Chapter 4

4. Comparison of BCR amplification and sequencing methods

4.1. Introduction

For immune repertoire sequencing to be useful, it is therefore vital that sample preparation and sequencing approaches give reproducible, unbiased and sensitive representations of BCR repertoires. However, there is concern over the validity and biases of biological insights gained from the different BCR and TCR enrichment, amplification and sequencing methods, particularly whether the sequencing data truly represents the corresponding B-cell populations. As the B-cell receptor is highly diversified, there is potential for some immunoglobulin rearrangements to be preferentially captured and amplified, leading to biased sequencing data.

This chapter integrates both the theoretical and experimental frameworks for BCR sequencing to determine whether the B-cell sequencing data represents that expected theoretically. Then, the utilities, biases and reproducibilities of different sequencing depths, sequencing technologies, amplification methods, read lengths and starting material are assessed using samples of diverse B-cell populations from healthy peripheral blood (PB), clonal B-cell populations from lymphoblastoid cell lines (LCL) and PB from chronic lymphocytic leukaemic (CLL) patients.

4.2. Results

4.2.1. Generation of BCR sequencing datasets for comparative studies

Experimental BCR sequencing datasets were generated through the amplification of LCLs and PB B-cell BCRs from healthy individuals, and CLL patients by the three main BCR amplification methods; multiplex PCR, 5' Rapid amplification of cDNA ends (5'RACE), and RNA-capture, and sequenced by 454 Roche and Illumina MiSeq (summarised in Table 4.1). Each sample generated an average of 40,763 reads (summarised in Table 4.2). For each sample, reads were filtered for immunoglobulin similarity and length, and, where relevant, primer sequences were removed according to Methods (Section 2.2.4). IgHV classifications were performed on each BCR sequence by determining the best alignment to the ImMunoGeneTics (IMGT) database (Lefranc et al., 2009) using BLAST (Altschul et

al., 1990). For each sample, the reference IgHV gene frequencies and clonality measures developed in Chapter 3, namely the vertex and cluster Gini indices and maximum cluster sizes, were determined for the comparisons in this chapter. These diversity measures correspond to that seen in equivalent sample types in previous studies (Table 4.3, (Bashford-Rogers et al., 2013)).

Table 4.1. Samples used in this study for each amplification method.

Sample type*	ID	Multiplex (454)	Multiplex (MiSeq)	5' RACE (MiSeq)	RNA capture (MiSeq)
CLL	Sample 1	Y	Y	Y	Y
CLL	Sample 2	Y	Y	Y	-
CLL	Sample 3	Y	Y	-	-
CLL	Sample 4	Y	Y	Y	-
CLL	Sample 5	Y	Y	Y	-
CLL	Sample 6	Y	Y	Y	-
CLL	Sample 7	Y	Y	Y	-
CLL	Sample 8	Y	Y	Y	-
Healthy	Sample A	Y	-	Y	-
Healthy	Sample B	Y	-	Y	-
Healthy	Sample C	Y	Y	Y	-
Healthy	Sample D	Y	-	Y	-
Healthy	Sample E	Y	Y	Y	-
Healthy	Sample F	Y	Y	-	-
Healthy	Sample G	Y	Y	-	-
Healthy	Sample H	Y	Y	-	-
Healthy	Sample I	Y	Y	-	Y
LCL	LCL 1	Y	-	-	-
LCL	LCL 2	Y	-	-	-
LCL	LCL 3	Y	-	-	-
LCL	LCL 4	Y	-	-	-
LCL	LCL 5	Y	-	-	-
LCL	LCL 6	Y	-	-	-
LCL	LCL 7	Y	-	-	-
LCL	LCL 8	Y	-	-	-
LCL	LCL 9	Y	-	-	-
LCL	LCL 10	Y	-	-	-

* Abbreviations: CLL = chronic lymphocytic leukaemia, healthy = PBMC from healthy blood donor, LCL = human lymphoblastoid cell line.

Table 4.2. Mean and standard deviation of read depths per sample.

Technology	Mean read depth per sample (after filtering)	Average number of multiplexed samples per lane/run	Average % BCR sequences after filtering*
Multiplex PCR (454)	33,413	12	76.10
Multiplex PCR (MiSeq)	31,118	50	60.30
5' RACE (MiSeq)	72,586	95	55.09
RNA capture (MiSeq)	58,015	2	1.53**

* Percentage of reads after filtering for open reading frames rearranged BCR sequences from the whole read set.
** RNA capture has lower percentage of filtered reads due to the designed simultaneous capture of immunoglobulin heavy and light chains as well as T-cell receptors.

Table 4.3. Mean diversity measures for each sample type.

Sample type	Mean maximum cluster size (% of total BCR sequences)	Mean vertex Gini Index	Mean cluster Gini Index
Healthy	0.581	0.182	0.047
Chronic lymphocytic leukaemia (CLL)	95.117	0.931	0.612
Human lymphoblastoid cell line (LCL)	65.205	0.934	0.790

4.2.2. Theoretical framework for sampling and sequencing BCR repertoires

As exhaustive sampling of total B-cells is not possible in humans, the "true" extent of the total BCR repertoire in humans can only be estimated. To understand the BCR sequencing data properly, it is first important to estimate the types of B-cells sampled, and the proportion of total B-cells from a patient in each sample. This will give a theoretical estimate for the expected percentage of BCRs to be shared between technical repeats and the expected sampling stochasticity in any given BCR sample, which will then be tested experimentally.

