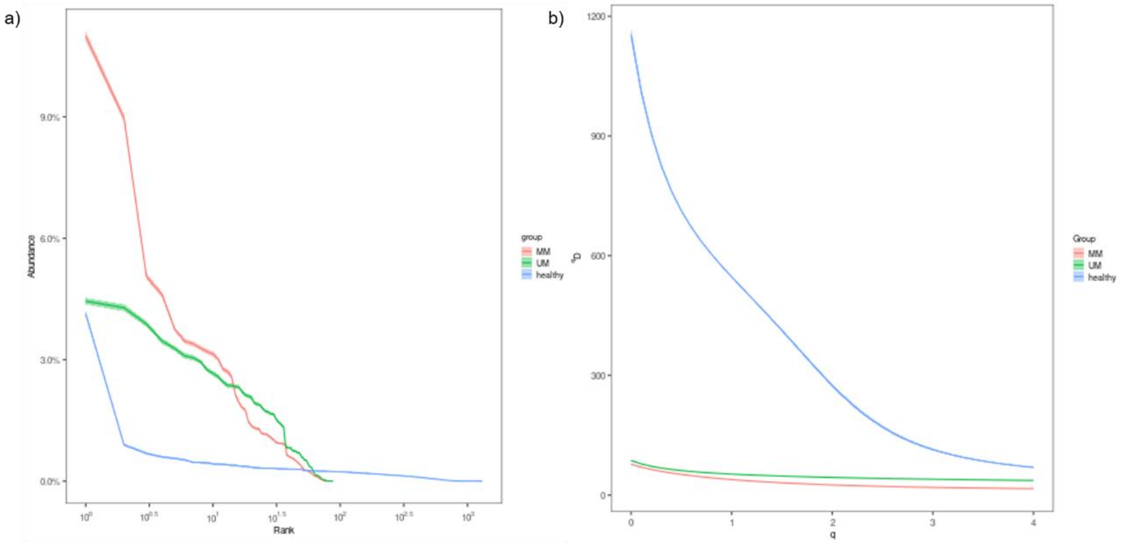


Figure 4.26. Diversity quality control performed with Alakazam calcAbundance and caclDiversity functions. *In a) higher abundance can be observed in MM-CLL repertoires with respect to UM-CLL b) Diversity differences between UM-MM and healthy donors are more notable when giving more weight to species richness (q close to 0), than evenness (q >1).*

4.6 Testing and validation of the primary pipeline with the CLL dataset

General steps to evaluate the performance of BMyRepCLL pipeline were exemplified in section 4.3. However, the different steps of the pipeline were tuned over successive sequencing experiments in order to correctly determine the mutational status and clonal profile of samples from real CLL patients. During the process, discordant results were manually inspected as we encountered exceptions to the rules implemented in the analysis pipeline. Once optimized, we analyzed 300 CLL samples sequenced with the experimental design chosen (4.1, 4.2), for the validation of the methods.

4.6.1 Quality control

After the preprocessing steps performed by BMyRepCLL, quality scores improved over Q30 (Figure 4.27), the quality score desired for performing accurate B cell clone determinations. 314 CLL samples selected for validation passed sequence quality filters after preprocessing, with an average number of paired reads of 28628.44 ± 33759.26. Samples surpassed 1000 reads assigned to the major rearrangement. The average percentage of reads mapped against IMGT *IGHV* alleles was 80.03 ± 17.63%.

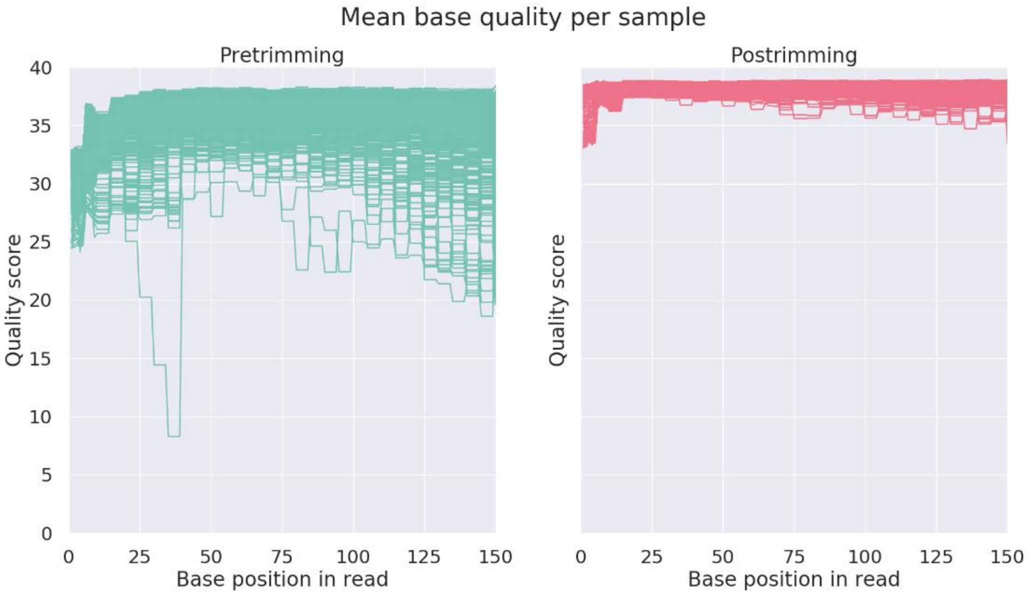


Figure 4.27. Quality scores for preprocessed and postprocessed FASTQ files *(both R1 and R2 per sample are included). Quality values per position in read are parsed from the output of FastQC program.*

4.6.2 Artifact rearrangements filtering

Output table from BMyRepCLL module 1 contained 37002 Ig raw rearrangements for the total 314 CLL samples dataset employed for the test and validation. On average, there are more than 100 rearrangements per sample when mapping reads directly to IMGT *IGHV* alleles and counting how many reads matched each one of them. The number of rearrangements was reduced to 14554 (2.5-fold) when joining rearrangements with common *IGHV* genes sharing 95% of their consensus sequences, and to 11087 (3.3-fold) after subsequently repeating this selection with rearrangements sharing *IGHV* gene families. The next filtering step, reallocates rearrangements supported by unique fragments and reports a rearrangement per *IGHV* gene, indicating the major productive allele (if there is any) within that group. As a consequence, the final output main table contained 1229 total Ig rearrangements. 35773 rearrangements were filtered out, making the use of these steps necessary to avoid tedious manual curations. 955/1229 rearrangements were UM and 274/1229, MM, represented with 59 different *IGHV* genes (Figure 4.28). Some *IGHV* genes, such as *IGHV1-69* and *IGHV4-34/IGHV3-7*, were found mainly as UM and MM clones, respectively.

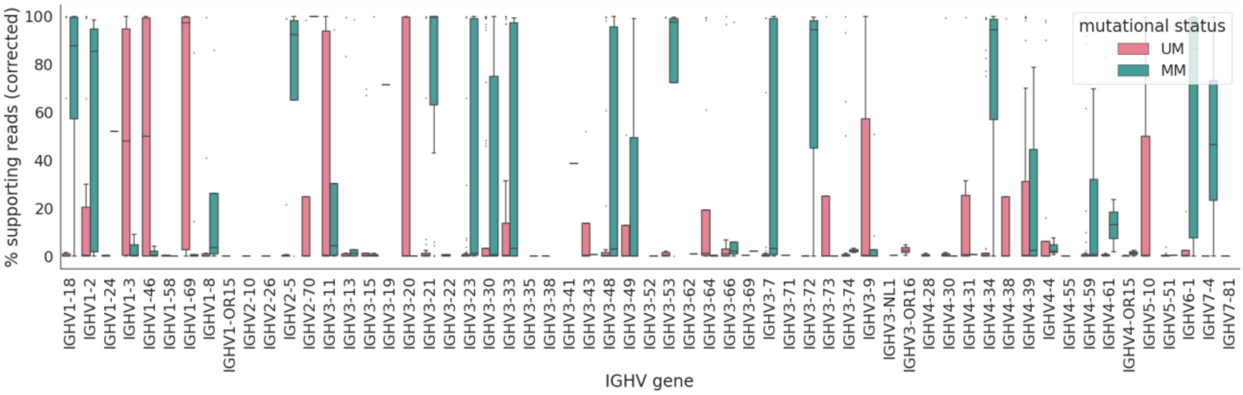


Figure 4.28. Percentage of reads supporting IGHV genes in the 314 CLL samples, grouped by clone mutational status.

4.6.3 Tuning of clonal threshold

After artifact-filtering steps, curated clones were represented by different clonal percentages, including proportions below 10%. In order to classify these clones into clonal and subclonal to make a prioritization of rearrangements in the final report that would serve for clinical determinations, a method for tuning the clonal threshold was advisable.

Tuning threshold and the obtaining of accurate B cell clone characterizations by BMyRepCLL was tested with an initial group of 20 healthy donors and 34 clonal CLL samples. Clonal CLL samples were divided into groups following the number of clones detected by SSeq: 24 single clone cases and 10 cases of double clone profiles. Regarding the 3 groups of samples, the difference ratios between consecutive clones encountered within a sample were calculated, and the maximum difference among them (MAX_DIFF parameter), was used to define a threshold for determining a sample as clonal (polyclonal, otherwise). Average MAX_DIFF value was 137 and 70-fold higher in the 1CLONE and 2CLONE groups compared to the polyclonal group, respectively (Table 4.7). Despite the fact that the minimum MAX_DIFF value found in a clonal sample was 8.19 in a 1CLONE sample, differences between the polyclonal samples group and both the 1CLONE samples and 2CLONE samples after Mann-Whitney-Wilcoxon test were highly significant (p. values with Bonferroni correction: 4.931e-08 and 3.603e-05, respectively), whereas between the clonal groups (1 CLONE and 2 CLONE), differences were not significant (3.243e-01) (Figure 4.29). In Figure 4.30, differences in the MAX_DIFF parameter can be observed between the 3 clonality profiles (a: healthy, b: 1CLONE, c: 2CLONE). The maximum difference (longest negative Y axis bar) is placed in the last clonal rearrangement detected (the second in the case of the 2CLONE examples; Figure 4.30c).

Clonal cutoff was set as MAX_DIFF ≥ 5. If the MAX_DIFF value within a sample was greater than or equal to 5, the sample was considered to have B cell expanded clone(s). If the sample was determined clonal, the same MAX_DIFF value was used to determine the clonal cutoff, so as to distinguish the predominant clone(s) and consider the fraction below it as subclonal.

	average MAX_DIFF	maximum MAX_DIFF	minimum MAX_DIFF
1CLONE	272.26	996.52	8.19
2CLONE	140.46	418.82	34.03
polyclonal	1.98	2.51	1.53

Table 4.7. Test MAX_DIFF values. *Average, maximum and minimum values for the maximum clonal difference within a sample (MAX_DIFF) in the 3 groups tested (1CLONE, 2CLONE and polyclonal).*

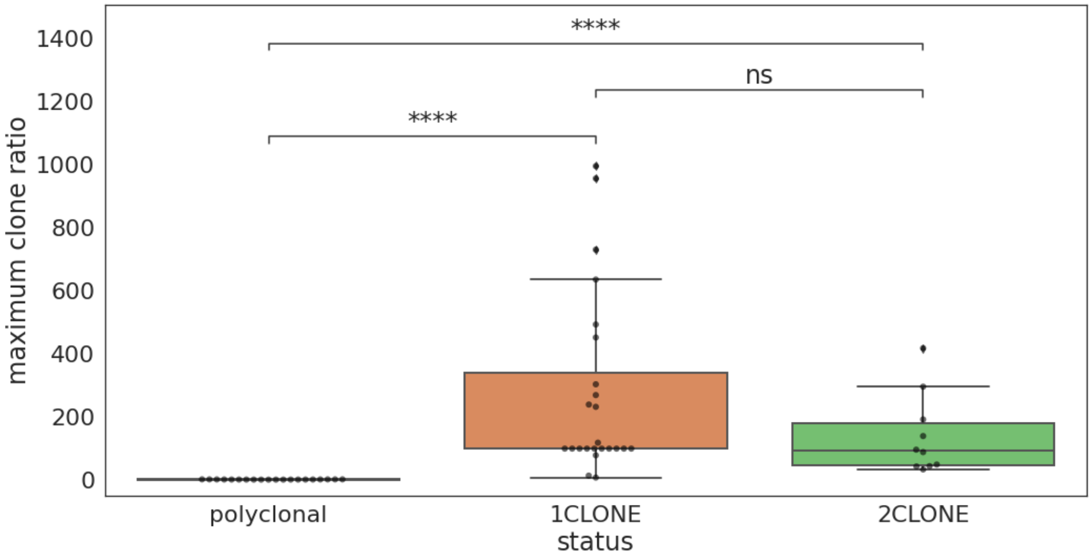


Figure 4.29. Boxplot for MAX_DIFF values per sample grouped by polyclonal, 2CLONE and 1CLONE. *After Mann-Whitney U test, corrected Bonferroni p.values are annotated to show differences between group distributions (4.931e-08 and 3.603e-05, respectively in the comparisons polyclonal-1CLONE and polyclonal-2CLONE; and 3.243e-01 between the 1CLONE and 2CLONE group). (ns: p <= 1.00e+00; *: 1.00e-02 < p <= 5.00e-02; **: 1.00e-03 < p <= 1.00e-02; ***: 1.00e-04 < p <= 1.00e-03; ****: p <= 1.00e-04).*

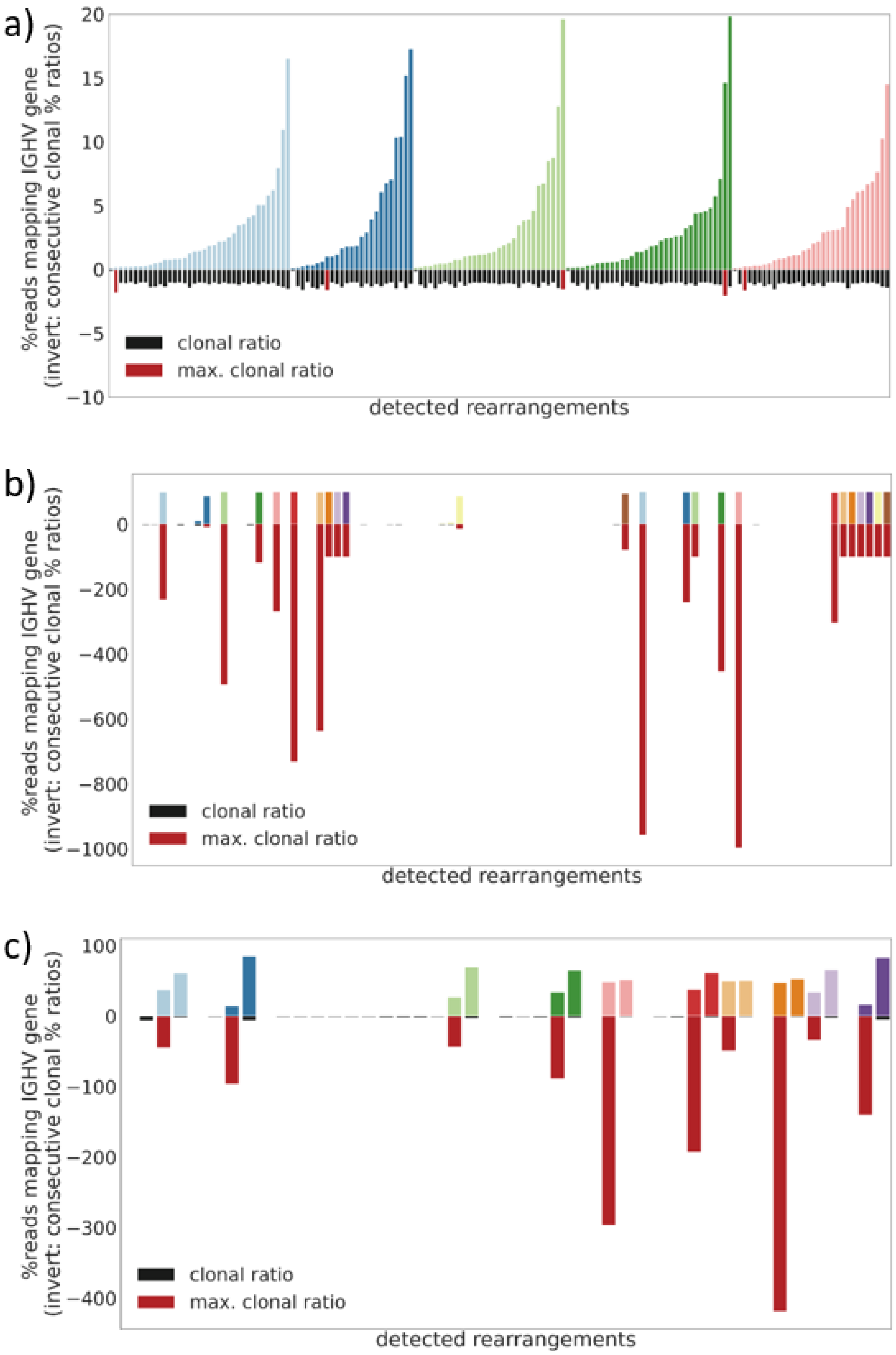


Figure 4.30. Clonal percentage ratios representation in 3 different clonality profiles. *Bars in the positive Y axis represent the clonal percentage of the different Ig rearrangements detected per individual; colors are used to differentiate rearrangements between individuals. Bars in the negative Y axis represent the ratios of the % between the consecutive clones of a sample ordered by abundance. The maximum value within these ratios is colored in red whereas the rest of clonal ratios are colored in black. a) 5/20 healthy donor samples used for the test. b) 24 samples with a single predominant clone used in the test and c) 10 samples with double predominant rearrangements used for the test.*

4.6.3.1 Test clonal classifications

Using the MAX_DIFF parameter to discriminate between the presence of any clonal rearrangements firstly (if the sample was polyclonal or clonal; MAX_DIFF > 5), and secondly, if the first condition was met, how many clonal rearrangements were present (maximum MAX_DIFF value found within a sample as cut-off), a specificity and sensitivity of 100% was obtained in the test, after the comparison of the clonal profile determined by the gold standard SSeq and the predicted status of the NGS pipeline.

Regarding clonal profiles and number of clones, 100% of the test samples were correctly classified (20 samples classified as polyclonal, 24 as 1CLONE and 10 as 2CLONE). VDJ genes, mutational status and CDR3 amino acid sequences were compared in the 44 rearrangements present in these 34 samples. 1/10 secondary rearrangements could not be compared because SSeq sequence was not valid for analysis with IMGT/V-QUEST, even though *IGHV3-21* gene was identified. In 43/43 rearrangements, the same clonal *IGHV* gene and mutational status were found by NGS and Sanger (100%). There was an agreement of 90.7% in the case of the *IGHJ* gene and 93% for CDR3.

These results were promising and therefore, were reproduced in a larger set of samples with different clonal statuses.

4.6.4 Validation of the clonal threshold

The validation dataset contained 27 healthy donors and 280 clonal CLL samples (260 samples with a single SSeq-determined clone and 20 samples with multiple clones determined previously by SSeq).

The average MAX_DIFF value was 111 and 40-fold higher in the 1CLONE and multiple clone groups against the polyclonal group, respectively (Table 4.8). Samples were split into single and multiple CLL clones following the status determined with SSeq to assess the results obtained using NGS.

	average MAX_DIFF	maximum MAX_DIFF	minimum MAX_DIFF
1CLONE	231.34	991.47	8.41
multiple clones	84.06	253.34	6.05
polyclonal	2.08	3.77	1.50

Table 4.8. Validation MAX_DIFF values*. Average, maximum and minimum values for the maximum clonal difference within a sample (MAX_DIFF) in the 3 groups tested in the validation dataset (1CLONE, multiple clones and polyclonal).*

4.6.4.1 Single clone group

244 out of 260 samples were classified into the 1CLONE group exactly as the SSeq sequence, whereas various cases of additional rearrangements were reported by NGS. As described in the methods section, additional rearrangements detected after the NGS approach, (except for those with coexisting equal *IGHV* families in the same sample), were subjected to SSeq *IGHV* family-directed approach, finding the same rearrangement as in NGS in 7 cases, and those were reclassified as 2CLONE for the SSeq counterpart (therefore, not counted in this section but in 4.6.4.2).

After this checkpoint, 16/260 samples remained as cases with additional rearrangements. 9 out of 16 were confirmed using GeneScan analysis (Table 4.9; Appendix Figure 8.1-Figure 8.9), 8 of them with coexisting rearrangements from the same family. The resting case was a sample whose additional rearrangement could be amplified with the *IGHV* family primers but the consensus sequence could not be discerned correctly with SSeq (additional clones sample n.8; Appendix Figure 8.8).

Secondary clones and their relative abundance were compared with CLL Immcantation (Table 4.9). Fold changes in clonal percentages equal or below 1.5 were considered matches. Cases of additional clones 2, 4, 6, 8 and 9 matched in clonal percentage (maximum difference 3%), between both pipelines (fold changes between 1.02 and 1.35). In the remaining 4 cases, secondary rearrangements were identified with the two pipelines, but notable differences in clonal percentages were observed, with fold changes from 2.35 to 13. In cases 1, 5, and 7 the unproductive rearrangement was detected at low proportions with CLL Immcantation (2%, 1% and 6%, respectively). In case 3, the productive rearrangement was the one reported at a low proportion (4% vs 93% for the unproductive rearrangement). Given the more even proportions shown with BMyRepCLL (50%-48%), and since the productive rearrangement was the unique rearrangement detected using SSeq, the case was probably biallelic. A similar scenario occurred with 2 of the previous cases (57%-40% in case 1, 64%-31% in case 7), where the Immcantation pipeline reported lower abundances for the unproductive clone.

Regarding the cases not-matching in clonal percentages (>1.5 fold change difference) between CLL Immcantation and BMyRepCLL, samples with secondary clones were analyzed with a third program: MiXCR (Table 4.9). The aforementioned 4 cases with discordant clonal percentages, matched BMyRepCLL determinations when analyzed with MiXCR (fold changes ranging from 1 to 1.33).

Sample	BMyRepCLL	Clonal%	Considered Clonal?	SSeq	Mutational Status	Productivity	Immcantation Results	MiXCR Results
Additional clones 1	IGHV1-69*13_IGHJ3*02	57.4	YES	IGHV1-69*13_IGHJ6*02	UM	productive	IGHV1-8*01; 96.54%	IGHV1-69D*0; 16%
	IGHV1-8*01_IGHJ3*02	40.7	YES		UM	unproductive	IGHV4-34*01; 2%	IGHV1-8*0; 54%
Additional clones 2	IGHV1-18*01_IGHJ5*02	65	YES		UM	unproductive	IGHV1-18*01; 63%	IGHV1-18*0; 50%
	IGHV1-69*13_IGHJ6*02	34	YES	IGHV1-69*13_IGHJ6*02	UM	productive	IGHV1-69*01; 36%	IGHV1-69D*0; 33%
Additional clones 3	IGHV1-69*13_IGHJ6*02	50	YES	IGHV1-69*13_IGHJ6*01	UM	productive	IGHV1-69*01; 4%	IGHV1-69D*0; 33%
	IGHV1-3*02_IGHJ5*02	48	YES		UM	unproductive	IGHV1-3*01; 93%	IGHV1-3*0; 41%
Additional clones 4	IGHV1-3*01_IGHJ5*02	48	YES		UM	unproductive	IGHV1-3*01; 49%	IGHV1-3*0; 39%
	IGHV1-69*13_IGHJ6*03	50	YES	IGHV1-69*13_IGHJ6*03	UM	productive	IGHV1-69*01; 47%	IGHV1-69D*0; 46%
Additional clones 5	IGHV4-34*02_IGHJ6*04	85	YES	IGHV4-34*02_IGHJ6*03	MM	productive	IGHV4-34*01; 98.48%	IGHV4-34*0; 70%
	IGHV4-61*04_IGHJ6*02	13	YES		MM	unproductive	IGHV4-61*01; 1%	IGHV4-61*0; 15%
Additional clones 6	IGHV4-4*02_IGHJ4*02	14	YES		UM	unproductive	IGHV4-4*02; 19%	IGHV4-4*0; 18%
	IGHV4-34*02_IGHJ6*02	84	YES	IGHV4-34*02_IGHJ6*02	UM	productive	IGHV4-34*01; 76%	IGHV4-34*0; 68%

Additional clones 7	IGHV3-74*01_IGHJ4*02	64	YES	IGHV3-74*01_IGHJ4*02	UM	productive	IGHV3-74*01; 92%	IGHV3-74*0; 55%
	IGHV3-33*01_IGHJ5*02	31	YES		UM	unproductive	IGHV3-33*01; 6%	IGHV3-33*0; 31%
Additional clones 8*	IGHV3-23*01_IGHJ4*02	16.7	YES		MM	productive	IGHV3-23*01; 19%	IGHV3-23*0; 14%
	IGHV4-34*02_IGHJ6*02	79.8	YES	IGHV4-34*02_IGHJ6*02	UM	productive	IGHV4-34*01; 76%	IGHV4-34*0; 70%
Additional clones 9	IGHV3-30*01_IGHJ4*02	46	YES	IGHV3-30*01_IGHJ4*03	UM	productive	IGHV3-30*01; 45.46%	IGHV3-30*0; 41%
	IGHV3-43D*01_IGHJ5*02	52	YES		UM	unproductive	IGHV3-43D*04; 52.56%	IGHV3-43*0; 45%

Table 4.9. Additional clones detected with NGS against SSeq.

4.6.4.2 Multiple clones group

A count of 19 out of 20 multiple-clone samples were grouped into the same number of clones detected with SSeq (Table 4.10). The remaining was included in the false positives (FP) group (FP1; Table 4.11). Coincident cases included 7 samples reclassified from 1CLONE to the 2CLONE group after detecting additional rearrangements by NGS and confirming them by direct SSeq (described in 4.6.4.1).

The confirmation with CLL Immcantation was used in these 19 coincident samples as a double check for clonal percentages. 5 rearrangements were not detected: unproductive *IGHV3-23*01_IGHJ6*0*2 in case 5, productive *IGHV1-69*13_IGHJ6*02* in case 6, unproductive *IGHV3-23*01_IGHJ6*02* in case 9 and productive *IGHV3-49*03_IGHJ4*02* in case 18. In case 13, CLL Immcantation reported 2 rearrangements but not *IGHV4-34*02* (unproductive), confirmed by BMyRepCLL and SSeq (even though with different *IGHJ* gene assignments). Clonal percentages matched in 8/14 cases 1, 3, 4, 10, 11, 12, 15 and 19 (fold changes from 1 to 1.5). Case 8, with 1.6 fold change difference between both pipelines, surpassed the 1.5 threshold chosen but various different *IGHV2-5*02* clones were reported by Immcantation (13%, 8%, 4%, 3%), and only the most predominant was used to calculate the difference. The resting 5 cases did not match in clonal percentages as well (fold changes from 2 to 75.7). In case 2, the productive, UM, *IGHV4-31*03_IGHJ5*02* rearrangement, was detected in 17% reads by Immcantation vs 34% with BMyRepCLL. In case 7, the unproductive MM *IGHV3-41*02_IGHJ4*02* rearrangement, was detected at a lower proportion (10% vs 38%). The same occurs in case 14 for the unproductive MM rearrangement *IGHV7-4-1*01_IGHJ6*02* (1% vs 46%). In cases 16 and 17, the second, unproductive rearrangements were detected at very low proportions (0.3% and 1.5%), compared to BMyRepCLL (22% and 33%).

