Chapter 5

5. Minimal residual disease in B-acute lymphoblastic leukaemia

5.1. Introduction

Many of the therapies for ALL cause significant toxicities and carry the potential of long-term complications including secondary malignancies. Additionally, the majority of treatment failures occur as a result of disease relapse occurring either during or after completion of treatment. Therefore, improved detection and monitoring of minimal residual disease in B-cell ALL (B-ALL) is of great clinical importance, particularly for tailoring therapeutic dosing and strategies (Biondi and Masera, 1998). Here, the BCR repertoire was sequenced in a set of B-ALL patients to determine whether BCR sequencing can be used to (a) monitor B-ALL residual disease load and (b) decipher the ontogeny and B-cell population dynamics in relapse patients?

5.2. Results

5.2.1. BCR sequencing of longitudinal samples from B-ALL patients

Longitudinal samples from six B-ALL patients over the course of therapy were analysed for the presence of residual leukaemic cells by both a routine clinical MRD monitoring method of quantifying qPCR transcript levels of fusion genes associated with individual leukaemias (performed by the molecular diagnostic laboratory of the Karaiskakio Foundation), and also by sequencing the BCR repertoire and mining leukaemia-specific BCR sequences. For each patient, a "primary sample" was studied with high leukemic load, as indicated by a qPCR T/C transcript (T/C) ratio greater than 1.66; a ratio which was reduced to zero in subsequent samples taken over the course of therapy (summarised in Table 5.1). Additionally, BCR sequencing was performed on peripheral blood samples from 18 healthy individuals within the range of 20-75 years of age. BCR sequencing yielded 124,302 to 2,972,494 filtered BCR sequences per sample (Table 5.1). BCR network analysis was applied to the sequencing datasets, to identify clusters representing groups of highly related BCR sequences (Bashford-Rogers et al., 2013). Clonality was observed in all B-ALL primary samples, as indicated by largest cluster sizes greater than 2.796% of the total

BCR repertoire. In comparison, the largest cluster sizes from the 18 healthy individuals averaged 0.618% (standard deviation of 0.641%, range 0.14-2.5%).

Table 5.1. B-ALL patient sample information.

Patient ID	Sample ID	Time since first sample (days)	Target/ control transcript ratio	Total BCR sequences in sample	Target transcript type**	Sample source*	Largest cluster (% of BCR sequences)
527	527_A	0	13.95	124,302	E2A-PBX1	BM	43.733
527	527_B	8	0.02	270,572	E2A-PBX1	BM	0.696
527	527_C	15	0.00	756,674	E2A-PBX1	BM	0.320
527	527_D	30	0.00	698,592	E2A-PBX1	BM	0.097
527	527_E	109	0.00	2,320,485	E2A-PBX1	BM	6.818
527	527_F	889	0.00	2,301,914	E2A-PBX1	BM	0.580
859	859_A	0	1.66	454,071	TEL-AML1	PB	2.796
859	859_B	7	0.03	786,283	TEL-AML1	BM	0.179
859	859_C	84	0.00	737,736	TEL-AML1	BM	0.738
859	859_D	374	0.00	1,929,858	TEL-AML1	BM	2.159
859	859_E	1241	0.00	2,025,955	TEL-AML1	BM	0.219
1592	1592_A	0	34.60	259,439	E2A-PBX1	BM	26.600
1592	1592_B	12	12.98	264,698	E2A-PBX1	BM	26.105
1592	1592_C	33	0.02	216,356	E2A-PBX1	BM	0.192
1592	1592_D	554	0.00	129,923	E2A-PBX1	PB	1.040
1611	1611_A	0	35.04	189,634	E2A-PBX1	BM	27.843
1611	1611_B	12	0.00	264,128	E2A-PBX1	BM	0.448
1611	1611_D	510	0.00	284,526	E2A-PBX1	BM	0.175
1611	1611_F	944	0.00	346,134	E2A-PBX1	PB	1.751
1703	1703_A	0	0.12	2,972,494	TEL-AML1	PB	0.390
1703	1703_B	18	0.00	2,209,688	TEL-AML1	BM	1.049
1703	1703_C	336	0.00	1,861,228	TEL-AML1	BM	0.131
1703	1703_D	567	0.00	1,475,750	TEL-AML1	BM	0.353
1703	1703_E	567	3.12	1,237,270	TEL-AML1	CSF	3.833
3243	3243_A	0	1.75	297,165	BCR-ABL	BM	10.196
3243	3243_B	20	0.02	372,194	BCR-ABL	BM	0.194
3243	3243_C	31	0.01	340,706	BCR-ABL	BM	0.331
3243	3243_D	56	0.00	315,718	BCR-ABL	BM	0.320
3243	3243_E	91	0.00	319,850	BCR-ABL	BM	0.451

Samples highlighted in orange denote the primary samples for each patient.

* Abbreviations: BM is bone marrow, PB is peripheral blood and CSF is cerebrospinal fluid.

** E2A-PBX1: gene fusion between the transcription factor *E2A* with the homeodomain protein *PBX1*. TEL-AML1: gene fusion between the transcription factor *TEL* with the transcription factor *AML1*. BCR-ABL: gene fusion between the "breakpoint cluster region" (*BCR*), a 5.8 kbp region of DNA on chromosome 22 (22q11), with the tyrosine kinase *ABL1*.

5.2.2. Comparison of ALL and CLL repertoires

B-ALL is thought to arise from a malignant transformation of immature hematopoietic progenitor at one of several stages of early B-cell development. The BCR repertoire in B-ALL has been shown to be distinct from that of later-stage B-cell leukaemias, such as chronic lymphocytic leukaemia (CLL), where the preferential IgHV-J gene usage in B-ALL has been shown to reflect early B-cell repertoires in B-ALL (Duke et al., 2003). To determine whether B-ALL B-cells have undergone less maturation and diversification than both the mutated and unmutated subtypes of CLL, we compared the BCR clusters in the B-ALL samples to those seen in 9 CLL patients from Chapter 3.

Firstly, the mutational distance of the dominant leukaemic BCR sequences in each CLL and B-ALL patient were compared to confirm that the B-ALL sequences represent an earlier stage of B-cell development than CLL. For each patient the dominant BCR sequence of the B-ALL or CLL cluster was aligned to the IMGT reference database and the percentage sequence identity to reference IgHV genes was determined using IgBLAST (Ye et al., 2013) (**Figure 5.1**A). The dominant BCR sequences in each B-ALL patient were either identical or within 3bp from a reference germline sequence (mean 99.52% of nucleotides identical to reference), supporting the hypothesis that B-ALL arises from B-cells that have not undergone somatic hypermutation (SHM). The sequences that were not identical to the reference germline sequences may be accounted for by allelic variation of the IgHV locus not present in the reference BLAST database. CLL can be defined into two subtypes, where two different mutational statuses of CLL patients are thought to be derived from two different stages of B-cell ontology, with the unmutated CLL cases corresponding to pre-antigenic stimulation, and the mutated cases corresponding to post-antigenic stimulation (Hamblin et al., 1999, Damle et al., 1999). Therefore, the CLL patients were subgrouped into unmutated (where the dominant BCR had >98% sequence similarity with reference germline IgHV-D-J sequences) or mutated CLL (dominant BCR <98% sequence similarity with reference germline IgHV-D-J sequences). The unmutated subtype CLL patients exhibited no significant difference in sequence similarity to the reference IgHV BLAST database (mean 98.9% identical to reference, p-value=0.478), whereas the mutated subtype CLL patients had

significantly lower sequence similarity to the reference IgHV BLAST database (mean 92.5% identical to reference).

