

Chapter 3

3. Developing computational methods for assessing B-cell receptor populations from next-generation sequencing

3.1. Introduction

To date next-generation sequencing (NGS) of BCRs have primarily focused on classifying the IgHV, D and J recombination frequencies to understand the diversity of the BCR repertoire (Boyd et al., 2009, Campbell et al., 2008, Maletzki et al., 2012, Lev et al., 2012, Jager et al., 2012, Weinstein et al., 2009). However, computational assignment of V-D-J sequences to reference databases results in many incompletely identified IgHV, D and J genes even when the germline alleles are known (Weinstein et al., 2009). This is most likely due to somatic hypermutation (SHM) masking the identity of the germline genes present in the NGS, or the existence of new diverse IgH gene alleles not present in the reference database. Further, investigation of V-D-J gene usage frequencies utilises only part of the BCR sequence diversity with important information about the V-D-J joining regions and somatic hypermutations not considered.

This chapter describes the development of analysis methods for BCR sequence data using the full BCR V-D-J sequence variation and that does not rely on prior V-D-J gene classification. It was previously shown that zebrafish BCR repertoire diversity can be interpreted through full V-D-J genotype diversity using BCR networks, and that these are an intuitive way for understanding B-cell repertoires (Ben-Hamo and Efroni, 2011). In such networks, the lowest level of organisation in a population of B-cells, namely unique B-cells, are represented by sparse networks whereas highly developed (connected) networks most likely result from clonal expansions of B-cells, arising through antigenic exposure or B-cell malignancies (Ben-Hamo and Efroni, 2011). However, such analyses have never been applied to mammals, or during infection or disease. In this chapter, network methods were developed to provide a robust framework for analysing vast NGS sequencing repertoires from B-cell populations. This chapter aimed to distinguish between

diverse B-cell populations and clonal B-cell populations both qualitatively and quantitatively.

3.2. Results

3.2.1. Next-generation sequencing of IgH variable genes

RT-PCR amplification of the expressed rearranged IgHV-D-J loci from mRNA from human B-cell populations was performed using the consensus IgHJ primer and FR1 or FR2 IgHV family primers (**Figure 2.1** and Table 2.1) (van Dongen et al., 2003). Peripheral blood (PB) samples from thirteen healthy individuals, eleven CLL patients, and eight LCLs yielded PCR products of expected sizes (310-360bp for FR1 and 250-295bp for FR2 primed samples) and were 454 sequenced (Table 3.1). Samples yielded an average of 42,324 sequencing reads after filtering for quality and presence of IgH sequence (Table 3.2). Briefly, only reads were retained with median base quality Phred scores of greater than 32, with significant similarity to reference IgHV genes (E-value $<1 \times 10^{-10}$), and with identifiable primer sequences. Two additional samples from CLL patient A (pre and post treatment) were sequenced on the MiSeq platform (Table 3.2). The BCR 454 sequence datasets from *Boyd et al.* (Boyd et al., 2009) were also analysed, which includes three healthy individuals and five patients with clonal blood disorders (Table 3.3).

Table 3.1. Patient sample information.

Sample	Patient type	Age, years	Gender	Time since CLL diagnosis, years
CLL 1	CLL	77	Male	7
CLL 2	CLL	58	Male	2
CLL 3	CLL	78	Male	1.5
CLL 4	CLL+HCC	77	Male	2.5
CLL 5	CLL	59	Female	1.25
CLL 6	CLL	67	Male	2
CLL7	CLL	69	Male	13
CLL 8	CLL	64	Male	4.5
CLL 9	CLL	77	Male	5.25
CLL 10	CLL	81	Male	8
CLL 11	CLL	81	Male	10
Healthy 1	Age matched control 1	74	Female	-
Healthy 2	Age matched control 2	62	Female	-
Healthy 3	Age matched control 3	75	Female	-
Healthy 4	Age matched control 4	67	Female	-
Healthy 5	Age matched control 5	68	Female	-
Healthy 6	Healthy 6	55	Male	-
Healthy 7	Healthy 7	23	Male	-
Healthy 8	Healthy 8	23	Male	-
Healthy 9	Healthy 9	25	Male	-
Healthy 10	Healthy 10	24	Female	-
Healthy 11	Healthy 11	24	Female	-
Healthy 12	Healthy 12	24	Female	-
Healthy 13	Healthy 13	24	Female	-

* Abbreviations: CLL=chronic lymphocytic leukemia, HCC=Hepatocellular carcinoma.

Table 3.2. Sample information and number of sequencing reads.

Primer	Type*	ID	Platform	Number of reads	Number of reads (after filtering**)	Average read length (bp)	Multiplex
FR1	CLL	CLL 1	454	58700	51311	290.4	Multiplex half plate C
FR1	CLL	CLL 2	454	54937	31694	290.6	Multiplex half plate C
FR1	CLL	CLL 3	454	46657	26828	310.2	Multiplex half plate C
FR1	CLL	CLL 4	454	45632	27126	287.9	Multiplex half plate C
FR1	CLL	CLL 5	454	40780	26086	294.6	Multiplex 7/8 plate D
FR1	CLL	CLL 6	454	59847	54761	310.6	Multiplex 7/8 plate D
FR1	CLL	CLL 7	454	22036	18273	303.9	Multiplex 7/8 plate D
FR1	CLL	CLL 8	454	44079	37208	308.5	Multiplex 7/8 plate D
FR1	CLL	CLL 9	454	34139	29401	305.9	Multiplex 7/8 plate D
FR1	CLL	CLL 10	454	55331	51018	311.9	Multiplex 7/8 plate D
FR1	CLL	CLL 11	454	33950	27650	301.8	Multiplex 7/8 plate D
FR1	Healthy	Healthy 1	454	56105	28638	288.4	Multiplex half plate A
FR1	Healthy	Healthy 2	454	77698	40556	288.5	Multiplex half plate A
FR1	Healthy	Healthy 3	454	45539	23848	286.6	Multiplex half plate A
FR1	Healthy	Healthy 4	454	132359	59456	286.5	Multiplex half plate A
FR1	Healthy	Healthy 5	454	53350	40435	315.9	Multiplex half plate C
FR1	Healthy	Healthy 6	454	60637	41878	292.5	Multiplex half plate C
FR1	Healthy	Healthy 7	454	50600	35852	291.8	Multiplex half plate C
FR1	Healthy	Healthy 8	454	35163	25454	296.5	Multiplex half plate C
FR1	Healthy	Healthy 9	454	34796	26849	289.3	Multiplex half plate C
FR1	Healthy	Healthy 10	454	44991	34248	291.4	Multiplex half plate C
FR1	Healthy	Healthy 11	454	33085	25083	291.9	Multiplex half plate C
FR1	Healthy	Healthy 12	454	45134	36828	299.2	Multiplex 7/8 plate D
FR1	Healthy	Healthy 13	454	40984	33792	296.2	Multiplex 7/8 plate D
FR1	LCL	LCL 1	454	65182	58117	290.5	Multiplex whole plate B
FR1	LCL	LCL 2	454	64483	53894	305	Multiplex whole plate B
FR1	LCL	LCL 3	454	24473	17285	302	Multiplex whole plate B
FR1	LCL	LCL 4	454	101156	82317	295.4	Multiplex whole plate B
FR1	LCL	LCL 5	454	53964	45325	298.3	Multiplex whole plate B
FR1	LCL	LCL 6	454	47691	40233	301	Multiplex whole plate B
FR1	LCL	LCL 7	454	43047	32340	290.4	Multiplex whole plate B
FR1	LCL	LCL 8	454	59503	50617	308.1	Multiplex whole plate B
FR2	Healthy	Healthy 1	454	43209	33628	229.3	Multiplex half plate A
FR2	Healthy	Healthy 2	454	27379	20904	228.3	Multiplex half plate A
FR2	Healthy	Healthy 3	454	23379	19009	228.1	Multiplex half plate A
FR2	Healthy	Healthy 4	454	36846	27756	226.9	Multiplex half plate A
FR2	LCL	LCL 1	454	81271	55741	239.4	Multiplex whole plate B
FR2	LCL	LCL 2	454	106236	88253	257.3	Multiplex whole plate B
FR2	LCL	LCL 3	454	117359	107230	247.4	Multiplex whole plate B
FR2	LCL	LCL 4	454	96943	88771	249.6	Multiplex whole plate B
FR2	LCL	LCL 5	454	69621	61840	240.1	Multiplex whole plate B
FR2	LCL	LCL 6	454	55010	48408	234.2	Multiplex whole plate B
FR2	LCL	LCL 7	454	57697	50834	222.1	Multiplex whole plate B
FR2	LCL	LCL 8	454	50789	45501	250.9	Multiplex whole plate B
FR1	CLL	Patient A Pre-treatment	MiSeq	56864	40414	264.3	Multiplex 1/74 lane
FR1	CLL	Patient A Post-treatment	MiSeq	42053	36197	265.4	Multiplex 1/74 lane

* Abbreviations: LCL = Human lymphoblastoid cell line, CLL = chronic lymphocytic leukemia.

** Reads were filtered for complete primer sequences and length (where reads shorter than 255bp are removed for FR1 primed samples, and reads shorter than 195bp for FR2 primed samples).

Table 3.3. Sample information and number of sequencing reads from the Boyd et al. dataset.

Primer	Type* **	ID	Number of reads (after filtering)	Average read length (bp)
FR2	Healthy donor 1, time 0	Healthy 12 a1	12316	228
FR2	Healthy donor 1, time 0	Healthy 12 a2	17943	227.9
FR2	Healthy donor 1, time 14 months	Healthy 12 b1	13189	227.3
FR2	Healthy donor 1, time 14 months	Healthy 12 b2	10361	227.6
FR2	Patient 1; CLL/SLL time 0 months	CLL/SLL 1a	2774	216.4
FR2	Patient 1; CLL/SLL time 3 months	CLL/SLL 1b	2353	213.9
FR2	Patient 2; FL	FL1	11293	228.4
FR2	Patient 3; FL and SLL in Lymph node	FL/SLL	30391	215.5
FR2	Patient 4; CLL/SLL	CLL/SLL 2	31201	227.2
FR2	Healthy donor 2	Healthy 13	24545	226
FR2	Patient 6; CLL	CLL 12	17438	225.9
FR2	Healthy donor 3	Healthy 14	29883	223
FR2	Patient 6 CLL diluted 1:10	CLL 12 1:10	13362	223.2
FR2	Patient 6 CLL diluted 1:100	CLL 12 1:100	26966	222.9
FR2	Patient 6 CLL diluted 1:1000	CLL 12 1:1000	22063	222.9
FR2	Patient 6 CLL diluted 1:10000	CLL 12 1:10000	26464	222.7
FR2	Patient 6 CLL diluted 1:100000	CLL 12 1:100000	26635	222.8

* Abbreviations: CLL = chronic lymphocytic leukemia, SLL =Small lymphocytic lymphoma, FL= Follicular lymphoma.

**From (Boyd et al., 2009).

3.2.2. Next-generation sequencing error rate

Firstly, the 454 NGS error rate was determined to assess the number of sequencing errors to expect in BCR sequencing. To achieve this, reverse transcription and PCR was performed to amplify universally expressed genes, beta-actin, beta-globin and GAPDH, from two healthy individuals (amplicon sizes of 150bp, 150bp and 340bp respectively, **Table 3.4**). After sequencing by either 454 or MiSeq, the same read quality filtering was performed as with the BCR sequences. The sequence representing the majority of the reads for each sample was classified as the ‘true’ gene sequence for that individual to account for individual allelic variation. Any differences between this sequence and the reads were considered to be RT-PCR and/or sequencing error and classified as homopolymeric indels (occurring in a region of two or more consecutive identical bases), non-homopolymeric indels, or mismatches. The distribution of mismatches and indels was random across the genes. By counting the base-pair differences between the true gene sequence and sequence variants, the combined per-base error-rate for the RT-PCR and sequencing process for the 454 platform was 1.74×10^{-4} (**Table 3.5**, of which homopolymeric indels and non-homopolymeric errors accounted for 59.7% (1.04×10^{-4}) and 40.3% (7.04×10^{-5}) of the total error-rate respectively). These error rates were consistent between repeats of the same genes. Of note is the high homopolymeric error-rate, which has been previously reported with 454 sequencing at similar levels (Luo et al., 2012, Wang et al., 2007, Boyd et al., 2009, Gall et al., 2013). Similarly the combined per-base error-rate for RT-PCR and MiSeq sequencing was 1.70×10^{-4} (**Table 3.6**), where, again, the error rates are consistent between repeats of the same genes and similar to previously reported error rates of 5.9×10^{-4} (Lou et al., 2013). This means for every 4,070bp sequenced, there is a 50% chance that there is at least one sequencing error using MiSeq sequencing, and for every 3,980bp sequenced there is a 50% chance that there is at least one sequencing error using 454 sequencing.

Table 3.4. Sample information and number of sequencing reads for control genes.