From these 6 samples not matching in clonal percentages, 2 of them matched between MiXCR and BMyRepCLL (16 and 17; fold changes between 1 and 1.5), and 2 of them matched between MiXCR and Immcantation. Fold change in case 2 was 1 and 3 in case 14 (1% Immcantation vs 3% MiXCR), whereas the in-house pipeline reported 46% clonal for the same rearrangement in case 14. The resting cases (7 and 8) were not considered equal among any pair of pipelines.

About the 5 rearrangements not detected with Immcantation, 4/5 were detected with MiXCR, with matching fold changes between 1.17 to 1.3 with respect to BMyRepCLL. The remaining, was a *IGHV4-34* rearrangement undetected by both Immcantation and MiXCR (case 13) but detected by SSeq as commented previously.

Sample	BMyRepCLL	Clonal%	Considered Clonal?	SSeq	Mutational Status	Productivity	Immcantation Results	MiXCR Results
1	IGHV1-2*02_IGHJ6*02	65	YES	IGHV1-2*02_IGHJ6*02	UM	Productive	IGHV1-2*02; 58%	IGHV1-2*0; 51%
	IGHV3-9*01_IGHJ5*02	34.7	YES	IGHV3-9*01_IGHJ5*02	UM	unproductive	IGHV3-9*01; 39%	IGHV3-9*0; 33%
2	IGHV3-21*03_IGHJ3*02	65.5	YES	IGHV3-21*03_IGHJ3*02	MM	unproductive	IGHV3-21*01; 80%	IGHV3-21*0; 66%
	IGHV4-31*03_IGHJ5*02	33.9	YES	IGHV4-31*03_IGHJ5*02	UM	Productive	IGHV4-31*01; 17%	IGHV4-31*0; 17%
3	IGHV4-34*01_IGHJ6*02	54.8	YES	IGHV4-34*01_IGHJ6*02	MM	Productive	IGHV4-34*01; 53%	IGHV4-34*0; 45%
	IGHV3-9*01_IGHJ6*02	45%	YES	IGHV3-9*01_IGHJ6*02	MM	unproductive	IGHV3-9*01; 44.9%	IGHV3-9*0; 11%
4	IGHV3-15*07_IGHJ6*03	67.7	YES	IGHV3-15_IGHJ6*03	UM	Productive	IGHV3-15*07; 62.04%	IGHV3-15*0; 47%
	IGHV1-2*02_IGHJ4*02	29.4	YES	IGHV1-2*02_IGHJ4*02	UM	productive	IGHV1-2*02; 28.51%	IGHV1-2*0; 26%
5	IGHV3-23*01_IGHJ6*02	70	YES	IGHV3-23*01_IGHJ6*02	UM	unproductive		IGHV3-23*0; 54%
	IGHV4-31*03_IGHJ5*02	29.4	YES	IGHV4-31*03_IGHJ5*02	UM	productive	IGHV4-31*02; 97.8%	IGHV4-31*0; 23%
6	IGHV1-69*13_IGHJ6*02	29.4	YES	IGHV1-69_IGHJ6*02	UM	productive		IGHV1-69D*0; 24%
	IGHV3-30-5*02_IGHJ5*02	70	YES	IGHV3-30_IGHJ5*02	UM	unproductive	IGHV3-30*02; 95.15%	IGHV3-33*0; 40%
7*	IGHV4-34*01_IGHJ5*02	58	YES	IGHV4-34*01_IGHJ5*02	MM	productive	IGHV4-34*01; 84%	IGHV4-34*0; 50%

	IGHV3-41*02_IGHJ4*02	38	YES	IGHV3-41*02_IGHJ4*02	MM	unproductive	IGHV3-41; 10%	IGHV3-41*0; 23%
8*	IGHV2-5*02_IGHJ6*02	21	YES	IGHV2-5*02_IGHJ6*02	UM	unproductive	IGHV2-5*02^	IGHV2-5*0; 8%
	IGHV3-11*01_IGHJ1*01	78	YES	IGHV3-11*01_IGHJ1*01	UM	productive	IGHV3-11*01; 65.92%	IGHV3-11*0; 66%
9*	IGHV3-23*01_IGHJ6*02	29	YES	IGHV3-23*01_IGHJ6*02	UM	unproductive		IGHV3-23*0; 34%
	IGHV4-39*01_IGHJ3*02	70	YES	IGHV4-3*01_IGHJ3*02	UM	productive	IGHV4-39*01; 99%	IGHV4-39*0; 51%
10*	IGHV1-69*12_IGHJ4*02	13	YES	IGHV1-69*13_IGHJ4*02	MM	productive	IGHV1-69*01; 19%	IGHV1-69D*0; 9%
	IGHV4-34*02_IGHJ4*02	61	YES	IGHV4-34*02_IGHJ4*02	MM	productive	IGHV4-34*01; 68.5%	IGHV4-34*0; 70%
11*	IGHV3-7*01_IGHJ6*02	67	YES	IGHV3-7*01_IGHJ6*02	UM	productive	IGHV3-7*01; 75%	IGHV3-7*0; 55%
	IGHV1-46*01_IGHJ4*02	2	YES	IGHV1-46*01_IGHJ4*02	UM	productive	IGHV1-46*01; 4%	IGHV1-46*0; 2%
	IGHV6-1*02_IGHJ5*02	19	YES	IGHV6-1*02_IGHJ5*02	UM	unproductive	IGHV6-1*01; 6.5%	IGHV6-1*0; 20%
	IGHV4-39*01_IGHJ5*02	10	YES	IGHV4-39*01_IGHJ5*02	MM	productive	IGHV4-39*01; 8%	IGHV4-39*0; 6%
12	IGHV4-59*01_IGHJ4*02	57	YES	IGHV4-59*01_IGHJ4*02	UM	unproductive	IGHV4-59*01; 67%	IGHV4-59*0; 53%
	IGHV1-69*06_IGHJ3*02	42	YES	IGHV1-69*06_IGHJ3*02	UM	productive	IGHV1-69*01; 38%	IGHV1-69D*0; 33%
13	IGHV4-34*02_IGHJ5*02	29	YES	IGHV4-34*01_IGHJ3*01	MM	unproductive	IGHV3-53*01; 2%	IGHV3-53*0; 1.5%
	IGHV1-18*04_IGHJ5*02	68	YES	IGHV1-18*04_IGHJ5*02	UM	productive	IGHV1-18*01; 97%	IGHV1-18*0; 89%
14	IGHV4-34*02_IGHJ6*04	53	YES	IGHV4-34*01_IGHJ6*02	MM	productive	IGHV4-34*01; 98%	IGHV4-34*0; 86%
	IGHV7-4-1*01_IGHJ6*02	46	YES	IGHV7-4-1*01_IGHJ6*02	MM	unproductive	IGHV7-4-1*01; 1%	IGHV7-81*0; 3%
15	IGHV3-21*04_IGHJ6*02	30	YES	IGHV3-21*04_IGHJ6*02	MM	productive	IGHV3-21*01; 36%	IGHV3-21*0; 27%
	IGHV4-34*02_IGHJ6*04	47	YES	IGHV4-34*01_IGHJ6*02	MM	unproductive	IGHV4-34; 58%	IGHV4-34*0; 57%
16	IGHV1-3*02_IGHJ4*02	22.7	YES	IGHV1-3*02_IGHJ4*02	UM	unproductive	IGHV1-3; 0.3%	IGHV1-3*0; 15%
	IGHV6-1*02_IGHJ6*02	77.3	YES	IGHV6-1*01_IGHJ6*02	UM	productive	IGHV6-1; 98%	IGHV6-1*0; 66%
17	IGHV1-24*01_IGHJ6*02	32.9	YES	IGHV1-24*01_IGHJ6*02	UM	unproductive	IGHV1-24*01; 1.5%	IGHV1-24*0; 33%
	IGHV4-39*07_IGHJ6*02	58.5	YES	IGHV4-39*07_IGHJ6*02	UM	productive	IGHV4-39*07; 96%	IGHV4-39*0; 50%
18*	IGHV3-49*03_IGHJ4*02	50	YES	IGHV3-49*03_IGHJ4*02	UM	productive		IGHV3-49*0; 40%
	IGHV1-3*01_IGHJ4*02	48	YES	IGHV1-3*01_IGHJ4*02	UM	unproductive	IGHV1-3*01; 97%	IGHV1-3*0; 44%
19*	IGHV1-69*12_IGHJ4*02	28	YES	IGHV1-69*01_IGHJ4*02	UM	productive	IGHV1-69*01; 27.22%	IGHV1-69D*0; 26%
	IGHV3-19*01_IGHJ4*02	71	YES	IGHV3-19*01_IGHJ4*02	UM	unproductive	IGHV3-19*01; 72.29%	IGHV3-35*0; 59%

Table 4.10. Multiple clones confirmed NGS-SSeq. ** detected SSeq after NGS validation. ^various clones IGHV2-5*02 13%, 8%, 4%, 3%...*

Sample	BMyRepCLL	Clonal %	Considered Clonal?	SSeq	Mutational Status	Productivity	Immcantation Results	MiXCR Results
FP1	IGHV4-61*04_IGHJ5*01	24	YES	IGHV4-61_IGHJ5*01	MM	-	IGHV4-59; 99%	IGHV4-59*0 56% and 22% (different clones)
	IGHV1-69*12_IGHJ6*02	6	YES	IGHV1-69_IGHJ6*02	UM	CARGGMATMMFDSW	-	IGHV1-69D*0; 5%
	IGHV4-59*08_IGHJ5*01	68	YES		MM	CSSTSCPVGWYYYYYGMDVW	-	-
FP2	IGHV4-34*02_IGHJ5*02	86	YES	IGHV4-34*02_IGHJ4*02	MM	CARGRTGWYPPGSW	IGHV4-34*01; 94.7%	IGHV4-34*0; 83%
	IGHV1-3*02_IGHJ4*02	7	YES		MM	CARDDL_RWLAEVDYW	IGHV1-3*02 3.74%	IGHV1-3*0; 2%
	IGHV4-4*06_IGHJ5*02	5	YES		MM	CARGRTGWYPPGSW	-	-
FP3	IGHV3-48*04_IGHJ6*02	24	YES		UM	CARDANGMDVW	-	-
	IGHV3-21*02_IGHJ6*02	75	YES	IGHV3-21*02_IGHJ6*02	MM	CARDANGMDVW	IGHV3-21; 99%	IGHV3-21*0; 85%
FP4	IGHV3-33*05_IGHJ5*02	16	YES		MM	CVKDSWRQHDSSGYSPFGFW	-	-
	IGHV3-30-5*01_IGHJ4*02	83	YES	IGHV3-30-5*01_IGHJ4*02	MM	CVKDSWRQHDSSGYSPFGFW	IGHV3-30*18; 99%	IGHV3-30*0; 86%
FP5	IGHV4-31*01_IGHJ2*01	19	YES	IGHV4-31*01_IGHJ2*01	UM	CARGGPGWYRQYWYFDLW	IGHV4-31*01; 16%, 5.8%	IGHV4-31*01; 19%
	IGHV4-34*01_IGHJ4*02	79	YES		UM	-	IGHV4-34*01; 71.12%, 6.64%	IGHV4-34*01; 69%
FP6	IGHV4-59*04_IGHJ5*02	79	YES		MM	CARAMSDSGWHFDSW	-	-
	IGHV4-39*07_IGHJ4*02	17	YES	IGHV4-39*07_IGHJ4*02	MM	CARAMSDSGWHFDSW	IGHV4-39*07; 99%	IGHV4-4*0; 89%
FP7	IGHV3-7*01_IGHJ6*02	5	YES		MM	-	IGHV3-7*01; 6%	IGHV3-7*0; 5.7%
	IGHV3-72*01_IGHJ6*02	94	YES	IGHV3-72*01_IGHJ3*01	MM	-	IGHV3-73*01; 92.76%	IGHV3-72*0; 82%
FP8	IGHV3-72*01_IGHJ5*02	90	YES	IGHV3-72*01_IGHJ5*02	MM	CARGNNYGDYMLGWFDPW	IGHV3-72*01; 98.77%	IGHV3-72*0; 86%
	IGHV3-11*05_IGHJ5*02	8	YES		MM	CARGNNYGDYMLGWFDPW	-	

Table 4.11. False positive clones detected with NGS and confirmed as artifacts reported by the in-house pipeline BMyRepCLL.

4.6.4.3 Sensitivity and Specificity

The minimum MAX_DIFF value established in the test was successful in the analysis of the validation samples grouping them into polyclonal and clonal groups with a specificity and sensitivity of 100%: 27/27 polyclonal samples were determined as polyclonal and 280/280 clonal samples were determined as clonal.

In 8 samples, 9 additional rearrangements from both groups (single and multiple clones) were subjected to GeneScan validation and categorized as false positives, as the presence of secondary rearrangements could not be confirmed (Table 4.11; Appendix Figure 8.10-Figure 8.17). From those additional rearrangements reported by the pipeline, in 5/9 cases the FP rearrangement shared CDR3 with a confirmed true rearrangement, evidencing artifacts from unspecific mapping against *IGHV* genes.

These samples were also checked using CLL Immcantation and MiXCR pipelines to see whether these FP clones were also reported. FP1 had 2 clones confirmed by SSeq and NGS. The second, *IGHV1-69* (productive, 6% clonal), was not detected with the Immcantation pipeline where a single *IGHV4-59* rearrangement is detected (MiXCR did detect the 2 true clones and not the FP). In cases FP3, FP4, FP6 and FP8, neither the Immcantation pipeline or MiXCR reported FP rearrangements. However, in FP2, 1/2 FP clones reported with BMyRepCLL was also found with MiXCR and CLL Immcantation. In FP5 and FP7, the same FP clones were also found with the three analysis methods, with matching clonal percentages.

In conclusion, even though the pipeline reported 9 extra rearrangements in 8 samples as clonal, and 3 of them (FP2.1, FP5 and FP7) have been detected by the 3 tested pipelines, we have classified them as false clones, which probably derive from noise of the most dominant Ig rearrangements. Except those that were found with three analysis methods, only one case was detected with true different CDR3s.

Therefore, the specificity for the number of rearrangements detected for BMyRepCLL in the validation was 97.15% (VN/VN+FP=273/281), and 100% in the case of the sensitivity, as there were no false negative (FN) cases in the validation dataset.

4.6.4.4 Clonal percentages in the whole set of secondary rearrangements

Regarding the cases not-matching in clonal percentages (>1.5 fold change difference) between CLL Immcantation and BMyRepCLL, samples with secondary clones were analyzed with a third program: MiXCR. Clonal percentages were compared with the previous ones reported by the other two pipelines and can be found annotated in Table 4.9, Table 4.10 and Table 4.11.

Even though more similarities are seen between MiXCR and BMyRepCLL in the discordant cases between BMyRepCLL and Immcantation commented in 4.6.4.1, 4.6.4.2 and 4.6.4.3, significant differences in clonal percentages of the whole set of secondary clones (including the test samples) were found only between CLL Immcantation and BMyRepCLL (Mann-Whitney-Wilcoxon p.values with Bonferroni correction: MiXCR vs Immcantation=1; BMyRepCLL vs Immcantation=0.048; BMyRepCLL vs MiXCR=0.05). No significant differences were found between MiXCR and BMyRepCLL or Immcantation, suggesting similar determination of clonal percentages (Figure 4.31a). However, the significant p.value is not very low. To simplify results, clonal percentage matches regarding the 1.5 fold change threshold of the whole set of secondary rearrangements are schematized in a Venn Diagram (Figure 4.31b). A total number of 35 rearrangements matched in clonal percentage between MiXCR and BMyRepCLL, whereas the coincidences Immcantation-BMyRepCLL are 21 and 19 the coincidences between Immcantation-MiXCR. MiXCR has the best rate of overlap with the other tools (84% for MiXCR, 76% for BMyRepCLL and 56% for Immcantation).

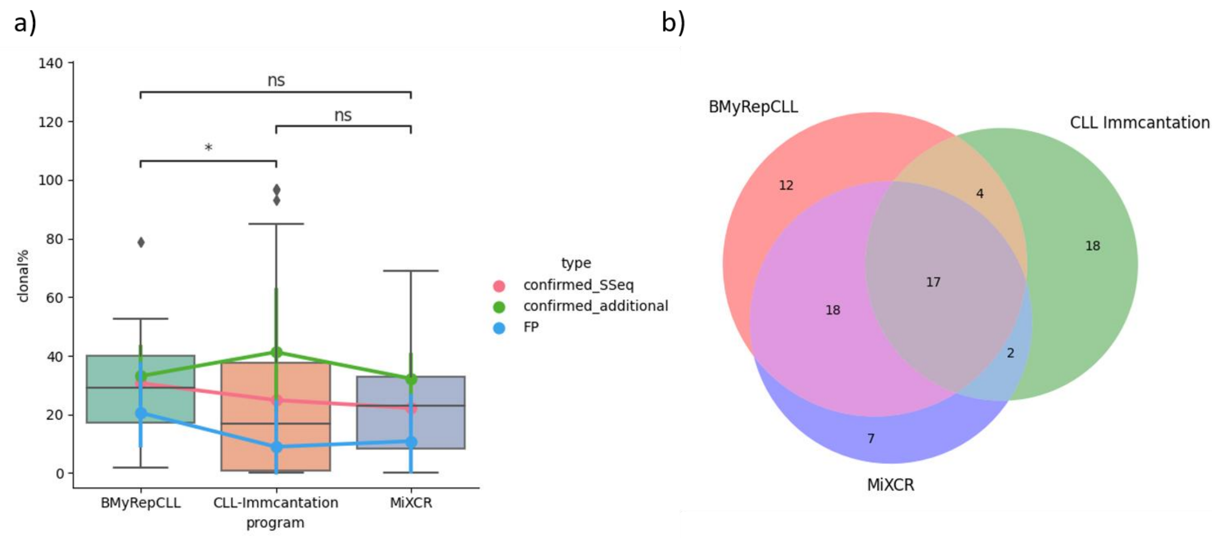


Figure 4.31. Secondary clones percentages comparison among pipelines. *a) Clonal percentages for the secondary rearrangements obtained with CLL Immcantation, BMyRepCLL and MiXCR. Connected lines represent average values split by group (confirmed SSeq, confirmed additional and FP). b) Venn diagram of secondary rearrangement concordance (<=1.5 fold change) among the samples with multiple CLL rearrangements (including confirmed with SSeq, additional rearrangements and false positives), using BMyRepCLL, CLL Immcantation and MiXCR.*

4.6.4.5 Clone characterization

To test the ability of characterizing Ig rearrangements by BMyRepCLL, confirmed SSeq Ig rearrangements for each of these samples were compared to the rearrangements obtained by NGS in means of *IGHV* and *IGHJ* genes, mutational status and CDR3. 300 out of 300 *IGHV* genes detected (280 predominant rearrangements and 20 secondary) were equal to SSeq (100%). In the case of *IGHJ*, 289 out of 300 total *IGHJ* sanger alleles (including differences in allele) (96.33%). Also, CDR3 amino acid sequence was equally identified in 270/290 rearrangements in those CDR3s characterized using SSeq (93.10%) (minor differences in CDR3 were found in 17 samples, involving 1 amino acid change; 6% of cases considered equal).

The mutational status obtained for each of these 300 rearrangements was also compared after their classification into UM or MM following the clinical guidelines: 297 out of 300 rearrangements were equally classified (99%). In addition, the identity percentages obtained with both techniques were compared (Pearson correlation r-squared 0.862; p.value 2.093e-129) (Figure 4.32).

4.6.4.6 Mutational status differences

Regarding mutational status differences, 3 samples differed between SSeq and BMyRepCLL. The difference was due to point mutations that could not be detected in the alignment of the consensus sequences against IMGT *IGHV* alleles due to the primers design employed in the NGS protocol, as they were upstream the FR1 fragment or falling in the FR1 primer region. These represent 0.9% of the total number of samples used in the validation and their statuses vary from BD (97.92, 97.59 and 97.57 identity percentages) to UM, with 98 identity percentage in the 3 cases (1 mutation difference on each case).

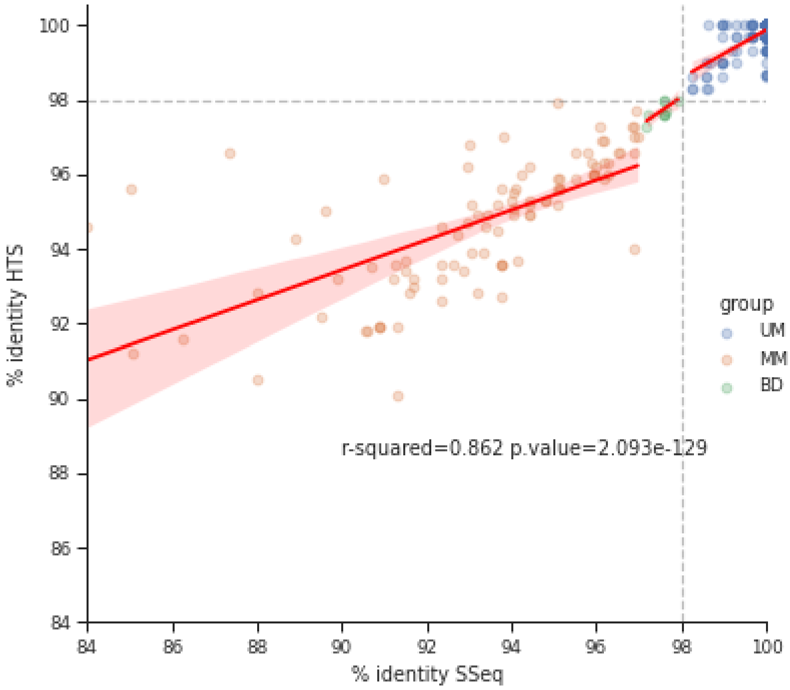


Figure 4.32. BMyRepCLL-SSeq mutational status comparison. *Comparison of mutational status determination (identity percentages against germline IGHV alleles) by BMyRepCLL against gold-standard SSeq. Points in the regression are grouped by SSeq mutational status.*

4.6.5 Evaluation of the set of filtering steps

Efficiency of each of the filter rounds integrated in the in-house pipeline (2nd module of BMyRepCLL; 3.6.1.2) was assessed at three levels: rearrangement ranks, CDR3 sequences uniqueness and *IGHV* genes usage. At the first step, raw rearrangements reported at allele level after mapping read counts, represented a total of 37002 rearrangements, and presented 10 clone ranks and above (shown up to 10 clone ranks), with high point density in clonal percentages below 50 in rank 1 (Figure 4.33a). In subsequent steps (Figure 4.33b and Figure 4.33c), the number of rearrangements decreased 2.5-fold and 3.3-fold, respectively, as they augmented in abundance (rearrangement rank1 reached higher density in clonal percentages close to 100 and the rest of rearrangement ranks contained less clones). This achievement happened when joining rearrangements with consensus sequences surpassing 95 identity percentages and correcting fragment biases. Lastly, Figure 4.33d gathers rearrangement ranks after unraveling the clonal fraction on each patient, and therefore, only clonal rearrangements are shown. The number of clonal ranks decreased to rank 4, being the predominant samples with a single B cell clone, whereas secondary clonal rearrangements were represented in a smaller fraction of patients (12%).

In the same way, Figure 4.34 represents the prioritization of clones along filtering steps using the number of *IGHV* genes as a measure of the number of rearrangements. These should be correlated with the number of unique CDR3 sequences detected. Until the last step, Figure 4.34a-c shows uncorrelated count of *IGHV* alleles or genes compared to junction sequences. In contrast, Figure 4.34d shows a lineal tendency whose values are detailed in Table 4.12, where 270 occurrences corresponded to unique *IGHV* genes paired with unique CDR3 sequences. 20 reflected double rearrangements, as they contained 2 unique CDR3 sequences for 2 *IGHV* gene rearrangements, and 1 case with 3 rearrangements. In 23 cases the number of *IGHV* genes surpassed the number of unique CDR3s for clonal rearrangement and the resting case contained no CDR3 sequences, probably not being characterized. In Figure 4.35, the same process occurred involving *IGHV* genes represented among the samples tested. In Figure 4.35a, raw rearrangements were represented by more than 70 different *IGHV* genes, compared to 41 in the last step (Figure 4.35d), where *IGHV* gene rearrangements were joined into the major rearrangements considering them the same clonotype, and clonal percentages are therefore, higher.