Secondly, the diversification of the malignant clusters in B-ALL and CLL were compared to determine whether B-ALL has a lower propensity to diversify than CLL. For each patient, all BCR sequences within the malignant cluster were aligned and for the sequences not identical to the dominant vertex of the cluster, the numbers of mutations away from this highest-observed BCR sequence were determined (**Figure 5.1**B). The B-ALL malignant clusters showed a lower mutational distances from the dominant BCR sequence than CLL (means distances of 2.017bp, 2.277 and 2.694bp for B-ALL unmutated CLL and mutated CLL respectively, p-values<0.005), suggesting lower levels of SHM within the B-ALL B-cell population compared to both CLL mutational subtypes. Together, this data supports the idea that B-ALL arises from earlier stages of B-cell differentiation than the mutated CLL subtype, and as such displays lower, albeit detectable, levels of SHM and clonal diversification than both CLL mutational subtypes.

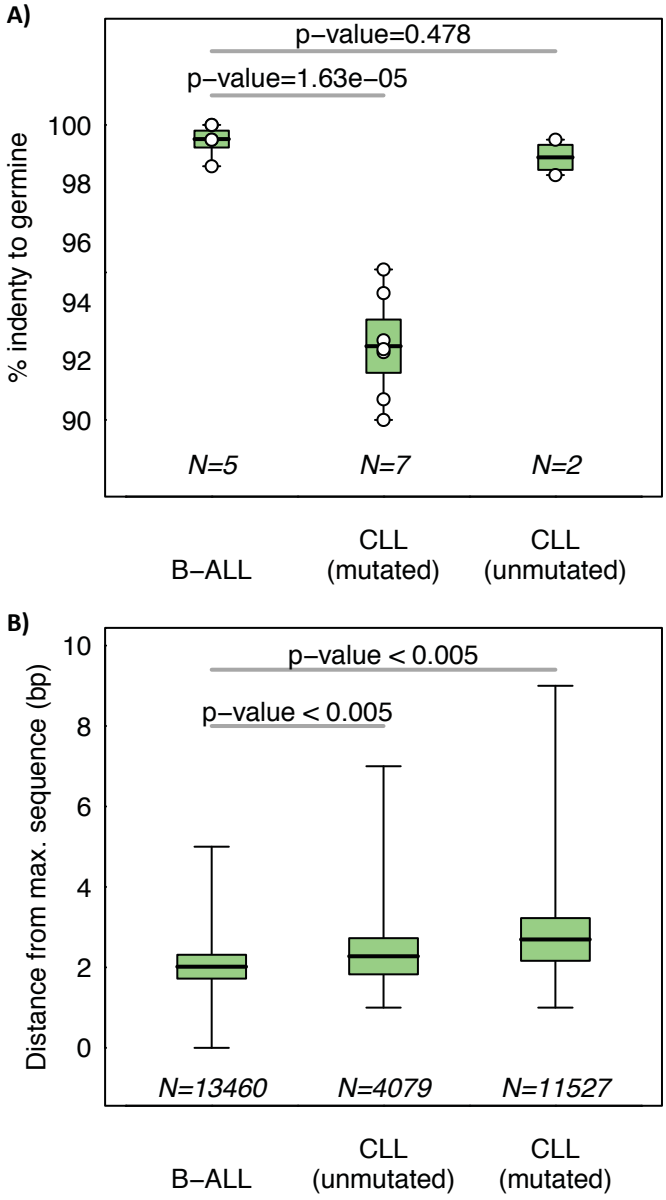


Figure 5.1. Comparing the B-cell repertoire in B-ALL with CLL.

A) The percentage sequence identity of the dominant clonal sequence compared to reference germline sequences from B-ALL and CLL patients with either unmutated CLL (dominant BCR >98% sequence similarity with reference germline IgHV-D-J sequences) or mutated CLL (dominant BCR <98% sequence similarity with reference germline IgHV-D-J sequences) and **B)** the distribution of base-pair distances of all unique sequences in the malignant clusters away from the dominant clonal sequence from B-ALL and unmutated CLL or mutated CLL patients. P-values for comparisons between the distributions are indicated (two-sided T-test).

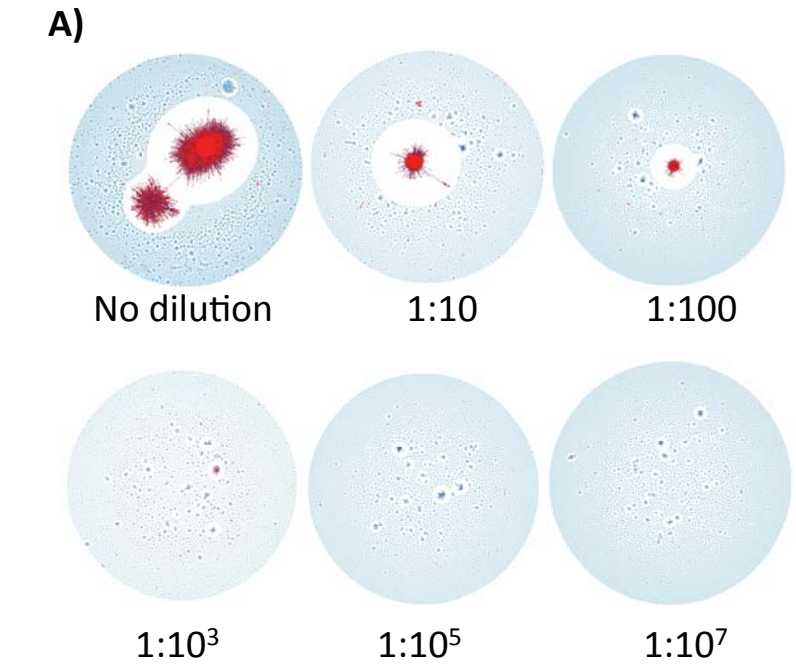
5.2.3. BCR sequencing sensitivity to detect B-ALL clones

Minimal residual disease detection by BCR sequencing requires the accurate detection of B-ALL- associated BCR sequences within a large sequencing dataset. Therefore a Python code, named MRD Assessment and Retrieval Code in pYthon (MRDARCY), was developed to identify malignant BCR sequences from a diagnostic B-ALL sample represented by sequences in the largest cluster, and to search samples at later time points for identical or related BCRs allowing for a specified number of base-pair mismatches (here a threshold of 8 bp is used).