Sample name**	Gene	Platform	Number of gene specific reads	Number of reads after filtering*	% of original reads retained after filtering	Multiplexing
Healthy 1	Beta-Actin	454	7673	7671	99.97	Multiplex 1/8th plate D
Healthy 2	Beta-Actin	454	2109	2105	99.81	Multiplex 1/8th plate D
Healthy 1	Beta-Globin	454	6983	4871	69.76	Multiplex 1/8th plate D
Healthy 2	Beta-Globin	454	5361	3387	63.18	Multiplex 1/8th plate D
Healthy 1	GAPDH	MiSeq	93213	89821	96.36	Multiplex 1/74 Lane
Healthy 2	GAPDH	MiSeq	50386	48242	95.74	Multiplex 1/74 Lane
Healthy 1	Beta-Actin	MiSeq	21551	13696	63.55	Multiplex 1/74 Lane
Healthy 2	Beta-Actin	MiSeq	86899	56909	65.49	Multiplex 1/74 Lane
Healthy 1	GAPDH	MiSeq	187923	179831	95.69	Multiplex 1/74 Lane
Healthy 2	GAPDH	MiSeq	181150	172914	95.45	Multiplex 1/74 Lane

* Reads were filtered for homology with the corresponding target gene and subsequently filtered for intact primer sequences and for complete primer sequences and length (reads shorter than 150bp for beta-actin, and shorter than 340bp for beta-globin were removed).

High read filtering in the beta-globin samples are due to non-specific PCR amplifications.

** Amplification of beta-actin, beta-globin and GAPDH genes from two healthy individual samples.

Table 3.5. Estimated average per-base 454 error frequencies by type.

Sample	Gene	Number of reads	Average read length, bp	Overall non-homopolymeric error rate	Type			Insertions and deletions	
					Insertions	Deletions	Mismatches	Non-homopolymeric	Homopolymeric
Healthy 1	Beta Actin	7671	105	3.97E-05	5.09E-05	3.73E-06	2.11E-05	1.86E-05	3.60E-05
Healthy 2	Beta Actin	2105	104	3.17E-05	1.36E-05	4.52E-06	3.17E-05	0.00E+00	1.81E-05
Healthy 1	Beta Globin	4871	297.9	1.03E-04	1.43E-04	1.18E-04	6.00E-05	4.34E-05	2.17E-04
Healthy 2	Beta Globin	3385	294	1.06E-04	1.75E-04	1.00E-05	6.63E-05	4.02E-05	1.45E-04
Average				7.03E-05	9.55E-05	3.40E-05	4.48E-05	2.56E-05	1.04E-04

*Amplicon lengths were 104bp for beta-actin, and 295bp for beta-globin.

Table 3.6. Estimated average per-base MiSeq error frequencies.

Sample	Gene	Overall error rate
Healthy 1	GAPDH	2.78E-04
Healthy 2	GAPDH	2.71E-04
Healthy 1 (repeat)	GAPDH	2.80E-04
Healthy 2 (repeat)	GAPDH	2.81E-04
Healthy 1	Beta-Actin	6.34E-05
Healthy 2	Beta-Actin	6.30E-05
Average		1.70E-04

3.2.3. Percentage of identical BCR reads between samples

The percentage of reads identical to the most abundant BCR sequence in each sample was determined to make an initial assessment of the differences in B-cell clonality of the PB and LCL samples. The percentage of reads corresponding to the most abundant BCR sequence in each of the CLL and LCL samples (range 39.3%-87.8% and 35.2%-78.7% respectively) were significantly higher than that of PB from healthy individuals (range 0.10% -14.0%) with a p-value <0.001 (**Figure 3.1**). There was no significant difference in the percentage of identical reads between the LCL and CLL patient samples (p-value=0.0594). Therefore, the healthy individuals represent diverse BCR populations, whereas the LCL and CLL samples represent more restricted or clonal BCR populations. Sanger and MiSeq sequencing confirmed that the dominant clonal sequences from the CLL samples were identical to that from 454 sequencing (excluding homopolymeric indels) indicating that there are no significant differences between sequencing platforms for high abundance sequences.

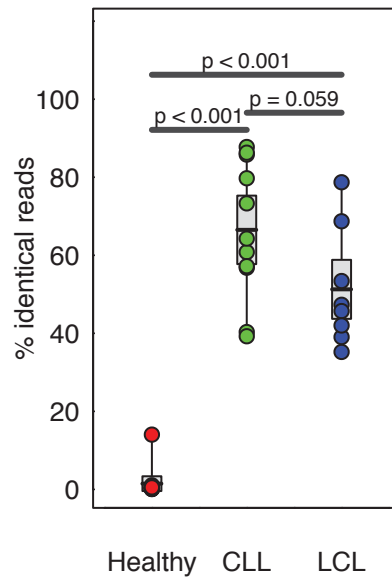


Figure 3.1. Sequencing of B-cell receptor repertoires.

The percentage of reads corresponding to the most abundant B-cell receptor sequence for each sample, separated into sample type: healthy individuals, CLL patients and LCLs. Two-sided t-tests were performed between the sample subsets, with the p-values indicated above.

3.2.4. Limitations of V-D-J gene classification

To determine the proportion of BCRs that cannot be classified in terms of IgHV, D and J gene usage, each BCR sequence was aligned to the germline sequences from the ImMunoGeneTics database (IMGT) (Lefranc et al., 2009) by BLAST (**Figure 3.2**). Due to the difference in length of the different gene families, different BLAST E-value thresholds were used for the IgHV, IgHD, and IgHJ-genes (10^{-70} , 10^{-3} and 10^{-20} respectively). The majority of sequences could be classified to their most closely related reference sequences for IgHV and IgHJ genes (an average of 99.8% and 96.1% of BCR sequences were classified respectively). Substantially fewer IgHD were identifiable (average of 40.5%) due to the shorter sequence length and potential insertions and deletions within the joining regions between the V-D-J boundaries, which has been noted in previous studies (Weinstein et al., 2009). Incomplete V-D-J gene classification may be due to SHM masking the identity of the germline genes present in individuals and/or the existence of allelic variants of reference IgH (Boyd et al., 2010a). There was no significant difference between the percentage of classified V, D and J genes of our dataset compared to that of *Boyd et al. (2009)* (**Figure 3.2**). To overcome the limitations of IgH V-D-J gene classification, the use of sequence-based network analysis is proposed next. Network analysis makes use of complete V-D-J sequence information and mutational relationships. Without the requirement of reference gene classification, network analysis is proposed to be more informative and robust framework for BCR repertoire analysis.

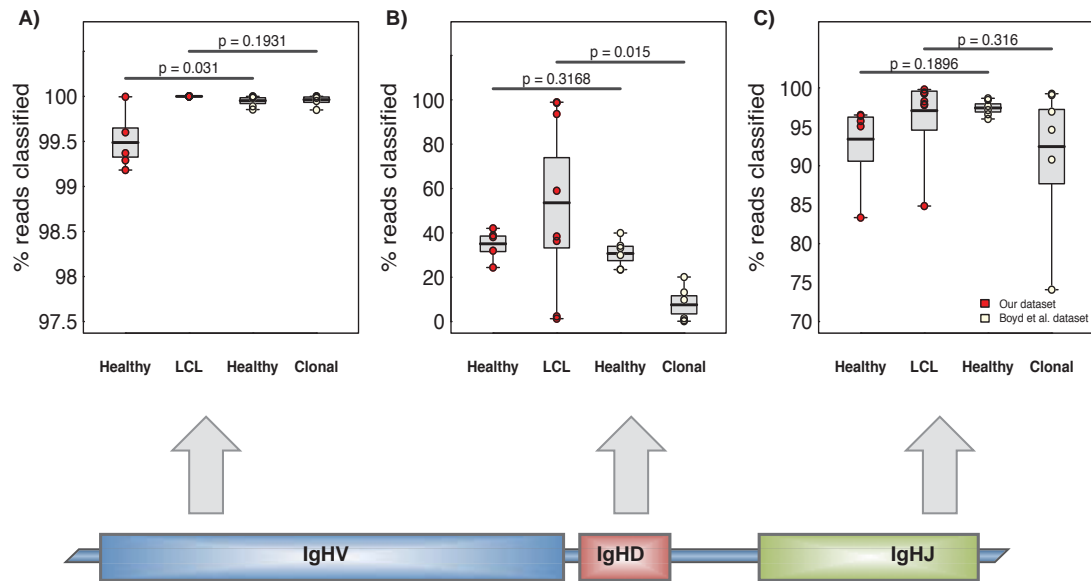


Figure 3.2. Percentage of reference sequences matched to 454 reads.

For **A)** IgHV, **B)** IgHD and **D)** IgHJ genes. Samples from healthy individuals or human lymphoblastoid cell lines from this study (red) and the dataset *Boyd et al. (2009)* (white). The clonal samples in *Boyd et al.*'s dataset refer to patients with CLL, small lymphocytic lymphoma and/or follicular lymphoma. The BLAST e-value thresholds for the IgHV, IgHD, and IgHJ-genes were 10^{-20} , 10^{-3} and 10^{-4} respectively. P-values were calculated using two-sided t-test in R.

3.2.5. BCR sequences organise into networks based on sequence diversity

It is reasonable to consider each different BCR sequence as a distinct product from amplification of a rearranged BCR from a B-cell. Therefore the B-cell repertoire can be represented as a network representing BCR sequence space. Networks are powerful tools for understanding the overall structure of large multidimensional datasets, where information is represented in the form of vertices and edges between vertices. Here, networks are able to represent the BCR sequence repertoire in the following way: a vertex represents a different sequence, and the number of identical BCR sequences defines the vertex size. Edges are created between vertices that differ by one nucleotide, i.e. highly related BCR sequences. Clusters are groups of interconnected vertices, where any two vertices in a cluster are related by the set of point mutations indicated by the edge-path between them (**Figure 3.3**). Therefore, code was developed here in python to analyse high-throughput BCR sequencing data to generate the networks.

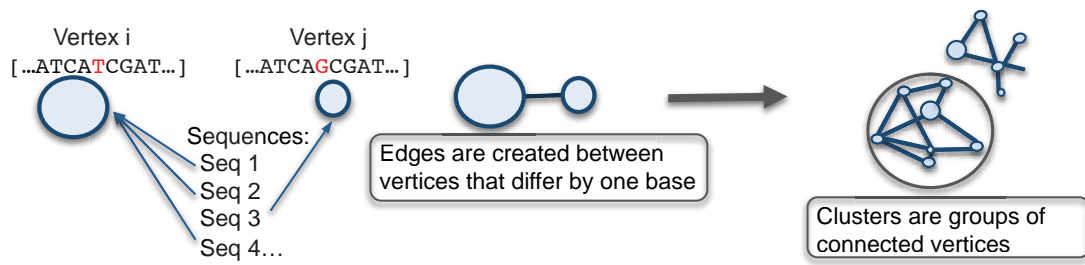


Figure 3.3. Generation of B-cell receptor sequence networks.

Schematic diagram showing the method by which the sequencing networks are generated: each vertex represents a unique sequence, where the relative size of the vertex is proportional to the number of sequencing reads that were identical to the vertex sequence. Edges are created between vertices that differ by one base (indel or substitution).

To test the analysis of high-throughput BCR sequencing data by networks, filtered and trimmed 454 or MiSeq sequences for each sample were used directly to generate a sequence network (**Figure 3.4**). Differences in network architectures are clearly seen by comparing B-cell populations from healthy individuals and LCLs. In LCLs, the majority of BCR sequences fall within a small number of clusters (greater than 40% of all sequencing reads form the largest cluster in each sample), as these samples are predominantly comprised of a small number of large B-cell clone types (**Figure 3.4Ai**). In contrast, healthy individuals have sparsely connected networks where most sequences are unique, thus yielding small vertices indicative of high overall BCR sequence diversity in the sampled repertoire (**Figure 3.4Aii**). From healthy individuals, the largest cluster representing 16.7% (4023 reads) of the total population in healthy individual 10.

The maximum CLL vertex sizes differ between samples (39.2-99.5% of total sequences) suggesting that large but variable-sized subsets of B-cells express the predominant BCR sequence(s), surrounded by BCR variants (including total process errors) of the dominant sequence. Of note, the extent of cluster size diversity is different between CLL samples, with some displaying extensive clonal enlargement (**Figure 3.4Bi**) whereas others have more limited clonal expansion (**Figure 3.4Bii**) or expansion of two dominant clusters (**Figure 3.4C**). Although the clinical relevance of dual clonal expansions are not known, previous studies have shown that the presence of two expanded IgH rearrangements can be either due to multiple productive gene rearrangements or the co-existence of two expanded clones with the CLL phenotype (Plevova et al., 2014). Therefore, the magnitude of connectivity of different samples varies between individual patients with CLL. However, in all cases, the CLL sequence networks are clearly distinct from the sparsely connected age-matched healthy individual BCR networks.