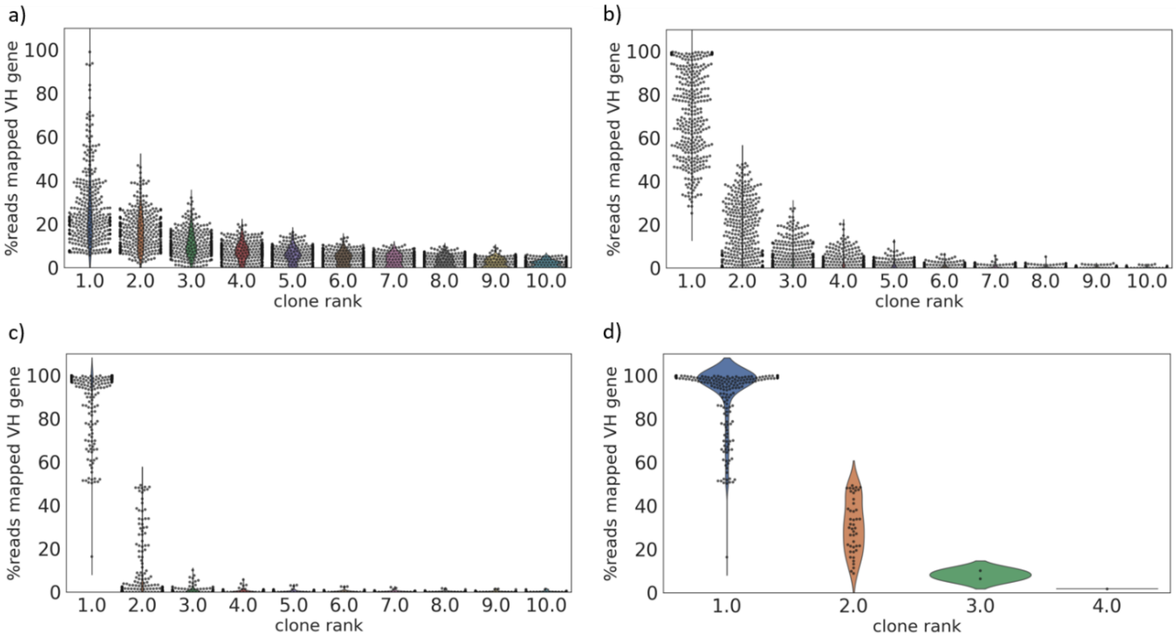


Figure 4.33 Rearrangement ranks at different important filtering checkpoint levels in the in-house pipeline BMyRepCLL (up to 10 rank number). *a) raw rearrangements reported in the initial table (rearrangements are based in IGHV alleles). b) After percentage identity filters (rearrangements whose consensus sequences sharing >=95% identity are joined into the most abundant IGHV allele first, and afterwards by gene), we observe that secondary rearrangements are less abundant and the primary rearrangement is represented in higher percentages than in a). c) the decrease in secondary rearrangements is further observed after correcting fragment biases, mostly from FR3 which is the shortest fragment. d) Only rearrangements considered clonal are shown after using the clonal background threshold (subclonal rearrangements are also kept for their study). The greatest rank considered clonal is rank 4 and the predominant rearrangement is present at higher percentages, many of the points being 100% or close.*

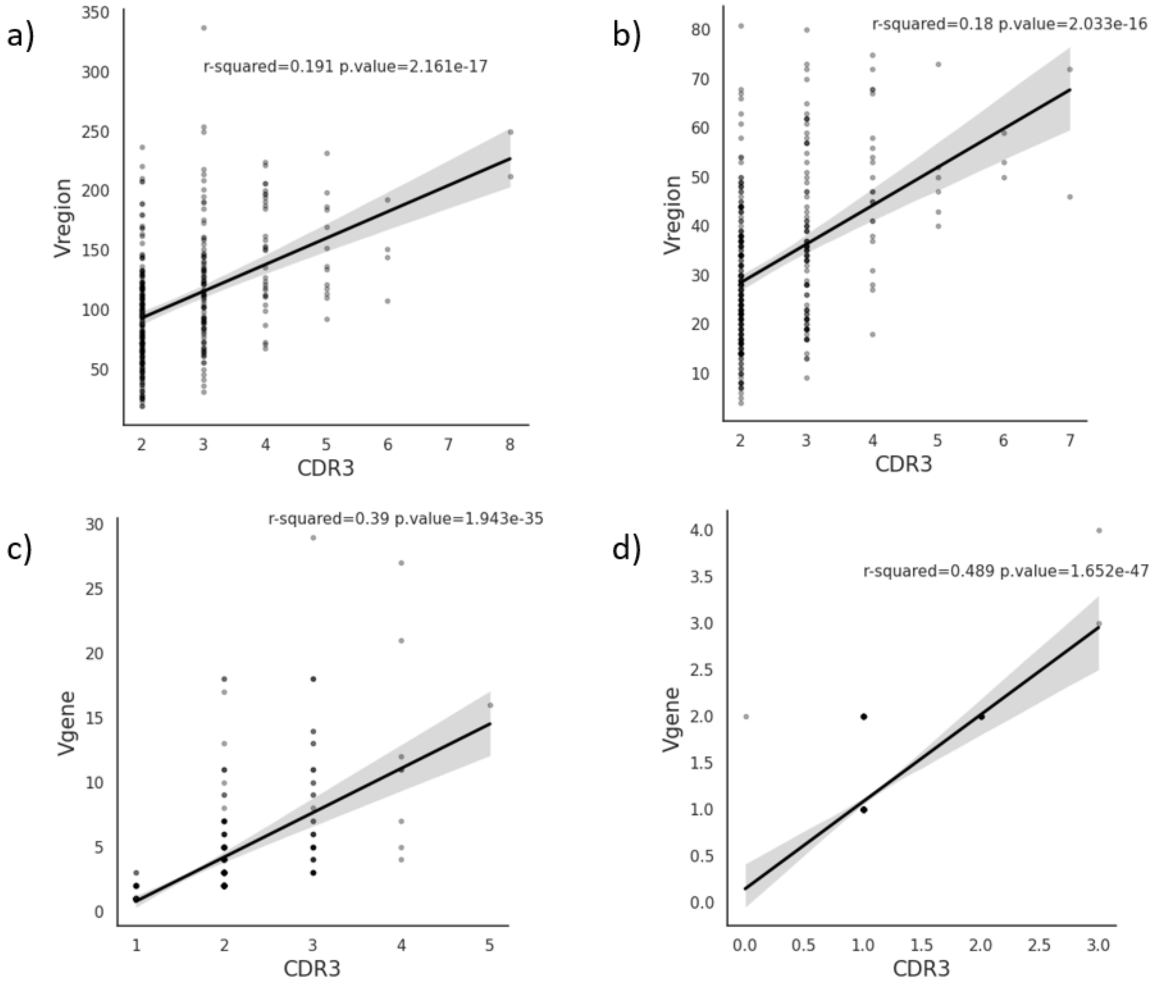


***Figure 4.34. Number of IGHV genes/alleles represented per unique CDR3 sequences**. a) IGHV alleles vs unique CDR3 sequences in the raw rearrangements table. b) IGHV alleles vs unique CDR3 sequences after identity filters. c) IGHV genes vs CDR3 sequences after fragment compensation and summary. d) IGHV genes vs CDR3 sequences after clonal threshold correction.*

count	n. unique cdr3	n. IGHV gene (rearrangements)
270	1	1
22	1	2
20	2	2
1	3	3
1	3	4
1	0	2

Table 4.12. Number of IGHV genes present per unique CDR3 sequence after filtering steps in B-MyRepCLL.

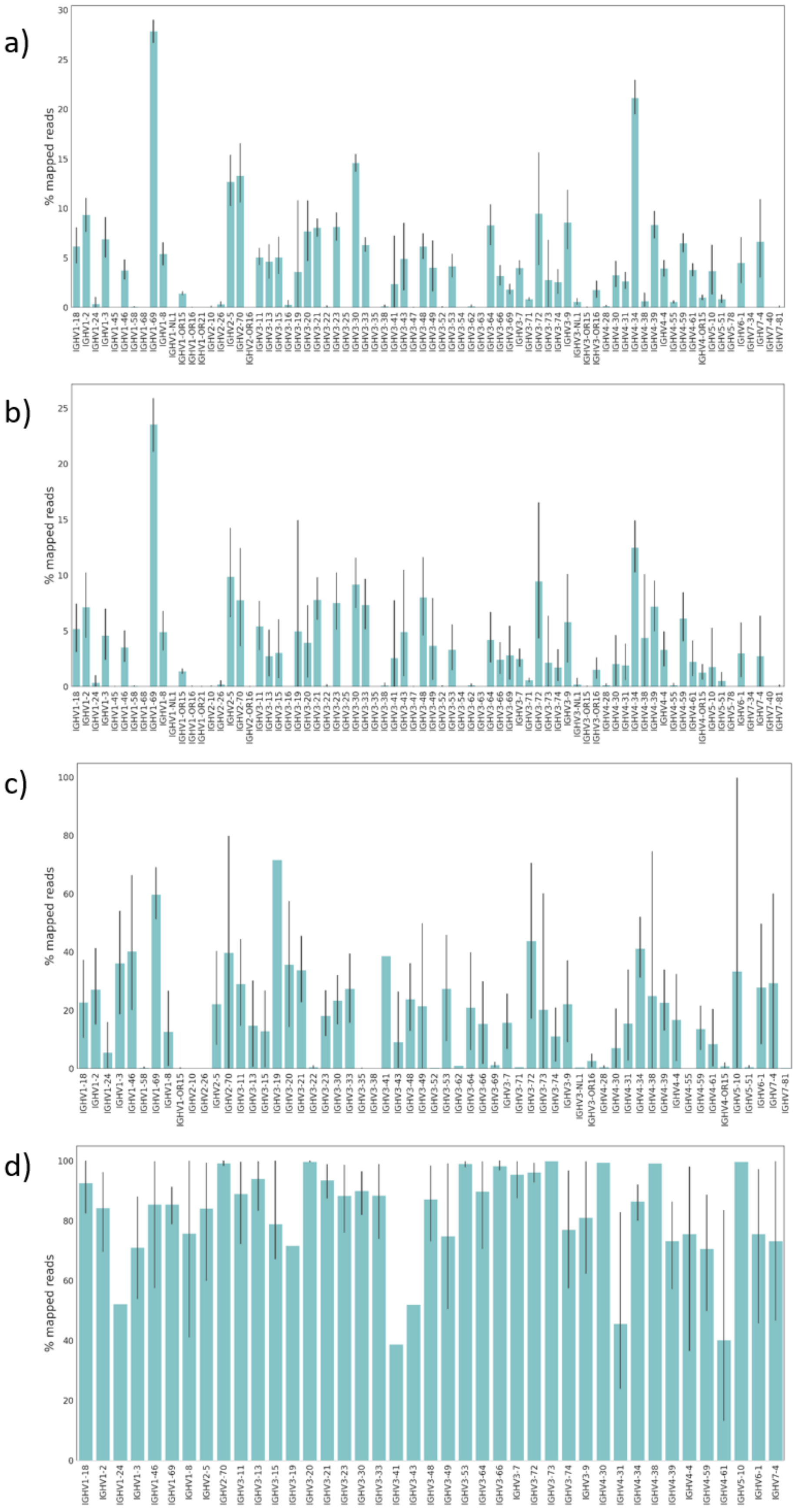


Figure 4.35. Number of different IGHV genes represented among validation samples across the filters employed in pipeline BMyRepCLL second module. *a) 74 different IGHV genes are represented in raw rearrangements. b) Identity filters do not reduce the number of IGHV genes detected. c) 50 unique IGHV genes represented after fragment compensation filters and d) 41 different IGHV genes represented in clonal CLL rearrangements.*

4.6.6 Coverage and gene usage

362 rearrangements were considered clonal after applying MAX_DIFF threshold. The rest of the rearrangements (867) are tagged as subclonal, with clonal percentages varying from 0.1 to 9.1 of the total reads assigned to Ig rearrangements.

The average percentage coverage breadth above 500 reads for Ig clonal rearrangements characterized was 85% (clonal percentages ranging from 2-100%) (Figure 4.36). Number of supporting reads for the different *IGHV* genes encountered by fragment is represented in Figure 4.37, where it can be observed that most *IGHV* genes were supported by the three amplicons employed (59 different *IGHV* genes represented among the samples including subclonal rearrangements in Figure 4.37a, and 41 considering only clonal rearrangements; Figure 4.37b). On the 362 rearrangements considered clonal, Figure 4.37c represents the distribution of coverage breadth percentage above 500X along the different *IGHV* genes, and differences that are present regarding that genes are represented at different clonal percentages on each sample.

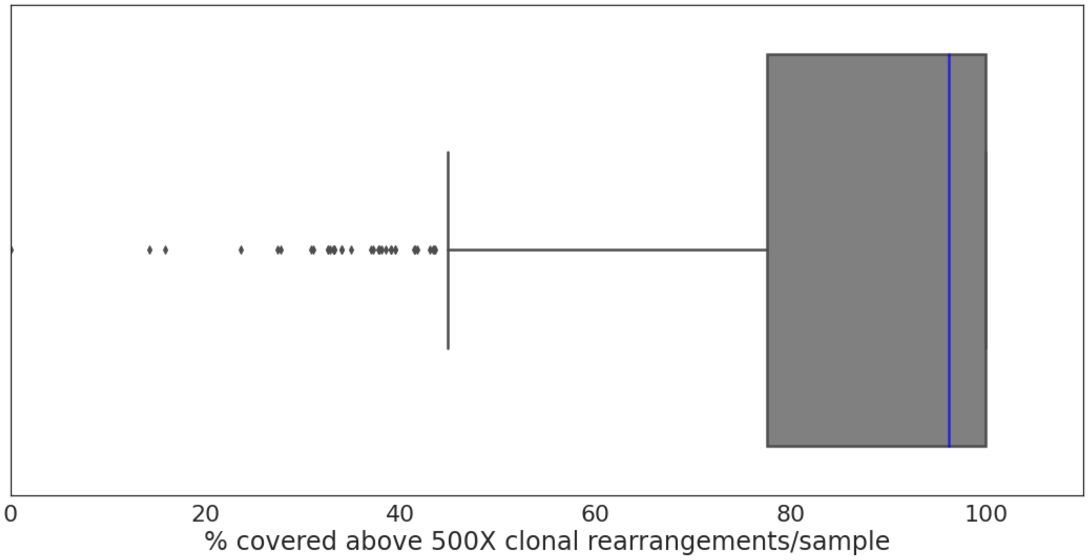


Figure 4.36. Coverage breadth over 500 reads in clonal rearrangements from validation CLL samples.

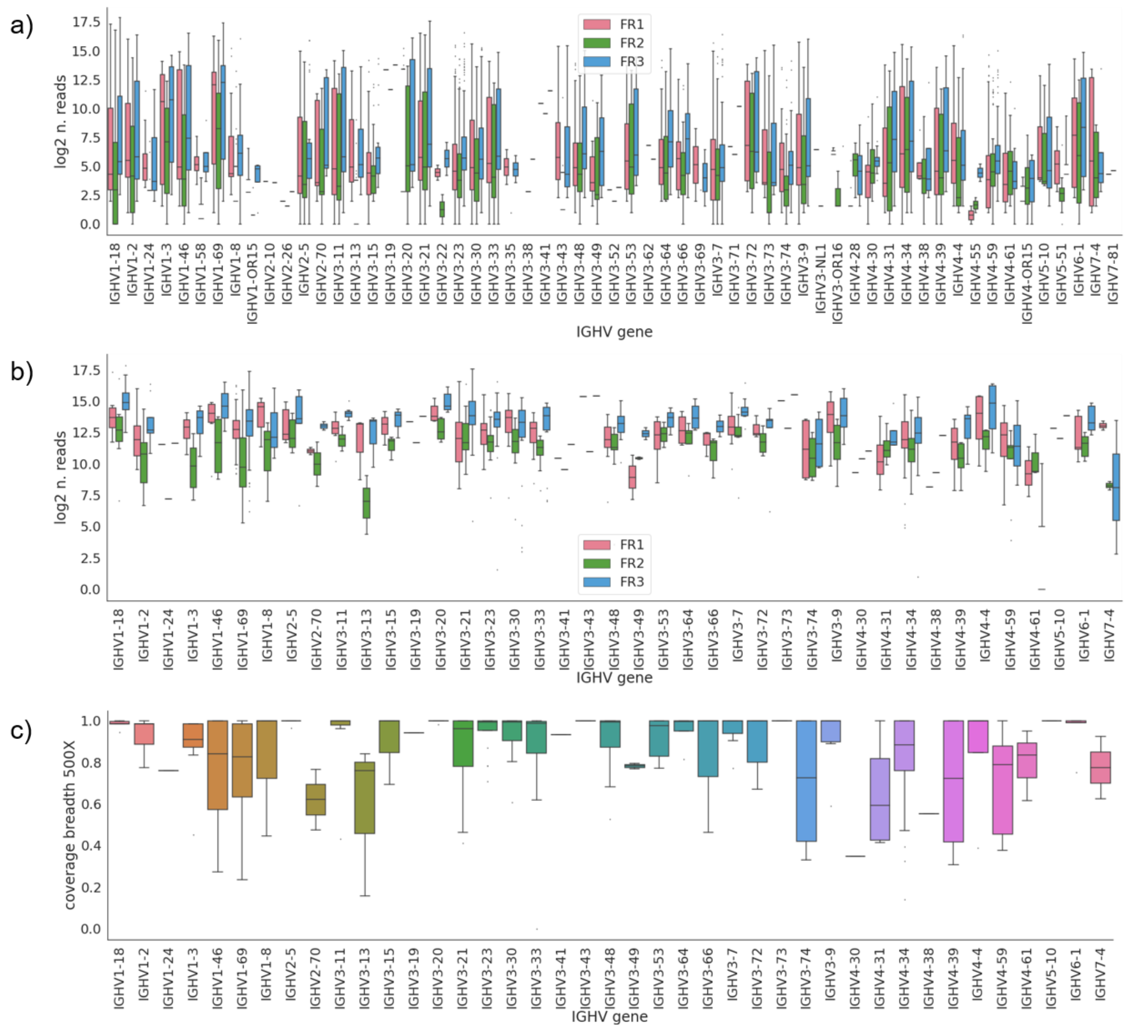


Figure 4.37. Amplicon reads distribution among IGHV genes and coverage. *a) Amplicon number of supporting reads (log2) among the different IGHV genes in all their appearances among the samples tested (above 1 read). FR amplicons are distinguished by color. b) Amplicon number of supporting reads (log2) among the different IGHV genes in clonal rearrangements above 10% in all their appearances among the samples tested. c) Coverage breadth % above 500X in the represented IGHV genes.*

4.6.7 Comparison of the assessment of mutational status with BMyRepCLL and CLL Immcantation

To evaluate the match in mutational status calculation between CLL Immcantation and SSeq, Pearson correlation was calculated between the identity percentages of SSeq sequences and the average mutation frequency in the major clone determined by CLL Immcantation. This method approximates accurately to the determination made using the gold-standard (r-squared 0.935; p.value 1.021e-158) (Figure 4.38).

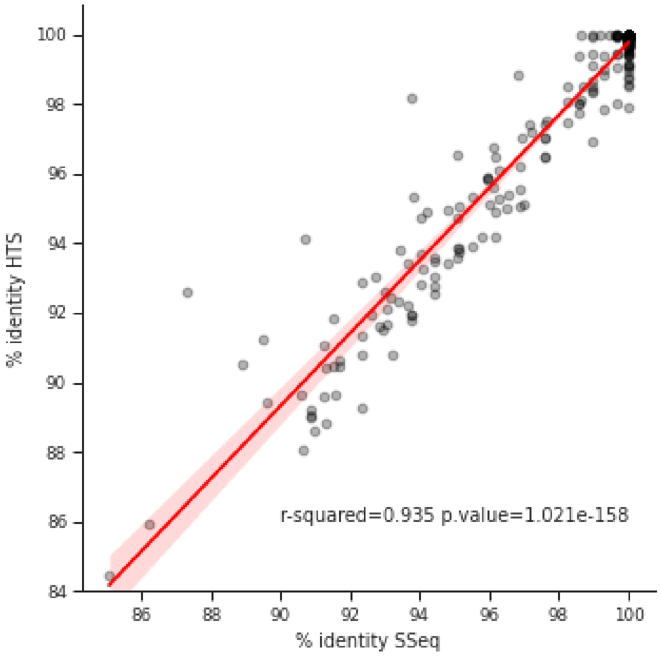


Figure 4.38. ***CLL Immcantation-SSeq mutational status comparison.*** *Comparison of mutational status determination (identity percentages against germline IGHV alleles) by CLL Immcantation against gold-standard SSeq.*

There were 6 discordances in the mutational status between SSeq and Immcantation (Table 4.13). They corresponded to 5 cases reclassified from UM to MM (4/5 borderline), and the opposite in the resting case (MM to UM). Some of the cases (1, 2, 5), had a wide distribution of mutations in the major clone (example with case 2 in Figure 4.39). Distribution of mutations in the major clone in Figure 4.39a is plotted in Figure 4.39b, showing both MM and UM groups of sequences. Intraclonal mutations were present as shown in Figure 4.39c and the inspection of the BAM rearrangement file obtained with BMyRepCLL (Figure 4.39d). In case 3, the same number of mutations was reported with both pipelines (Figure 4.40 a and b), but mutation frequencies appear to be borderline in the case of Immcantation (between 0.02 and 0.03) (Figure 4.40 c and d). The fourth case corresponded to a double rearrangement sharing *IGHV* family (*IGHV1-69* and *IGHV1-18*), whose reads were assigned indistinctly to both of them and appear to be variable (Figure 4.41). Mutation frequency distribution was wide in both predominant clones (Figure 4.41a), and even though most sequences were unmutated, there were highly mutated sequences represented at lower proportions (Figure 4.41b). Both *IGHV* alleles were paired with 2 different *IGHJ* genes (Figure 4.41c) and both predominant clones had reads assigned to *IGHV1-69* and *IGHV1-18* (Figure 4.41d).

Sample	Sseq	CLL immcantation	BMyRepCLL
1	99.29 (UM)	97.83 (BD*)	99.7 (UM)
2	98.97 (UM)	96.89 (MM)	99 (UM)
3	98.26 (UM)	97.48 (BD*)	98.3 (UM)
4	100 (UM)	97.89 (BD*)	99.7 (UM)
5	98.6 (UM)	97.72 (BD*)	98.6 (UM)
6	96.83 (MM)	98.84 (UM)	97.3 (BD*)

Table 4.13 Discordant mutational status samples between CLL immcantation and SSeq. ** BD (Borderline) samples are considered MM, for clinical decisions.*

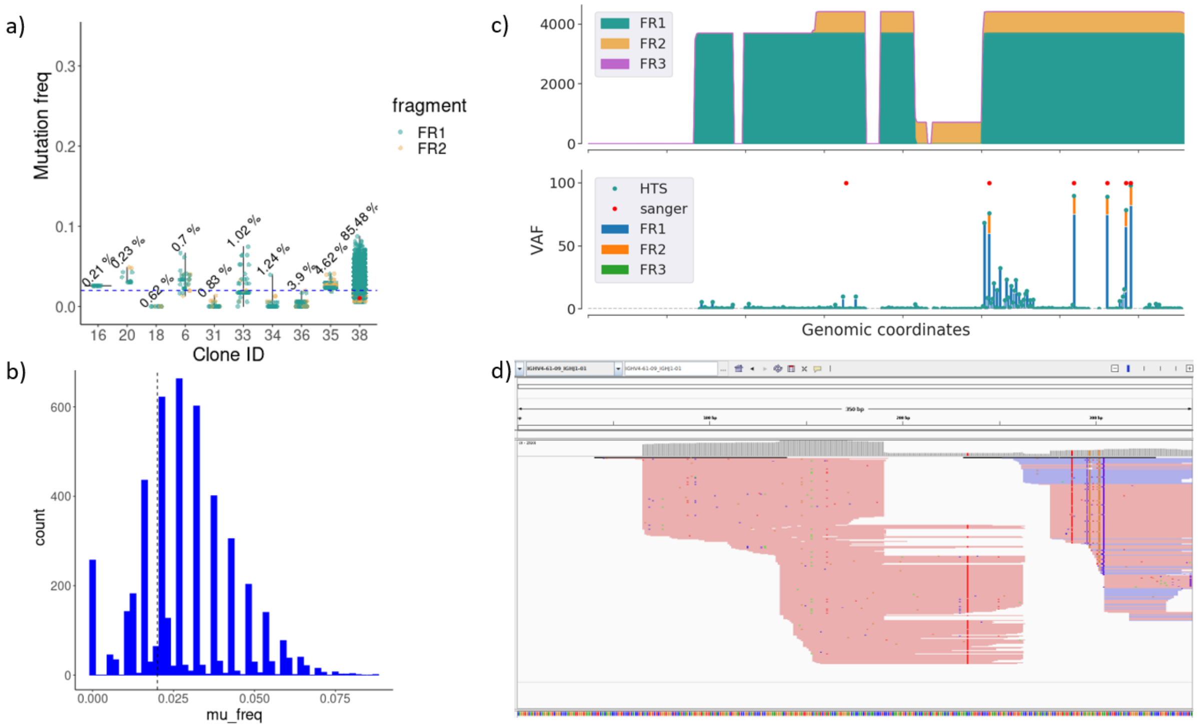


Figure 4.39. Discordant mutational status case 2. *a) Top 10 clones. b) Mutation frequencies histogram. c) Coverage and variants plot including the support of FR1 and FR2 fragments reads. d) Visualization of variants in the rearrangement-specific BAM file from BMyRepCLL using IGV.*

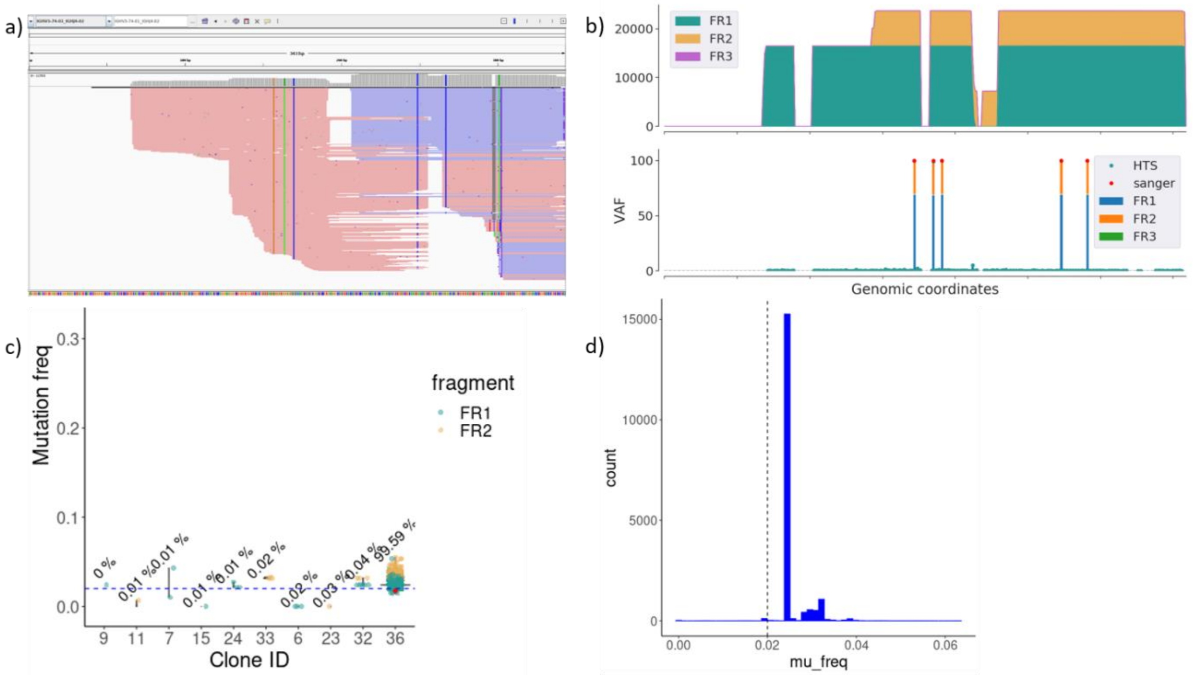


Figure 4.40. Discordant mutational status case 3. *a) Visualization of variants in the rearrangement-specific BAM file from BMyRepCLL using IGV. b) Coverage and variants plot including the support of FR1 and FR2 fragments reads. c) Top clones plot. d) Mutation frequencies histogram.*

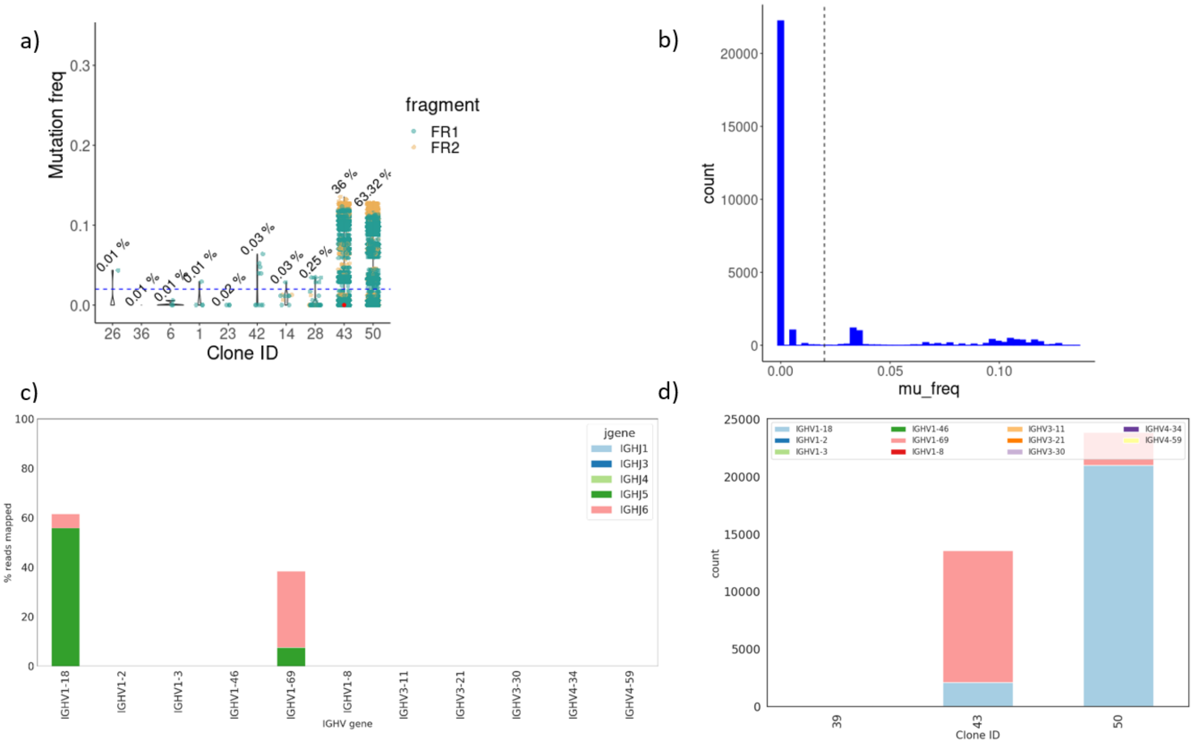


Figure 4.41. Discordant mutational status case 4. *a) Top clones plot. b) Mutation frequencies histogram. c) Gene usage regarding % of mapped reads. d) IGHV gene counts per clone.*

The ratios of coincidence in BD samples between NGS and SSeq methods were 77.8% for CLL Immcantation (7/9), and 66.7% for BMyRepCLL (6/9) (Table 4.14).

