Chapter 2

2. Materials and methods

2.1. Samples

Peripheral blood mononuclear cells (PBMCs) were isolated from 10ml of whole blood from healthy volunteers and CLL patients using Ficoll gradients (GE Healthcare), summarised in Table 2.1. For the B-ALL peripheral blood samples, DNA and RNA extraction was performed by incubation with erythrolysis buffer for 15 minutes, centrifugation, discarding supernatant (repeated twice), and resuspension of cells in PBS. Total RNA was isolated using TRIzol® and purified using RNeasy Mini Kit (Qiagen) including on-column DNase digestion according to manufacturer's instructions. Total RNA was also isolated from $1x10^6$ cells from Human lymphoblastoid cell lines (LCLs) from the HapMap project (Frazer et al., 2007), where the number of passages was unknown. DNA extraction was isolated using TRIzol® and the MiniPrep kit (Qiagen) according to manufacturer's instructions. CLL and B-ALL samples were approved by the relevant institutional review boards and ethics committees (07/MRE05/44 and ΕΕΒΚ/ΕΠ/2014/15 respectively).

Table 2.1. Table of samples used.

Sample ID in chapter	Sample ID in chapter 4	Patient type	Age, years	Gender	Time since CLL	Genomic abnormality	Phenotype
CLL 1	-	CLL	77	Male	7	13q deletion	Unknown
CLL 2	-	CLL	58	Male	2	Stable trisomy 12	Atypical: CD23 negative
CLL 3	-	CLL	78	Male	1.5	No FISH performed	CD38+
CLL 4	-	CLL + HCC	77	Male	2.5	No FISH performed	Unknown
CLL 5	Pat. 2	CLL	59	Female	1.25	No abnormalities seen	Unknown
CLL 6	-	CLL	67	Male	2	11q deletion	Unknown
CLL7	-	CLL	69	Male	13	No FISH performed	Recurrent haemolysis in the past
CLL 8	-	CLL	64	Male	4.5	No FISH performed	Unknown
CLL 9	-	CLL	77	Male	5.25	No FISH performed	Unknown
CLL 10	-	CLL	81	Male	8	13q deletion	Lambda light chain restricted, CD38-
CLL 11	Pat. 6	CLL	81	Male	10	No FISH performed	Unknown
-	Pat. 1	CLL + prostate carcinoma	67	Male	8	No FISH performed	Unknown
-	Pat. 3	CLL	80	Female	7	13q deletion	Unknown
-	Pat. 4	CLL	82	Female	5	No FISH performed	Unknown
-	Pat. 5	CLL	82	Male	0.75	No FISH performed	Unknown
-	Pat. 7	CLL	72	Female	4.5	13q14.3 deletion	Unknown
-	Pat. 8	CLL	64	Male	9.5	13q14.3 deletion	Unknown
-	Pat. 9	CLL	71	Male	8	No FISH performed	Unknown
-	Pat. 10	CLL	80	Male	5	No FISH performed	Unknown
-	Pat. 11	CLL	56	Male	0.75	No FISH performed	Unknown
-	Pat. 12	CLL	81	Male	1	13q deletion	CD5 negative
-	Pat. 13	CLL	63	Male	1.8	No FISH performed	Unknown
Healthy 1	Healthy 1	Age matched control 1	74	Female	-	Mix of t(11;14), 11q deletion, 13q addition	Anaemia
Healthy 2	Healthy 2	Age matched control 2	62	Female	-	NA	NA
Healthy 3	Healthy 3	Age matched control 3	75	Female	-	NA	NA
Healthy 4	Healthy 4	Age matched control 4	67	Female	-	NA	NA
Healthy 5	Healthy 5	Age matched control 5	68	Female	-	NA	NA
Healthy 6	Healthy 6	Healthy 6	55	Male	-	NA	NA
Healthy 7	Healthy 7	Healthy 7	23	Male	-	NA	NA
Healthy 8	Healthy 8	Healthy 8	23	Male	-	NA	NA
Healthy 9	Healthy 9	Healthy 9	25	Male	-	NA	NA
Healthy 10	Healthy 10	Healthy 10	24	Female	-	NA	NA
Healthy 11	Healthy 11	Healthy 11	24	Female	-	NA	NA
Healthy 12	Healthy 12	Healthy 12	24	Female	-	NA	NA
Healthy 13	Healthy 13	Healthy 13	24	Female	-	NA	NA