To assess the sensitivity of BCR sequencing for detecting specific B-cell clones from RNA, we performed a titration experiment using serial 10-fold dilutions of a known clonal B-ALL PB sample RNA (sample 1592_A) into normal peripheral blood RNA. The IgH multiplex PCR primer set used in the PCR amplification of the B-cell repertoire consists of six primers, with each primer binding to a different subset of IgHV genes. Therefore it was hypothesized that the primer from this set that binds best to the leukaemic BCRs in the dilution series, denoted IgHV-specific primer, would amplify the leukaemic BCRs preferentially, thus increasing sensitivity of detecting leukaemic BCR compared to the multiplex approach. To test this, multiplex PCR amplification and singleplex IgHV-specific PCR amplification were performed on these samples. Each dilution series sample yielded an average of 125,642 filtered BCR sequences (range of 18,970-294,354, **Figure 5.2**A-B). 31.41% of all BCR sequences in the undiluted sample are related to the leukaemic cluster as identified by MRDARCY, where the percentages of leukaemic BCRs detected approximated to a log-log correlation with dilution. Leukaemia-specific BCR sequences were detected in dilutions as low as 1 in 10^7 RNA molecules for both the multiplex and singleplex IgHV-specific PCR strategies (**Figure 5.2**B when the BCR identity of the tumour clone was known *a priori*). In contrast to this, qPCR has been shown to have a sensitivity of 1 cell in 10^5-10^6 (Campana, 2010). Interestingly, there was an increase in sensitivity of an average of 13.57x using the singleplex IgHV-specific PCR strategy across the dilution range, suggesting that this patient specific MRD monitoring approach, where multiplex BCR sequencing is used on the initial sample and followed by specific clonotypic IgHV primer, could be adapted into a powerful clinical MRD monitoring tool. In fact, with this sensitivity, if only 1 B-ALL cell is present in a typical 5ml blood sample containing ~$1.5x10^6$ B-cells, a read depth of

only $4.5x10^6$ is required to give a >95% probability of detection (using the Poisson distribution and assuming that all BCRs were amplified). Therefore, BCR sequencing has unparalleled sensitivity to capture specific sequences with an important application in MRD monitoring of B-ALL and potentially other B-cell leukaemias.

However, when the leukemic cluster BCR sequences are unknown, detection of expanded clones relies on detecting the maximum cluster size that is significantly different from that of healthy individuals, i.e. when there the leukaemic B-cell population represents 1 in 100-500 RNA molecules (light green line, **Figure 5.2**B), and is consistent with the dilution series in Section 3.2.9. Therefore, the sequencing of BCR repertoires at diagnosis of B-ALL may be critical to the subsequent detection and tracking of small clonal lymphoid populations in a background of polyclonal cells.


A)
No dilution
1:10
1:100
1:10^3
1:10^5
1:10^7

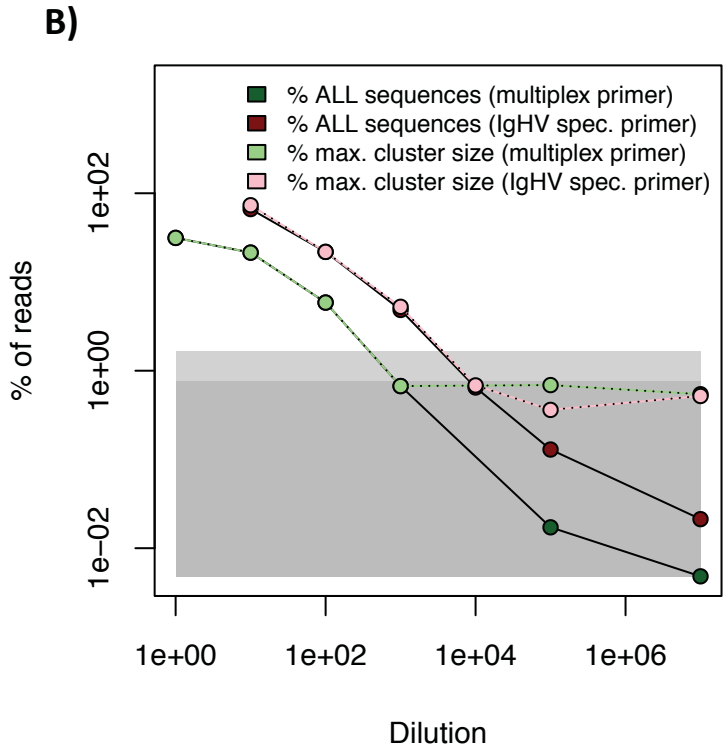

B)
% ALL sequences (multiplex primer)
% ALL sequences (IgHV spec. primer)
% max. cluster size (multiplex primer)
% max. cluster size (IgHV spec. primer)
% of reads
1e+02
1e+00
1e−02
1e+00
1e+02
1e+04
1e+06
Dilution

Figure 5.2. BCR sequencing sensitivity.

RNA from a clonal B-ALL patient sample was mixed with RNA from healthy peripheral blood PBMCs at different ratios. BCR sequencing using the full set of multiplex primers or a single PCR primer chosen from the multiplex primer set with the best alignment (i.e. annealing potential) to the malignant B-ALL BCR sequence (IgHV specific primer). **A)** Network diagrams showing sequential dilution of B-ALL BCR population into healthy blood using the multiplex primers, where vertices within 5bp sequence similarity to the B-ALL cluster are marked in red at each dilution, otherwise coloured blue. **B)** The percentages of BCR sequences corresponding to the B-ALL BCR population at each dilution into healthy RNA using multiplex primers (dark-green) and IgHV specific primer (dark-red). Overlaid with the percentage of BCR sequences in the largest cluster for multiplex primers (light-green) and IgHV specific primer (light-red).

It is possible, but unlikely, that the same IgHV-D-J rearrangement and joining regions can be generated by chance in independent B-cell clones, particularly as the B-ALL clonal BCRs are typically unmutated. To determine the false positive-rate for B-ALL BCR sequence detection, MRDARCY was used to detect B-ALL BCR sequences from the 6 B-ALL patients in 13 unrelated healthy BCR sequencing datasets using the same parameters (Table 5.2). A total of 23,480,661 BCR sequences were tested from unrelated samples, with only a single BCR match to a B-ALL cluster in B-ALL patient 5. This sequence was unmutated with short non-template additions (4bp) with 100% identity to a minor BCR clone in the B-ALL cluster (observed 219 times on the B-ALL patient). Therefore the presence of unrelated sequences matching the B-ALL-specific BCR sequence by chance occurs at a rate of 1 in $2x10^7$ BCR sequences/cells.

Table 5.2. False positive rate for detecting B-ALL MRD.

B-ALL patient	Number of unrelated healthy BCRs tested against B-ALL cluster	Number of reads matched*
B-ALL 1	3,730,269	0
B-ALL 2	4,098,690	0
B-ALL 3	4,097,093	0
B-ALL 4	3,836,054	0
B-ALL 5	3,922,068	1
B-ALL 6	3,796,487	0
Total	23,480,661	1

* By matching of B-ALL BCR sequences in 13 unrelated healthy sample BCR datasets.