Sample	SSeq	CLL immcantation	BMyRepCLL	status_sanger	status_NGS_immcantation	status_NGS_inhouse
1	97.64	97.51	97.7	BD	BD	BD
2	97.57	97.02	98	BD	BD	**UM**
3	97.57	96.51	97.6	BD	**MM**	BD
4	97.59	97.43	98	BD	BD	**UM**
5	97.92	97.58	98	BD	BD	**UM**
6	97.22	97.16	97.6	BD	BD	BD
7	97.57	96.49	97.6	BD	**MM**	BD
8	97.18	97.42	97.3	BD	BD	BD
9	97.57	97.02	97.6	BD	BD	BD

Table 4.14. Comparison of the determination of BD samples with both analysis methods (BMyRepCLL and CLL Immcantation).

4.6.8 Subclones confirmation by two analysis methods

We found productive low frequency rearrangements tagged as subclonal with BMyRepCLL, and tested their presence with CLL Immcantation. Since some of them are below 1% mapped reads, it is necessary to use a double check to consider them potential subclones. In 11 samples, these subclones have been confirmed with both methods (Table 4.15), with CDR3 sequences being different to the clonal rearrangements. None of them have been assigned to stereotyped subsets using ARREST/AssingSubsets tool.

sample	BMyRepCLL	Clonal%	MS	MS in majoritary clone(s)	CDR3	Immcantation confirmation
1	IGHV3-23D*01_IGHJ6*02	6.8	**UM**	**MM**	CAKDRSTNYCYYGMDVW	IGHV3-23D*01; 2.8%; UM
2	IGHV4-4*07_IGHJ6*02	3	UM	UM	CARMMGGSFPRSHYYYGMDVW	IGHV4-4*07; 5.8%; UM
3	IGHV3-48*02_IGHJ4*02	1	**UM**	**MM**	CARTTPFDYW	IGHV3-48*02; 0.7%; MM
4	IGHV6-1*02_IGHJ6*02	7.8	MM	MM	CARGYPGLNLW	IGHV6-1*01; 1%; MM
5	IGHV3-64D*06_IGHJ6*02	5.5	UM	UM	CVNLGRDGYNKWIYYGMDVW	IGHV3-64D*06; 4%; UM
6	IGHV3-48*04_IGHJ3*01	3	MM	MM	CARPNWEDGFDLW	IGHV3-48*04; 17%; MM
7	IGHV4-39*08_IGHJ5*02	1.4	UM	UM	CARRIGYSSSWYAKDNWFDPW	IGHV4-39*01; 1.55%; UM
8	IGHV3-9*01_IGHJ4*02	8.7	MM	MM	CAKGRRWGHYVPNFDHW	IGHV3-9*01 FR1 only; 12%; MM
9	IGHV3-11*01_IGHJ6*02	0.6	UM	UM;UM	CARDIVVVVTAPNYYYDMDVW	IGHV3-11*01; 0.67%; UM
10	IGHV3-23D*01_IGHJ4*02	1.8	UM	MM;UM	CATTVKRSDYW	IGHV3-23*01; 1.8%; UM
11	IGHV3-48*02_IGHJ6*03	1.6	**UM**	**MM**	CARTQEWLNHYYYMDVW	IGHV3-48*01; 0.9%; UM

Table 4.15. Subclonal rearrangements encountered by BMyRepCLL and CLL Immcantation. *MS stands for "Mutational Status". In bold, cases with UM subclone whereas the predominant rearrangement is MM.*

5 Discussion

5.1 Motivation

For both library preparation and bioinformatic analysis in immune *repertoires*, many methods have been described in the past years, as well as commercial kits. However, the library design described herein is not conventional, due to the fact that the region is covered partially with different DNA fragments with the purpose of using short-read sequencing. The methods have been validated alongside the clinical team at the Hematology department of the Clinical University Hospital of Valencia (HCUV), sticking to their needs and planning possible scaling of the same methods to other hospitals and clinical laboratories dedicated to B cell neoplasms, focusing especially on the bioinformatics part.

We adapted two analysis methods: an in-house pipeline constructed with open-source bioinformatics software, and other using a suite of tools developed by a group with expertise in immunoinformatics. For the second workflow, created in collaboration with the Kleinstein´s lab from Yale University Pathology Department, specific adaptation to the library design and lymphoid neoplasms was made by adding modules and scripts. Even though these methods have been tested with CLL patients for being one of the most common diseases among B cell neoplasms, the same approach would be applicable to other diseases where expanded B cell clones are present.

5.1.1 Amplification methods

Some studies describe the multiplex PCR method as the most reliable option to obtain mutational information of immune *repertoires* and high coverage rates (164). Other amplification methods commonly used by the community of immune *repertoires* sequencing, such as 5´RACE RNA (template switching methods), circumvents the primer biases of multiplexed PCR, and would be beneficial on highly somatically mutated samples (107). However, full-length sequencing of *IGHV* region is compromised (89) and therefore, we discarded that option primarily for clinical determinations.

5.1.2 Choice of the library preparation method

For sequencing of the IGH locus, 3 different library preparation approaches were tested. The first protocol (library protocol A; 3.4.1), consisted of amplification with Leader-JH primers and v3 300bpx2 sequencing kit. The advantage is that the region is covered with a single fragment, but 300bpx2 sequencing kit can only be used in the MiSeq platform as no other Illumina machine has compatibility with this kit (maximum is 250bpx2 in some HiSeq models). It was of utmost interest that this determination could be done in parallel with other clinical gene panels, as routine determinations of CLL patients would not cover in many cases complete MiSeq runs (1run/week). Short-read 150bpx2 sequencing is compatible with the sequencing of other gene panels, due to the fact that amplicons are commonly designed to be 100-200bp long (165). Therefore, having more alternatives apart from 300bpx2 sequencing was a necessity, and for that purpose, library protocols B and C were tested. Both imply using short-read sequencing with 150bpx2 cycles kit, compatible with all Illumina sequencing machines, being thus, more flexible and with the possibility of augmenting sequencing capacity. To summarize, the advantages obtained with the use of short-read 150bpx2 sequencing compared to full-length 300bpx2 Illumina sequencing kit are:

1. Improved quality values (approximately 10%) compared to larger run kits (vendor quality specifications: Q30>80% for v2 151bpx2, Q30>75% for v2 251bpx2 and Q30>70% for v3 301bpx2). The extension of the VDJ region, has converted 300bpx2 MiSeq kit in one of the most used for sequencing immune *repertoires*. However, this kit only exists for the MiSeq platform, and decreases quality scores due to increased accumulation of phasing errors within the Illumina sequencing by synthesis technology (166).

2. Sequencing experiments can be economized by including other gene panels for somatic mutations, (for example, TP53). In fact, 150bpx2 is the most used for clinical gene panels, and also have greater turnaround time for being half the cycles than 300bpx2. Ion Torrent PGM allows longer sequencing than Illumina (~400bp) but the error rate is 10-fold higher (134). As the regions of interest are highly variable, the method has to be the lowest error prone as possible to avoid miscalls due to sequencing artifacts (134). Using 150bpx2 sequencing, there is no need to use different sequencing kits for different determinations and sequencing experiments can be adapted to the hospital necessities, with the capability of including many determinations together and therefore, not having to wait to take full advantage of sequencing runs solely with Ig determination.

To cope with quality decrease related to the increase of sequencing cycles, research groups have used DNA tagmentation after amplification of the complete region using Leader primers (133). Similarly, Illumina Nextera tagmentation protocol was proven in this work as an alternative method to sequence the VDJ region using short 150bpx2 reads (library protocol B; 3.4.2), and given the fact that we obtained similar and valid results with the in-house multiplexed protocol (library protocol C; 3.4.3), we decided to continue with the last approach due to lower costs, not depending on commercial kits. For the multiplex reaction with Leader or FR1 primers, regarding SHM determination, there are also a few full commercial kits which increase the costs of these experiments and limit the number of samples introduced (134).

On the other hand, the use of different primer sets augments the probabilities of amplification of the region. Due to its inner variability, SHM can impede sometimes the correct hybridization of the degenerate oligonucleotides during PCR steps.

5.1.3 DNA or RNA?

The first ~50 bps of *IGHV* exon 2 are not covered with the in-house protocol (3.4.3) using gDNA as genetic material. Therefore, the cDNA method including Leader primers set has also been adjusted, in order to sequence the whole region. cDNA method will allow to refine the annotation of *IGHV* alleles and the mutational status determination will be more accurate.

On the other hand, RNA can produce a biased clone quantification due to differential expression of surface antibodies. For instance, high levels of expression occur in antibody-secreting cells, whereas using gDNA would maintain a ratio of 1 DNA copy/cell (167), allowing to quantify clonality in the samples studied. Also, both productive and unproductive (not expressed) rearrangements can be detected with the use of gDNA (147).

For that reason, we encourage further validation experiments using cDNA with a wider volume of samples, as we did with gDNA. The validation of this study was performed with the gDNA protocol due to the availability of the samples.

5.1.4 FR1 or Leader primers?

The guidelines from the ERIC CLL consortium are consistent with the use of primer sets for the determination of the mutational status in the IGH locus. FR1 primers will be used only if the amplification from Leader region is technically impossible (148). Even though many works have reported non-significant differences between the use of FR1 and Leader primers (149,168), and better amplification efficiency in the case of FR1 primers (169), the discordant mutational statuses in BMyRepCLL against SSeq corresponded to 3 cases, reclassified from BD to UM, and even representing a very low percentage of the samples (0.9%), mutational status changes influence clinical decisions. For that reason, we encourage the use of Leader primers (cDNA protocol). For the purpose of characterizing accordingly the mutational status and the relative abundance of the clones, both assays could be paired to take advantage of their combination (147). Another advantage of the 150bpx2 short-read method is that both gDNA-cDNA libraries could be combined in the same sequencing experiment.

Commercial kits, such as Lymphotrack (Invivoscribe), are based on conventional 2x300 whenever using Leader, and 250bpx2 for FR1. The Ion Torrent OncoMine LR kit (TermoFisher), is designed for sequencing the IGH locus from FR1 region to the Constant region (reverse primers), with an amplicon of 400bp (to our knowledge, the design does not include Leader primers).

5.2 Design of the in-house pipeline

The most renowned published tools for the annotation of VDJ genes are IMGT/V-QUEST, IgBlast and MiXCR (147). At the beginning of this thesis work, the use of one or another was studied. IMGT/V-QUEST is the standard method employed and recommended by the ERIC guidelines to characterize IGH locus, as it provides accurate alignment results, productivity, detects mutations (insertions, deletions, stop codons…). However, the use of this web portal requires submitting sequences one by one, as it is designed for the standard clinical protocols employing SSeq. There is a version for NGS data (IMGT/HighV-QUEST) (150,170), but it has sequence limit and requires a license. The use of IgBlast is highly spread for being open-source and supporting large NGS files, ensuring fast results as it is based on the BLAST algorithm (171). Pipelines such as IgGalaxy (172) and the Immcantation Framework, wrap IgBlast to perform gene annotation. Unlike IMGT/V-QUEST and IgBlast, MiXCR conforms and end-to-end pipeline which includes preprocessing steps and pair read merging, annotation, and clustering of sequences into clones (141). It has different options with specific workflows for multiplexed-PCR and 5´RACE, and also corrects PCR errors internally, as it does not support UMIs. Academic use requires a free license, which is easy to obtain and the documentation makes its use straightforward. At the same time, the alignment algorithm employed is very fast (144). In fact, the performance of IMGT/HighV-QUEST, IgBlast and MiXCR has been compared, and MiXCR was the most efficient tool regarding running time (173). Moreover, the fact that MiXCR supports both full and partial profiling of Ig and TCR *repertoires* encouraged us to test it with our library method in the early stages of this work (data not shown). However, its annotation accuracy is low compared to IMGT/HighV-QUEST and IgBlast, and it does not report annotations at allele level, which will be critical for *IGHV* and *IGHJ* genes annotation in clinical determinations, as alignment against the closest *IGHV* allele is not performed to determine precise identity percentages (173,174). LymAnalyzer developers were also aware of the compromised accuracy of MiXCR and improved it in their own open-source tool (175). The use of LymAnalyzer was also studied in early stages of this thesis work (data not shown). Even though it provides accurate gene annotations and germline alignment, preprocessing steps are not included and the final clones are grouped by VJ genes and CDR3 sequence, not performing further clustering steps, and therefore, the decision was to create an in-house workflow.

The creation of an in-house bioinformatics workflow specifically designed for our library design data (yielding partial reads of the IGH locus), is complemented with the pipeline adapted from the Immcantation Framework suite. Both approaches are different and therefore, their use was intended as a benchmarking for BMyRepCLL and cross validation of the results. The main strategy in BMyRepCLL is mapping reads against V and J IMGT alleles independently and generating a consensus sequence from the overlap of these reads, whereas the Immcantation Framework wraps IgBlast, meaning that gene annotation is performed on each read separately. This is followed by the performance of hierarchical clustering, which connects to other downstream analyses. The Immcantation Framework is very flexible and has solutions for many different library preparation methods and for bulk and single-cell sequencing of immune *repertoires*. Both workflows were adapted to the detection of abnormally expanded B cell clones, so as to make them specific for B cell neoplasms.

Even though the clinical application is not yet as feasible, pipelines performing reconstruction of the IGH locus from partial reads have been applied to WGS and WES data. In this case, the pipeline developed by Nadeu et al, and named IgCaller, integrates the analysis of the IGH locus with the detection of somatic genetic aberrations (176). As we mentioned before, our method can be integrated as well with the targeted sequencing of B cell neoplasms gene panels, offering better sequencing qualities than the common protocols. In single-cell RNA-seq analyses, pipelines such as BRACER and TRUST4 are being used to reconstruct VDJ region sequences from split reads, to map gene expression information and immune *repertoires* in the same experiment (some are also applicable to bulk RNA-seq, but pairing of gene expression and the BcR/TcR receptors information is not obtained) (177–180).

5.3 Performance of BMyRepCLL

5.3.1 Quality control

The importance of preprocessing steps relies on the necessity to improve read quality and other parameters which will facilitate the correct performance of downstream steps in bioinformatic workflows. Trimming reads under a quality Phred score of 30 (implying estimated probability of an incorrect basecall in 1/1000 bases) has been recommended for adaptive immune *repertoires*, for minimizing sequencing errors (181,182). In particular, it improved sequencing quality in the raw FASTQ files even when high quality sequencing experiments were obtained (Figure 4.8). Quality values obtained with the commercial samples were replicated with CLL patients, obtaining very high-confidence reads, as average Q-scores were above 30 in the whole set of samples (Figure 4.27). The report of artifacts has been assessed firstly with a commercial clonal sample, and a small proportion of artifact rearrangements was measured (0.88%). In this case, the example with a clonal control gives proof of the report of low levels of background noise alleles, being highly specific in reporting rearrangements.

In addition, these are the main control points in the in-house analysis that can be inspected by the user:

1) Importing the BAM files from V and J genes independently or inspecting the specific-rearrangement BAM file on IGV to validate low-coverage cases (reported as coverage breadth in the final table report), and variants.
2) Inspecting IGHV-IGHJ gene usage plots.

3) Using the consensus sequences to confirm results with the IMGT/V-QUEST web portal.

5.3.2 VDJ gene annotation

Annotation of *IGHV* and *IGHJ* alleles after the mapping strategy employed with BMyRepCLL, was achieved with 100% and 96.33% concordance with SSeq results (annotated with IMGT/V-QUEST), respectively, in the primary and secondary rearrangements considered clonal from the validation dataset. However, CDR3 concordance was lower (93.10%), since BMyRepCLL does not use IMGT/V-QUEST algorithms to delineate CDR3 sequence. For that reason, when CDR3 sequence is not encountered or it is potentially unproductive but the sequence is not shown in the final report, the consensus sequence obtained by NGS for that rearrangement should be inserted on IMGT/V-QUEST portal to check these results. These cases represent a 6.9% proportion of the samples evaluated and therefore, saves time with respect to other pipelines that require inserting the consensus sequence in IMGT/V-QUEST portal always. On the *IGHD* segment counterpart, BMyRepCLL reports the 3 alleles with the highest alignment scores. *IGHD* alleles comprise short sequences and therefore the annotation is more prone to errors. Furthermore, nucleotide deletions and P/N additions at the junction are responsible for additional untemplated diversity (136,183,184). For that reason, the concordance NGS-SSeq of D segments was not calculated.

5.3.3 Artifact filtering and rearrangement prioritization

Different studies and analysis frameworks use different definitions of a clonotype (185). In a work by Soto and collaborators, clonotypes are defined as sequences with the same *IGHV*, *IGHJ* genes (also their homologues in the light chain), and CDR3 amino acids sequence (186). On the other hand, CellRanger pipeline for analyzing BcR *repertoires* after VDJ 10X single-cell sequencing (10x Genomics, California; EEUU), defines clonotypes as groups of cells sharing the same CDR3 sequence in the variable regions from both heavy and light chains. Regardless of these conceptual differences, the important fact is that it is highly improbable that two different B cells possess the same V-J-CDR3 sequence combinations, or just the same CDR3 sequence, as it is a sequence ~13 amino acids long, and subjected to enormous variation (187). As Heavy and Light chains paired information extraction is not straightforward for bulk sequencing (188) (there is a recent work which approximates this pairing from bulk sequencing: (189)), in most studies it is sufficient to employ the heavy chain information alone to define B cell clonotypes (19,190). The importance of clonotype definition is to represent the minimum information to set a fingerprint that can be tracked across different immune *repertoire* sequencing methods (191).

Regarding development stages of B cells, there are several points where the definition of a "clone" can be drawn (heavy chain, paired light-heavy chain, SHM between cells originated from the same unique receptor rearrangement…). In a naïve B cell *repertoire*, the number of clonotypes is consistent with the number of different receptors (as with T cell receptors, that do not suffer SHM), but after SHM and clonal expansion, cells with the same original rearrangement can be grouped into different clonotypes, depending on the degree of SHM and how it changed the original sequence. Therefore, even though NGS provided with the possibility of sequencing numerous B cell receptors, dissecting which B cells arise from the same naïve B cell of origin, represents a challenging task (146). For instance, longer CDR3s and higher mutation degrees need lower clonal thresholds (191).

After analyzing the CLL *repertoire*, the first goal was to ensure that the rearrangements encountered previously with SSeq for these patients were correctly characterized in means of VDJ genes, mutational status, CDR3 sequence, etc. However, distinguishing only the major rearrangement left behind much more information that was being reported in the output tables for each of these patients.

The purpose of the second module of the pipeline was to perform a prioritization of rearrangements so that false positive clones are avoided (more than 100 clones are reported on average before performing these filters). Otherwise, the predominant rearrangement could be selected among the rest, but it will involve including many false positives as cases of multiple rearrangements can also be present.

Due to the mapping strategy used in BMyRepCLL, different reads from the same clone can align with high confidence against different *IGHV* genes and alleles. Clonotype information, composed in this case by V-J and CDR3 is used to detect these kind of artifacts (not the V in this case). Following this basis, rearrangements that have been split into different *IGHV* genes or alleles, are joined using the similarity methods described for filtering in the second module of the pipeline. Similar filtering steps have been reviewed in (192).

In results section 4.6.5, the aforementioned filtering steps were evaluated. In Figure 4.34a-c, we see uncorrelated count of *IGHV* alleles or genes with respect to the number junction sequences (which approximates to 1 clone/1 CDR3 sequence), meaning that until the last filtering step (Figure 4.34d) the number of rearrangements was being overestimated. Even though the regression coefficient, does not reach a strong correlation after the clonal threshold step (r-square=0.4) (Figure 4.34d), this is due to 24/314 points that do not fit the regression (7.6%) (Table 4.12). All 24 cases harbored a superior number of *IGHV* gene rearrangements than CDR3 sequences. These cases can also be due to various clones with unproductive CDR3s (for example: 2 different rearrangements but only one of them was determined productive by the pipeline, and these sometimes appear as "None").

The rearrangement information was curated throughout each of the prioritization approaches, and it has been demonstrated observing rearrangement ranks and the unique number of *IGHV* genes as well (Figure 4.33, Figure 4.35). The number of unique *IGHV* genes is reduced from the first to the last filtering step, resembling more to the CLL *repertoire* (Figure 4.35). As it has previously been reported, CLL harbors a biased BcR *repertoire* with usage of recurrent *IGHV* genes (*IGHV1-69*, *IGHV3-23, IGHV3-7*, *IGHV3-21*, *IGHV4-34*, among others). These genes are present in the majority of cases either as UM or MM. For instance, *IGHV1-69* is reported with lack or presence of very few mutations, whereas genes *IGHV4-34*, *IGHV3-7* and *IGHV3-23* contain a high load of mutations (Figure 4.28)

(53,62,63,66,87,193). In the normal B cell *repertoire*, genes such as *IGHV3-23*, *IGHV3-7*, *IGHV3-30* are highly represented (194).

Even after these steps, artifacts can remain, but the purpose has been to minimize them to a feasible extent, and then setting the clonal threshold, as various clones are reported per sample. BMyRepCLL validation numbers, with a 100% sensitivity and 97.15% specificity in reporting CLL clonal rearrangements, reflects the successful application of different types of filters: sequence similarity joining, fragment-biases reduction and clonal threshold.

5.4 Comparison between the analysis methods

5.4.1 Mutational status

Differences have been described between bioinformatics tools performing VDJ call annotation, regarding the inner methodology used by each software (173). Using CLL Immcantation improved the correlation of identity percentages between SSeq and NGS with respect to BMyRepCLL. Igblast is a well-known tool developed for accurate annotation of VDJ alleles and it is highly standardized (195). Even though Pearson correlation shows higher r-squared values for the comparison Immcantation-SSeq, the number of samples correctly classified is higher in BMyRepCLL, with 99.05% of concordance, being discordant in 3 Sanger borderline cases. 6 cases with mutational status discordances Immcantation-SSeq (Table 4.13), were correctly determined using BMyRepCLL. The concordance Immcantation-SSeq is 98.1%.

On the other hand, differences in BD classification with CLL Immcantation did not reclassify mutational status, as those 2 samples were categorized as MM, and the ratio of BD cases correct delineation was higher in CLL Immcantation than in BMyRepCLL (77.8 vs 66.7%, respectively). As commented previously in 1.2.7, there is no specific stratification for the marginal BD interval, and these patients are considered M-CLL for clinical decisions. Some works declare that survival is equivalent to those M-CLL, and others report mixture of indolent and aggressive courses (110,196), so these should be correctly identified and treated carefully.

The pipeline constructed with the Immcantation framework is more sensitive to intraclonal variations as it uses all the set of sequences and not a consensus, and therefore the definition of the mutational status is more precise but there are more samples with SSeq-discordant mutational status than with BMyRepCLL.

We compared CDR3 lengths regarding the mutational status. CDR3 length is commonly determined after analyzing immune repertoires. A diverse repertoire is expected to have a Gaussian distribution when representing CDR3 lengths, which does not occur with highly clonally expanded, abundant clones (197) or when *IGHV* replacement takes place (198).

CDR3 lengths in BD-CLL have been described to be more similar to MM-CLL clones. We obtained significant differences in pairwise comparison between the 3 groups and log2 fold change was lower in MM vs BD with respect to comparisons against the UM group (Figure 4.23b), but a lower sample was available in the case of BD patients (5 cases), and this comparison should be further evaluated with a larger sample. However, the differences in CDR3 lengths between UM-CLL and MM-CLL clones have been described as longer for the UM groups as we observe with the dataset analyzed using CLL Immcantation, and attributed to certain gene usage (62,64,65,87).

5.4.2 Clustering methods

Clonal lineage inference is a recurrent step in the analysis of adaptive immune *repertoires* (199,200). The implementation of methods for this purpose can be found in many different works with diversity in the approaches employed (144,160,187,201,202). Even probabilistic methods have been described for inferring clonally related sequences, but they are computationally expensive (203,204).

CLL *repertoire* is characterized by a low diversity for having one or a few predominant clones (55). Low *repertoire* diversity is also related to ageing, making adaptive immune response less efficient among older population (205–207). This highly skewed *repertoire* made the difference when performing clonal threshold definition with CLL Immcantation, not obtaining a bimodal distribution of sequence distances as it has been described for B cell *repertoires* (187,208), and therefore, the aim of finding a threshold specific per patient could not be fulfilled. The same distribution was obtained when plotting sequence distances against an intersample background (Figure 4.16). Even though Sanger sequences clustered within the predominant clone accordingly (85% of samples had SSeq sequence grouped in the predominant NGS clone), 5 rearrangements confirmed using SSeq were not detected and therefore, in future works the application of this method should be revisited to find a more specific cut-off for clonal relationships.