2.2. B-cell methods

2.2.1. RT-PCR

RT-PCR reagents were purchased from Invitrogen. The FR1 and FR2 primer sets used (supplied by Sigma Aldrich) are described by Van Dongen *et al.* (van Dongen et al., 2003) and in Table 2.1. Reverse transcription was performed using 500ng of total RNA mixed with 1µl JH reverse primer (10µM), 1µl dNTPs (0.25mM), and RNase free water added to make a total volume of 11µl. This was incubated for 5 minutes at 65°C, and 4µl First strand buffer, 1µl DTT (0.1M), 1µl RNaseOUT™ Recombinant Ribonuclease Inhibitor and 1µl SuperScript™ III reverse transcriptase (200units/µl) was added. RT was performed at 50°C for 60 minutes before heat-inactivation at 70°C for 15 minutes (**Figure 2.1**). PCR amplification of cDNA (5µl of the RT product) was performed with the JH reverse primer and the FR1 or FR2 forward primer set pools (0.25 µM each), using 0.5µl Phusion® High-Fidelity DNA Polymerase (Finnzymes), 1µl dNTPs (0.25mM), 1µl DTT (0.25mM), per 50µl reaction. For multiplex PCR amplification of DNA, 30ng of DNA was mixed with the JH reverse primer and the FR1 forward primer set (0.25 µM each), using 0.5µl Phusion® High-Fidelity DNA Polymerase (Finnzymes), 1µl dNTPs (0.25mM), 1µl DTT (0.25mM), per 50µl reaction. The following PCR program was used: 3 minutes at 94°C, 35 cycles of 30 seconds at 94°C, 30 seconds at 60°C and 1 minute at 72 °C, with a final extension cycle of 7 minutes at 72 °C on an MJ Thermocycler.

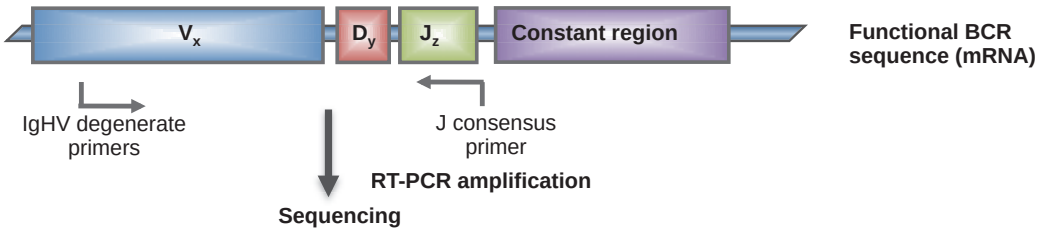

Figure 2.1. Sequencing of B-cell receptor repertoires.

Representation of the genomic rearrangement process during V-D-J recombination to generate the heavy chain B-cell receptor. B-cell receptor amplification was performed by reverse transcription on total RNA by single J region primer, and subsequent multiplex PCR amplification.

Table 2.1. Human B-cell receptor PCR primers.

Primer	Sequence	
JH reverse	CTTACCTGAGGAGACGGTGACC	
VH1-FR1 forward	GGCCTCAGTGAAGGTCTCCTGCAAG	
VH2-FR1 forward	GTCTGGTCCTACGCTGGTGAAACCC	
VH3-FR1 forward	CTGGGGGGTCCCTGAGACTCTCCTG	FR1 primer set*
VH4-FR1 forward	CTTCGGAGACCCTGTCCCTCACCTG	
VH5-FR1 forward	CGGGGAGTCTCTGAACATCTCCTGT	
VH6-FR1 forward	TCGCAGACCCTCTCACTCACCTGTG	
VH1-FR2 forward	CTGGGTGCGACAGGCCCCTGGACAA	
VH2-FR2 forward	TGGATCCGTCAGCCCCCAGGGAAGG	
VH3-FR2 forward	GGTCCGCCAGGCTCCAGGGAA	
VH4-FR2 forward	TGGATCCGCCAGCCCCCAGGGAAGG	FR2 primer set*
VH5-FR2 forward	GGGTGCGCCAGATGCCCGGGAAAGG	
VH6-FR2 forward	TGGATCAGGCAGTCCCCATCGAGAG	
VH7-FR2 forward	TTGGGTGCGACAGGCCCCTGGACAA	
B-actin forward	CGCCTTTGCCGATCCGCCG	
B-actin reverse	CTTCTCGCGGTTGGCCTTGGG	
GAPDH forward	GAAGGTGAAGGTCGGAGTC	
GAPDH reverse	GAAGATGGTGATGGGATTTC	
B-globin forward	CTGCCGTTACTGCCCTGTGGG	
B-globin reverse	GGACAGCAAGAAAGCGAGCTTAGTG	

* The FR1 primers were pooled to give the FR1 primer set, and the FR2 primers were pooled to give the FR2 primer set. Primer JH reverse was used to prime cDNA synthesis. The expected amplicon sizes for the IgHV PCR products for JH reverse/FR1 primer set is 310-360bp, and the expected size ranges for the IgHV PCR products for JH reverse/FR2 primer set is 260-295bp. The expected amplicon sizes for beta-actin and beta-globin PCR products are 150bp and 340bp respectively.

2.2.2. RNA capture for sequencing BCR repertoires

Total RNA was initially processed for target enrichment using the NEBNext kit (NEB) according to manufacturers protocol. Briefly, mRNA was isolated by polyA+ selection and converted to cDNA. cDNA at 0.3 to 0.7ng/µl was fragmented to 200bp (Covaris), ligated to sequencing adaptors (Illumina) and size selected at 200bp (Life Technologies E-Gel). Samples were then indexed for pre-capture pooling (NEBNext Multiplex Oligos for Illumina Index Primers 1 to 12). A pre-capture library was generated using 12 cycles of PCR (KAPA Biosystems Library Amplification Kit). Libraries were pooled and hybridised to biotinylated RNA-capture baits (custom design (Fisher et al., 2014), full protocol available on request), Agilent SureSelect) at 65°C for 24 h. Hybridised fragments were selected using streptavidin magnetic beads, washed and eluted for multiplexed sequencing on Illumina Miseq.