A typical PB sample (10-20ml) accounts for ~0.4% of the total PB (average of 5L of blood in a healthy adult), from which only a fraction is used in current BCR sequencing methods with approximately 0.012% of all B-cells being represented in the material that is sequenced, Table 4.4. The healthy peripheral blood B-cell population contains approximately 80% naïve B-cells and 20% memory B-cells (Tangye and Good, 2007). As naïve B-cells are antigen inexperienced, each naïve B-cell BCR is often considered to be unique. This means that sequencing BCRs from only naïve B-cells theoretically result in a diverse BCR population, with all BCRs represented with equal probability. Therefore the distribution of BCR frequencies should follow approximately a binomial distribution, where parameters depend on the number of RNA molecules per cell, the number of B-cells represented and efficiencies of the RT-PCR, PCR and sequencing steps. Under the assumption that each naïve B-cell is unique, it therefore is theoretically impossible to resample identical BCRs from a population of naïve cells. The memory B-cell population, however, consists of cells that have undergone proliferation and potentially somatic hypermutation. This means that it is possible to sample multiple memory B-cells exhibiting identical BCR sequences or highly related BCRs after somatic mutation that originate from the same pre-B-cell.

Table 4.4. Estimation of number and percentage of sampled peripheral blood B-cells.

	Number of cells	% of total B-cell repertoire	Notes
Total B-cells in blood*	1,500,000,000	100	Average adult has 5L of blood
B-cells in sample	3,000,000	0.4	10 ml blood sample
RNA extraction	3,000,000	0.4	Assume 100 % efficiency
RT-PCR**	300,000	0.04	Average 10ug RNA extracted per sample, 1000ng RNA used in RT-PCR
PCR	90,000	0.012	6ul out of 20ul of RT-product used in PCR

* (www.stemcell.com)

** Average of 10ug RNA extracted per healthy PB sample.

To achieve a theoretical estimation of the percentage of BCRs that should overlap between any two samples from the same peripheral blood aliquot from a single individual, simulations were generated as follows: the total B-cells in the sample is *N*, from which each RT-PCR sample taken contains *n* B-cells equivalent. The proportion of memory B-cells in the peripheral blood is given by p_m, and the proportion of naïve B-cells is *1-* p_m (assuming no plasma B-cell in the blood). Therefore the number of memory B-cells in the total population is $N^* p_m$. The number of memory B-cells per unique BCR can be modelled as a normal distribution $N(\mu, \sigma)$, with a mean μ and standard deviation σ. Each simulation draws a random sample of $N^* p_m$ BCRs where the probability of resampling a specific BCR follows $N(\mu, \sigma)$. From this, two random samples from this simulated total B-cell populations are drawn, and the percentage overlap between the samples is determined.

Three such simulations were generated, where the parameters are summarised in Table 4.5. All three simulations begin with the estimation of total and sampled B-cells from Table 4.4 and **Figure 4.1**. Simulations A and B assume that the proportion of memory B-cells is 20%, whereas simulation C models the proportion of memory B-cells as a normal distribution with mean of 20% and standard deviation of 8% to reflect inter-individual differences in the memory-to-naïve B-cell ratios (Tangye and Good, 2007). The percentage of overlapping BCRs between samples in simulations A, B and C determined for 1000 simulation repeats was 6.185%, 18.29%, and 19.71% respectively (**Figure 4.2**, blue box plots). The lower overlap in simulation A is

explained by the lower number of memory B-cells per BCR, therefore a lower probability of re-sampling B-cells with the same BCR. The higher variance of overlapping BCR percentages in simulation C is explained by the higher variance of percentage of memory B-cells in the PB.

Table 4.5. B-cell sampling simulation parameters.

Parameter	Simulation A	Simulation B	Simulation C
N	3000000	3000000	3000000
n	90000	90000	90000
p_m	0.2	0.2	*N(20, 8)*
μ	1	1	1
σ	10	100	100
Number of simulation repeats	1000	1000	1000

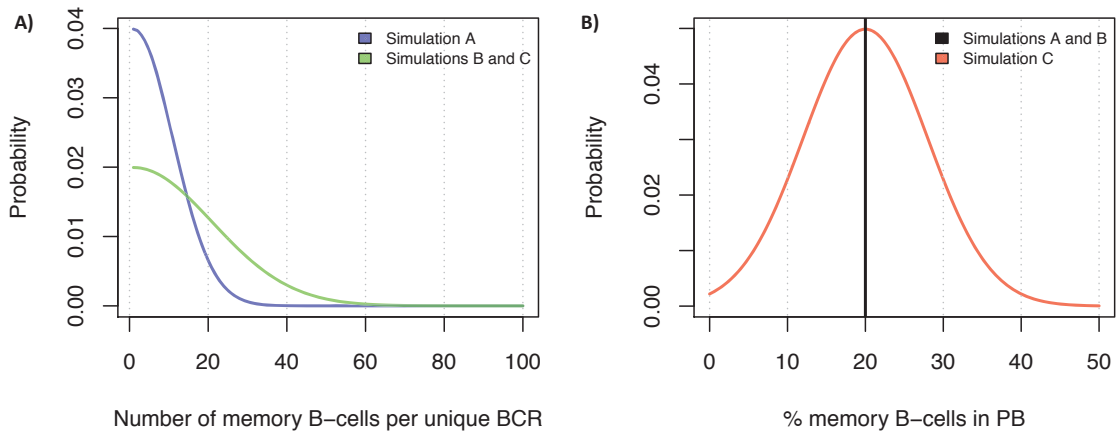

Figure 4.1. Simulation distributions.

Simulation distributions for **A)** the number of memory B-cells per unique BCR sequence (varying σ) and **B)** the percentage of memory B-cells in the peripheral blood (PB) (varying p_m). The colours of the distributions correspond to the simulations indicated in the key.