BMyRepCLL and CLL Immcantation use clonal grouping methods based on sequence similarity. BMyRepCLL performs grouping of clones using comparison of the identity percentages between their consensus sequences, whereas CLL Immcantation employs hierarchical clustering with the hamming distance at a sequence level, measuring nucleotide distances in the junction sequence. Both methods are challenged to be biased by the start-point element, and those are minimized when performing sequence grouping priorly (191). *defineClones* function in the Immcantation framework uses *IGHV*, *IGHJ* calls and CDR3 length to subdivide sequences as in (127,208–214) and only after that, hamming distance is applied. To implement this method, they tested different distance metrics and linkage methods (single, average and complete) to assess clonal relationships, and found single-linkage hierarchical clustering with the hamming distance to be the best performance in means of sensitivity and specificity, surpassing models that take into account SHM biases (187). Similarly, the second module of BMyRepCLL, which performs gene prioritization, groups together rearrangements with the same *IGHV* gene/family. This grouping performed in two steps makes the process less computationally expensive.

5.4.3 Multiple rearrangements

The NGS approach described here allowed the detection of additional rearrangements that were initially under-appreciated with SSeq in a simple, unbiased manner. Detection of multiple rearrangements by SSeq is more tedious and less straightforward, as each *IGHV* family has to be amplified uniquely, using different primers and sequenced separately. Cases with multiple rearrangements from the same *IGHV* family are hardly discernible and unconsciously sequenced together with SSeq. 8/9 (88.9%) cases of the additional rearrangements detected only using NGS are due to this limitation, none of them being both considered productive. Regardless of the productivity, detecting various rearrangements from the same family can cause noise after SSeq, and distort the number of mutations detected.

Clone determination is especially relevant for cases with multiple CLL rearrangements. In this work, the majority of FP clones reported with BMyRepCLL were not being reported by Immcantation. On the other hand, MiXCR outperformed Immcantation in the detection of SSeq clones (4 out of 5 abovementioned cases not detected with Immcantation but confirmed with SSeq were reported with MiXCR), and the behavior was equal with the FPs determined by BMyRepCLL. 3 FP clones were encountered with the three methods, being dependent of the methodology inherent to each pipeline. Even though GeneScan results did not confirm the presence of a secondary clone in FP2.1, FP5 and FP7 (Figure 8.11, Figure 8.14, Figure 8.16), they could be present for being highly similar in length, but this would have to be proven designing allele-specific primers and using them for amplification to prove NGS results. On the other hand, in 5/9 cases the FP rearrangement detected with BMyRepCLL shared CDR3 with a confirmed true rearrangement, giving a hint for categorizing them as possible artifacts. Therefore, even though the pipeline reported them as real rearrangements, the comparison between the rearrangements found on each patient confirmed that they are not potential CLL clones, but derive from noise of a predominant Ig rearrangement.

In some cases, both CLL Immcantation and MiXCR reported various clones with the same VDJ rearrangements, in different proportions. However, to annotate percentage results, only the major in abundance was used, so as to not influence the results by manual inspection. This is precisely what we intended to avoid with BMyRepCLL, and therefore, joined rearrangements that were potentially the same clone. Following this basis, the results obtained with this pipeline are already curated and are not open to interpretations depending on the human expert.

MiXCR would have been a great option to determine the number of clones as it is fast and easy to use, with an extensive documentation and many support publications (215). However, the fact that it reports *IGHV* alignments with less confidence hinders its application for clinical determination of the mutational status.

5.4.4 Is it worth including FR3 amplicon?

Depending on the analysis methods, FR3 fragment can be valuable or just prone to artifactual *IGHV* calls. In this case, in BMyRepCLL we integrated these fragment reads or the information with the others, removing possible artifacts priorly. In the CLL Immcantation pipeline, the approach is very different and these reads are split from the rest of fragments in the steps of the workflow (FR1, FR2 and FR3). For that reason, when treated independently, FR3 information reports untrustworthy results mainly because of *IGHV* allele/gene miscalls, that provoked clustering into different clones (Figure 4.17; Example 2), or distorted mutation frequencies inside the major clones (Figure 4.17; Examples 1-3). As a consequence, in this pipeline we chose to remove those reads, whereas in BMyRepCLL they are used with the awareness that they need a careful integration with the rest of sequence data.

Currently, FR3 fragment alone is being used for clonotyping immune *repertoires* samples in B cell neoplasms, as CDR3 sequence can be characterized, but not the whole *IGHV* region (129,130). It is also useful for FFPE (formalin-fixed paraffin embedded) samples, where DNA can be notably degraded (129). With the methods implemented herein, our intention is to perform clonality and mutational status determination in a single workflow.

5.5 Advantages of the use of NGS for clinical applications in B cell neoplasms

The genetic landscape of CLL is extensively heterogeneous. AIRR-seq (Adaptive Immune Receptor Repertoire sequencing) is helping to understand antigen-driven selection of B cell lymphocytes, which derives in biased gene usage and stereotyped receptors, and the degree of SHM found in the variable region of the Ig heavy chain (IGH locus) (82,216–218). On the other hand, the impact of chromosome aberrations and other small scale mutations where different biological signaling pathways are involved, are also being further studied using NGS (46,219,220).

In the latest iwCLL guidelines, the important prognostic markers detailed are *TP53* alterations and *IGHV* mutational status (48). However, with the new NGS studies, these subgroups could be refined further, with harmonization of gene panels regarding IGH locus and genes found to be recurrently mutated (95). Using ultra-deep sequencing technologies (digital PCR as well), treatment resistances and relapses will be detected in earlier stages, and stratification regarding genetic markers will be refined.

With the efforts of many groups in the last decade, now it is understood, due to the inner complexity of CLL in means of *repertoire* and genetic background in general, that NGS represents an improvement in clinical determinations, due to the intrinsic and evident limitations of the use of SSeq, that have already been discussed for allowing a better understanding of adaptive immune *repertoires* (104,105). However, the standardization is the challenging part, as the additional information obtained using NGS has to be managed. That includes coping with a great initial number of rearrangements that, in the case of BMyRepCLL, were tackled by removing mapping biases, and therefore, after removing these artifacts, the clonal threshold was studied. Even though the MAX_DIFF value chosen for the decision between clonal and not-clonal, was arbitrary given the data used in this work (below 5 is considered polyclonal, and above, clonal), the threshold between the clonal and subclonal fraction within a patient depends solely on the *repertoire* architecture of that sample and not generally, or randomly chosen for all the patients, as it can be variable. Prior to performing this filter, we studied using fixed background in clonal percentages, as other works have used (107). This would not work, as we find subclonal rearrangements represented in different percentages among patients: the maximum clonal percentage in a rearrangement considered subclonal is 9%, whereas there are clonal rearrangements below that value.

5.5.1 Intraclonal diversity

Even though BMyRepCLL is adapted more thoughtfully to current clinical practices for being a consensus-oriented analysis, in the near future, as NGS determinations become more frequent and standardized, tools like the Immcantation Framework will be highly applicable to dissect the whole picture in CLL clones. Even though a trend compliant with the overall mutational status is maintained in mutational frequencies distribution (Figure 4.23), some patients have more variability inside the predominant clone that are not reflected calculating the average value among those sequences (Figure 4.21, Figure 4.22). On the other hand, we also found cases with discordant mutational status due to evidences of variability inside the major clone (Figure 4.39). All of the above will require inspection with the suitable tools. Moreover, CLL Immcantation allows to perform many downstream analyses that can be of utmost interest. For example, describing AID enzyme hotspots or constructing lineage trees to unravel clonal drift, could allow to extract more exhaustive profiles of CLL patients and broaden the scope of personalized medicine.

With the rise of NGS, intraclonal diversity is gaining importance in CLL as reviewed in (104,105), and studies have evidenced presence of intraclonal diversity with NGS (221). There is a tool developed recently for characterizing Ig intraclonal diversity (IgIDiVa) (222). The use of a consensus sequence with BMyRepCLL helps to avoid sequencing artifacts, and filtering low frequency variants has been implemented within CLL Immcantation to achieve the same. Combining these approaches with high-coverage experiments can help to characterize inner diversity. However, to characterize variations at very low proportions, special care is needed since experimental methods to be able to distinguish sequencing artifacts from SHM events such as UMIs, would have to be integrated in the protocol.

Methods for correcting sequencing errors computationally have been described for DNA-seq in general and concretely for antibodies and different sequencing platforms (223–229).

5.5.2 Subclonal architecture

Evidences point to adverse prognosis similar to U-CLL in cases with double productive rearrangements (108). Subclones need special attention, as clonal drift can make them become dominant. Studies in the last decade, point to strikingly adverse outcome in patients with presence of *TP53* subclones at diagnosis, comparable to cases with *TP53* aberrations (230–232). In addition, resistance to ibrutinib treatment has been related to the emergence of subclones harboring *BTK* and *PLCG2* mutations. This supports the need of a continuous follow-up of disease evolution (93).

A novel study, tracked multi-omics profiles of Richter-transformed CLL cells from bulk and single cell experiments, and discovered that these profiles matched subclones appearing even at the time of CLL diagnosis, augmenting the importance of the detection of subclonal rearrangements in early disease stages (233).

In the near future, NGS will be a powerful tool to determine the diversity in the CLL architecture of patients and provide new clinical stratification groups. For that reason, it is important that IGH rearrangements can be determined along with target gene panels to make the determination more straightforward. To identify somatic mutations, high depth and quality score values are needed and thus, shorter reads to avoid accumulating many phasing errors with Illumina sequencing, are the most recommendable approach. Characterizing the subclonal fraction at the beginning of the treatment could anticipate whether this patient will suffer relapse (234). For that reason, *TP53* alterations screening at the beginning each line of therapy is advisable to detect subclones ahead (235).

Subclonal rearrangements encountered with BMyRepCLL and supported by CLL Immcantation (Table 4.15), show potential capabilities in these workflows to determine the subclones present on each patient. However, for this work the specific validation of this part has not been performed as for the moment it is not included in the clinical guidelines but could serve as valuable information in the near future. To determine whether they are real subclones or not, it is important that CDR3 sequence is defined and not present in other rearrangements from the same sample (they can represent noise from the predominant clones) or even other samples (cross-contaminations), and advisable to detect them using various techniques and analysis methods.

5.6 Future prospects

BMyRepCLL circumvents many limitations of the gold-standard for determination of the mutational status in the IGH locus of CLL patients (SSeq), and simplifies the results obtained so as to overcome augmented information provided by NGS and inner variability. The results shown in this work, validate the pipeline for its potential use in the clinical area, as it is an automated and inexpensive process, employing 150bpx2 Illumina sequencing and being therefore, more feasible in terms of turnaround times, quality, and integration with other gene panels (also paired DNA-cDNA determinations). A multicentric study could validate these results further. On the other hand, to make the use of this pipeline user-friendly, it can be developed as a web app in the future.

The whole validation of the methods presented in this thesis work, employed DNA. However, we also adapted the methods to perform the same determinations from cDNA, and believe on the other hand, that this could present advantages to cover the entire IGH locus and refine mutational status determination. We encourage the use of these methods for a validation employing paired cDNA-DNA sequencing experiments.

5.6.1 Immcantation clustering methods

The clustering methods employed within CLL Immcantation workflow should be revisited to find a threshold specific within each patient. A reliable alternative would be testing the package *SCOPER* on our dataset, which employs a spectral clustering method for measuring specific thresholds, in cases where a bimodal distribution between nearest-neighbor sequence distances is not present (201,202,209,236,237). The great potential of the Immcantation suite will allow to perform many downstream analyses after the refinement of clonal threshold, discussed in 5.6.2.

Another option could be employing an alignment-free clustering method that does not rely on the initial gene assignments or junction length, solving cases prone to ambiguous V and J genes annotation and indels (small-scale insertions and deletions) occurring in the junction (238).

5.6.2 Downstream analyses

DIVERSITY

Diversity measures depict *repertoire* architecture and they can be used to determine disease profiles (162) and therefore, can also be used as a quality control measure to check if samples within each group are compliant with the expected repertoire diversity. In Figure 4.25 and Figure 4.26, we compared the profiles of healthy donors and CLL patients using alpha diversity and abundance calculations from package Alakazam (Immcantation Framework). The use of these measures could be further explored to set quality control methods and explore the architecture of the adaptive immune *repertoire* of B cell neoplasms.

MUTATION ANALYSIS

SHM entails the introduction of point mutations along the variable region of IGH and IGL/IGK chains at a rate of ~1/1000 bp/cell division. This is 10^6 times higher than mutation rates occurring in other parts of the genome (239–241). CDRs are subjected to more selective pressure than FWRs, as they contain higher missense/silent mutation rates (242). This fact evidences T-cell dependent antigen recognition by M-CLL during GC reactions (63). Since both M-CLL and U-CLL have been proven to derive from mature activated B cells, U-CLL could undergo antigen/autoantigen recognition independently of T cells and GCs or be selected with none or a scarce presence of mutations. Antigen influence in this scenario is supported by the fact that there are also high missense/silent mutations ratio in CDR3 of minimally mutated U-CLL clones (243).

SHM patterns can be studied to identify hotspots and constructing lineage trees representing different time-points or immunological events, which can help to dissect clonal evolution processes and affinity maturation in CLL B cell clones (141,244).

There are also tools designed for the discovery of new alleles of the V(D)J genes from TcR and BcR data. Allele databases are incomplete and it is a challenge to differentiate whether mutations are due to SHM, or by contrary, novel alleles can be present (245).

5.6.3 MRD monitoring

Minimum residual disease is conventionally assessed with the use of flow cytometry, validated by ERIC consortium (246,247). MRD analysis has been approved by the EMA (European Medicines Agency), in randomized clinical trials, for treatment efficacy assessment (248). Two preliminary studies involving MRD detection in B cell neoplasms were conducted in the Hematology Unit of the Clinical University Hospital of Valencia, employing the methods developed in this thesis:

1) 69 samples from 19 CLL patients under treatment were sequenced and analyzed using BMyRepCLL after gDNA amplification with the multiplexed FR regions method. 44 samples were MRD positive by flow cytometry and NGS, and 3 were negative by both techniques as well. 3 samples were negative by flow cytometry whereas the clones were detected using NGS (249).
2) Liquid biopsy samples not only from CLL, but different B cell lymphomas were sequenced and analyzed with this method (ctDNAs extracted during and after receiving treatment) to identify residual tumoral cells. Residual IGH clones after treatment could be detected in 80% of the samples tested, where

patients in remission and those refractory or with relapse risk are identifiable (250).

The method has been used for detecting MRD along with the mutational status and clonality characterization of the samples, and thus, presents a potential alternative for the paired SSeq-flow cytometry use in the conventional clinical routines. The use of NGS will provide with more straightforward determinations whereas the conventional methods can be used for validation of unclear results.

6 Conclusions

The final remarks of this thesis work are:

1. The multiplexed in-house library preparation method has been proven to be the optimal method for NGS sequencing of the IGH locus in CLL patients, being easily adaptable to clinical routines, for being cost-effective and improving turnaround times with respect to the other two methods tested. The in-house method employs the 3 sets of Framework Regions primers (FR1, FR2, FR3) from the BIOMED-2 standard primers design and 150bpx2 cycles Illumina kit, shorter than the method commonly employed (300bpx2), and compatible with Illumina platforms of higher capacities. Moreover, the same method from cDNA has been optimized as well, including the Leader primers set to cover the IGH locus region entirely.
2. A specific bioinformatic pipeline to reconstruct VDJ genes from the partial reads obtained from the aforementioned library design, has been designed and programmed from scratch, integrating scripts with open-source bioinformatics software. Apart from detecting the predominant CLL clones, the pipeline provides with direct distinction of the clonal and subclonal fraction after exhaustive adjustments of analysis steps implemented for artifact removal and prioritization of rearrangements.
3. A second bioinformatic pipeline has been developed from tools designed by experts in the analysis of adaptive immune repertoires. Since this pipeline employs computational methods which are specific for in-depth studies of B cell repertoires, it allowed to observe variability in the mutation frequencies within the predominant clones of some patients.
4. The validation against the gold-standard techniques, demonstrated that the methods developed herein for sequencing and bioinformatics analysis of the IGH locus, are highly robust in the annotation of characteristics of VDJ rearrangements and the report of potential

expanded clones, with a sensitivity of 100% and a specificity of 97%. Last but not least, the mutational status has been characterized equally to SSeq in 99% of the patients studied.

7 References

1. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes Immun. 2012 Jul;13(5):363–73.

2. Jung D, Giallourakis C, Mostoslavsky R, Alt FW. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. Annu Rev Immunol. 2006;24(1):541–70.

3. Teng G, Papavasiliou FN. Immunoglobulin somatic hypermutation. Annu Rev Genet. 2007;41(1):107–20.

4. Early P, Huang H, Davis M, Calame K, Hood L. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. Cell. 1980 Apr;19(4):981–92.

5. Early P, Huang H, Davis M, Calame K, Hood L. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. 1980. J Immunol. 2004 Dec 1;173(11):6503–14.

6. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W34-40.

7. Lucas JS, Murre C, Feeney AJ, Riblet R. The structure and regulation of the immunoglobulin loci. In: Molecular Biology of B Cells. Elsevier; 2015. p. 1–11.

8. Hengeveld PJ, Levin M-D, Kolijn PM, Langerak AW. Reading the B-cell receptor immunome in chronic lymphocytic leukemia: revelations and applications. Exp Hematol. 2021 Jan;93:14–24.

9. Hozumi N, Tonegawa S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. Proc Natl Acad Sci U S A. 1976 Oct;73(10):3628–32.

10. Bassing CH, Swat W, Alt FW. The mechanism and regulation of chromosomal V(D)J recombination. Cell. 2002 Apr;109 Suppl:S45-55.

11. Tonegawa S. Somatic generation of antibody diversity. In: Immunology. Elsevier; 1995. p. 145–62.

12. Hesslein DG, Schatz DG. Factors and forces controlling V(D)J recombination. Adv Immunol. 2001;78:169–232.

13. Hardy RR, Carmack CE, Shinton SA, Kemp JD, Hayakawa K. Resolution and characterization of pro-B and pre-pro-B cell stages in normal mouse bone marrow. J Exp Med. 1991 May 1;173(5):1213–25.

14. Hendrickson EA, Qin XQ, Bump EA, Schatz DG, Oettinger M, Weaver DT. A link between double-strand break-related repair and V(D)J recombination: the scid mutation. Proc Natl Acad Sci U S A. 1991 May 15;88(10):4061–5.

15. van Gent DC, McBlane JF, Ramsden DA, Sadofsky MJ, Hesse JE, Gellert M. Initiation of V(D)J recombinations in a cell-free system by RAG1 and RAG2 proteins. Curr Top Microbiol Immunol. 1996;217:1–10.

16. Roth DB, Zhu C, Gellert M. Characterization of broken DNA molecules associated with V(D)J recombination. Proc Natl Acad Sci U S A. 1993 Nov 15;90(22):10788–92.

17. Schlissel M, Constantinescu A, Morrow T, Baxter M, Peng A. Double-strand signal sequence breaks in V(D)J recombination are blunt, 5′-phosphorylated, RAG-dependent, and cell cycle regulated. Genes Dev. 1993 Dec;7(12B):2520–32.

18. Little AJ, Matthews A, Oettinger M, Roth DB, Schatz DG. The mechanism of V(D)J recombination. In: Molecular Biology of B Cells. Elsevier; 2015. p. 13–34.

19. Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. Immunity. 2000 Jul;13(1):37–45.

20. Morea V, Tramontano A, Rustici M, Chothia C, Lesk AM. Conformations of the third hypervariable region in the VH domain of immunoglobulins. J Mol Biol. 1998 Jan 16;275(2):269–94.

21. Larimore K, McCormick MW, Robins HS, Greenberg PD. Shaping of human germline IgH repertoires revealed by deep sequencing. J Immunol. 2012 Sep 15;189(6):3221–30.

22. Schatz DG, Spanopoulou E. Biochemistry of V(D)J recombination. Curr Top Microbiol Immunol. 2005;290:49–85.

23. Grawunder U, Leu TM, Schatz DG, Werner A, Rolink AG, Melchers F, et al. Down-regulation of RAG1 and RAG2 gene expression in preB cells after functional immunoglobulin heavy chain rearrangement. Immunity. 1995 Nov;3(5):601–8.

24. Abbas AK, Lichtman AH, Pillai S, Baker DL. Lymphocyte development and the rearrangement and expression of antigen receptor genes. In: Cellular and Molecular Immunology. Elsevier; 2010. p. 153–87.

25. Ghia P, Scielzo C, Frenquelli M, Muzio M, Caligaris-Cappio F. From normal to clonal B cells: Chronic lymphocytic leukemia (CLL) at the crossroad between neoplasia and autoimmunity. Autoimmun Rev. 2007 Dec;7(2):127–31.

26. Moser M, Leo O. Key concepts in immunology. Vaccine. 2010 Aug 31;28 Suppl 3:C2-13.

27. Calis JJA, Rosenberg BR. Characterizing immune repertoires by high throughput sequencing: strategies and applications. Trends Immunol. 2014 Dec;35(12):581–90.

28. Nussenzweig MC, Alt FW. Antibody diversity: one enzyme to rule them all. Nat Med. 2004 Dec;10(12):1304–5.

29. Marks C, Deane CM. How repertoire data are changing antibody science. J Biol Chem. 2020 Jul 17;295(29):9823–37.

30. Matsuda F, Ishii K, Bourvagnet P, Kuma K i., Hayashida H, Miyata T, et al. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. J Exp Med. 1998 Dec 7;188(11):2151–62.

31. Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucleic Acids Res. 2015 Jan;43(Database issue):D413-22.

32. Giudicelli V, Chaume D, Lefranc M-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D256-61.

33. Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol. 2003 Jan;27(1):55–77.

34. Lefranc M-P. IMGT Unique Numbering. In: Encyclopedia of Systems Biology. New York, NY: Springer New York; 2013. p. 952–9.

35. Lefranc M-P, Lefranc G. IGHV. In: The Immunoglobulin FactsBook. Elsevier; 2001. p. 107–240.

36. Romo-González T, Vargas-Madrazo E. Structural analysis of substitution patterns in alleles of human immunoglobulin VH genes. Mol Immunol. 2005 May;42(9):1085–97.

37. Liu L, Lucas AH. IGH V3-23*01 and its allele V3-23*03 differ in their capacity to form the canonical human antibody combining site specific for the capsular polysaccharide of Haemophilus influenzae type b. Immunogenetics. 2003 Aug;55(5):336–8.

38. Hallek M. Chronic lymphocytic leukemia: 2015 Update on diagnosis, risk stratification, and treatment. Am J Hematol. 2015 May;90(5):446–60.

39. Dores GM, Anderson WF, Curtis RE, Landgren O, Ostroumova E, Bluhm EC, et al. Chronic lymphocytic leukaemia and small lymphocytic lymphoma: overview of the descriptive epidemiology. Br J Haematol. 2007 Dec;139(5):809–19.

40. Chiorazzi N, Chen S-S, Rai KR. Chronic Lymphocytic leukemia. Cold Spring Harb Perspect Med. 2021 Feb 1;11(2):a035220.

41. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, et al. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A. 2002 Nov 26;99(24):15524–9.

42. Klein U, Lia M, Crespo M, Siegel R, Shen Q, Mo T, et al. The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia. Cancer Cell. 2010 Jan 19;17(1):28–40.

43. Lähdesmäki A, Kimby E, Duke V, Foroni L, Hammarström L. ATM mutations in B-cell chronic lymphocytic leukemia. Haematologica. 2004 Jan;89(1):109–10.

44. Schaffner C, Stilgenbauer S, Rappold GA, Döhner H, Lichter P. Somatic ATM mutations indicate a pathogenic role of ATM in B-cell chronic Lymphocytic leukemia. Blood. 1999 Jul 15;94(2):748–53.

45. Zenz T, Benner A, Döhner H, Stilgenbauer S. Chronic lymphocytic leukemia and treatment resistance in cancer: the role of the p53 pathway. Cell Cycle. 2008 Dec 15;7(24):3810–4.

46. Puente XS, Pinyol M, Quesada V, Conde L, Ordóñez GR, Villamor N, et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature. 2011 Jun 5;475(7354):101–5.

47. Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nat Genet. 2011 Dec 11;44(1):47–52.

48. Hallek M, Cheson BD, Catovsky D, Caligaris-Cappio F, Dighiero G, Döhner H, et al. iwCLL guidelines for diagnosis, indications for treatment, response assessment, and supportive management of CLL. Blood. 2018 Jun 21;131(25):2745–60.

49. Hallek et al. Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute–Working Group 1996 guidelines. Blood. 2008;111:5446-5456. Blood. 2008 Dec 15;112(13):5259–5259.

50. Brown CJ, Lain S, Verma CS, Fersht AR, Lane DP. Awakening guardian angels: drugging the p53 pathway. Nat Rev Cancer. 2009 Dec;9(12):862–73.

51. Campo E, Cymbalista F, Ghia P, Jäger U, Pospisilova S, Rosenquist R, et al. TP53 aberrations in chronic lymphocytic leukemia: an overview of the clinical implications of improved diagnostics. Haematologica. 2018 Dec;103(12):1956–68.

52. Malcikova J, Tausch E, Rossi D, Sutton LA, Soussi T, Zenz T, et al. ERIC recommendations for TP53 mutation analysis in chronic lymphocytic leukemia—update on methodological approaches and results interpretation. Leukemia. 2018 May;32(5):1070–80.

53. Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. Blood. 1999 Sep 15;94(6):1848–54.

54. Damle RN, Wasil T, Fais F, Ghiotto F, Valetto A, Allen SL, et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. Blood. 1999 Sep 15;94(6):1840–7.

55. van Dongen JJM, Langerak AW, Brüggemann M, Evans PAS, Hummel M, Lavender FL, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia. 2003 Dec;17(12):2257–317.

56. Langerak AW, Davi F, Ghia P, Hadzidimitriou A, Murray F, Potter KN, et al. Immunoglobulin sequence analysis and prognostication in CLL: guidelines from the ERIC review board for reliable interpretation of problematic cases. Leukemia. 2011 Jun;25(6):979–84.

57. Kikushige Y, Ishikawa F, Miyamoto T, Shima T, Urata S, Yoshimoto G, et al. Self-renewing hematopoietic stem cell is the primary target in pathogenesis of human chronic lymphocytic leukemia. Cancer Cell. 2011 Aug 16;20(2):246–59.

58. Rosenwald A, Alizadeh AA, Widhopf G, Simon R, Davis RE, Yu X, et al. Relation of gene expression phenotype to immunoglobulin mutation genotype in B cell chronic lymphocytic leukemia. J Exp Med. 2001 Dec 3;194(11):1639–47.

59. Klein U, Tu Y, Stolovitzky GA, Mattioli M, Cattoretti G, Husson H, et al. Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. J Exp Med. 2001 Dec 3;194(11):1625–38.

60. Damle RN, Ghiotto F, Valetto A, Albesiano E, Fais F, Yan X-J, et al. B-cell chronic lymphocytic leukemia cells express a surface membrane phenotype of activated, antigen-experienced B lymphocytes. Blood. 2002 Jun 1;99(11):4087–93.