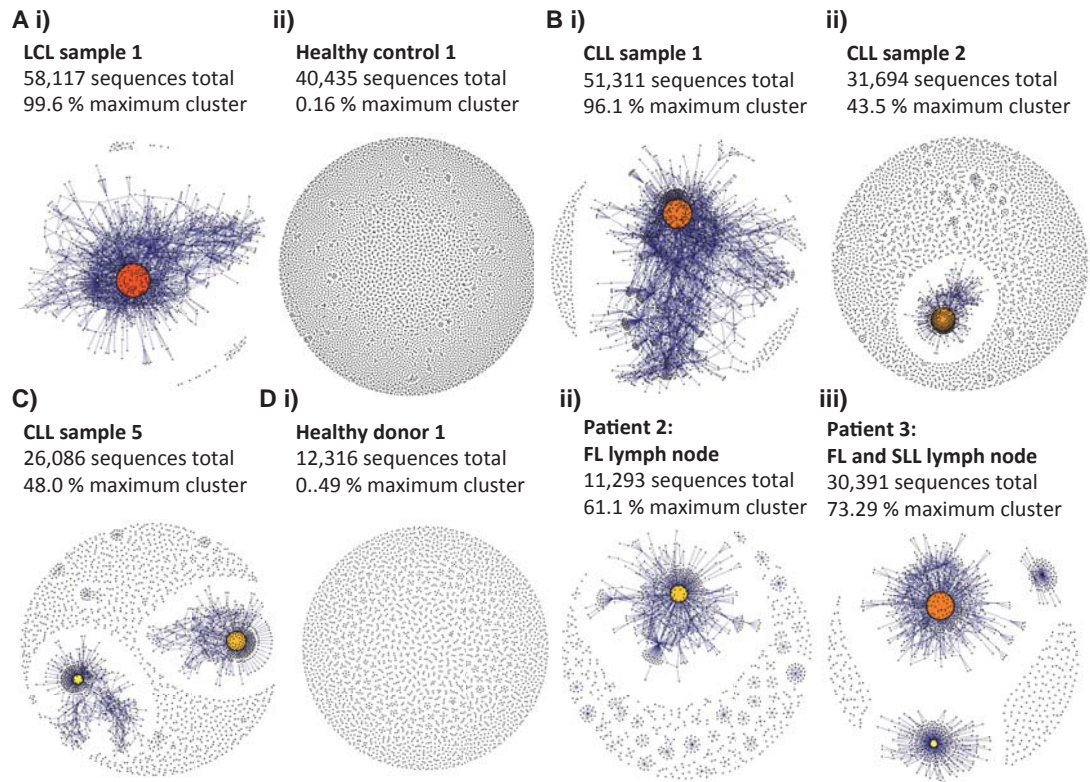


Figure 3.4. B-cell receptor repertoires from different samples.

A) Comparison of BCR sequence networks between i) a typical LCL sample and ii) a typical healthy individual. **B)** BCR sequence networks of CLL patients with i) extensive clonal enlargement and ii) limited clonal expansion. **C)** BCR sequence networks of CLL patient 5 showing expansion of two dominant clusters. **D)** Networks generated from sequencing dataset from *Boyd et al.* (Boyd et al., 2009) of i) healthy donor 1, ii) patient 2 with follicular lymphoma (FL) and iii) patient 3 with FL and small lymphocytic lymphoma (SLL). The vertex colors correspond to the relative abundance of the corresponding sequences, where red, orange and yellow indicates observation of a sequence in >90%, between 40-90% and <40% of the reads in the sample respectively.

It is proposed that sequences within a cluster are most likely related to a single rearranged, unique IgHV-D-J BCR progenitor that has undergone proliferation and somatic hypermutation, but also potentially contains BCRs with sequencing error(s). Somatic hypermutation has been found previously to preferentially occur within the CDRs of the BCR compared to the FRW regions (Lin et al., 1997), whereas sequencing errors would be distributed randomly along the length of the BCR. This could be due to either preferential AID targeting to the CDRs, or to selection of B-cells with fewer mutations in the FRW regions, such as those that would negatively change the BCR structure. To test this, sequences within clusters were aligned, and the distribution of base-pair changes was determined (**Figure 3.5**). Although base-pair differences are distributed along the length of the 454 sequences (**Figure 3.5A-C**), in all the healthy individual samples, mutations significantly occur within the CDR regions, known to be hotspots for somatic hypermutation (Lin et al., 1997), compared to the FWR regions (p-value=0.000338, **Figure 3.5D**), suggesting that these are a result of SHM rather than errors.

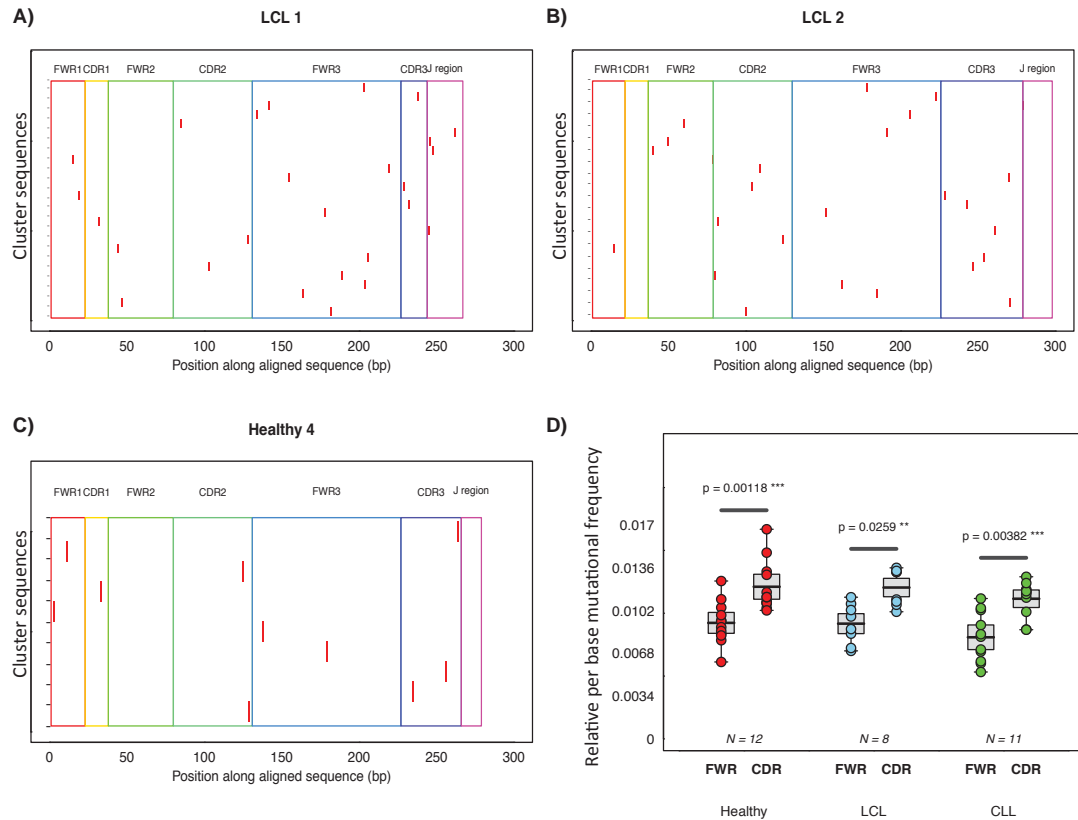


Figure 3.5. Distribution of mutations between connected vertex sequences.

Locations of mutational differences between vertices for **A-B)** two LCL samples and **C)** a representative healthy individual sample. For each sample, the sequence representing the highest connectivity was aligned with up to 25 randomly selected sequences to which it is connected, and the spatial distribution of mismatches was calculated. Each row on the figure represents a 454 sequence, and the red lines mark positions where the sequence differs in terms of mismatches. These are overlaid with the different structural regions of the BCR, as defined by IMGT/V-Quest (Giudicelli et al., 2011). **D)** The relative proportion of mutations found in either the FWR or CDR regions for all the 12 healthy, 8 LCL and 11 CLL samples. Mutations were determined by aligning all sequences separated by a single edge within each network. P-values calculated by paired T-test.

3.2.6. Population measures capture network and sample diversity

Several parameters were investigated to describe the quantitative features of the sequenced BCR repertoire from B-cell populations, including the Gini Index, maximum and second maximum cluster sizes. The Gini index is an unevenness measure, that can take a value between 0 and 1. A Gini index of 0 reflects complete equality and Gini index values close to 1 indicates high inequality or unevenness. When applied to the vertex size distribution for a given sample, these measures quantify the overall clonal nature of a sample. When the Gini index is applied to the cluster size distributions, these measures quantify the overall clustering of the sample. As shown in the previous section (Section 3.2.5), clustered sequences represent highly related BCRs with the hallmarks of SHM. Therefore, the cluster Gini index relates to the overall SHM of a sample. A high cluster Gini index is indicative of some clusters with high numbers of connected and related vertices, whereas a cluster Gini index suggests that all the clusters have more equal and lower numbers of connected and related vertices. The maximum cluster size measure is the percentage of reads corresponding to the largest cluster and indicates the degree of clonal expansion of a sample. To assess the possibility of dual clonal expansions, a measure of the second maximum cluster size as a percentage of reads in a sample was also included.

The LCL samples, due to the more restricted BCR repertoires and highly connected clusters yield high cluster and vertex Gini Indices (averages of 0.94 and 0.80, range 0.91-0.97 and 0.62-0.91 respectively) (**Figure 3.6A**) showing high unevenness of the size distributions. By contrast, B-cell networks of healthy individuals occupy a distinct region of Gini Index vertex and cluster space (averages of 0.21 and 0.05, range 0.10-0.39 and 0.03-0.11 respectively). The low vertex Gini indices shows that the healthy samples have more even distributions of vertex sizes, where each unique BCR sequence is observed a small number of times, and no BCR sequences dominate the repertoire. The low vertex Gini indices shows that the healthy samples have no clusters that dominate the repertoire. The CLL samples occupy a spatial range between healthy individuals and LCL B-cell population extremes with low vertex (between 0.62 and 0.97), and cluster Gini Indices (between 0.15 and 0.83), due to their B-cell clonal expansions. There is however considerable variation between the cluster Gini Indices, with CLL patients 1, 10 and 11 having low cluster Gini Indices, indicative of a highly expanded dominant cluster or dominant clones.

Of note, one healthy individual (healthy individual 10) has a more developed network as defined by an increase in connectivity and vertex sizes resulting in higher vertex and cluster Gini Indices (**Figure 3.6A** point (a)). This increased clonality was verified by independent sequencing using the BIOMED-2 FR2 primer set (strong linear correlation between BIOMED-2 FR1 and FR2 primed samples, R^2 -value>0.996, **Figure 3.7** and **Figure 3.8**). These strong correlations also indicate no significant primer amplification bias, which has been the major caution of PCR based approaches. Further, the highest expressed BCR sequence for healthy individual 10 has 90.6% sequence identity with the closest germline IgHV gene (16 mismatches in 243bp of alignment) suggesting that this B-cell clone has undergone SHM, therefore could potentially be antigen driven.

Networks were generated from the sequences derived from *Boyd et al.* (Boyd et al., 2009) to validate these population measures on independent BCR sequence data. This showed that the clonal populations of the patients with CLL, small lymphocytic lymphoma (SLL) and/or follicular lymphoma (FL) are distinct from the diverse populations of healthy individuals (**Figure 3.6B**), occupying equivalent regions of the cluster and vertex Gini Index graphs to CLL samples within this study. Therefore, the Gini Index population measure robustly separates distinct B-cell populations into different regions based on the clonal nature of the sample and is applicable to data from other laboratories.

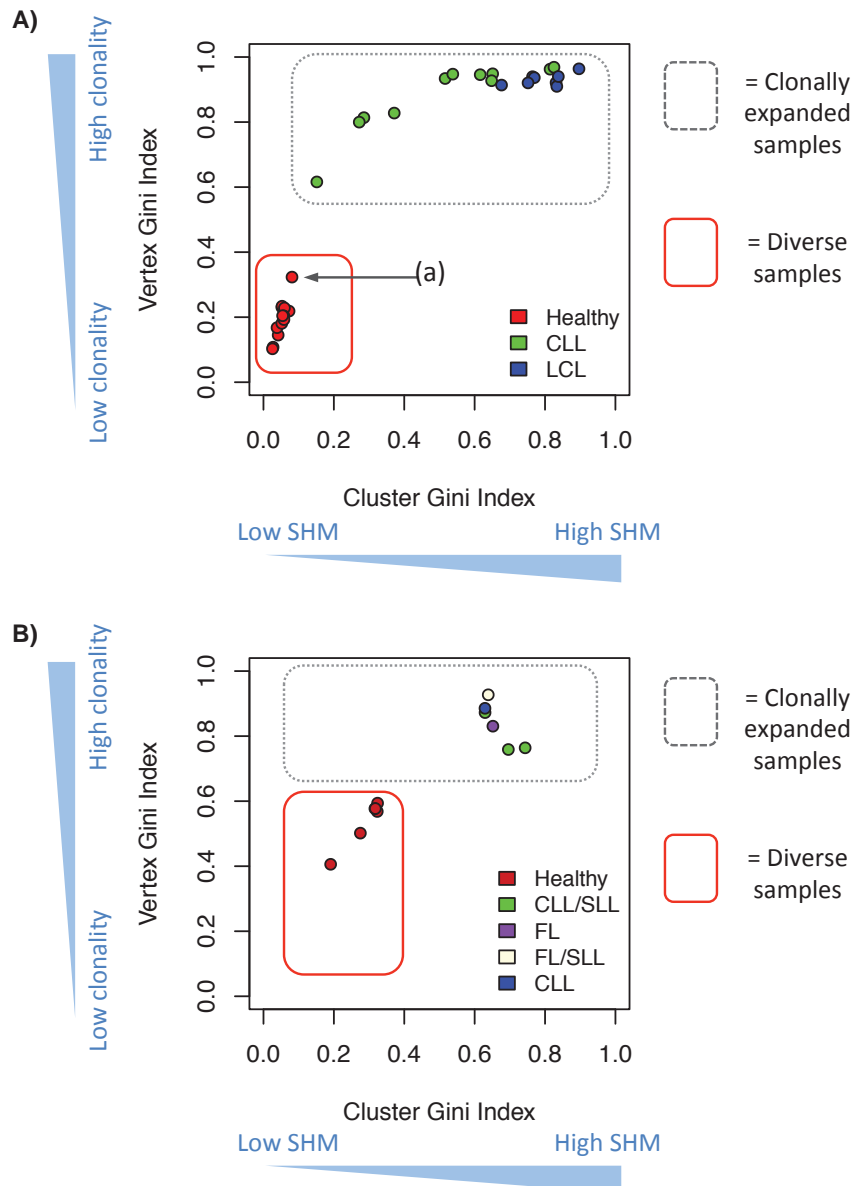


Figure 3.6. Measures differentiating between B-cell receptor populations.