5.2.4. Detecting B-ALL BCRs in clinical samples

Having shown the sensitivity of BCR sequencing, it was hypothesised that B-ALL clonal sequences will be detected in all the samples that were defined as qPCR T/C ratio MRD positive. Therefore, for each B-ALL patient, MRDARCY was used to identify BCR sequences in the largest cluster in the primary qPCR positive samples (highlighted in Table 5.1) and the percentage of matched BCR sequences in longitudinal samples was determined (allowing a maximum of 8 bp mismatches, **Figure 5.3**). Each of the six patients' samples showed a strong correlation between the fusion qPCR transcript levels (blue lines, **Figure 5.3**) and the frequencies of B-ALL sequences related to the largest cluster, known as clonotypic sequences (red lines, **Figure 5.3**), where the Pearson product-moment correlation coefficients between the percentage of B-ALL BCRs matched per sample and T/C ratios are >0.87 (Table 5.3). All samples that were qPCR positive were also positive for B-ALL BCR sequences. As the BCR sequencing sensitivity for detecting BCR sequences in RNA is greater than 1 in 107 and the qPCR result was very low, the lack of detection of B-ALL in patient 859 day 84 is likely to be due to the lack of sampling a B-ALL cell in the BM RNA aliquot used for PCR rather than failure of RNA detection. To determine whether detecting low-level B-ALL sequences is subject to sampling stochasticity, the low-level B-ALL RNA samples were re-amplified and re-sequenced (Table 5.4). Detection of B-ALL sequences were reproducible in samples where the number of B-ALL matched sequences was greater than 0.0016% of the BCR repertoire confirming that MRD above this level can be reliably detected using BCR sequencing. However, below this level detection of very low-level B-ALL sequences was subject to sampling stochasticity. Furthermore, some patient samples were positive for B-ALL BCR sequences where MRD was undetected using qPCR, such as patient 527 (day 15), indicating that the sensitivity of the BCR sequencing method equal to or better than that of qPCR, with the additional advantage that BCR sequencing based MRD monitoring can be done without gene fusion knowledge.

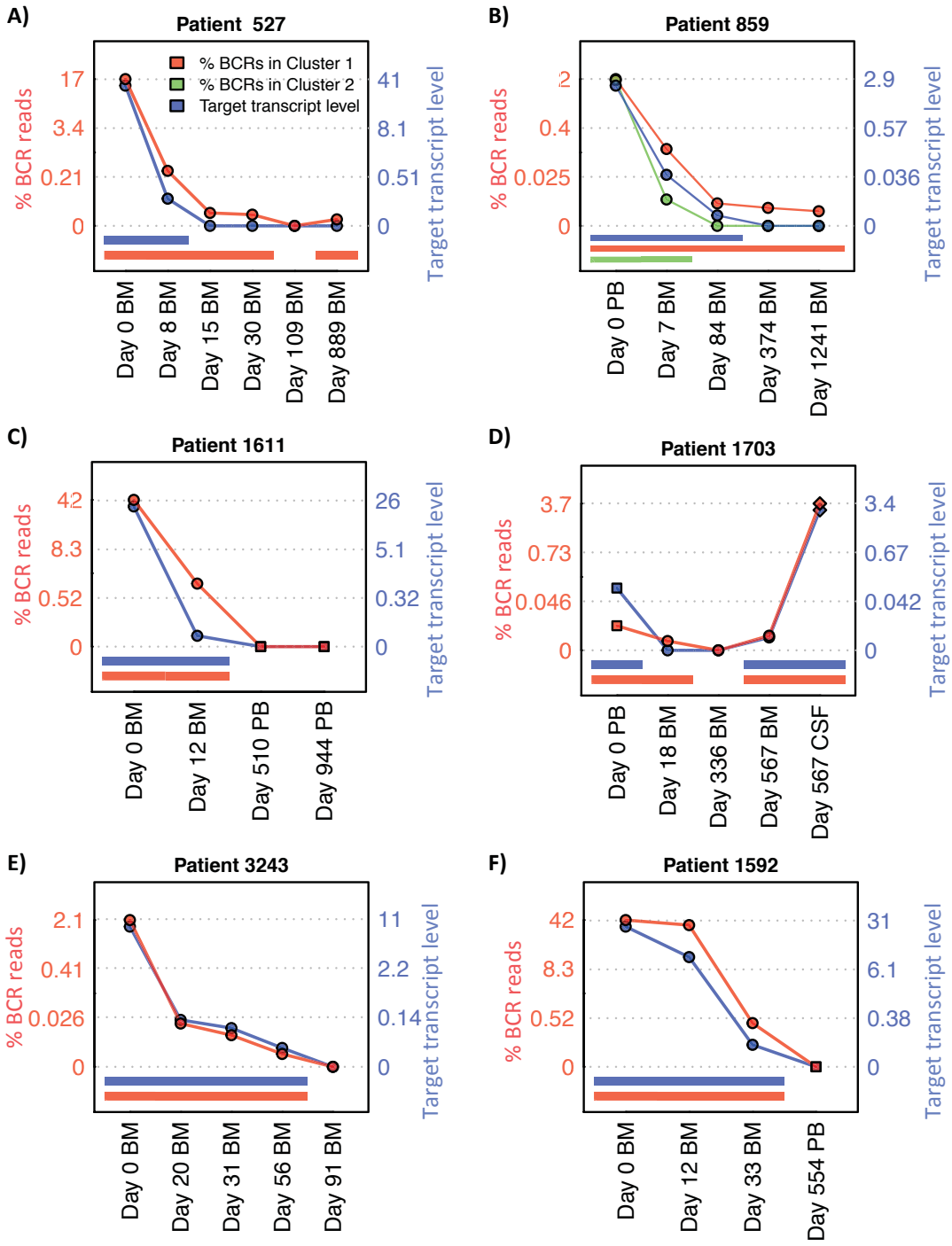

A)
Patient 527
% BCRs in Cluster 1
% BCRs in Cluster 2
Target transcript level
% BCR reads
17
3.4
0.21
0
41
8.1
0.51
0
Target transcript level
Day 0 BM
Day 8 BM
Day 15 BM
Day 30 BM
Day 109 BM
Day 889 BM
B)
Patient 859
% BCR reads
2
0.4
0.025
0
2.9
0.57
0.036
0
Target transcript level
Day 0 PB
Day 7 BM
Day 84 BM
Day 374 BM
Day 1241 BM
C)
Patient 1611
% BCR reads
42
8.3
0.52
0
26
5.1
0.32
0
Target transcript level
Day 0 BM
Day 12 BM
Day 510 PB
Day 944 PB
D)
Patient 1703
% BCR reads
3.7
0.73
0.046
0
3.4
0.67
0.042
0
Target transcript level
Day 0 PB
Day 18 BM
Day 336 BM
Day 567 BM
Day 567 CSF
E)
Patient 3243
% BCR reads
2.1
0.41
0.026
0
11
2.2
0.14
0
Target transcript level
Day 0 BM
Day 20 BM
Day 31 BM
Day 56 BM
Day 91 BM
F)
Patient 1592
% BCR reads
42
8.3
0.52
0
31
6.1
0.38
0
Target transcript level
Day 0 BM
Day 12 BM
Day 33 BM
Day 554 PB

Figure 5.3. B-ALL BCR populations.