61. Bosch F, Dalla-Favera R. Chronic lymphocytic leukaemia: from genetics to treatment. Nat Rev Clin Oncol. 2019 Nov;16(11):684–701.

62. Duke VM, Gandini D, Sherrington PD, Lin K, Heelan B, Amlot P, et al. V(H) gene usage differs in germline and mutated B-cell chronic lymphocytic leukemia. Haematologica. 2003 Nov;88(11):1259–71.

63. Fais F, Ghiotto F, Hashimoto S, Sellars B, Valetto A, Allen SL, et al. Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. J Clin Invest. 1998 Oct 15;102(8):1515–25.

64. Donisi PM, Di Lorenzo N, Riccardi M, Paparella A, Sarpellon C, Zupo S, et al. Pattern and distribution of immunoglobulin VH gene usage in a cohort of B-CLL patients from a Northeastern region of Italy. Diagn Mol Pathol. 2006 Dec;15(4):206–15.

65. Karan-Djurasevic T, Palibrk V, Kostic T, Spasovski V, Nikcevic G, Srzentic S, et al. Mutational status and gene repertoire of IGHV-IGHD-IGHJ rearrangements in Serbian patients with chronic lymphocytic leukemia. Clin Lymphoma Myeloma Leuk. 2012 Aug;12(4):252–60.

66. Ghia P, Stamatopoulos K, Belessi C, Moreno C, Stella S, Guida G, et al. Geographic patterns and pathogenetic implications of IGHV gene usage in chronic lymphocytic leukemia: the lesson of the IGHV3-21 gene. Blood. 2005 Feb 15;105(4):1678–85.

67. Murray F, Darzentas N, Hadzidimitriou A, Tobin G, Boudjogra M, Scielzo C, et al. Stereotyped patterns of somatic hypermutation in subsets of patients with chronic lymphocytic leukemia: implications for the role of antigen selection in leukemogenesis. Blood. 2008 Feb 1;111(3):1524–33.

68. Chiorazzi N, Rai KR, Ferrarini M. Chronic lymphocytic leukemia. N Engl J Med. 2005 Feb 24;352(8):804–15.

69. Arvaniti E, Ntoufa S, Papakonstantinou N, Touloumenidou T, Laoutaris N, Anagnostopoulos A, et al. Toll-like receptor signaling pathway in chronic lymphocytic leukemia: distinct gene expression profiles of potential pathogenic significance in specific subsets of patients. Haematologica. 2011 Nov;96(11):1644–52.

70. Stevenson FK, Krysov S, Davies AJ, Steele AJ, Packham G. B-cell receptor signaling in chronic lymphocytic leukemia. Blood. 2011 Oct 20;118(16):4313–20.

71. Agathangelidis A, Ntoufa S, Stamatopoulos K. B cell receptor and antigens in CLL. Adv Exp Med Biol. 2013;792:1–24.

72. Marcatili P, Ghiotto F, Tenca C, Chailyan A, Mazzarello AN, Yan X-J, et al. Igs expressed by chronic lymphocytic leukemia B cells show limited binding-site structure variability. J Immunol. 2013 Jun 1;190(11):5771–8.

73. Ten Hacken E, Gounari M, Ghia P, Burger JA. The importance of B cell receptor isotypes and stereotypes in chronic lymphocytic leukemia. Leukemia. 2019 Feb;33(2):287–98.

74. Hervé M, Xu K, Ng Y-S, Wardemann H, Albesiano E, Messmer BT, et al. Unmutated and mutated chronic lymphocytic leukemias derive from self-reactive B cell precursors despite expressing different antibody reactivity. J Clin Invest. 2005 Jun;115(6):1636–43.

75. Lanemo Myhrinder A, Hellqvist E, Sidorova E, Söderberg A, Baxendale H, Dahle C, et al. A new perspective: molecular motifs on oxidized LDL, apoptotic cells, and bacteria are targets for chronic lymphocytic leukemia antibodies. Blood. 2008 Apr 1;111(7):3838–48.

76. Burger JA, Ghia P, Rosenwald A, Caligaris-Cappio F. The microenvironment in mature B-cell malignancies: a target for new treatment strategies. Blood. 2009 Oct 15;114(16):3367–75.

77. Ghia P, Circosta P, Scielzo C, Vallario A, Camporeale A, Granziero L, et al. Differential effects on CLL cell survival exerted by different microenvironmental elements. Curr Top Microbiol Immunol. 2005;294:135–45.

78. Tobin G, Thunberg U, Karlsson K, Murray F, Laurell A, Willander K, et al. Subsets with restricted immunoglobulin gene rearrangement features indicate a role for antigen selection in the development of chronic lymphocytic leukemia. Blood. 2004 Nov 1;104(9):2879–85.

79. Bomben R, Dal Bo M, Capello D, Forconi F, Maffei R, Laurenti L, et al. Molecular and clinical features of chronic lymphocytic leukaemia with stereotyped B cell receptors: results from an Italian multicentre study. Br J Haematol. 2009 Feb;144(4):492–506.

80. Ghiotto F, Fais F, Valetto A, Albesiano E, Hashimoto S, Dono M, et al. Remarkably similar antigen receptors among a subset of patients with chronic lymphocytic leukemia. J Clin Invest. 2004 Apr;113(7):1008–16.

81. Darzentas N, Hadzidimitriou A, Murray F, Hatzi K, Josefsson P, Laoutaris N, et al. A different ontogenesis for chronic lymphocytic leukemia cases carrying stereotyped antigen receptors: molecular and computational evidence. Leukemia. 2010 Jan;24(1):125–32.

82. Agathangelidis A, Darzentas N, Hadzidimitriou A, Brochet X, Murray F, Yan X-J, et al. Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. Blood. 2012 May 10;119(19):4467–75.

83. Messmer BT, Albesiano E, Efremov DG, Ghiotto F, Allen SL, Kolitz J, et al. Multiple distinct sets of stereotyped antigen receptors indicate a role for antigen in promoting chronic lymphocytic leukemia. J Exp Med. 2004 Aug 16;200(4):519–25.

84. Darzentas N, Stamatopoulos K. The significance of stereotyped B-cell receptors in chronic lymphocytic leukemia. Hematol Oncol Clin North Am. 2013 Apr;27(2):237–50.

85. Murray F, Thorselius M, Krober A, Thunberg U, Tobin G, Buhler A, et al. Strikingly homologous immunoglobulin gene rearrangements and poor outcome in VH3-21-utilizing chronic Lymphocytic leukemia independent of geographical origin and mutational status. Blood. 2005 Nov 16;106(11):175–175.

86. Tobin G, Thunberg U, Johnson A, Thörn I, Söderberg O, Hultdin M, et al. Somatically mutated Ig V(H)3-21 genes characterize a new subset of chronic lymphocytic leukemia. Blood. 2002 Mar 15;99(6):2262–4.

87. Stamatopoulos K, Belessi C, Moreno C, Boudjograh M, Guida G, Smilevska T, et al. Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: Pathogenetic implications and clinical correlations. Blood. 2007 Jan 1;109(1):259–70.

88. Potter KN, Mockridge CI, Neville L, Wheatley I, Schenk M, Orchard J, et al. Structural and functional features of the B-cell receptor in IgG-positive chronic lymphocytic leukemia. Clin Cancer Res. 2006 Mar 15;12(6):1672–9.

89. Agathangelidis A, Chatzidimitriou A, Chatzikonstantinou T, Tresoldi C, Davis Z, Giudicelli V, et al. Immunoglobulin gene sequence analysis in chronic lymphocytic leukemia: the 2022 update of the recommendations by ERIC, the European Research Initiative on CLL. Leukemia [Internet]. 2022 May 25; Available from: http://dx.doi.org/10.1038/s41375-022-01604-2

90. Dühren-von Minden M, Übelhart R, Schneider D, Wossning T, Bach MP, Buchner M, et al. Chronic lymphocytic leukaemia is driven by antigen-independent cell-autonomous signalling. Nature. 2012 Sep 13;489(7415):309–12.

91. López-Oreja I, Playa-Albinyana H, Arenas F, López-Guerra M, Colomer D. Challenges with approved targeted therapies against recurrent mutations in CLL: A place for new actionable targets. Cancers (Basel). 2021 Jun 24;13(13):3150.

92. Fabbri G, Dalla-Favera R. The molecular pathogenesis of chronic lymphocytic leukaemia. Nat Rev Cancer. 2016 Mar;16(3):145–62.

93. Woyach JA, Furman RR, Liu T-M, Ozer HG, Zapatka M, Ruppert AS, et al. Resistance mechanisms for the Bruton's tyrosine kinase inhibitor ibrutinib. N Engl J Med. 2014 Jun 12;370(24):2286–94.

94. Baliakas P, Hadzidimitriou A, Sutton L-A, Rossi D, Minga E, Villamor N, et al. Recurrent mutations refine prognosis in chronic lymphocytic leukemia. Leukemia. 2015 Feb;29(2):329–36.

95. Rossi D, Rasi S, Spina V, Bruscaggin A, Monti S, Ciardullo C, et al. Integrated mutational and cytogenetic analysis identifies new prognostic subgroups in chronic lymphocytic leukemia. Blood. 2013 Feb 21;121(8):1403–12.

96. Li Y, Li H, Weigert M. Autoreactive B cells in the marginal zone that express dual receptors. J Exp Med. 2002 Jan 21;195(2):181–8.

97. Plevova K, Francova HS, Burckova K, Brychtova Y, Doubek M, Pavlova S, et al. Multiple productive immunoglobulin heavy chain gene rearrangements in chronic lymphocytic leukemia are mostly derived from independent clones. Haematologica. 2014 Feb;99(2):329–38.

98. Kriangkum J, Motz SN, Mack T, Beiggi S, Baigorri E, Kuppusamy H, et al. Single-cell analysis and next-generation immuno-sequencing show that multiple clones persist in patients with chronic Lymphocytic leukemia. PLoS One. 2015 Sep 9;10(9):e0137232.

99. Küppers R, Klein U, Hansmann ML, Rajewsky K. Cellular origin of human B-cell lymphomas. N Engl J Med. 1999 Nov 11;341(20):1520–9.

100. Sutton L-A, Kostareli E, Hadzidimitriou A, Darzentas N, Tsaftaris A, Anagnostopoulos A, et al. Extensive intraclonal diversification in a subgroup of chronic Lymphocytic leukemia patients with stereotyped IGHV4-34/IGKV2-30 B cell receptors: Implications for ongoing interactions with antigen. Blood. 2009 Nov 20;114(22):2337–2337.

101. Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, et al. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc Natl Acad Sci U S A. 2008 Sep 2;105(35):13081–6.

102. Logan AC, Gao H, Wang C, Sahaf B, Jones CD, Marshall EL, et al. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. Proc Natl Acad Sci U S A. 2011 Dec 27;108(52):21194–9.

103. Klinger M, Zheng J, Elenitoba-Johnson KSJ, Perkins SL, Faham M, Bahler DW. Next-generation IgVH sequencing CLL-like monoclonal B-cell lymphocytosis reveals frequent oligoclonality and ongoing hypermutation. Leukemia. 2016 May;30(5):1055–61.

104. Gupta SK, Viswanatha DS, Patel KP. Evaluation of somatic hypermutation status in chronic Lymphocytic leukemia (CLL) in the era of next generation sequencing. Front Cell Dev Biol. 2020 May 19;8:357.

105. Davi F, Langerak AW, de Septenville AL, Kolijn PM, Hengeveld PJ, Chatzidimitriou A, et al. Immunoglobulin gene analysis in chronic lymphocytic leukemia in the era of next generation sequencing. Leukemia. 2020 Oct;34(10):2545–51.

106. Karan-Djurasevic T, Pavlovic S. Somatic hypermutational status and gene repertoire of immunoglobulin rearrangements in chronic Lymphocytic leukemia. In: Lymphocyte Updates - Cancer, Autoimmunity and Infection. InTech; 2017.

107. Stamatopoulos B, Timbs A, Bruce D, Smith T, Clifford R, Robbe P, et al. Targeted deep sequencing reveals clinically relevant subclonal IgHV rearrangements in chronic lymphocytic leukemia. Leukemia. 2017 Apr;31(4):837–45.

108. Sanchez M-L, Almeida J, Gonzalez D, Gonzalez M, Garcia-Marcos M-A, Balanzategui A, et al. Incidence and clinicobiologic characteristics of leukemic B-cell chronic lymphoproliferative disorders with more than one B-cell clone. Blood. 2003 Oct 15;102(8):2994–3002.

109. Ghia P, Stamatopoulos K, Belessi C, Moreno C, Stilgenbauer S, Stevenson F, et al. ERIC recommendations on IGHV gene mutational status analysis in chronic lymphocytic leukemia. Leukemia. 2007 Jan;21(1):1–3.

110. Raponi S, Ilari C, Della Starza I, Cappelli LV, Cafforio L, Piciocchi A, et al. Redefining the prognostic likelihood of chronic lymphocytic leukaemia patients with borderline percentage of immunoglobulin variable heavy chain region mutations. Br J Haematol. 2020 Jun;189(5):853–9.

111. Guièze R, Wu CJ. Genomic and epigenomic heterogeneity in chronic lymphocytic leukemia. Blood. 2015 Jul 23;126(4):445–53.

112. Nadeu F, Clot G, Delgado J, Martín-García D, Baumann T, Salaverria I, et al. Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. Leukemia. 2018 Mar;32(3):645–53.

113. Nadeu F, Diaz-Navarro A, Delgado J, Puente XS, Campo E. Genomic and epigenomic alterations in chronic Lymphocytic leukemia. Annu Rev Pathol. 2020 Jan 24;15(1):149–77.

114. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.

115. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001 Feb 16;291(5507):1304–51.

116. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016 May 17;17(6):333–51.

117. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. Hum Immunol. 2021 Nov;82(11):801–11.

118. Bewicke-Copley F, Arjun Kumar E, Palladino G, Korfi K, Wang J. Applications and analysis of targeted genomic sequencing in cancer studies. Comput Struct Biotechnol J. 2019 Nov 7;17:1348–59.

119. Samorodnitsky E, Jewell BM, Hagopian R, Miya J, Wing MR, Lyon E, et al. Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing. Hum Mutat. 2015 Sep;36(9):903–14.

120. Mertes F, Elsharawy A, Sauer S, van Helvoort JMLM, van der Zaag PJ, Franke A, et al. Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief Funct Genomics. 2011 Nov;10(6):374–86.

121. Levy SE, Boone BE. Next-generation sequencing strategies. Cold Spring Harb Perspect Med. 2019 Jul 1;9(7):a025791.

122. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022 Apr;376(6588):44–53.

123. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015 May 21;58(4):586–97.

124. Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif). 2013;6(1):287–303.

125. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood. 2009 Nov 5;114(19):4099–107.

126. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. 2009 Oct;19(10):1817–24.

127. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med. 2009 Dec 23;1(12):12ra23.

128. Arnaout RA, Prak ETL, Schwab N, Rubelt F, Adaptive Immune Receptor Repertoire Community. The future of blood testing is the immunome. Front Immunol. 2021 Mar 15;12:626793.

129. Scheijen B, Meijers RWJ, Rijntjes J, van der Klift MY, Möbs M, Steinhilber J, et al. Next-generation sequencing of immunoglobulin gene rearrangements for clonality assessment: a technical feasibility study by EuroClonality-NGS. Leukemia. 2019 Sep;33(9):2227–40.

130. van den Brand M, Rijntjes J, Möbs M, Steinhilber J, van der Klift MY, Heezen KC, et al. Next-generation sequencing-based clonality assessment of ig gene rearrangements: A multicenter validation study by EuroClonality-NGS. J Mol Diagn. 2021 Sep;23(9):1105–15.

131. He L, Sok D, Azadnia P, Hsueh J, Landais E, Simek M, et al. Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. Sci Rep. 2014 Oct 27;4(1):6778.

132. Vergani S, Korsunsky I, Mazzarello AN, Ferrer G, Chiorazzi N, Bagnara D. Novel method for high-throughput full-length IGHV-D-J sequencing of the immune repertoire from bulk B-cells with single-cell resolution. Front Immunol. 2017 Sep 14;8:1157.

133. Cole C, Volden R, Dharmadhikari S, Scelfo-Dalbey C, Vollmers C. Highly accurate sequencing of full-length immune repertoire amplicons using Tn5-enabled and molecular identifier-guided amplicon assembly. J Immunol. 2016 Mar 15;196(6):2902–7.

134. Chaudhary N, Wesemann DR. Analyzing Immunoglobulin Repertoires. Front Immunol [Internet]. 2018 Mar 14;9. Available from: http://dx.doi.org/10.3389/fimmu.2018.00462

135. Teraguchi S, Saputri DS, Llamas-Covarrubias MA, Davila A, Diez D, Nazlica SA, et al. Methods for sequence and structural analysis of B and T cell receptor repertoires. Comput Struct Biotechnol J. 2020 Jul 17;18:2000–11.

136. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology. 2012 Mar;135(3):183–91.

137. Langerak AW, Brüggemann M, Davi F, Darzentas N, van Dongen JJM, Gonzalez D, et al. High-throughput immunogenetics for clinical and research applications in immunohematology: Potential and challenges. J Immunol. 2017 May 15;198(10):3765–74.

138. Liu H, Pan W, Tang C, Tang Y, Wu H, Yoshimura A, et al. The methods and advances of adaptive immune receptors repertoire sequencing. Theranostics. 2021 Aug 19;11(18):8945–63.

139. Greiff V, Miho E, Menzel U, Reddy ST. Bioinformatic and statistical analysis of adaptive immune repertoires. Trends Immunol. 2015 Nov;36(11):738–49.

140. Lefranc MP. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. 2001 Jan 1;29(1):207–9.

141. López-Santibáñez-Jácome L, Avendaño-Vázquez SE, Flores-Jasso CF. The pipeline repertoire for Ig-Seq analysis. Front Immunol. 2019 Apr 30;10:899.

142. Vander Heiden JA, Marquez S, Marthandan N, Bukhari SAC, Busse CE, Corrie B, et al. AIRR Community standardized representations for annotated immune repertoires. Front Immunol. 2018 Sep 28;9:2206.

143. Rubelt F, Busse CE, Bukhari SAC, Bürckert J-P, Mariotti-Ferrandiz E, Cowell LG, et al. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. Nat Immunol. 2017 Nov 16;18(12):1274–8.

144. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods. 2015 May;12(5):380–1.

145. Giudicelli V, Duroux P, Kossida S, Lefranc M-P. IG and TR single chain fragment variable (scFv) sequence analysis: a new advanced functionality of IMGT/V-QUEST and IMGT/HighV-QUEST. BMC Immunol [Internet]. 2017 Dec;18(1). Available from: http://dx.doi.org/10.1186/s12865-017-0218-8

146. Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. Genome Med. 2015 Nov 20;7(1):121.

147. Kim D, Park D. Deep sequencing of B cell receptor repertoire. BMB Rep. 2019 Sep;52(9):540–7.

148. Rosenquist R, Ghia P, Hadzidimitriou A, Sutton L-A, Agathangelidis A, Baliakas P, et al. Immunoglobulin gene sequence analysis in chronic lymphocytic leukemia: updated ERIC recommendations. Leukemia. 2017 Jul;31(7):1477–81.

149. René C, Prat N, Thuizat A, Broctawik M, Avinens O, Eliaou J-F. Comprehensive characterization of immunoglobulin gene rearrangements in patients with chronic lymphocytic leukaemia. J Cell Mol Med. 2014 Jun;18(6):979–90.

150. Alamyar E, Duroux P, Lefranc M-P, Giudicelli V. IMGT® Tools for the Nucleotide Analysis of Immunoglobulin (IG) and T Cell Receptor (TR) V-(D)-J Repertoires, Polymorphisms, and IG Mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. In: Methods in Molecular Biology^{TM}. Totowa, NJ: Humana Press; 2012. p. 569–604. (Methods in molecular biology (Clifton, N.J.)).

151. Langerak AW, Molina TJ, Lavender FL, Pearson D, Flohr T, Sambade C, et al. Polymerase chain reaction-based clonality testing in tissue samples with reactive lymphoproliferations: usefulness and pitfalls. A report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia. 2007 Feb;21(2):222–9.

152. Tange O. GNU Parallel - The Command-Line Power Tool. The USENIX Magazine. 2011 Feb;36:42–7.

153. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. arXiv [q-bio.GN]. 2013. Available from: http://arxiv.org/abs/1303.3997

154. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience [Internet]. 2021 Feb 16;10(2). Available from: http://dx.doi.org/10.1093/gigascience/giab008

155. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

156. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing [Internet]. arXiv [q-bio.GN]. 2012. Available from: http://arxiv.org/abs/1207.3907

157. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

158. Rice P, Longden I, Bleasby A. EMBOSS: The European molecular biology open software suite. Trends Genet. 2000 Jun;16(6):276–7.

159. Vander Heiden JA, Yaari G, Uduman M, Stern JNH, O'Connor KC, Hafler DA, et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics. 2014 Jul 1;30(13):1930–2.

160. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics. 2015 Oct 15;31(20):3356–8.

161. Hill MO. Diversity and evenness: A unifying notation and its consequences. Ecology. 1973 Mar;54(2):427–32.

162. Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. Genome Med. 2015 May 28;7(1):49.

163. Chao A, Gotelli NJ, Hsieh TC, Sander EL, Ma KH, Colwell RK, et al. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol Monogr. 2014 Feb;84(1):45–67.

164. Bashford-Rogers RJM, Palser AL, Idris SF, Carter L, Epstein M, Callard RE, et al. Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. BMC Immunol [Internet]. 2014 Dec;15(1). Available from: http://dx.doi.org/10.1186/s12865-014-0029-0

165. Rigolin GM, Saccenti E, Bassi C, Lupini L, Quaglia FM, Cavallari M, et al. Extensive next-generation sequencing analysis in chronic lymphocytic leukemia at diagnosis: clinical and biological correlations. J Hematol Oncol [Internet]. 2016 Dec;9(1). Available from: http://dx.doi.org/10.1186/s13045-016-0320-z

166. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 2015 Mar 31;43(6):e37.

167. Wardemann H, Busse CE. Novel approaches to analyze immunoglobulin repertoires. Trends Immunol. 2017 Jul;38(7):471–82.

168. McClure R, Mai M, McClure S. High-throughput sequencing using the Ion Torrent personal genome machine for clinical evaluation of somatic hypermutation status in chronic lymphocytic leukemia. J Mol Diagn. 2015 Mar;17(2):145–54.

169. Huet S, Bouvard A, Ferrant E, Mosnier I, Chabane K, Salles G, et al. Impact of using leader primers for IGHV mutational status assessment in chronic lymphocytic leukemia. Leukemia. 2020 Aug;34(8):2257–9.

170. Li S, Lefranc M-P, Gowans E, Li S, Giudicelli V. High throughout sequencing and IMGT/HighV-QUEST analysis of T cell receptor repertoire. Protoc Exch [Internet]. 2013 Sep 5; Available from: http://dx.doi.org/10.1038/protex.2013.072

171. Basic local alignment search tool (BLAST). In: Bioinformatics and Functional Genomics. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2009. p. 100–38.

172. Moorhouse MJ, van Zessen D, IJspeert H, Hiltemann S, Horsman S, van der Spek PJ, et al. ImmunoGlobulin galaxy (IGGalaxy) for simple determination and quantitation of immunoglobulin heavy chain rearrangements from NGS. BMC Immunol. 2014 Dec 13;15(1):59.

173. Smakaj E, Babrak L, Ohlin M, Shugay M, Briney B, Tosoni D, et al. Benchmarking immunoinformatic tools for the analysis of antibody repertoire sequences. Bioinformatics. 2020 Mar 1;36(6):1731–9.

174. Mandric I, Rotman J, Yang HT, Strauli N, Montoya DJ, Van Der Wey W, et al. Profiling immunoglobulin repertoires across multiple human tissues using RNA sequencing. Nat Commun. 2020 Jun 19;11(1):3126.

175. Yu Y, Ceredig R, Seoighe C. LymAnalyzer: a tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. Nucleic Acids Res. 2016 Feb 29;44(4):e31.

176. Nadeu F, Mas-de-Les-Valls R, Navarro A, Royo R, Martín S, Villamor N, et al. IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. Nat Commun. 2020 Jul 7;11(1):3390.

177. Song L, Cohen D, Ouyang Z, Cao Y, Hu X, Liu XS. TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. Nat Methods. 2021 Jun;18(6):627–30.

178. Lindeman I, Emerton G, Mamanova L, Snir O, Polanski K, Qiao S-W, et al. BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. Nat Methods. 2018 Aug;15(8):563–5.

179. Valkiers S, de Vrij N, Gielis S, Verbandt S, Ogunjimi B, Laukens K, et al. Recent advances in T-cell receptor repertoire analysis: Bridging the gap with multimodal single-cell RNA sequencing. ImmunoInformatics. 2022 Mar;5(100009):100009.

180. Rubio T, Chernigovskaya M, Marquez S, Marti C, Izquierdo-Altarejos P, Urios A, et al. A Nextflow pipeline for T-cell receptor repertoire reconstruction and analysis from RNA sequencing data. ImmunoInformatics. 2022 Jun;6(100012):100012.

181. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome Res. 2011 May;21(5):790–7.

182. Nguyen P, Ma J, Pei D, Obert C, Cheng C, Geiger TL. Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. BMC Genomics. 2011 Feb 11;12(1):106.

183. Feeney AJ, Victor KD, Vu K, Nadel B, Chukwuocha RU. Influence of the V(D)J recombination mechanism on the formation of the primary T and B cell repertoires. Semin Immunol. 1994 Jun;6(3):155–63.

184. Benedict CL, Gilfillan S, Thai T-H, Kearney JF. Terminal deoxynucleotidyl transferase and repertoire development. Immunol Rev. 2000 Jun;175(1):150–7.

185. Kotouza MT, Gemenetzi K, Galigalidou C, Vlachonikola E, Pechlivanis N, Agathangelidis A, et al. TRIP - T cell receptor/immunoglobulin profiler. BMC Bioinformatics. 2020 Sep 29;21(1):422.

186. Soto C, Bombardi RG, Branchizio A, Kose N, Matta P, Sevy AM, et al. High frequency of shared clonotypes in human B cell receptor repertoires. Nature. 2019 Feb;566(7744):398–402.

187. Gupta NT, Adams KD, Briggs AW, Timberlake SC, Vigneault F, Kleinstein SH. Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data. J Immunol. 2017 Mar 15;198(6):2489–99.

188. Hanamsagar R, Reizis T, Chamberlain M, Marcus R, Nestle FO, de Rinaldis E, et al. An optimized workflow for single-cell transcriptomics and repertoire profiling of purified lymphocytes from clinical samples. Sci Rep. 2020 Feb 10;10(1):2219.

189. Ralph DK, Matsen FA IV. Inference of B cell clonal families using heavy/light chain pairing information [Internet]. bioRxiv. 2022. Available from: http://dx.doi.org/10.1101/2022.03.22.485213

190. Zhou JQ, Kleinstein SH. Cutting edge: Ig H chains are sufficient to determine most B cell clonal relationships. J Immunol. 2019 Oct 1;203(7):1687–92.