The cluster Gini Index plotted against vertex Gini Index for **A)** thirteen healthy individual samples, eleven chronic lymphocytic leukemia (CLL), and eight human lymphoblastoid cell line (LCL) samples and **B)** six healthy individual samples, two samples from patients with CLL and small lymphocytic lymphoma (SLL), one sample from a patient with follicular lymphoma (FL), and one sample from a patient with FL and SLL using the dataset from *Boyd et al.* (**Boyd et al., 2009**). The red box and grey dashed box distinguish between the regions occupied between diverse and clonal populations respectively. Point (a) corresponds with healthy individual 10.

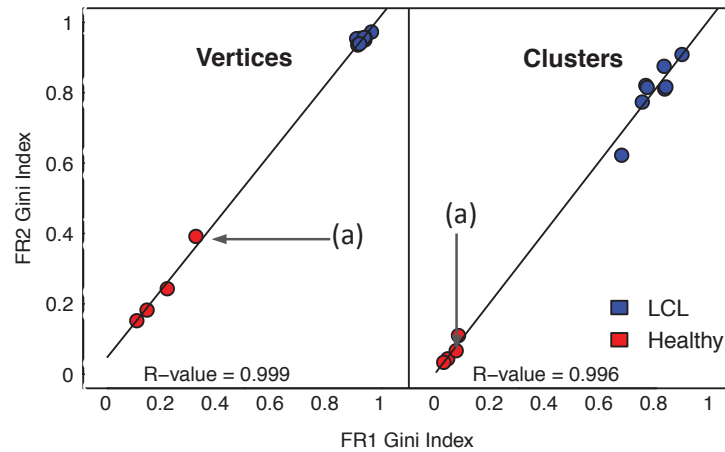


Figure 3.7. Comparison of diversities from FR1 and FR2 primer sets.

Correlation between the Gini Indices of BIOMED-2 FR1 or FR2 primed samples for vertex sizes and cluster sizes, with the corresponding Pearson R-value. LCL represented in blue and healthy individual samples in red. Point (a) corresponds with healthy individual 10.

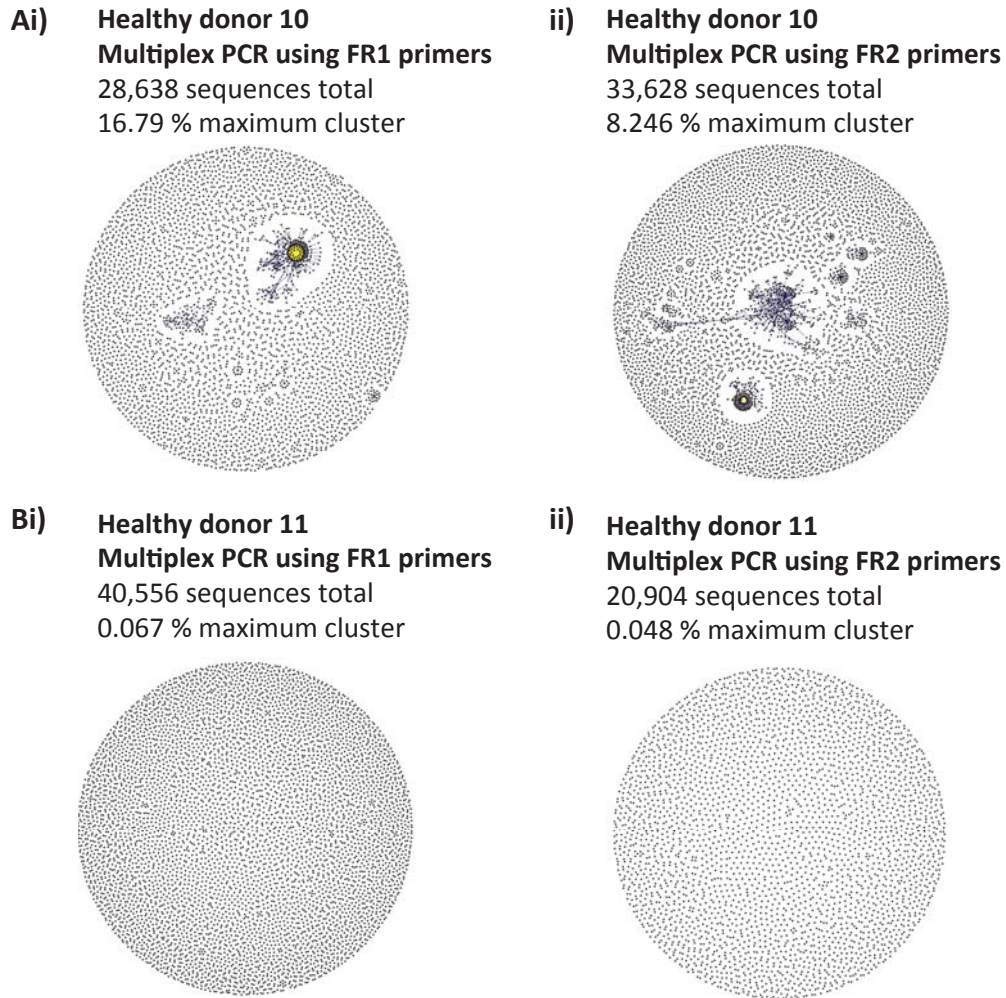


Figure 3.8. B-cell receptors networks for FR1 and FR2 primer amplified healthy donors.

A) B-cell receptor repertoires from healthy individual 10 amplified using **i)** the FR1 primer set and **ii)** the FR2 primer set. **B)** B-cell receptor repertoires from healthy individual 11 amplified using **i)** FR1 primer set and **ii)** the FR2 primer set. The vertex colors correspond to the relative abundance of the corresponding sequences, where red, orange and yellow indicates observation of a sequence in >90%, between 40-90% and <40% of the reads in the sample respectively.

Next, separation of monoclonal expansions, biclonal expansions and diverse B-cell populations was investigated using the maximum cluster sizes and second maximum cluster sizes (**Figure 3.9A**). The CLL and LCL samples have maximum cluster sizes >30% of the total reads compared to maximum cluster sizes of healthy individual samples of <20%. However, the LCLs and CLLs collectively occupy two distinct regions in this space. One group exhibits a single dominant clonal sequence (monoclonal), where all remaining clusters are <5% of the total reads (**Figure 3.9A** surrounded by the dashed line).

The second group of samples has two dominant clusters above 40% and 20% of the total reads respectively (bi-clonal). To determine whether the two dominant clusters are derived from the same B-cell lineage, alignments between the cluster sequences can be used. Firstly, if the two clusters derived from the same B-cell progenitor, they would exhibit the same IgHV-D-J rearrangement. If the two clusters came from different B-cell progenitors but have undergone the same IgHV-D-J rearrangement, the joining regions between the rearranged genes should be different. Therefore, to test whether the two dominant clusters in CLL patient 5 (**Figure 3.4D**) originate from the same B-cell progenitor the IgHV-D-J combinations were determined using IgBLAST. The two dominant clusters use different V-D-J genes ([IGHV3-66*03/IGHD6-19*01/IGHJ3*02] and [IGHV6-1*01/IGHD3-3*01/IGHJ4*02] respectively), and the alignment between the most abundant BCR sequences within these clusters show poor sequence similarity (**Figure 3.10**). Together this indicates that the two dominant clusters in CLL patient 5 originate from two different B-cell progenitors, or secondary rearrangements within the CLL clone. This could potentially be clarified by determining whether the B-cells from these two clusters share identical light chain sequences.

Limited polyclonal expansions were observed also in 5/8 of the LCL samples reflecting that EBV transformation of peripheral B-cells frequently results in polyclonal LCLs. Using the dataset from *Boyd et al.* (Boyd et al., 2009), the same phenomenon of polyclonal expansions in a subset of samples was shown (patients with CLL/SLL and FL/SLL, **Figure 3.4Eiii**) where the maximum cluster sizes are >35% and second maximum cluster sizes are >19% of the total reads (**Figure 3.9B**). Therefore the polyclonal status of the tumour samples can be determined using B-cell network reconstruction and analysis.

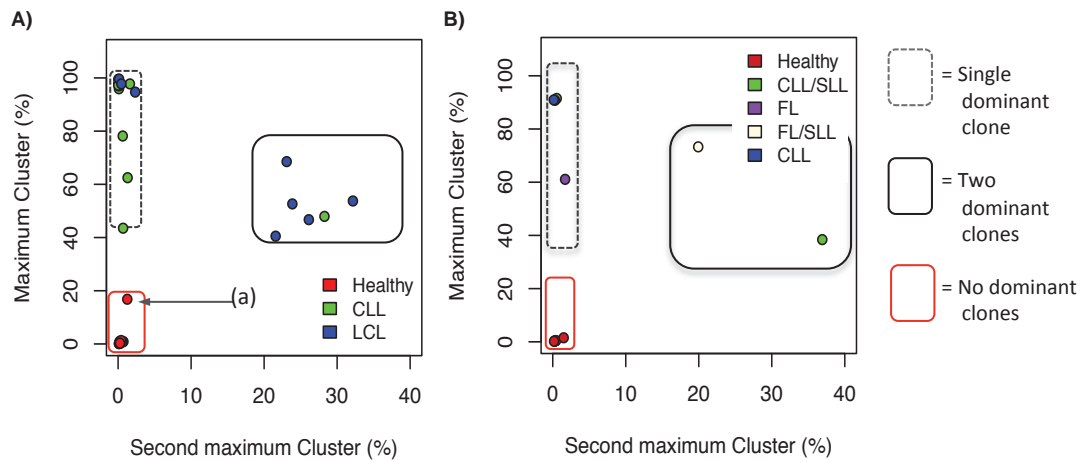


Figure 3.9. Measures differentiating between B-cell receptor dominant clusters.

The second maximum cluster sizes plotted against the maximum cluster sizes for **A)** thirteen healthy individual samples, eleven chronic lymphocytic leukemia (CLL), and eight human lymphoblastoid cell line (LCL) samples and **B)** six healthy individual samples, two samples from patients with CLL and small lymphocytic lymphoma (SLL), one sample from a patient with follicular lymphoma (FL), and one sample from a patient with FL and SLL using the dataset from *Boyd et al.* (**Boyd et al., 2009**). The red, grey dashed and black solid boxes distinguish between the regions occupied between unexpanded populations, monoclonal expanded populations and biclonally expanded populations respectively. Point (a) corresponds with healthy individual 10.