Variation of T/C qPCR transcript ratios (blue) and percentage of clonotypic B-ALL BCR reads over time for each patient (red and green for largest and second largest clusters respectively). The blue axis on the right of each plot corresponds to the T/C qPCR transcript ratios levels and the red axis on the left of each plot refers to the percentage of sequences in the corresponding clusters, both of which have a square scale to highlight lower frequency observations. Blue and red bars under each plot indicate time-points that are positive for B-ALL transcripts and B-ALL BCR reads respectively.

Table 5.3. Correlations between the percentage of B-ALL BCRs matched and qPCR levels.

Patient ID	Linear gradient between % BCRs matched and T/C ratio*	R^2-value*
527	0.3384	0.9997
859	0.5917	0.9997
1611	1.3216	0.9988
1703	0.9229	0.9986
3243	0.1627	1.0000
1592	0.8312	0.8782

* Linear gradients and Pearson product-moment correlation coefficients (R^2-values) between the percentage of B-ALL BCRs matched per sample and qPCR target to control transcript (T/C) ratios.

Table 5.4. Percentages of B-ALL clonotypic BCR sequences in repeated samples.

			BCR sequencing (initial sample)*	BCR sequencing (re-amplified)**
Patient ID	**qPCR T/C level**	**Time since first sample (days)**	**% of B-ALL sequences**	**% of B-ALL sequences**
527	13.9510	0	41.21494	-
527	0.0197	8	0.81457	-
527	0.0000	15	0.00249	0.00056
527	0.0000	30	0.00140	0.00000
527	0.0000	109	0.00000	0.00000
527	0.0000	889	0.00016	0.00000
859	1.6612	0	2.89096	-
859	0.0292	7	0.21739	0.18325
859	0.0001	84	0.00159	0.00028
859	0.0000	374	0.00065	0.00032
859	0.0000	1241	0.00029	0.00031
1592	34.6048	0	31.45017	-
1592	12.9828	12	27.33152	-
1592	0.0211	33	0.24774	-
1592	0.0000	554	0.00000	-
1611	35.0403	0	26.48259	-
1611	0.0013	12	0.90122	0.12890
1611	0.0000	19	0.06560	0.00000
1611	0.0000	33	0.00000	0.00000
1611	0.0000	510	0.00000	-
1611	0.0000	944	0.00000	-
1703	0.1211	0	0.00266	0.00329
1703	0.0000	18	0.00005	0.00000
1703	0.0000	336	0.00000	0.00000
1703	0.0002	567	0.00033	0.00148
1703	3.1218	567	3.38261	-
3243	1.7453	0	10.73040	-
3243	0.0219	20	0.08141	-
3243	0.0102	31	0.02319	-
3243	0.0006	56	0.00063	-
3243	0.0000	91	0.00000	-

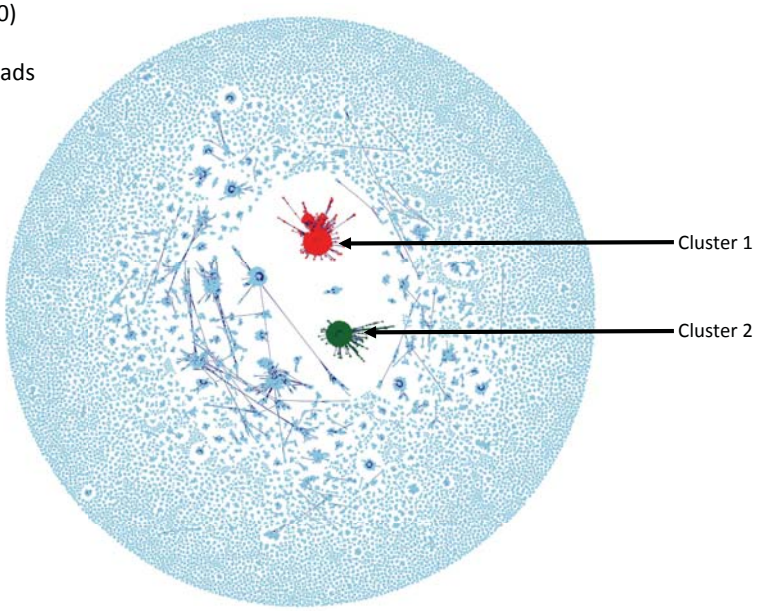
* The initial BCR sequencing dataset.

** Where the RNA was re-amplified and sequenced independently.

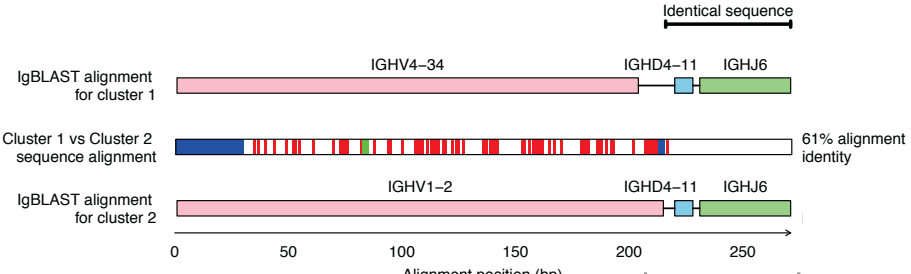
For patient 859, the two largest clusters had similar sizes (2.807% and 2.891% of total reads) corresponding to IgHV gene rearrangements of [IgHV4-34, IgHD4-11, IgHJ6] and [IgHV1-2, IgHD4-11, IgHJ6] (**Figure 5.3**E, red and green lines respectively). The identical IgHD-IgHJ gene usage may be indicative of IgH secondary rearrangements. Although the second largest cluster (indicated in green) became undetectable after day 7, the largest cluster (indicated in red) was never fully eradicated over the 1241 days of sampling. Ongoing IgHV rearrangements have been shown to occur as the result of either of two processes. Firstly an ancestral B-ALL clone may undergo partial *IgH* gene rearrangement firstly of the IgHD-J genes, with multiple B-cells in this clone able to recombine the IgHD-J with different IgHV segments to become fully rearranged, thus generating multiple IgHV-D-J combinations sharing the same IgHD-J region. Secondly, in a secondary rearrangement, an existing IgHV in a full IgHV-D-J rearrangement may be exchanged for a 5' germline IgHV while retaining the same IgHD-J region (Marshall et al., 1995, Steenbergen et al., 1993, Gawad et al., 2012, Choi et al., 1996, Liu et al., 2013). Therefore, to assess whether these clusters may have originated from secondary rearrangements of a single ancestral BCR, the most frequently observed BCR sequence from both clusters were aligned to each other (**Figure 5.4**B). Although there is only 61% alignment identity between the two BCR sequences representing the two clusters, the 55 nucleotides spanning the IgHD-IgHJ region and, notably, 3pb of the 3' end of the IgHV gene in the cluster 2 BCR sequence is identical to IgHV-D joining region (consisting of random nucleotide additions during *IgH* gene rearrangement) were identical, which is consistent with the hypothesis of secondary rearrangements. In addition, these BCR sequences show no mutations in the IgHV genes compared to the reference germline database, thus reinforcing the hypothesis that these two clonal B-ALL BCRs are indeed from the same progenitor B-ALL B-cells from early stages of B-cell differentiation that have not undergone SHM but where a secondary rearrangement of the IgHV has occurred. This could potentially be determined through the sequencing of the light chain BCR sequences.