191. Hershberg U, Luning Prak ET. The analysis of clonal expansions in normal and autoimmune B cell repertoires. Philos Trans R Soc Lond B Biol Sci. 2015 Sep 5;370(1676):20140239.

192. Heather JM, Ismail M, Oakes T, Chain B. High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. Brief Bioinform. 2017 Jan 10;bbw138.

193. Mauerer K, Zahrieh D, Gorgun G, Li A, Zhou J, Ansén S, et al. Immunoglobulin gene segment usage, location and immunogenicity in mutated and unmutated chronic lymphocytic leukaemia. Br J Haematol. 2005 May;129(4):499–510.

194. Brezinschek HP, Foster SJ, Brezinschek RI, Dörner T, Domiati-Saad R, Lipsky PE. Analysis of the human VH gene repertoire. Differential effects of selection and somatic hypermutation on human peripheral CD5(+)/IgM+ and CD5(-)/IgM+ B cells. J Clin Invest. 1997 May 15;99(10):2488–501.

195. Lasabova Z, Plank L, Flochova E, Burjanivova T, Vanochova A, Mihok L, et al. Clinical Laboratory Method for detection of IGHV Mutation Status in patients with CLL Validated by IgBLAST and IMGT/V-QUEST. Acta medica Martiniana. 2011 Aug 1;11(2):17–25.

196. Hamblin TJ, Davis ZA, Oscier DG. Determination of how many immunoglobulin variable region heavy chain mutations are allowable in unmutated chronic lymphocytic leukaemia - long-term follow up of patients with different percentages of mutations. Br J Haematol. 2008 Feb;140(3):320–3.

197. Miqueu P, Guillet M, Degauque N, Doré J-C, Soulillou J-P, Brouard S. Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. Mol Immunol. 2007 Feb;44(6):1057–64.

198. Kleinfield R, Hardy RR, Tarlinton D, Dangl J, Herzenberg LA, Weigert M. Recombination between an expressed immunoglobulin heavy-chain gene and a germline variable gene segment in a Ly 1+ B-cell lymphoma. Nature. 1986;322(6082):843–6.

199. Lees WD. Tools for adaptive immune receptor repertoire sequencing. Curr Opin Syst Biol. 2020 Dec;24:86–92.

200. Marquez S, Babrak L, Greiff V, Hoehn KB, Lees WD, Luning Prak ET, et al. Adaptive immune receptor repertoire (AIRR) community guide to repertoire analysis. Methods Mol Biol. 2022;2453:297–316.

201. Ralph DK, Matsen FA 4th. Likelihood-based inference of B cell clonal families. PLoS Comput Biol. 2016 Oct;12(10):e1005086.

202. Nouri N, Kleinstein SH. A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data. Bioinformatics. 2018 Jul 1;34(13):i341–9.

203. Kepler TB. Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. F1000Res. 2013 Apr 3;2:103.

204. Ralph DK, Matsen FA IV. Likelihood-based inference of B-cell clonal families [Internet]. arXiv [q-bio.PE]. 2016. Available from: http://arxiv.org/abs/1603.08127

205. Wu Y-CB, Kipling D, Dunn-Walters DK. Age-related changes in human peripheral blood IGH repertoire following vaccination. Front Immunol. 2012 Jul 9;3:193.

206. Ademokun A, Wu Y-C, Martin V, Mitra R, Sack U, Baxendale H, et al. Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. Aging Cell. 2011 Dec;10(6):922–30.

207. Wang C, Liu Y, Xu LT, Jackson KJL, Roskin KM, Pham TD, et al. Effects of aging, cytomegalovirus infection, and EBV infection on human B cell repertoires. J Immunol. 2014 Jan 15;192(2):603–11.

208. Glanville J, Kuo TC, von Büdingen H-C, Guey L, Berka J, Sundar PD, et al. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proc Natl Acad Sci U S A. 2011 Dec 13;108(50):20066–71.

209. Stern JNH, Yaari G, Vander Heiden JA, Church G, Donahue WF, Hintzen RQ, et al. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. Sci Transl Med. 2014 Aug 6;6(248):248ra107.

210. Tsioris K, Gupta NT, Ogunniyi AO, Zimnisky RM, Qian F, Yao Y, et al. Neutralizing antibodies against West Nile virus identified directly from human B cells by single-cell analysis and next generation sequencing. Integr Biol (Camb). 2015 Dec;7(12):1587–97.

211. Jiang N, He J, Weinstein JA, Penland L, Sasaki S, He X-S, et al. Lineage structure of the human antibody repertoire in response to influenza vaccination. Sci Transl Med. 2013 Feb 6;5(171):171ra19.

212. Chen Z, Collins AM, Wang Y, Gaëta BA. Clustering-based identification of clonally-related immunoglobulin gene sequence sets. Immunome Res. 2010 Sep 27;6 Suppl 1(Suppl 1):S4.

213. Wu Y-C, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. Blood. 2010 Aug 19;116(7):1070–8.

214. Briney B, Le K, Zhu J, Burton DR. Clonify: unseeded antibody lineage assignment from next-generation sequencing data. Sci Rep [Internet]. 2016 Apr;6(1). Available from: http://dx.doi.org/10.1038/srep23901

215. Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, et al. Antigen receptor repertoire profiling from RNA-seq data. Nat Biotechnol. 2017 Oct;35(10):908–11.

216. Parikh SA, Strati P, Tsang M, West CP, Shanafelt TD. Should IGHV status and FISH testing be performed in all CLL patients at diagnosis? A systematic review and meta-analysis. Blood. 2016 Apr 7;127(14):1752–60.

217. Tobin G, Rosén A, Rosenquist R. What is the current evidence for antigen involvement in the development of chronic lymphocytic leukemia? Hematol Oncol. 2006 Mar;24(1):7–13.

218. Jaramillo S, Agathangelidis A, Schneider C, Bahlo J, Robrecht S, Tausch E, et al. Prognostic impact of prevalent chronic lymphocytic leukemia stereotyped subsets: analysis within prospective clinical trials of the German CLL Study Group (GCLLSG). Haematologica. 2020 Nov 1;105(11):2598–607.

219. González-Gascón-Y-Marín I, Muñoz-Novas C, Rodríguez-Vicente A-E, Quijada-Álamo M, Hernández-Sánchez M, Pérez-Carretero C, et al. From biomarkers to models in the changing landscape of chronic Lymphocytic leukemia: Evolve or become extinct. Cancers (Basel). 2021 Apr 8;13(8):1782.

220. Knisbacher BA, Lin Z, Hahn CK, Nadeu F, Duran-Ferrer M, Stevenson KE, et al. Molecular map of chronic lymphocytic leukemia and its impact on outcome. Nat Genet [Internet]. 2022 Aug 4; Available from: http://dx.doi.org/10.1038/s41588-022-01140-w

221. Gemenetzi K, Psomopoulos F, Carriles AA, Gounari M, Minici C, Plevova K, et al. Higher-order immunoglobulin repertoire restrictions in CLL: the illustrative case of stereotyped subsets 2 and 169. Blood. 2021 Apr 8;137(14):1895–904.

222. Zaragoza-Infante L, Junet V, Pechlivanis N, Fragkouli S-C, Amprachamian S, Koletsa T, et al. IgIDivA: immunoglobulin intraclonal diversification analysis. Brief Bioinform [Internet]. 2022 Aug 30; Available from: http://dx.doi.org/10.1093/bib/bbac349

223. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. Nat Methods. 2014 Jun;11(6):653–5.

224. Aita T, Ichihashi N, Yomo T. Probabilistic model based error correction in a set of various mutant sequences analyzed by next-generation sequencing. Comput Biol Chem. 2013 Dec;47:221–30.

225. Molnar M, Ilie L. Correcting Illumina data. Brief Bioinform. 2015 Jul;16(4):588–99.

226. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput Biol. 2013 Apr;9(4):e1003031.

227. Pattnaik S, Gupta S, Rao AA, Panda B. SInC: an accurate and fast error-model based simulator for SNPs, Indels and CNVs coupled with a read generator for short-read sequence data. BMC Bioinformatics. 2014 Feb 5;15(1):40.

228. Prabakaran P, Streaker E, Chen W, Dimitrov DS. 454 antibody sequencing - error characterization and correction. BMC Res Notes. 2011 Oct 12;4(1):404.

229. Sleep JA, Schreiber AW, Baumann U. Sequencing error correction without a reference genome. BMC Bioinformatics. 2013 Dec 18;14(1):367.

230. Malcikova J, Stano-Kozubik K, Tichy B, Kantorova B, Pavlova S, Tom N, et al. Detailed analysis of therapy-driven clonal evolution of TP53 mutations in chronic lymphocytic leukemia. Leukemia. 2015 Apr;29(4):877–85.

231. Rossi D, Khiabanian H, Spina V, Ciardullo C, Bruscaggin A, Famà R, et al. Clinical impact of small TP53 mutated subclones in chronic lymphocytic leukemia. Blood. 2014 Apr 3;123(14):2139–47.

232. Nadeu F, Delgado J, Royo C, Baumann T, Stankovic T, Pinyol M, et al. Clinical impact of clonal and subclonal TP53, SF3B1, BIRC3, NOTCH1, and ATM mutations in chronic lymphocytic leukemia. Blood. 2016 Apr 28;127(17):2122–30.

233. Nadeu F, Royo R, Massoni-Badosa R, Playa-Albinyana H, Garcia-Torre B, Duran-Ferrer M, et al. Detection of early seeding of Richter transformation in chronic lymphocytic leukemia. Nat Med. 2022 Aug;28(8):1662–71.

234. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714–26.

235. Stilgenbauer S, Schnaiter A, Paschka P, Zenz T, Rossi M, Döhner K, et al. Gene mutations and treatment outcome in chronic lymphocytic leukemia: results from the CLL8 trial. Blood. 2014 May 22;123(21):3247–54.

236. Nouri N, Kleinstein SH. Somatic hypermutation analysis for improved identification of B cell clonal families from next-generation sequencing data. PLoS Comput Biol. 2020 Jun;16(6):e1007977.

237. Meng W, Zhang B, Schwartz GW, Rosenfeld AM, Ren D, Thome JJC, et al. An atlas of B-cell clonal distribution in the human body. Nat Biotechnol. 2017 Sep;35(9):879–84.

238. Lindenbaum O, Nouri N, Kluger Y, Kleinstein SH. Alignment free identification of clones in B cell receptor repertoires. Nucleic Acids Res. 2021 Feb 26;49(4):e21.

239. Peled JU, Kuang FL, Iglesias-Ussel MD, Roa S, Kalis SL, Goodman MF, et al. The biochemistry of somatic hypermutation. Annu Rev Immunol. 2008;26(1):481–511.

240. Klein U, Goossens T, Fischer M, Kanzler H, Braeuninger A, Rajewsky K, et al. Somatic hypermutation in normal and transformed human B cells. Immunol Rev. 1998 Apr;162(1):261–80.

241. Kleinstein SH, Louzoun Y, Shlomchik MJ. Estimating hypermutation rates from clonal tree data. J Immunol. 2003 Nov 1;171(9):4639–49.

242. Kocks C, Rajewsky K. Stepwise intraclonal maturation of antibody affinity through somatic hypermutation. Proc Natl Acad Sci U S A. 1988 Nov;85(21):8206–10.

243. Weill J-C, Weller S, Reynaud C-A. Human marginal zone B cells. Annu Rev Immunol. 2009;27(1):267–85.

244. Yaari G, Benichou JIC, Vander Heiden JA, Kleinstein SH, Louzoun Y. The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. Philos Trans R Soc Lond B Biol Sci. 2015 Sep 5;370(1676):20140242.

245. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. Proc Natl Acad Sci U S A. 2015 Feb 24;112(8):E862-70.

246. Zalcberg I, D'Andrea MG, Monteiro L, Pimenta G, Xisto B. Multidisciplinary diagnostics of chronic lymphocytic leukemia: European Research Initiative on CLL - ERIC recommendations. Hematol Transfus Cell Ther. 2020 Jul;42(3):269–74.

247. Rawstron AC, Fazi C, Agathangelidis A, Villamor N, Letestu R, Nomdedeu J, et al. A complementary role of multiparameter flow cytometry and high-throughput sequencing for minimal residual disease detection in chronic lymphocytic leukemia: an European Research Initiative on CLL study. Leukemia. 2016 Apr;30(4):929–36.

248. Del Giudice I, Raponi S, Della Starza I, De Propris MS, Cavalli M, De Novi LA, et al. Minimal residual disease in chronic Lymphocytic leukemia: A new goal? Front Oncol. 2019 Aug 29;9:689.

249. Serrano A, Fuentes A, Ferrer Lores B, Lendinez V, Monzo C, Ivorra C, et al. Detection of immunoglobulin heavy chain gene clonality by high-throughput sequencing for minimal residual disease monitoring in Chronic Lymphocytic Leukaemia. Blood. 2019 Nov 13;134(Supplement_1):1747–1747.

250. López López V, Serrano A, Fuentes A, Ferrer Lores B, Chaves JF, Solano C, et al. Liquid biopsy and lymphoma monitoring in clinical practice. Blood. 2021 Nov 5;138(Supplement 1):4483–4483.

8 Appendix

8.1 Primer sequences employed

FORWARD

VH1_FR1_N _5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGCCTCAGTGAAGGTCTCCTGCAAG
VH2_FR1_N_5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTCTGGTCCTACGCTGGTGAAACCC
VH3_FR1_N _5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGGGGGGTCCCTGAGACTCTCCTG
VH4_FR1_N_5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTCGGAGACCCTGTCCCTCACCTG
VH5_FR1_N _5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGGGGAGTCTCTGAAGATCTCCTGT
VH6_FR1_N _5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCGCAGACCCTCTCACTCACCTGTG
VH1_FR2_N _5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGGGTGCGACAGGCCCCTGGACAA
VH2_FR2_N_5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGATCCGTCAGCCCCCAGGGAAGG
VH3_FR2_N _5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTCCGCCAGGCTCCAGGGAA
VH4_FR2_N_5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGATCCGCCAGCCCCCAGGGAAGG
VH5_FR2_N _5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGGTGCGCCAGATGCCCGGGAAAGG
VH6_FR2_N _5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGATCAGGCAGTCCCCATCGAGAG
VH7_FR2_N _5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTGGGTGCGACAGGCCCCTGGACAA
VH1_FR3_N _5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGAGCTGAGCAGCCTGAGATCTGA
VH2_FR3_N_5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAATGACCAACATGGACCCTGTGGA
VH3_FR3_N _5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCTGCAAATGAACAGCCTGAGAGCC
VH4_FR3_N_5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGCTCTGTGACCGCCGCGGACACG
VH5_FR3_N _5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGCACCGCCTACCTGCAGTGGAGC
VH6_FR3_N _5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTTCTCCCTGCAGCTGAACTCTGTG
VH7_FR3_N _5 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGCACGGCATATCTGCAGATCAG
VH1_leader TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTCACCATGGACTGGACCTGGAG
VH2_leader TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGGACATACTTTGTTCCAGGCTC
VH3_leader TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCATGGAGTTTGGGCTGAGCTGG
VH3-21_leader TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCATGGAACTGGGGCTCCGC
VH4_leader TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACATGAAACATCTGTGGTTCTTCC
VH5_leader TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTGGGTCAACCGCCATCCTCG
VH6_leader TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGTCTGTCTCCTTCCTCATCTTC

REVERSE

JH_3 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTTACCTGAGGAGACGGTGACC

8.2 GeneScan images for rearrangement validation

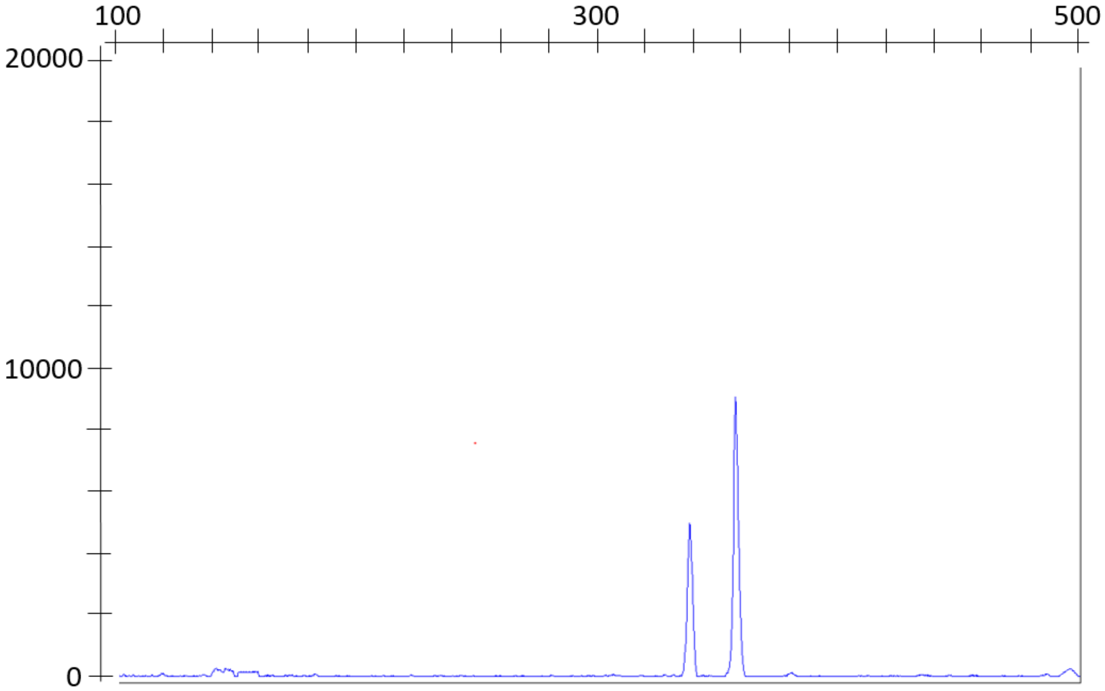


Figure 8.1. Additional clones 1 sample GeneScan IGH clonality analysis. Areas: 137886; 70431. Heights: 9107; 4989.

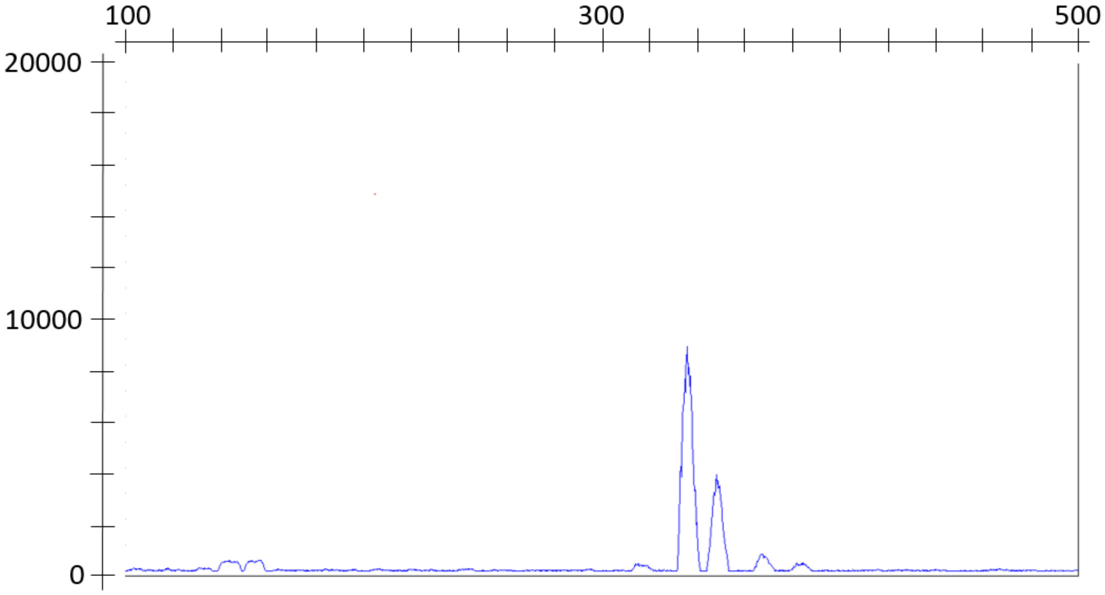


Figure 8.2. Additional clones 2 sample GeneScan IGH clonality analysis. Areas: 225176; 95782. Heights: 8721; 3753.

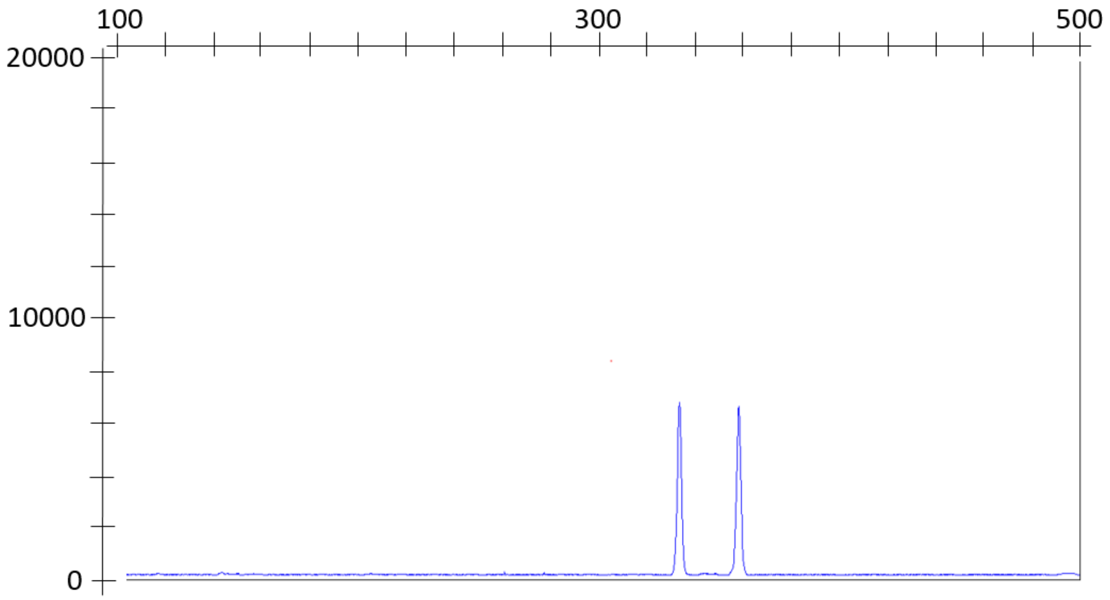


Figure 8.3. Additional clones 3 sample GeneScan IGH clonality analysis. *Areas: 86685; 91768 Heights: 6659; 6506.*

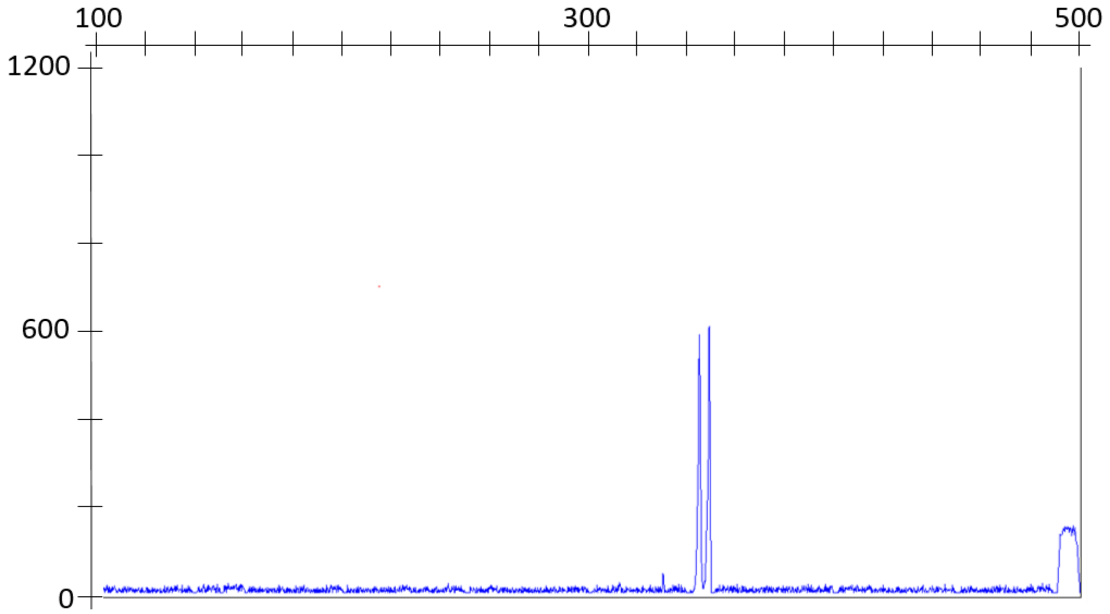


Figure 8.4. Additional clones 4 sample GeneScan IGH clonality analysis. *Areas: 4164; 4324. Heights: 584; 602.*

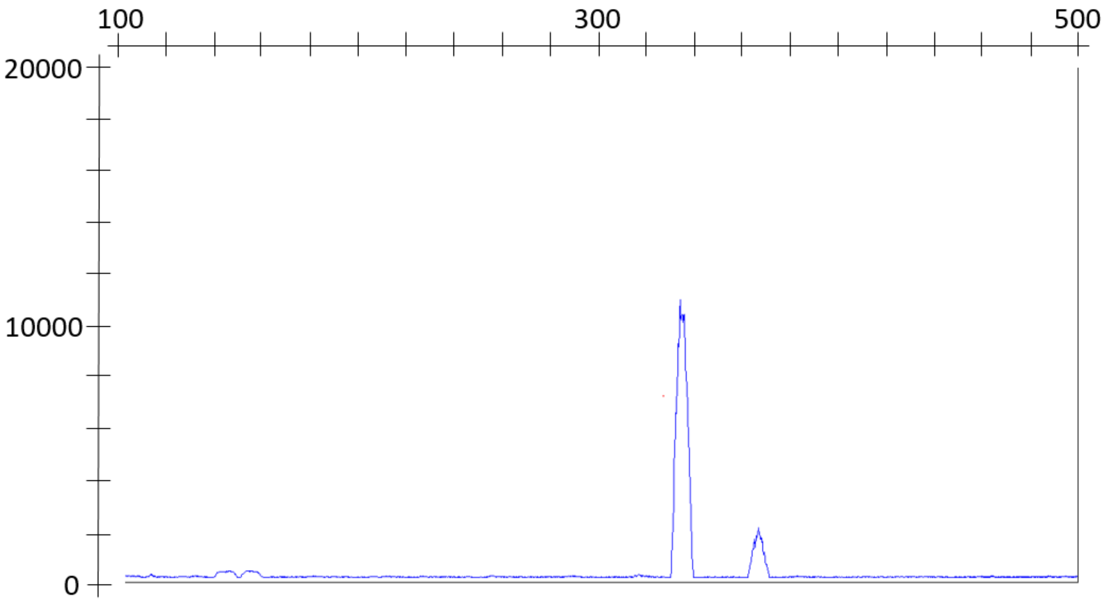


Figure 8.5. Additional clones 5 sample GeneScan IGH clonality analysis. *Areas: 308149;49458. Heights: 10727;1899.*