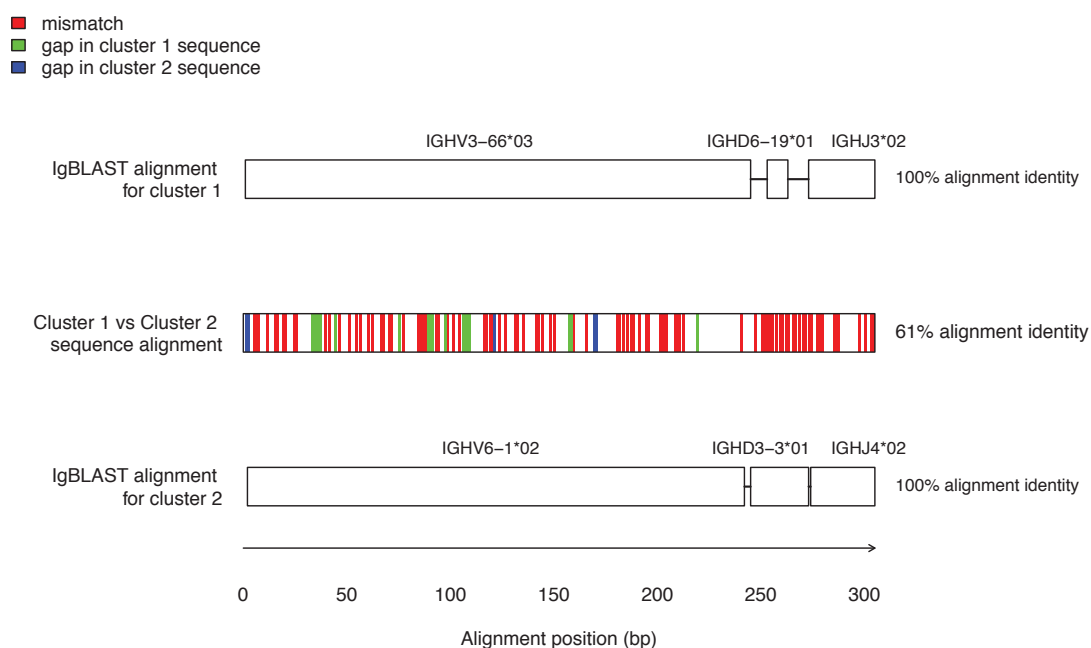


Figure 3.10. Comparison of cluster 1 and cluster 2 sequences for CLL patient 5.

Sequence alignment of the most highly expressed sequences in the two dominant clones for CLL patient 5 by ClustalW to reference IgHV genes and to each other. Cluster 1 and cluster 2 refer to the largest and second largest clonal clusters in the BCR network for CLL patient 5 respectively (representing 48% and 28.3% of the reads respectively). The cluster 1 and 2 sequences were aligned to each other, and the positions of mismatches and gaps are indicated by the coloured boxes in the corresponding alignment positions in the middle row. IgBLAST was used to identify the most similar reference IgHV, D and J genes to the cluster 1 and 2 sequences, shown in the corresponding rows, with the regions of alignments indicated by the boxes. These showed 100% sequence identity between the CLL cluster sequences and the reference germline sequences, as indicated by the text to the right of the alignments.

3.2.7. Network property sensitivity to sequencing depth and edge lengths

To robustly compare B-cell populations between samples, the population measures used must reflect differences in population structure rather than variations in depth of sequencing (scale invariant) and volume of PB sample. If a given diversity measure is scale invariant for B-cell networks then the network diversity measure should be the same regardless of the depth of sampled sequences, i.e. a subset of sequences should yield the same network diversity measure as the full set of sequences. All the proposed population measures were tested as a function of sequencing depth by randomly sampling different proportions of the sequence data for each sample followed by calculation of the corresponding network parameters for both the vertex and cluster size distributions for the LCL, CLL and healthy samples. All the proposed measures showed little variation at different sample sizes even when sub-sampling as low as 20% of the original total data (**Figure 3.11A-D**). Below 20%, small deviations in the Gini Index measures are seen, which is due to low sampling depth leading to higher relative sampling stochasticity. Therefore, this suggests a minimal read depth of ~8,000 for use in comparing populations. As these network measures had minimal standard deviation over all sub-sampling ranges, they are therefore robust parameters for inter-sample comparison.

Secondly, it was hypothesised that generating networks to allow edges to join BCR sequences with greater than one base-pair difference would not greatly influence the network architecture. This hypothesis is based on the assertion that any two B-cells derived from different progenitor B-cells would yield IgHV-D-J rearrangements or non-template insertions and deletions that would differ by only a few base-pairs. To test this, networks were generated to include edge lengths of up to 5 base-pair changes (**Figure 3.12**). It is shown that networks with edges between BCR sequences that differ by up to 5 base-pairs faithfully retain the network architecture for both the clonal and diverse samples (from LCLs and healthy individuals respectively).

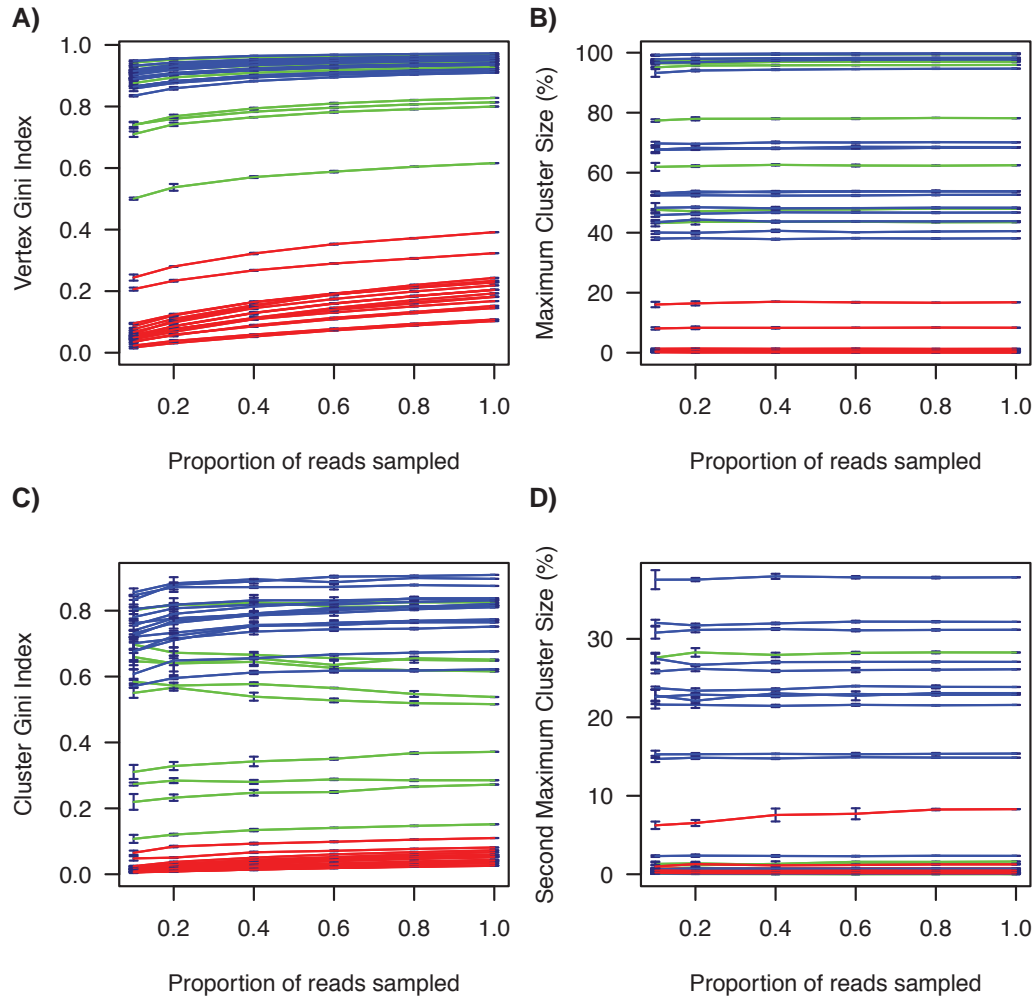


Figure 3.11. Variation of BCR population measures with sampling depth.

The 454 sequences were randomly sampled at a range of proportions of the overall number of reads for eight human lymphoblastoid cell line (LCL) (blue) and thirteen healthy individual samples (red) and eleven patients with chronic lymphocytic leukaemia (CLL) (green). The variation of the diversity measures against varying sequencing depth using, respectively, the Gini index for **A)** vertices and **B)** clusters, and **C)** maximum and **D)** second maximum cluster sizes. An average of 4 subsamples was taken at each proportion, and the error bars give the standard deviation from the mean for each network measure.

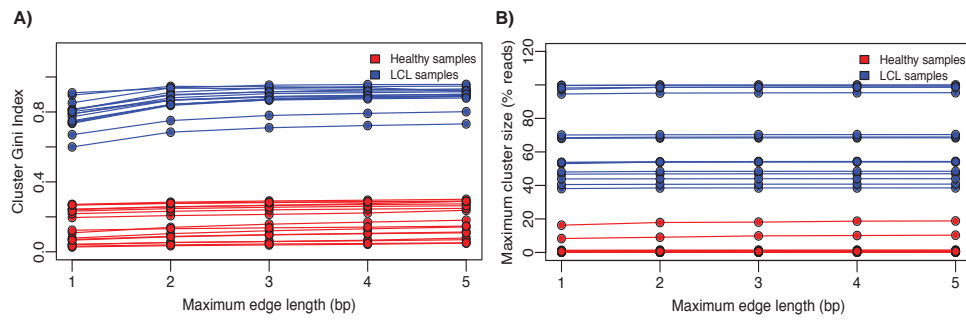


Figure 3.12. Network structure variation with edge length.

A) Cluster Gini Indices and **B)** maximum cluster sizes for networks with different maximum edge lengths. For each sample, networks were generated allowing generation of edges between vertices that differ by at most the corresponding number of mismatches (edge lengths). The corresponding network parameters were calculated for each network, and plotted.

3.2.8. Minimal effect of sequencing errors on network properties

Next, it was investigated whether the diversity of sequences within clusters were likely to be due to the process of somatic hypermutation or sensitive to or generated through sequencing error of a unique amplified BCR sequences. For a given BCR sequenced multiple times, such as when multiple B-cells express identical BCRs, the expected number of vertices comprising a cluster that could be due to sequencing error was estimated, given the experimentally derived PCR and sequencing error-rates (described in Section 2.12). All the samples have cluster sizes greater than that expected due to per-base error alone of 1.74×10^{-4} (Table 3.5), even at twice the measured error-rate (**Figure 3.13**). Therefore, the connectivity patterns of networks predominantly reveal differences in clonal expansions of B-cell populations rather than total sequencing errors. The clusters identified in BCR networks are derived from B-cells that share a common pro-B-cell progenitor with rearranged V-D-J that have subsequently expanded and diversified.

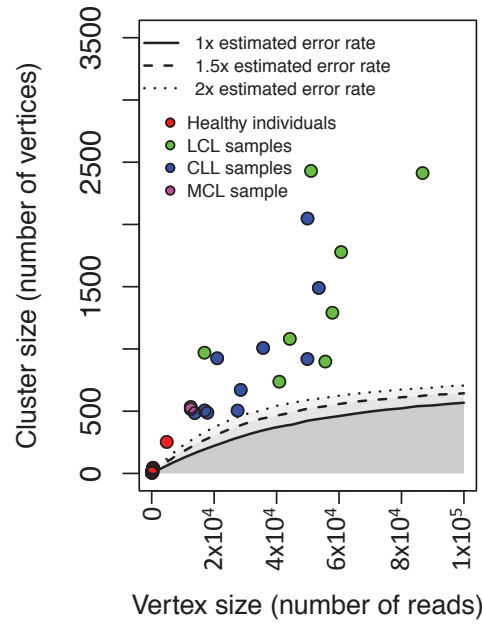


Figure 3.13. Assessment of error in BCR networks.

Comparison of the cluster sizes of the BCR networks with the expected cluster sizes that would result from sequencing and PCR error at three different error-rates. For a given BCR cDNA that is sequenced multiple times, the estimates of the number of vertices making up a cluster that may be due to sequencing error at a given PCR and sequencing error-rates, where the dotted, dashed and solid lines refer to 1x, 1.5x and 2x the experimentally determined non-homopolymeric 454 error rate of 7.04×10^{-5} , assuming the central BCR sequence is otherwise unconnected to any other vertices in the BCR network. The circles show the cluster sizes of the largest clusters in the networks for the different samples, where red, blue and green correspond to healthy individuals, CLL patients and LCL samples respectively.

3.2.9. BCR repertoire network parameters relate to CLL development

To assess the sensitivity of BCR sequencing using multiplex PCR amplification, the titration experiment from *Boyd et al.* (Boyd et al., 2009) in which serial 10-fold dilutions of a known clonal CLL PB sample into normal peripheral blood was used. 90.9% of all reads in the undiluted sample fall within the leukemic cluster (**Figure 3.14A-B**). Using these methods, the leukemic clonal sequences can be detected at dilutions as low as 1:100,000 when the sequence is known and pre-defined. (A MiSeq BCR dilution series was also performed in chapter 7 giving sensitivity of $>1:10^7$). When the leukemic cluster sequences are unknown, detection of expanded clones relies on detecting the maximum cluster size that is significantly different from that of healthy individuals. Significant increases in maximum cluster size were seen above that of the healthy individual in CLL dilutions of 1:100 or less.

The relationship between the BCR population measures and the CLL clinical information for each patient was next determined. Interestingly, there was a strong correlation between the length of time since CLL diagnosis with the vertex Gini Index (**Figure 3.15A**) and the maximum cluster size (**Figure 3.15B**). This suggests longer disease times lead to larger vertices representing larger tumor clonal populations, in agreement with previous studies (Hayes et al., 2010, Kelly et al., 2002).