A) Patient 859 (Day 0)

454,071 total BCR reads



B)



C)

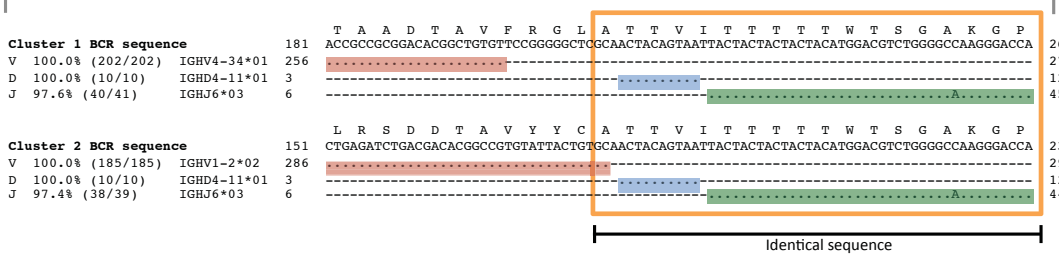


Figure 5.4. Bi-clonal B-cell expansion in B-ALL patient 859.

A) Network diagram for B-ALL patient 859 at day 0, where vertices within the largest cluster (Cluster 1) are coloured in red and vertices within the second largest cluster (Cluster 2) are coloured in green, otherwise vertices coloured in blue. **B)** BCR sequence alignment of the dominant sequences from the two dominant clusters in patient 859. Cluster 1 and cluster 2 refer to the largest and second largest clusters in the BCR sequence network for patient 859 respectively (representing 2.81% and 2.89% of BCRs respectively). The cluster 1 and 2 sequences were aligned to each other, and the positions of differences between sequences are indicated by the coloured boxes in the corresponding positions in the middle row, using red for mismatches, green for gaps in cluster 1 BCR and blue for gaps in cluster 2 BCR. The percentage identities of each alignment are indicated at the right of each sequence depiction. The cluster 1 and 2 sequences had 100% alignment identity with IgHV gene rearrangements of [IgHV4-34, IgHD4-11, IgHJ6] and [IgHV1-2, IgHD4-11, IgHJ6] respectively, where the red, blue and green boxes for IgHV, D and J genes mark the gene boundaries respectively. **C)** Alignments of Cluster 1 and cluster 2 BCR sequences with closest reference IgHV (highlighted in red), IgHD (highlighted in blue) and IgHJ (highlighted in green) genes, where . denotes alignment similarity between the cluster sequences and the reference genes, A/T/G/C denotes a different base to the reference, and - denotes the region outside of the gene alignments. The 55pg region of the BCR sequence that is identical between the cluster 1 and cluster 2 sequences is highlight in the orange box and yellow text.

In addition, these clusters display similar properties, including the mean distance from the most frequently observed BCR within each cluster (2.281bp and 2.135bp for clusters 1 and 2 respectively, Table 5.5). However, MRD was observed only for cluster 1 throughout the 1241 days of sampling, suggesting that these clusters were differentially affected by therapy. Therefore, BCR sequencing can detect multiple disease subclones irrespective of their composition of driver mutations and individual proliferative properties.

Table 5.5. Table of the properties of the largest two clusters in patient 859

	Cluster 1	Cluster 2
Cluster size (% of total sequences)	2.469	2.379
N reads	11211	10801
Number of unique sequences in cluster	2858	2037
IgHV gene	IGHV4-34*01	IGHV1-2*02
IgHJ gene	IGHJ6*03	IGHJ6*03
Number of sequences representing most frequently observed BCR	5625	6603
Mean distance from most frequently observed BCR	2.281	2.135

5.2.5. Detecting B-ALL BCRs in RNA and DNA

The BCR RNA expression in mature B-cells is greater than that of pre-B-cells or immature B-cells (Hoffmann et al., 2002). To account for the possibility that B-cell receptor expression in B-ALL cells/samples may be lower than in non-malignant mature B-cells, which may lead to the under-estimation of the number of malignant B-cells in a given sample, the DNA and RNA BCR repertoires were compared in three patient samples (**Figure 5.5**). For every patient time point, B-ALL-derived BCR sequences were detected in the DNA sample at a higher percentage of total BCR sequences compared to the percentage derived from studying the matched RNA sample. Therefore, although BCR sequencing is highly sensitive for the detection of B-ALL-derived sequences, the RNA BCR repertoire may be significantly underestimating the true percentage of B-ALL cells in the sample and the use of DNA repertoires in B-ALL may further increase the sensitivity for MRD detection. However, DNA is more stable in plasma than RNA so detection of plasma DNA may

be more indicative of lysed or dead cells, whereas plasma RNA is more readily degraded (El-Hefnawy et al., 2004, Garcia-Olmo et al., 2013). As the difference is very striking between the RNA and DNA clonotype frequencies in B-ALL samples, DNA BCR sequencing should be used as an MRD marker rather than RNA.

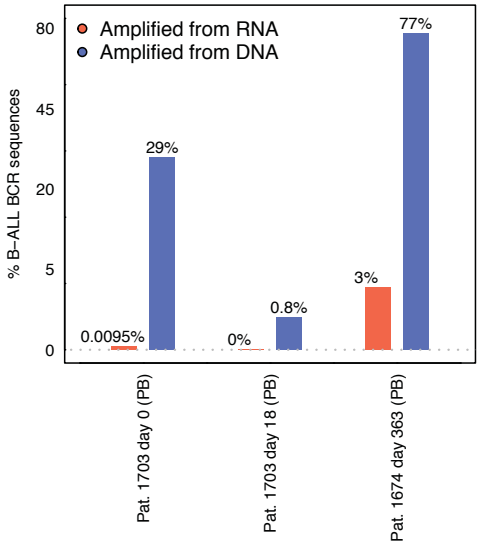


Figure 5.5. Detection of B-ALL BCR sequences in RNA and DNA samples.

Bar-graph showing the percentages of B-ALL sequences from BCR datasets generated from either the RNA or DNA from B-ALL patients (red and blue bars respectively).