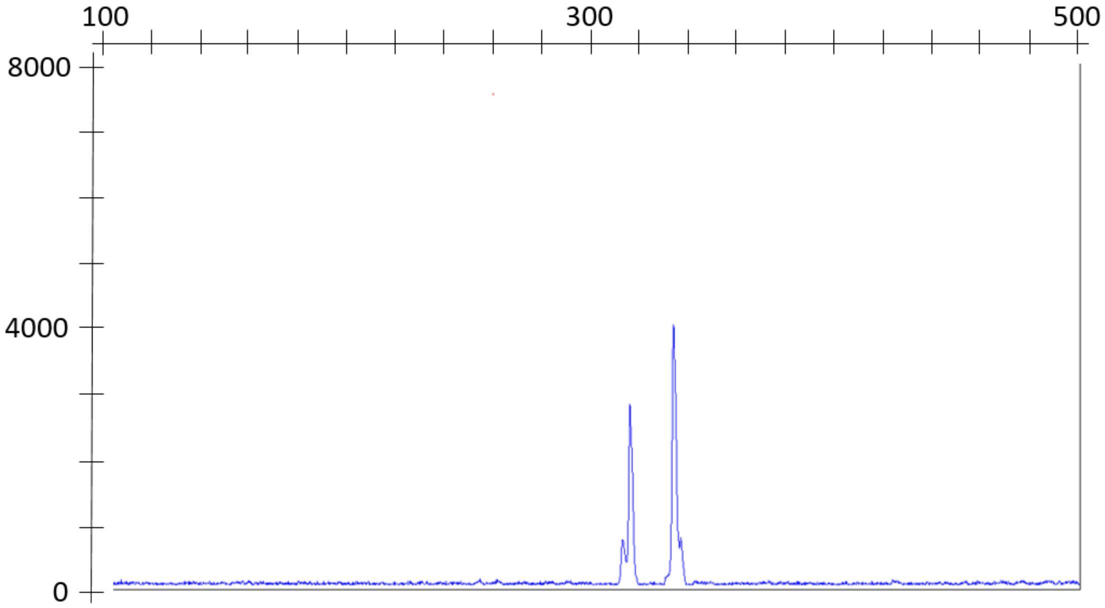


Figure 8.6. Additional clones 6 sample GeneScan IGH clonality analysis. *Areas: 30011; 45981. Heights: 2741; 3954.*

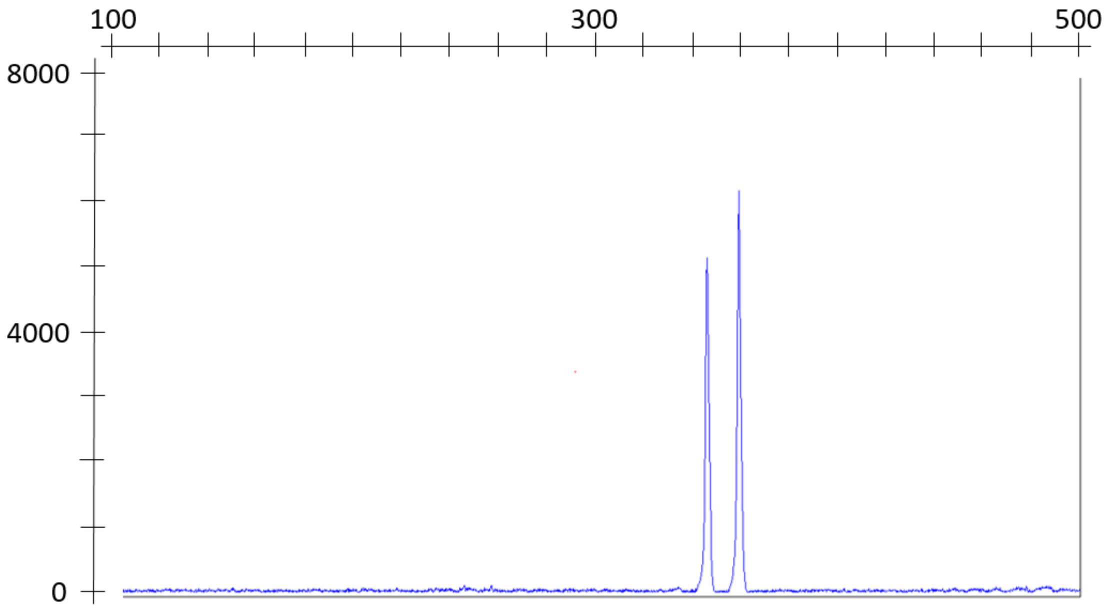


__Figure 8.7. Additional clones 7 sample GeneScan IGH clonality analysis.__ Areas:60302;73883. Heights:5143;6176.

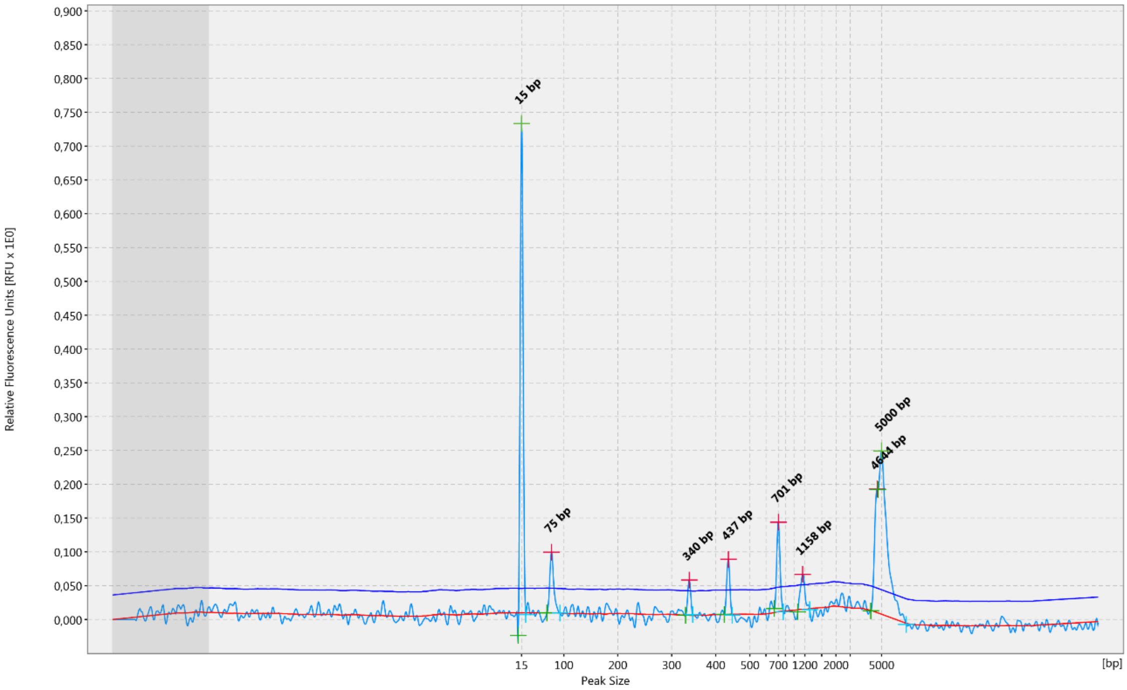


Figure 8.8. Additional clones 8 sample Qiaxcel DNA electrophoresis gel image after amplification of IGHV3 family with FR1-JH oligonucleotides.

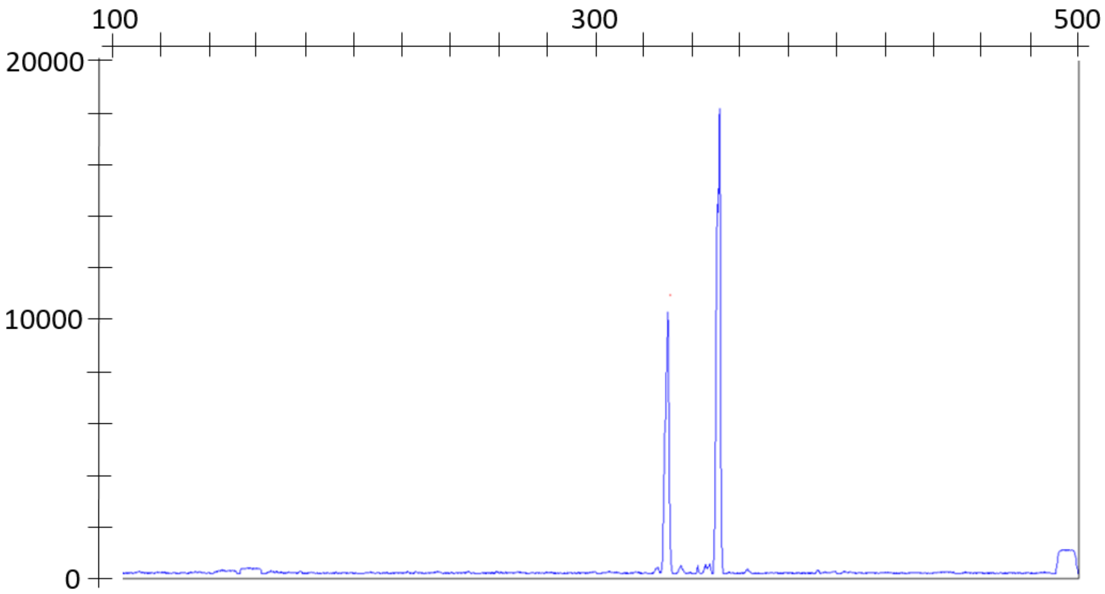


*Figure 8.9. **Additional clones 9 sample GeneScan IGH clonality analysis**. Areas:101333;203453. Heights: 10064;3704.*

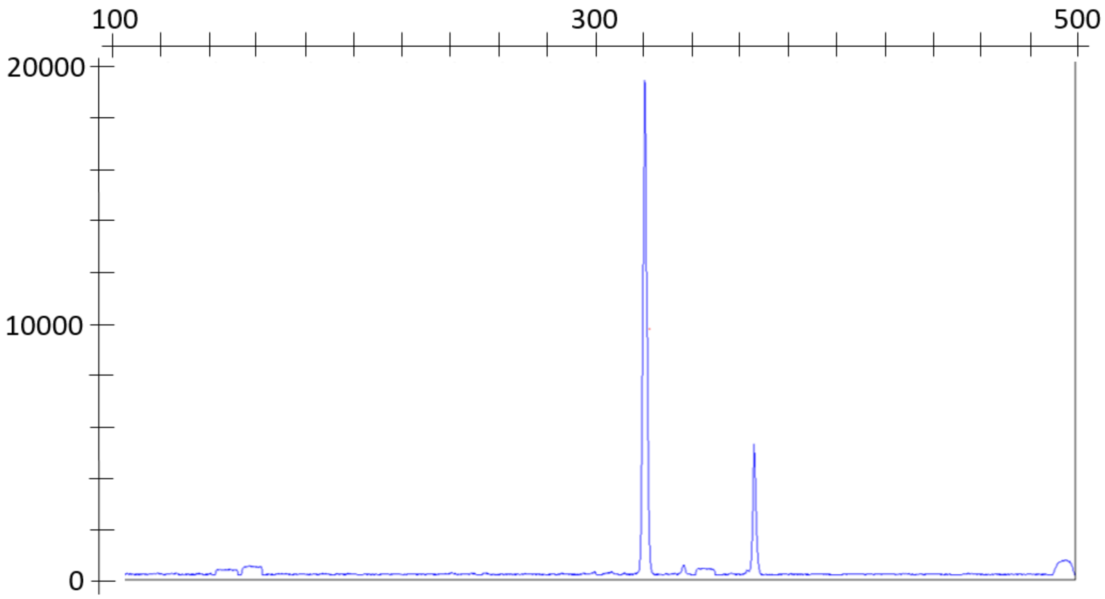


*Figure 8.10. **FP1 sample GeneScan IGH clonality analysis.** Areas:207630;50903. Heights:19166;5051.*

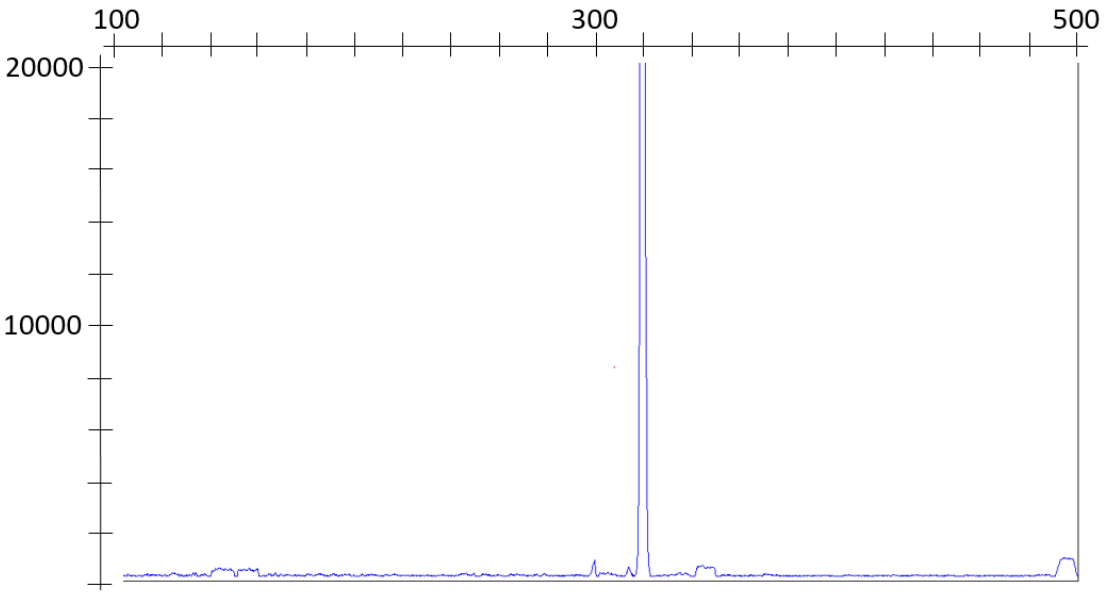


Figure 8.11. FP2 sample GeneScan IGH clonality analysis.

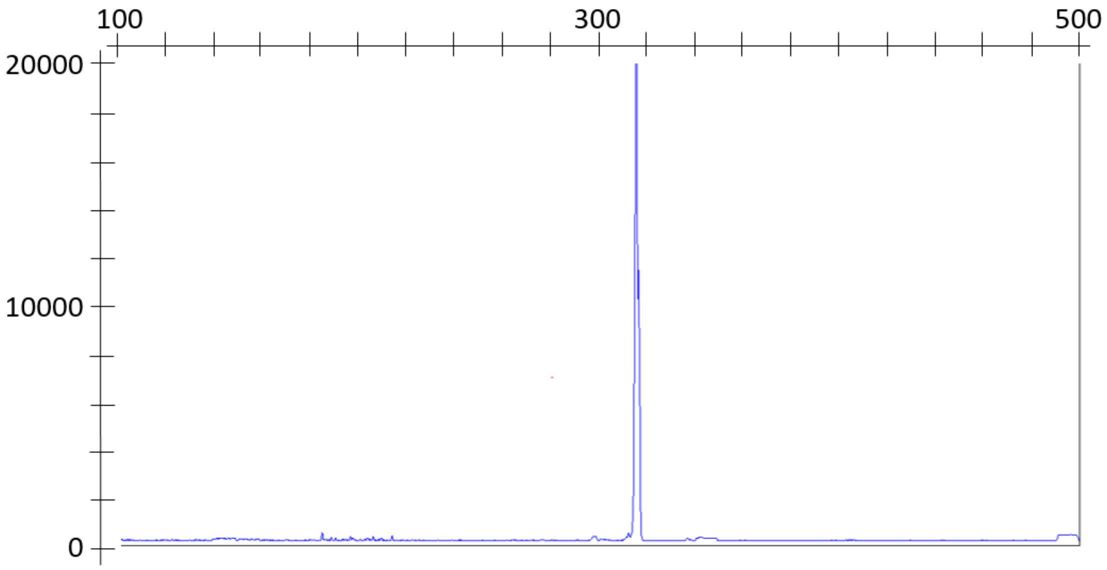


Figure 8.12. FP3 sample GeneScan IGH clonality analysis.

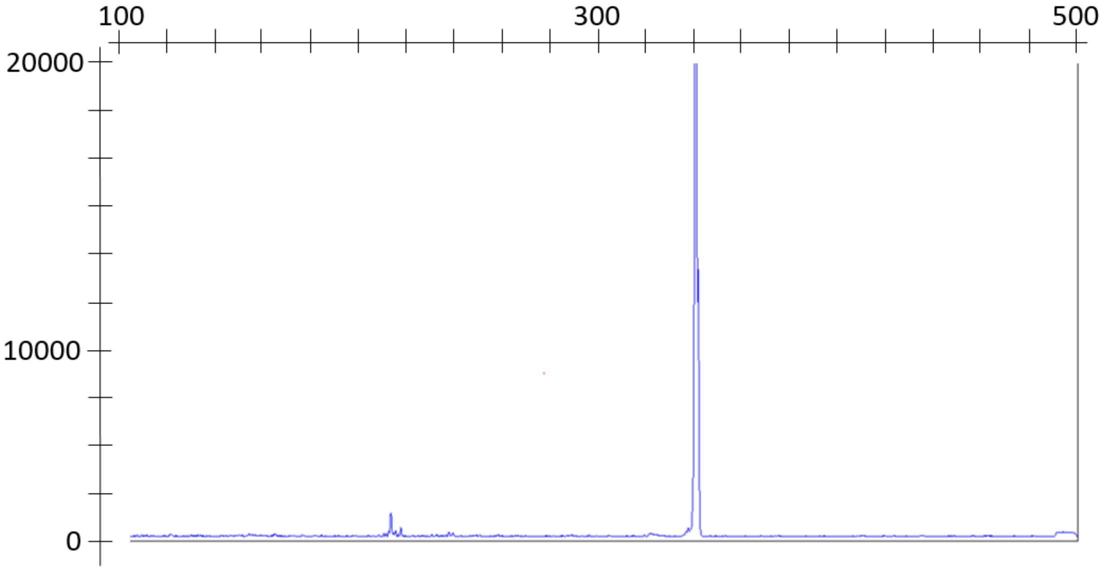


Figure 8.13. FP4 sample GeneScan IGH clonality analysis.

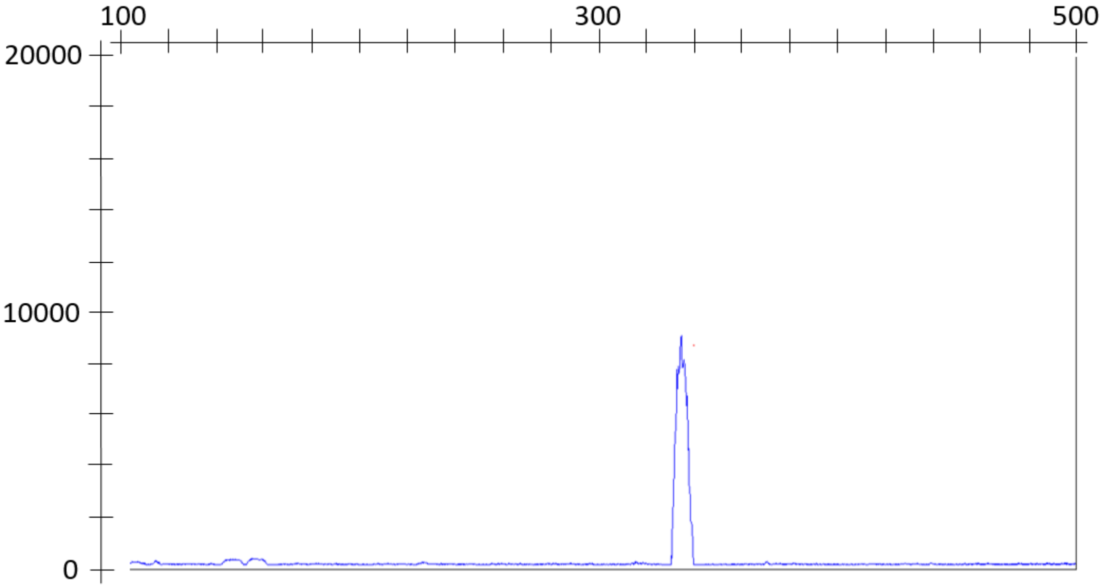


Figure 8.14. FP5 sample GeneScan IGH clonality analysis.

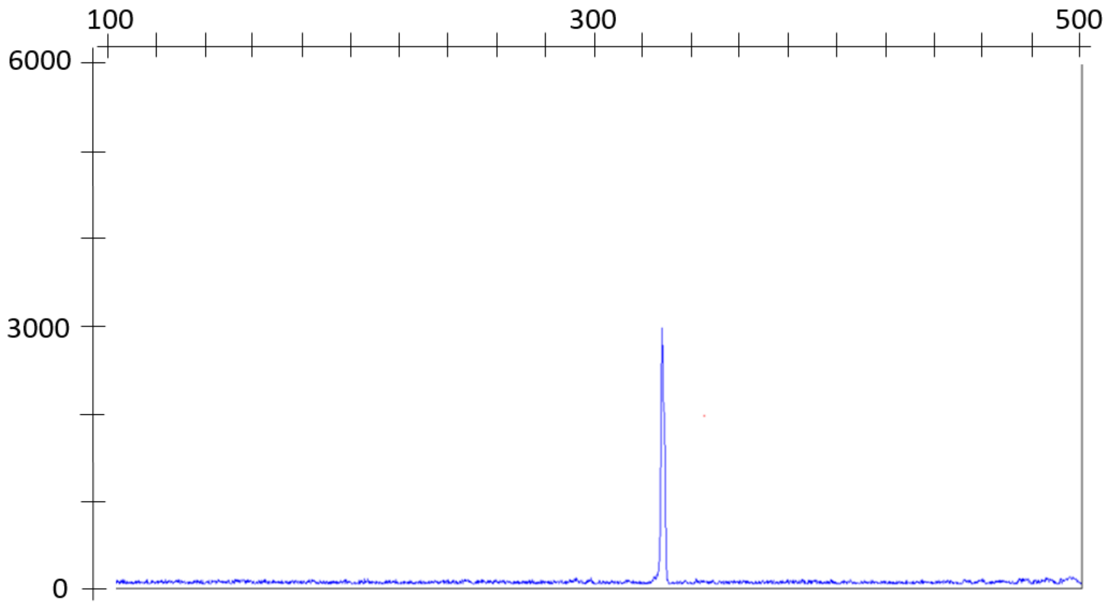


Figure 8.15. FP6 sample GeneScan IGH clonality analysis.

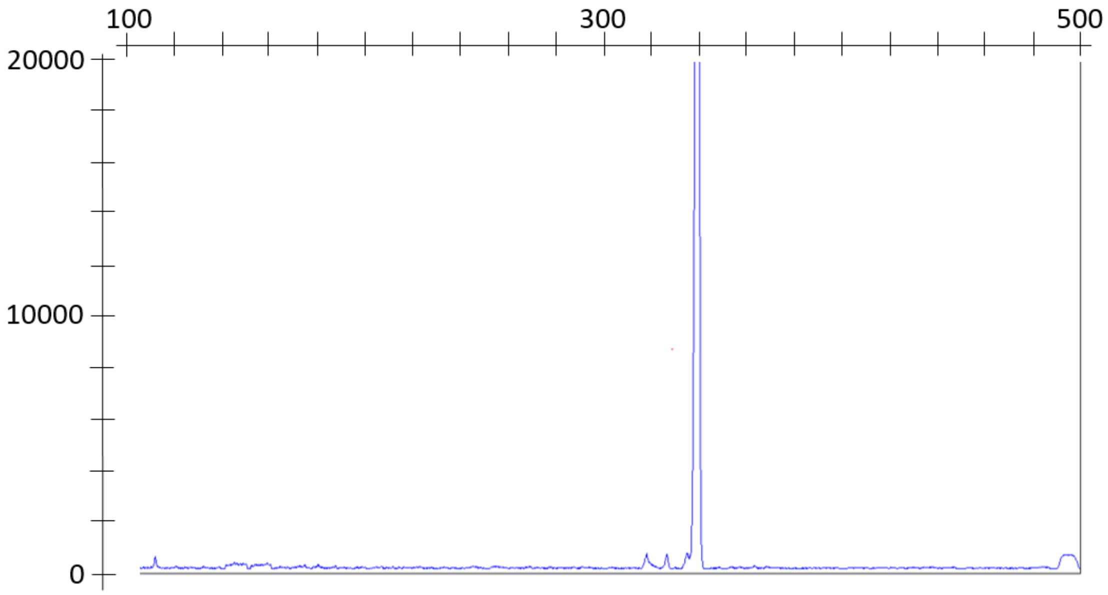


Figure 8.16. FP7 sample GeneScan IGH clonality analysis.

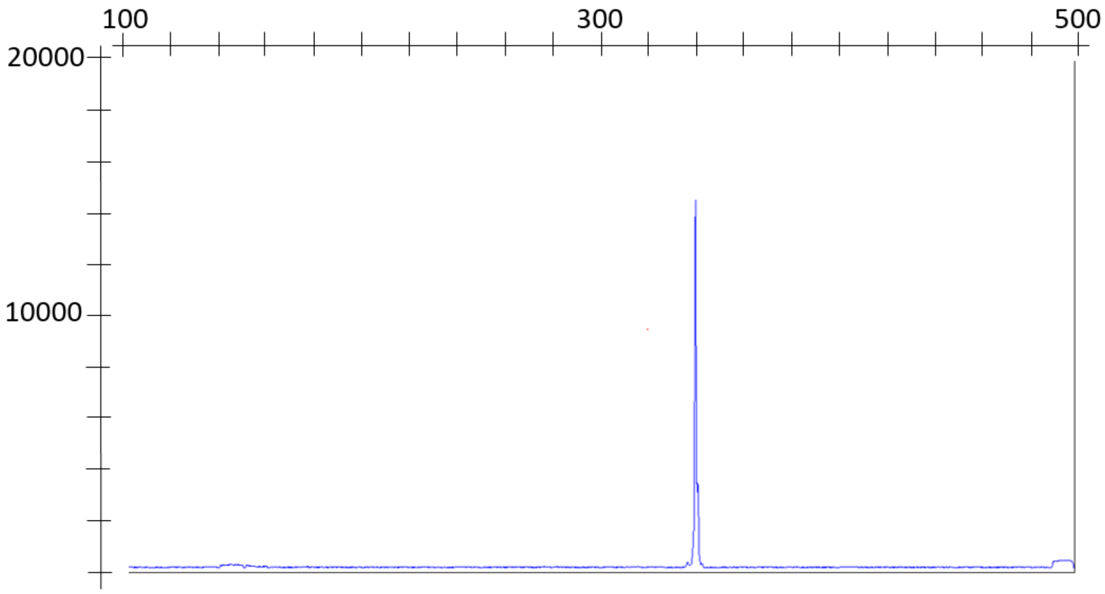


Figure 8.17. FP8 sample GeneScan IGH clonality analysis.