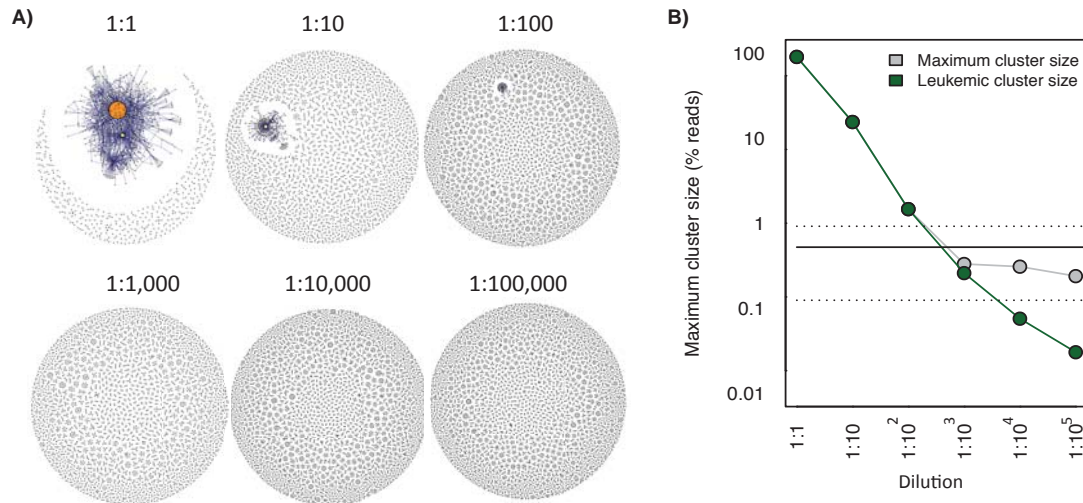


Figure 3.14. Variation of B-cell receptor populations.

A) B-cell receptor networks for the titration of a chronic lymphocytic leukemia clonal sample into healthy peripheral blood from the dataset from *Boyd et al. (2009)* and **B)** the corresponding number of reads corresponding to the leukemic clone (green) and the maximum cluster size of each dilution (grey). The solid horizontal line shows the mean maximum cluster size for healthy individuals from this dataset (0.52% of total reads), and the dashed horizontal lines show the mean \pm standard deviation of maximum cluster size for healthy individuals for this dataset.

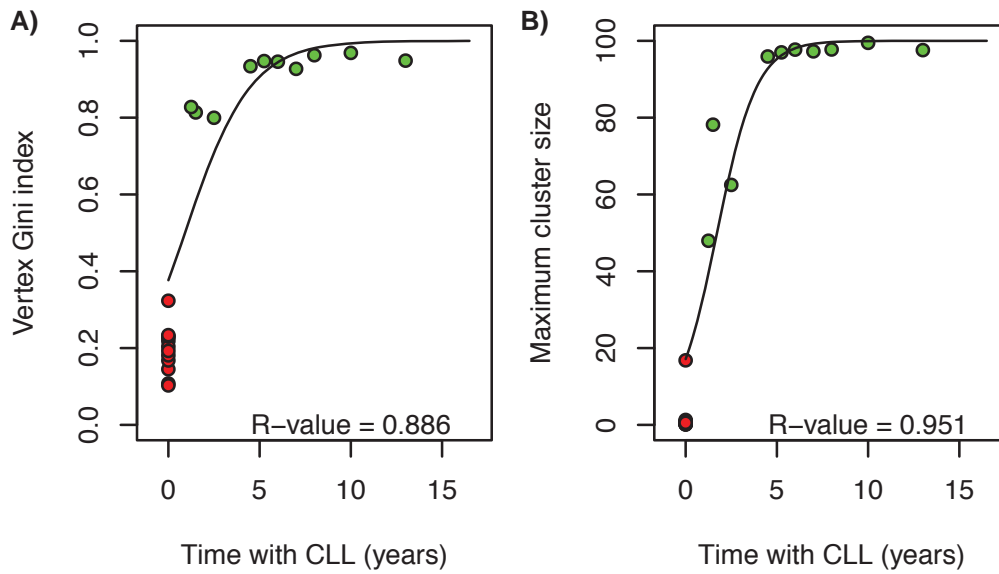


Figure 3.15. BCR diversity variation with time since CLL diagnosis.

Correlation between the **A)** vertex Gini Index and **B)** maximum cluster size with the length of time since chronic lymphocytic leukemia (CLL) diagnosis for each patient. The R^2 -value for the logistic regression is given. The red and green circles correspond to healthy individuals and CLL patients respectively.

3.2.10. Following malignant B-cell clonal dynamics by BCR sequencing

It was hypothesised that BCR sequencing can be used to follow the dynamics of B-cell clones in samples taken multiple time points from patients. To test this, samples were taken from patients separated by a period of time in which the patient had either (a) not undergone any treatment prior to or during sampling (12 samples, denoted no treatment samples), (b) before and after a round of Chlorambucil treatment (5 samples, denoted during treatment samples) or (c) patients who had previously undergone Chlorambucil treatment, but not treatment was given during sampling (4 samples, denoted after treatment samples), summarised in Table 3.7 and **Figure 3.16**. The BCRs from these samples were amplified by multiplex PCR and sequenced by MiSeq sequencing.

Table 3.7. Filtered BCR depths for temporal CLL patient samples.

ID	Number of filtered BCR reads		Clinical condition	Stage	Treatment history	Patient ID	Time between samples (days)
	Time 1	Time 2					
C1	85858	81882	No treatment	A	Untreated	Pat. 1	182
C2	149400	154339	No treatment	C	Untreated	Pat. 5	255
C3	45354	113142	No treatment	A	Untreated	Pat. 6	168
C4	113142	144814	No treatment	A	Untreated	Pat. 6	91
C5	132877	127714	No treatment	A	Untreated	Pat. 8	56
C6	127714	109907	No treatment	A	Untreated	Pat. 8	63
C7	109907	112753	No treatment	A	Untreated	Pat. 8	92
C8	112753	121805	No treatment	A	Untreated	Pat. 8	125
C9	221127	138806	No treatment	B	Untreated	Pat. 9	57
C10	138806	120070	No treatment	B	Untreated	Pat. 9	84
C11	161873	123582	No treatment	A	Untreated	Pat. 11	151
C12	112722	87626	No treatment	A	Untreated	Pat. 13	245
C13	58446	241118	During treatment	A	Chlorambucil	Pat. 2	623
C14	99029	91332	During treatment	B	Chlorambucil	Pat. 4	28
C15	91332	149131	During treatment	B	Chlorambucil	Pat. 4	91
C16	138208	154211	During treatment	B	Chlorambucil	Pat. 7	28
C17	90864	98690	During treatment	C	Chlorambucil	Pat. 10	140
C18	146188	175261	After treatment	Atypical	Chlorambucil	Pat. 3	182
C19	149131	123046	After treatment	B	Chlorambucil	Pat. 4	98
C20	126241	151777	After treatment	C	Chlorambucil, rituximab	Pat. 12	163
C21	151777	164127	After treatment	C	Chlorambucil, rituximab	Pat. 12	152

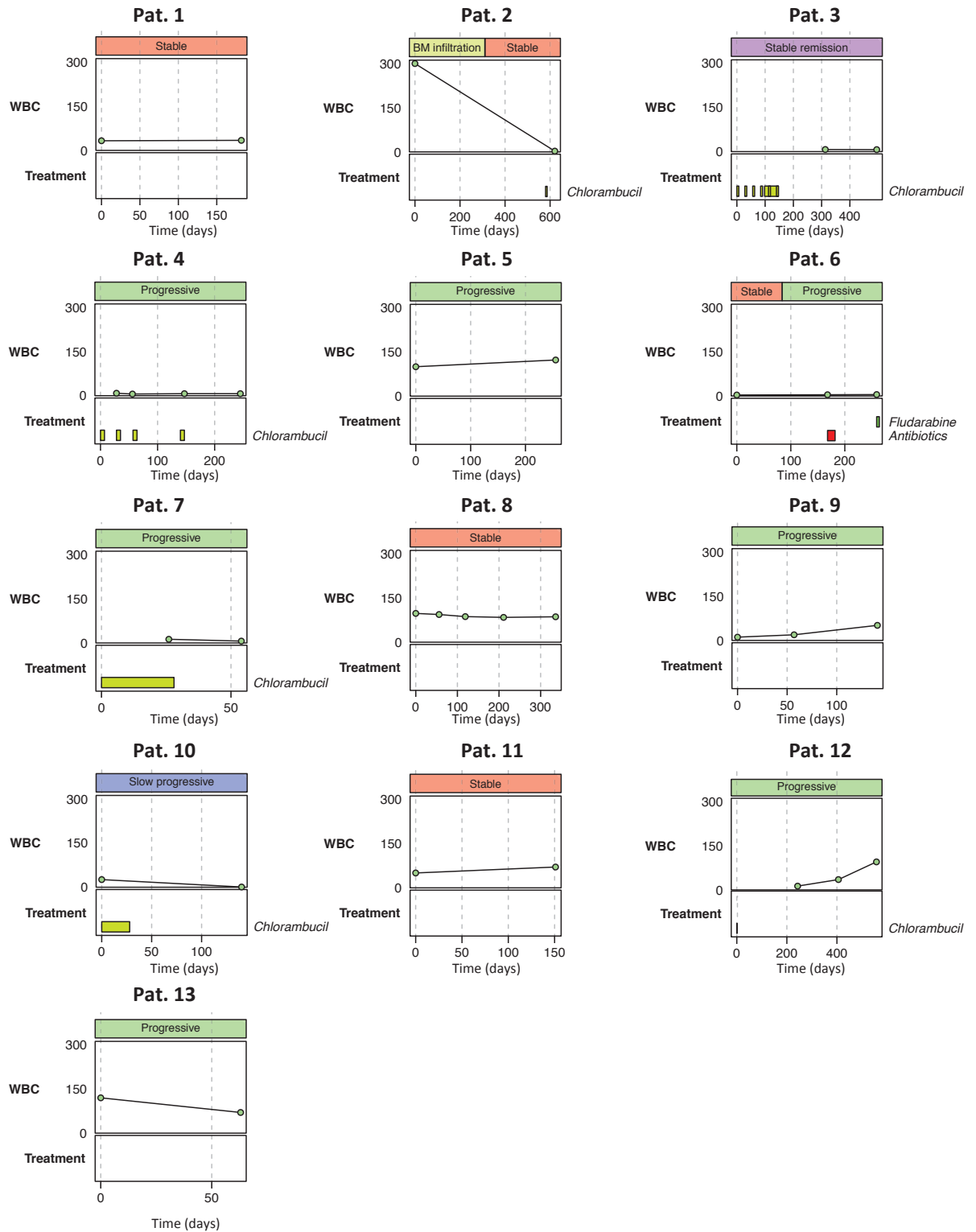


Figure 3.16. Treatment times and white blood cell count over time for temporal CLL samples.

Diagrams showing the relative sampling times with respect to treatment times and white blood cell count (WBC). In each panel, the top bar indicates the disease progression status of the patient at the corresponding time points, the middle section indicates the WBC at the sampled time points, which corresponded to the times that PBMC samples were taken for the BCR analyses, and the lower section indicated if and when treatment was given to the patients. Abbreviation: BM is bone marrow.

The largest (malignant) cluster was the same at each time point, where the higher frequency sequences are always retained between time points. However, the maximum (malignant) cluster size changes over the time intervals were distinct between clinical groups (range 2.52-95.77% of total sequences, **Figure 3.17Ai**). Notably, only the malignant clusters in samples taken during treatment significantly decreased in size over the time interval of sampling (p-value<0.005), whereas the no treatment and after treatment samples did not significantly increase or decrease in cluster sizes. The maximum cluster sizes from samples taken during treatment were significantly reduced compared to the no treatment samples (p-value=0.00445, **Figure 3.17Aii**). This reflects the white blood cell count (WBC) for these patients, where the WBC was significant reduced during treatment compared to the no treatment samples (p-value=0.00812, **Figure 3.17B**). Interestingly, the after therapy samples exhibited a mixed response to therapy, where the change in maximum cluster sizes range from a reduction by 10.14% to an increase by 10.60%, reflecting the change in WBC, which ranged from a reduction of -0.4×10^9 cells/L to an increase of 60.2×10^9 cells/L.

In addition, the vertex and cluster Gini indices derived in Section 3.2.6 that describe the B-cell clonalities of the samples were shown to change in a similar fashion (**Figure 3.18**). In patients that were undergoing active treatment, the vertex Gini indices significantly decreased over time, suggesting that the overall clonality of these patients were decreasing (p-value<0.005, **Figure 3.18Ai**). However, in patients that were not undergoing active treatment, the vertex Gini indices did not significantly increase or decrease (p-values>0.0759, **Figure 3.18Ai**) suggesting stable overall clonality in these patients. In fact, the vertex Gini indices from samples taken during treatment were significantly reduced compared to the no treatment samples (p-value= 1.71×10^{-5} , **Figure 3.18Aii**), suggesting significant changes in the overall B-cell population clonality during treatment. For samples taken after treatment (i.e. no active treatment given between sampling, p-value = 0.288, **Figure 3.18Ai**), there was no significant increase or decrease in vertex Gini index, and the changes were not significantly different from that of the no treatment samples (p-value=0.813, **Figure 3.18Aii**). This suggests that when active treatment is discontinued in CLL patients, the overall PB B-cell clonality remains stable.