5.2.6. Distinguishing between B-ALL and healthy samples

Increased clonality is observed in B-ALL samples with high levels of leukaemic load (i.e. when the qPCR T/C transcript ratio is greater than 1.66, Table 5.1). However, it is possible that the B-cell populations in B-ALL patients after therapy would still be distinct from healthy B-cell populations. If so, features of the B-cell repertoire would distinguish between B-ALL patient samples with high leukaemic loads (B-ALL high, T/C qPCR transcript ratio>1), B-ALL patient samples with low levels of leukaemic loads (B-ALL low, T/C qPCR transcript ratio<1), B-ALL patient samples with undetectable MRD after therapy (B-ALL undetectable, T/C qPCR transcript ratio=0) and healthy B-cell samples. Therefore, for each B-ALL sample and the 18 healthy individual samples, nine features of the B-cell sequencing data were calculated to distinguish between these different sample types, namely:

(a) *The vertex and cluster Gini index:* measurements of overall clonality and cluster size heterogeneity respectively.

(b) *The largest cluster size (as a percentage)*: to distinguish between samples with different maximum cluster sizes.

(c) *The sum of the largest two cluster sizes (as a percentage):* measurement to incorporate the second largest cluster size, which may distinguish between samples with secondary rearrangements.

(d) *The percentage of unique BCRs in largest cluster*: to distinguish between samples with different levels of SHM in the largest cluster.

(e) *The percentage of sequences representing the most frequently observed BCR sequence*: to distinguish between samples with or without dominant BCR sequences.

(f) The percentage of sequences representing the first and second most frequently observed IgHV-J rearrangement: measurement to distinguish between samples with specific rearrangements, irrespective of the largest cluster sizes.

(g) *The ratio of the number of unique CDR3 sequences to unique full length BCR sequences*: as the CDR3 length is shorter than the full length BCR sequence, but B-cells sharing the same CDR3 sequence are likely to originate from a single pre-B-all precursor, then lower ratios of unique

CDR3 sequences to unique full length BCR sequences suggests lower B-cell clonal complexity.

Each of these features describes different aspects of the B-cell repertoire. Linear discriminant analysis (LDA) was performed to find a linear combination of features that best separates sample types (**Figure 5.6**A) (Rindskopf, 1997). The first LDA dimension (LDA 1) separates the B-ALL high samples from the B-ALL low, B-ALL undetectable and healthy samples. The features that contribute most to distinguishing between these sample groups are the largest cluster size (contribution: -219.4), the sum of the largest two cluster sizes (denoted 1st + 2nd largest cluster (%), contribution: 302.8), the percentage of sequences corresponding the most frequently observed BCR sequence (denoted max. vertex, contribution: -189.0), and the percentage of sequences corresponding the first and second most frequently observed IgHV-J rearrangement (denoted Max. VJ Gene freq and 2nd max. VJ Gene freq contribution respectively: contributions of 204.7 and -165.4). The second LDA dimension (LDA 2) separates the healthy samples from the B-ALL low/undetectable samples. The features that contribute most to distinguishing between these sample groups are the vertex Gini index (contribution: 517.2), and cluster Gini index (contribution: -293.7), as indicated by the highest magnitude of the corresponding variable contributions for LDA2 (**Figure 5.6**B). Therefore, two-dimensional LDA successfully distinguishes B-ALL high, B-ALL low/undetectable and healthy samples.

To test whether the resulting LDA 1 and LDA2 linear combinations can be used as a linear classifier of sample type, hierarchical clustering was performed using the Euclidean distances between the LDA 1 and LDA 2 coordinates of each sample (as defined in Section 2.2.11, **Figure 5.6**C). This shows clear separation of B-ALL high samples (branch A, **Figure 5.6**C) from healthy samples (branch B, **Figure 5.6**C). The B-ALL low/undetectable samples were indistinguishable by these methods, but, interestingly, distinct from the other two groups (branch C, **Figure 5.6**C). 2 out of 18 healthy samples were misclassified into branch C, indicating that some healthy individuals may exhibit a range of B-cell repertoire features that can overlap with that of B-ALL low/undetectable.

These data show that patient B-ALL B-cell repertoires differ significantly from those of healthy individuals during maximum tumour burden, which is

unsurprising. Notably however, patient B-ALL B-cell repertoires during and after maximum tumour removal by therapy differ significantly from those of healthy individuals. Such a difference of low or undetectable B-ALL BCR repertoires may represent an effect of the prior presence of a large B-ALL clone or an effect of anti-leukaemic therapy, as patients remain on maintenance treatment for 2-3 years after diagnosis including lymphotoxic drugs such as corticosteroids and antimetabolites (e.g. Methotrexate). Overall, two-dimensional LDA in BCR sequencing repertoires successfully distinguished between B-ALL high samples, B-ALL low/undetectable samples and healthy samples and can effectively classify such samples. Whether subsets of the B-ALL low/undetectable clusters of patients, perhaps those without a "healthy" cluster member (branch C', **Figure 5.6**C), are more likely to relapse would be interesting to pursue. Alternatively, these "healthy" individuals that co-cluster with the B-ALL patients may have more clonal features of their B-cell repertoires for reasons that are unclear.