The cluster Gini index indicates the overall sample SHM, where an increase in the cluster index means a higher unevenness of the number of unique BCRs in

between the clusters in a sample. Therefore, the significantly increase in the cluster Gini indices for the patients who had never undergone therapy (p-value<0.005, **Figure 3.18Bi**), indicating that there is significant diversification in these patients even though the CLL clone is not significantly enlarging, as indicated by **Figure 3.17Ai** and **Figure 3.18Ai**. During treatment, the cluster Gini index typically reduces (p-value=0.0177, **Figure 3.18Bi**), most likely as a result of reducing the cluster size. However, this does not reach significance, therefore may be indicative of clonal diversification even when the clone is actively being reduced in size. However, the change in cluster Gini indices from samples taken during treatment were significantly reduced compared to the no treatment samples (p-value=0.0077, **Figure 3.18Bii**), suggesting that therapy significantly reduces CLL clonal. After therapy, 3/4 of the patients had stable cluster Gini indices, and only a single patient had 2-fold increase in cluster Gini index (**Figure 3.18Bi**), corresponding to an increase in WBC from 12.9×10^9 cells/L to 35.1×10^9 cells/L, thus further confirming a mixed post-therapy CLL response.

Together these data can be interpreted as, in the absence of treatment, the clone sizes are stable in frequency and undergo CLL clonal diversification. However, during active Chlorambucil treatment, the dominant CLL clone reduces in frequency, with suppressed diversification. This means that Chlorambucil treatment not only reduces the WBC, but also the proportion of the white blood cell population consisting of CLL cells. Once therapy is removed, there appears to be a mixed outcome, where some patients retain a stably low WBC and clonality, whereas others exhibit re-expansion of the CLL clone.

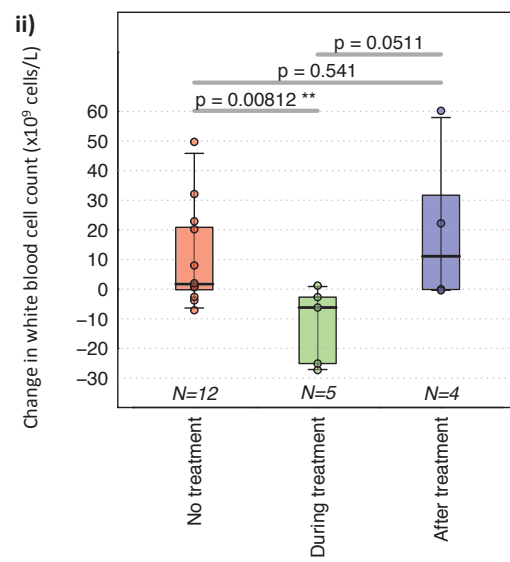
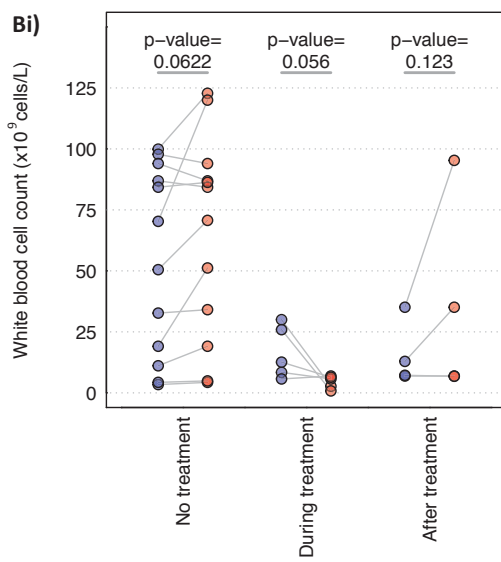
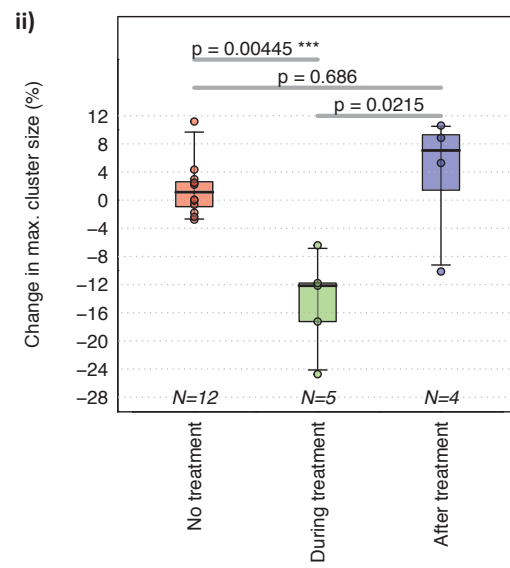
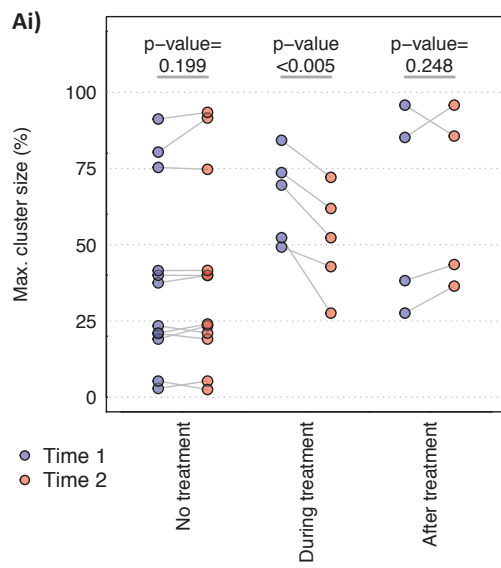


Figure 3.17. Dynamics of CLL BCR repertoires and white blood cell counts.

Samples were taken from patients separated by a period of time in which the patient had either (a) not undergone any treatment prior to or during sampling (12 samples, denoted no treatment samples), (b) before and after a round of Chlorambucil treatment (5 samples, denoted during treatment samples) or (c) patients who had previously undergone Chlorambucil treatment, but not treatment was given during sampling (4 patients, denoted after treatment samples). **Ai)** The plot of the maximum (malignant) B-cell cluster sizes for first and second samples taken, where the blue and red points represent the first and second samples respectively. The grey lines join together adjacent samples from the same patient, and two-side paired t-test p-values between the first and second samples are given above each group. **ii)** Boxplots of the changes in the maximum cluster sizes between the first and second samples for each patient, where the p-values of the significance of the changes between groups is given above, calculated by two-sided unpaired t-tests. **Bi)** The plot of the white blood cell counts (WBCs) for first and second samples taken, where the blue and red points represent the first and second samples respectively. The grey lines join together adjacent samples from the same patient, and two-side paired t-test p-values between the first and second samples are given above each group. **ii)** Boxplots of the changes in the WBCs between the first and second samples for each patient, where the p-values of the significance of the changes between groups is given above, calculated by two-sided unpaired t-tests.

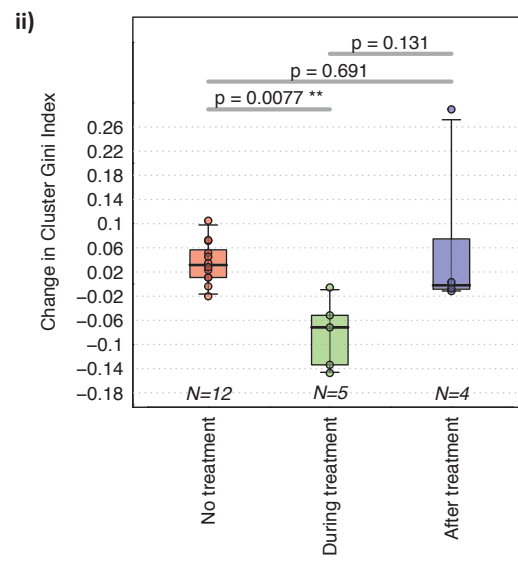
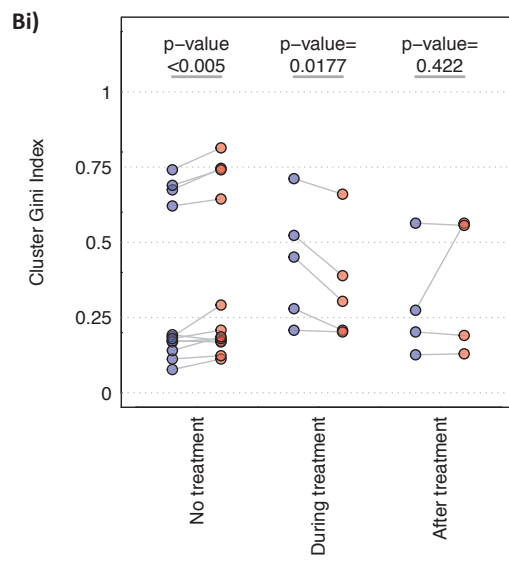
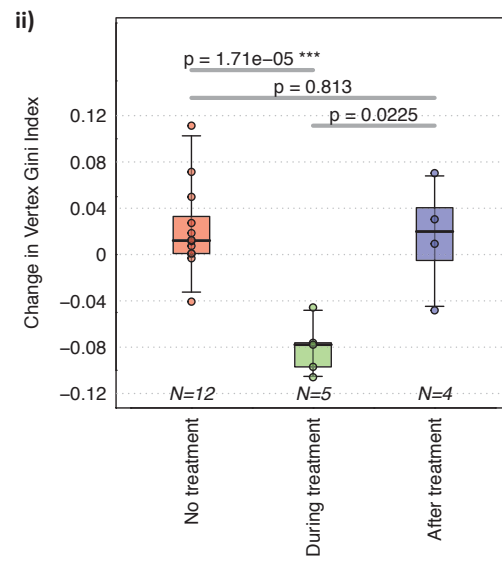
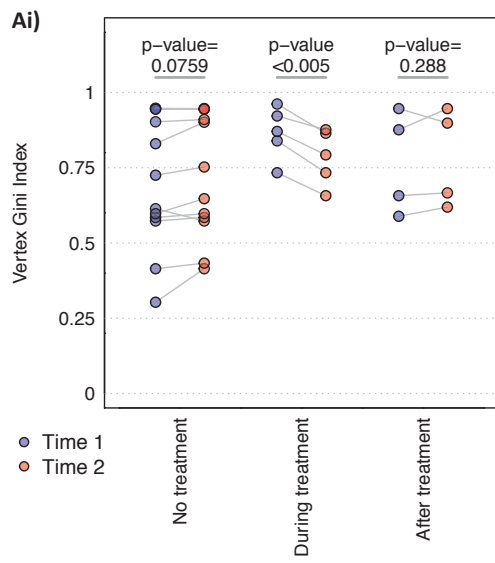


Figure 3.18. Dynamics of CLL BCR repertoires properties.

Samples were taken from patients separated by a period of time as in **Figure 3.17**.

Ai) The plot of the vertex Gini indices for first and second samples taken, where the blue and red points represent the first and second samples respectively. The grey lines join together adjacent samples from the same patient, and two-side paired t-test p-values between the first and second samples are given above each group. **ii)** Boxplots of the changes in the vertex Gini indices between the first and second samples for each patient, where the p-values of the significance of the changes between groups is given above, calculated by two-sided unpaired t-tests. **Bi)** The plot of the cluster Gini indices for first and second samples taken, where the blue and red points represent the first and second samples respectively. The grey lines join together adjacent samples from the same patient, and two-side paired t-test p-values between the first and second samples are given above each group. **ii)** Boxplots of the changes in the cluster Gini indices between the first and second samples for each patient, where the p-values of the significance of the changes between groups is given above, calculated by two-sided unpaired t-tests.

3.2.11. Phylogenetic analysis of B-cell clones

Clonal evolution in CLL as exemplified by the presence of mutations in the genome and by multiple BCRs related to the dominant CLL BCR sequence (Landau et al., 2013, Schuh et al., 2012). Mutations in the BCR may be used to infer the mutational route from a CLL B-cell ancestor to the rest of the leukaemic clone by phylogenetic analysis. Phylogenetic analysis can be used to reconstruct the evolutionary history of organisms (Pybus et al., 2002). However, to date, no B-cell specific evolutionary model of BCR diversification have been developed, hampered primarily by (a) a BCR evolutionary tree is not strictly bifurcating due to the expansion of multiple B-cells with identical BCRs, that can each independently diversify, (b) non-constant mutation rate, dependent on co-stimulation from multiple sources, such as T-cell activation, and (c) ongoing or secondary rearrangements can lead to the replacement of IgHV gene segment while retaining the same IgHD-J region, thus leading to different evolutionary histories in different regions in the BCR (Marshall et al., 1995, Steenbergen et al., 1993, Gawad et al., 2012, Choi et al., 1996).