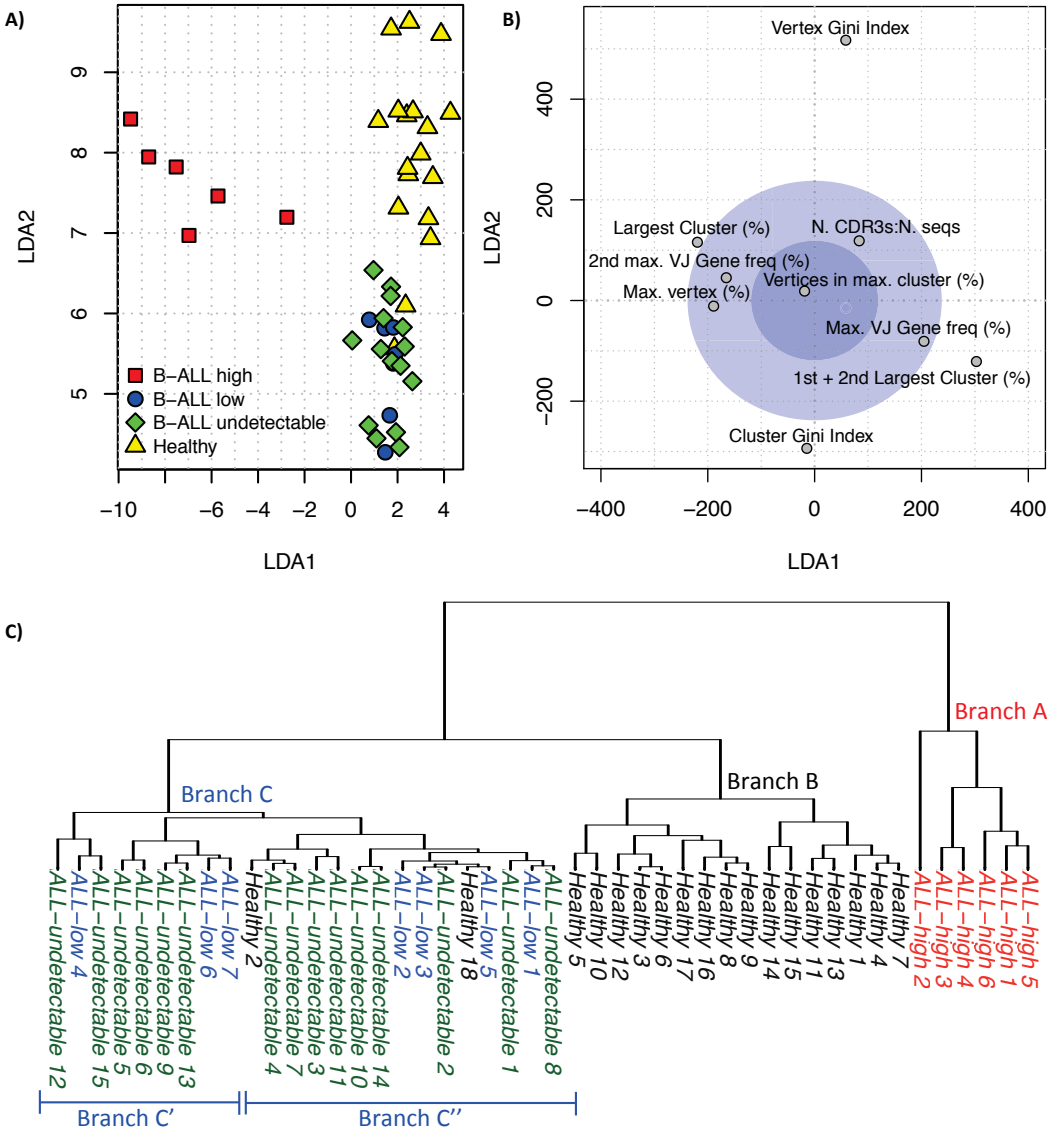

A)
LDA2
LDA1
B–ALL high
B–ALL low
B–ALL undetectable
Healthy
B)
Vertex Gini Index
Largest Cluster (%)
N. CDR3s:N. seqs
2nd max. VJ Gene freq (%)
Vertices in max. cluster (%)
Max. vertex (%)
Max. VJ Gene freq (%)
1st + 2nd Largest Cluster (%)
Cluster Gini Index
C)
Branch A
Branch B
Branch C
Branch C'
Branch C''

Figure 5.6. Distinguishing between B-ALL and healthy B-cell populations.

A) Linear discriminant analysis (LDA) performed on all samples to differentiate between diagnostic B-ALL samples (B-ALL-high, red (T/C qPCR transcript ratio>1)), B-ALL samples after treatment with detectable MRD (B-ALL-low, blue (T/C qPCR transcript ratio<1)), B-ALL samples with undetectable MRD (B-ALL-undetectable, green (T/C qPCR transcript ratio=0)), and healthy individuals (yellow). **B)** LDA variable contributions for the first two dimensions, where the blue circle indicates the mean scalar variable contributions for the first two dimensions, and variables outside this region indicate greatest contribution to the separation of classes. **C)** Hierarchical clustering tree of the patient samples using the distance measures derived from LDA 1 and 2. Branches (A), (B) and (C) refer to branches of the hierarchical tree corresponding to B-ALL-high, healthy and B-ALL-low/undetectable samples respectively, and where branches C' and C'' are sub-branches in branch C.

5.2.7. ALL Relapse: a case study of CSF relapse

One of the patients in this cohort, patient 1703, unfortunately developed CSF relapse after more than 2 years from initial therapy (summarised in). The sample taken on day 0 was taken more than one week after therapy had started and likely after a significant reduction in disease bulk, therefore the B-ALL qPCR T/C transcript level was relatively low. B-ALL was undetectable by day 18 (by qPCR MRD monitoring), but re-emerged at day 567 predominantly in the CSF, although it was also detectable in the BM. In this patient, both the B-cell were amplified and sequenced to understand the adaptive immune dynamics of relapse in B-ALL.

Table 5.6. Detection of B-ALL cells in patient 859.

Source*	Target/control transcript level**	% B-ALL BCR reads (from RNA)	% B-ALL BCR reads (from DNA)
Day 0, PB	0.121	0.00266	28.63019
Day 18, BM	0	5.42E-05	0.804093
Day 336, BM	0	0	-
Day 567, BM	0.000222	0.000332	-
Day 567, CSF	3.122	3.38	-

* Abbreviations: BM is bone marrow, PB is peripheral blood and CSF is cerebrospinal fluid.
** Target transcript: TEL/AML1 translocation.

The largest clone in the patient 1703 day 0 DNA sample, representing 28.63% of all BCR sequences, was identified as the B-ALL clone. This clone was detected as the largest cluster in the day 567 CSF sample (from RNA), representing 3.38% of BCR sequences (Table 5.6). However, 80% of cells in this sample resembled the leukaemia-associated immunophenotype of lymphoblasts ($CD10^{+}$, $CD19^{+}$, $CD45^{low/-}$) by flow cytometry. The reason for a low representation of B-ALL BCRs in the RNA sample compared to flow-cytometry is unclear but could be explained by the lower expression of immunoglobulin in B-ALL cells compared to mature B-cells (addressed in Section 5.2.5).

Clonal evolution has been observed in B-ALL as exemplified by the presence of tumour mutations in the genome (Mullighan, 2012) and by multiple BCRs related to the dominant B-ALL BCR sequence. Although expression of AID, which is

required for somatic hypermutation, has been detected only in some B-ALL patients (Feldhahn et al., 2007, Messina et al., 2011, Iacobucci et al., 2010, Hardianti et al., 2005), the accumulation of non-AID-mediated or mutations caused by low-level AID expression in these cells can result in clonal diversification in B-ALL (Jiao et al., 2014). These mutations may be used to infer the mutational route from a B-ALL B-cell ancestor to the rest of the leukaemic clone by phylogenetic analysis. To infer the phylogenetic relationships between B-ALL sequences before and after relapse, all the BCR sequences related to the B-ALL clone at day 0 derived by combining both RNA and DNA sequencing datasets and day 567 relapse (from RNA sequencing dataset) were identified (including identical or related BCRs within a threshold of 8 bp of the using MRDARCY) and 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. Bootstrapping was performed to evaluate the reproducibility of the trees, showing strong tree support (>95% certainty for all branches), and the tree tips were coloured according to whether the BCRs were observed at day 0 (BM) and/or day 567 (CSF) (**Figure 5.7**A). The tree has a star-like structure, suggesting that the original B-ALL BCR clone emerged from a single common ancestor (Martins and Housworth, 2002), represented by the central BCR, which was the most frequently observed BCR at day 0 (BM) (making up 40.0% and 74.6% of total related B-ALL sequences for BCR repertoires derived from RNA and DNA respectively) and day 567 (CSF) (40.0% and 63.0% of total related B-ALL sequences for BM and CSF respectively). Interestingly, there was high BCR sequence overlap between the day 0 (BM) and day 567 (CSF) samples (86.08%), even at distances of 7 nucleotides from the central BCR (**Figure 5.7**B). Furthermore, there is a strong linear correlation between the B-ALL BCR frequencies the day 0 (BM) and day 567 (CSF) samples (R^2-value=0.9993 for all BCRs), suggesting that some of the population structure of this B-ALL cluster is