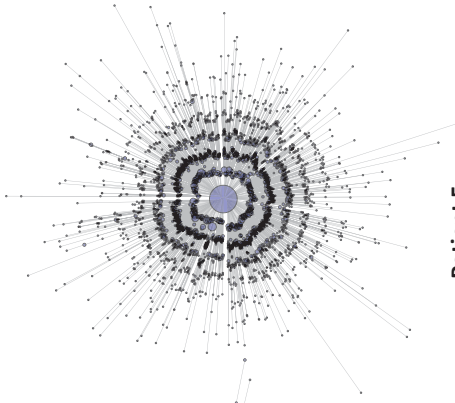
However, despite these drawbacks, the phylogenetic relationships between BCR sequences can inform about the process of B-cell clonal diversification, such as by the tree shape, such as star shapes of each of phylogenetic trees suggest unselected clonal expansion, compared to antigenic drift seen in, for example, influenza virus (Steinbruck and McHardy, 2012, Bedford et al., 2014). Therefore, using the patient samples in section 3.2.10, all the sequences from the dominant clusters were extracted, to determine the maximum parsimony phylogenetic tree structures of the leukaemic clone, and to infer the process of diversification. Maximum parsimony was chosen as the phylogenetic model as this imposes the fewest number of explicit assumptions on the data. For each patient, all the BCR sequences related to the CLL clone were aligned using Mafft (Katoh and Standley, 2013) and a maximum parsimony tree was fitted using Paup* (Wilgenbusch and Swofford, 2003). The branch lengths represent the evolutionary distance between BCR sequences and bootstrapping was performed to evaluate the reproducibility of the trees, showing strong tree support (>95% certainty for all branches). The majority of the trees from patients have a star-like structure (**Figure 3.19**), suggesting that the CLL clone emerged from a single common ancestor (Martins and Housworth, 2002), represented by the central BCR, which was the most frequently observed BCR. The concentric

rings of BCR variants represent incremental increases in base pair differences from the central dominant BCR sequence. However, the phylogenetic trees in patients 6, 10 and 13 show small outgrowths, suggesting potential growth and diversification advantages in the B-cells corresponding to these branches, potentially reflecting genomic variations in these B-cells.

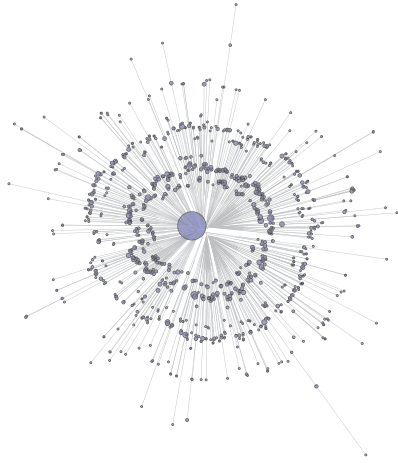
Patient 1
Cluster size: 81.57% of reads



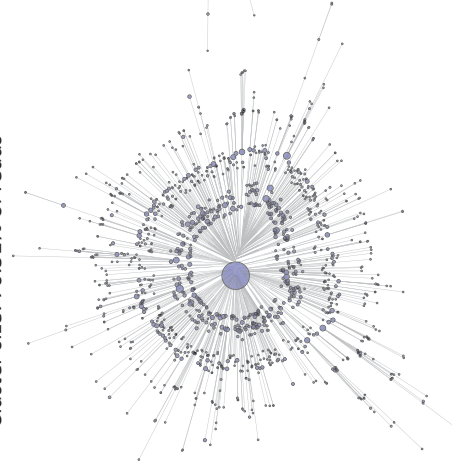
Patient 2
Cluster size: 49.00% of reads



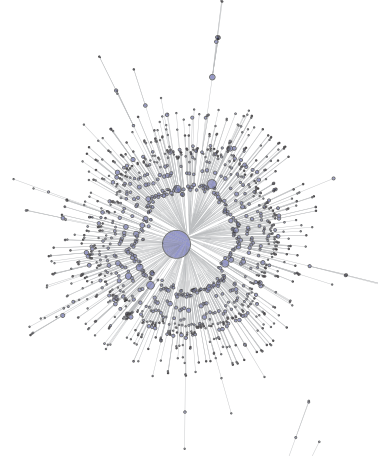
Patient 3
Cluster size: 27.78% of reads



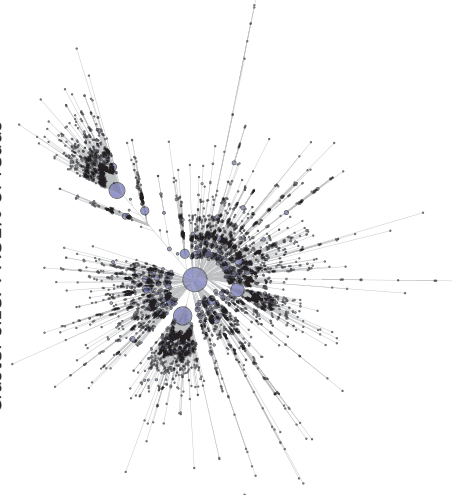
Patient 4
Cluster size: 70.51% of reads



Patient 5
Cluster size: 93.15% of reads



Patient 6
Cluster size: 77.51% of reads



Patient 7
Cluster size: 85.29% of reads



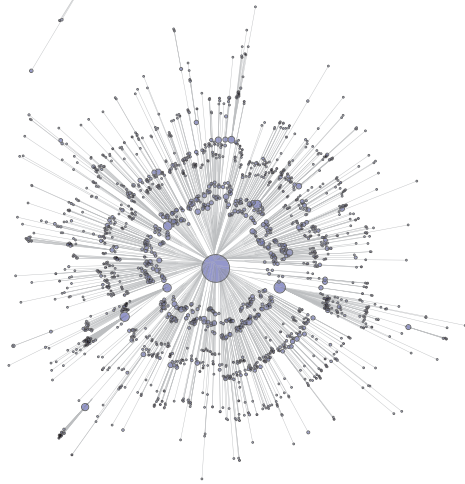
Patient 8
Cluster size: 21.14% of reads



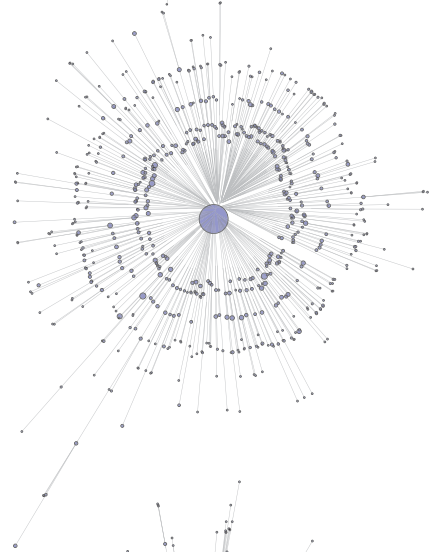
Patient 9
Cluster size: 38.04% of reads



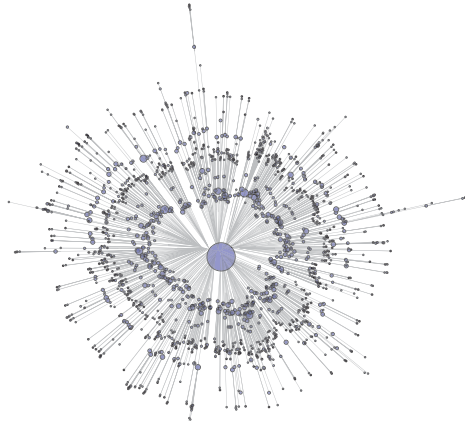
Patient 10
Cluster size: 69.78% of reads



Patient 11
Cluster size: 42.03% of reads



Patient 12
Cluster size: 86.41% of reads



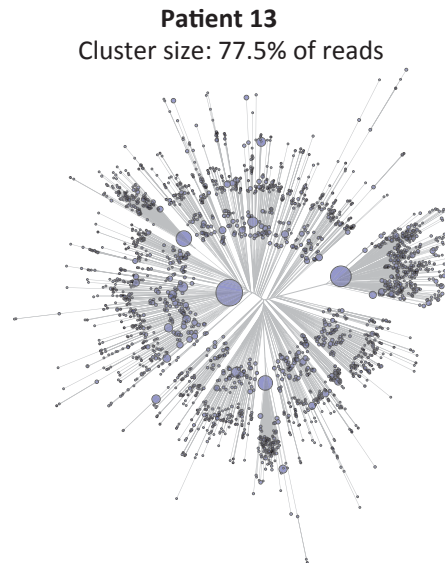


Figure 3.19. Unrooted maximum parsimony trees of the malignant CLL clusters.

For each patient, all sequences in the maximum cluster were aligned using Mafft (Kato and Standley, 2013) and a maximum parsimony tree was fitted using Paup* (Wilgenbusch and Swofford, 2003). The branch lengths represent the evolutionary distance between BCR sequences and bootstrapping was performed to evaluate the reproducibility of the trees, showing strong tree support (>95% certainty for all branches). The branch lengths are proportional to the number of varying bases (evolutionary distance) and the tips represent unique BCR sequences within the malignant CLL cluster. The sizes of the tips are proportional to the number of sequencing reads corresponding to the BCR.

3.3. Conclusions

The aim of this chapter is to discriminate between healthy and malignant B-cell expansions through BCR repertoire sequencing. To do this, methods must be robust to noise, such as PCR and sequencing error, as well as sequencing depth. The effects of amplification and sequencing error are often of concern for BCR deep sequencing. However, the strong linear correlations of the network parameters between samples that have been amplified using independent primer sets suggest limited amplification bias and no significant effect on the overall population structure. In all samples tested from healthy and haematological cancer patients, the cluster sizes are notably greater than that expected due to the process error alone, suggesting that the network structures represent the population structures of the B-cell sample.

The observation of frequent multiple identical BCR sequences in malignant B-cell samples and only low frequency identical BCR sequences from healthy individuals suggests that multiple identical RNA molecules from a single B-cell are rarely sequenced. Therefore, clusters of related sequences are likely to represent BCRs from clonal expansions of evolutionarily related B-cells, whereas naïve B-cell populations form singletons in sparsely connected networks. The probabilities of resampling BCRs from a population is revisited in more detail in Chapter 4.

If the B-cell network from limited sequencing is a random sample of the entire circulating peripheral blood BCR repertoire, then a scale invariant diversity measure should also capture the predominant structure of the unsampled network. Here, it has been shown that network structures, combined with these population measures discriminate between B-cell repertoires of different clonalities in health and disease. These measures are robust to variations in sequencing and sampling depth and different filtering strategies and are applicable to independently produced datasets (Boyd et al., 2009). Using different primer sets, sequencing depths and sequencing technologies, the samples still cluster according to the clonal nature of the samples, occupying the equivalent distinct regions of Gini Index and maximum/second maximum graphs. Therefore this analytical strategy is applicable to any BCR deep sequencing technology.

Deep sequencing of BCR repertoires potentially allows the detection of a clonal lymphoid population in a background of polyclonal cells without prior knowledge of the leukemic sequence (Sayala et al., 2007). Here, the limit of *de novo*

detection of malignant clonality is at least 1 in 100 dilution of CLL cells into healthy blood. In addition, the vertex Gini Index is strongly correlated with the time an individual has been living with CLL. This has potential applications in the detection of clonal B-cell disorders and malignancies, particularly as the early stages of these diseases are asymptomatic, such as in CLL. When there is prior knowledge of a BCR of interest, such as in leukaemia, the limit of detection is much greater (>1 in 10^5 cells). In practice, this has important potential uses in monitoring disease during therapy (addressed in Chapter 4) and minimal residual disease detection (addressed in Chapter 5).

An important result of this framework to assess B-cell repertoire structure is to understand the changes involved in a healthy immune repertoire, such as during vaccination, compared to malignant B-cell expansion. There was variation between the network-based diversity measures of a “normal” BCR repertoires between the healthy individuals, where a larger-scaled assessment of the primary immune response compared to early stage leukaemia could provide clinically important early diagnostic or prognostic information to patients. For example, one healthy individual (healthy individual 10) exhibited a more clonal BCR repertoire compared to the other healthy individuals, defined by an increase in connectivity. Further work could be performed to determine the likelihood of this clonality resulting from an antigen specific memory B-cell expansion or an undiagnosed malignant transformation in an otherwise asymptomatic individual.

Similarly, the presence of more than one BCR clonal expansion in CLL and other blood cancers has unknown clinical implications. These enlarged clusters representing BCRs with different V-D-J gene combinations may be due to either the expansion of two distinct malignant B-cell transformations, or separate antigen-stimulated B-cell clonal expansion unrelated to CLL. These methods used in time-series may allow the distinction between antigen-driven positive selections in CDRs compared to malignant-driven expansion.

B-cells form dynamic populations of cells. Here it is shown that these populations expand and potentially evolve over time. For the first time it is possible to observe the specifics of a short-term effect of therapy on the B-cell repertoire in CLL, and demonstrates how networks lend themselves to phylogenetic approaches. During therapy, there was a significant reduction in B-cell clonality and the percentage of BCRs relating to the malignant B-cell cluster. Work here therefore provides a

framework for analysing deep high-throughput BCR sequencing datasets to probe B-cell population changes between serial samples or individuals.