2.2.3. 5' Rapid amplification of cDNA ends (5'RACE) of B-cell receptors

5'RACE was performed using SMARTer™ Pico PCR cDNA Synthesis Kit (Clontech) according to Clontech protocols, using the JH-reverse primer (Table S3) and SMARTer 5' primer for PCR amplification.

2.2.4. Sequencing methods

454-libraries were prepared using standard Roche-454 Rapid Prep protocols incorporating 10-base multiplex identifier (MID) tags and sequenced using an FLX Titanium Genome Sequencer (Roche/454 Life Sciences). MiSeq libraries were prepared using Illumina protocols and sequenced by 250bp or 300bp paired-ended MiSeq (Illumina) as indicated. Raw 454 or MiSeq reads were filtered for base quality (median >32) using the QUASR program (http://sourceforge.net/projects/quasr/) (Watson et al., 2013). MiSeq forward and reverse reads were merged together if they contained identical overlapping region of >65bp, or otherwise discarded. The 250bp reads from the 5'RACE experiment were retained if they contained a JH-reverse primer sequence and orientated to begin with IgHV gene. Reads from RNA-capture were BLAST aligned to reference IgH genes, and trimmed if the reads extended outside the IgHV-D-J region, and filtered for length (>160bp). Non-immunoglobulin

sequences were removed and only reads with significant similarity to reference IgHV genes from the IMGT database (Lefranc et al., 2009) using BLAST (Altschul et al., 1990) were retained ($1x10^{-10}$ E-value threshold). Primer sequences were trimmed from the reads, and sequences retained for analysis only if both primer sequences were identified and if sequence lengths were greater than 255bp or 195bp for FR1 and FR2 primed samples respectively for 454, or both forward and reverse reads greater than 110 bp for MiSeq. FR1 primed PCR samples from CLL patients were also Sanger-sequenced.

2.2.5. Per-base error quantification

The same PCR protocol and read quality filtering was used to amplify beta-actin, beta-globin and GAPDH genes from two healthy individuals (amplicon sizes of 150bp, 340bp respectively). 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 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.

2.2.6. Reference-based V-D-J assignment

BLAST (Altschul et al., 1990) was used to align the 454 sequences against known BCR sequences from the ImMunoGeneTics (IMGT) database (Lefranc et al., 2009). 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).

2.2.7. Network assembly and analysis

The network generation algorithm is summarised in **Figure 2.2**. Briefly, each vertex represents a unique sequence, where the relative size of the vertex is proportional to the number of sequence reads identical to the vertex sequence. Edges were calculated between vertices that differed by single nucleotide non-indel differences. The network generation was performed using custom Python scripts

using CD-Hit (Li and Godzik, 2006) and analyses were performed using igraph implemented in R (http://igraph.sourceforge.net/index.html). The distribution of mismatches within a single network cluster were determined by aligning the sequence representing the largest vertex with the sequences to which it is connected and the positions of mismatches were determined along the sequences. Two-sided t-tests were performed in R.

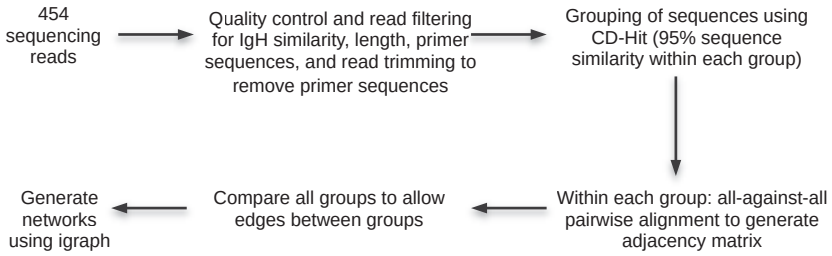

Figure 2.2. Outline of network generation method.

The sequencing reads were initially filtered for base quality (median >32) using QUASR quality control program (http://sourceforge.net/projects/quasr/). To filter for non-immunoglobulin sequences, only reads that had significant similarity to a reference IgHV gene from the IMGT database (Lefranc et al., 2009) using BLAST (Altschul et al., 1990) (E-value threshold of $1x10^{-10}$) were retained. The primer sequences were trimmed from the reads, and sequences were retained for analysis only if both primer sequences were identified and the sequence lengths were greater than 255bp for FR1 primed samples, or 195 bp for FR2 primed samples. Fast algorithms were used to cluster the reads into groups of similar sequences (with greater than 95% sequence identity) using the CD-HIT program (Li and Godzik, 2006). Within each group, all-against-all pairwise alignments were performed, using customized python scripts, to determine all intra-group edges within the network, after which each group was compared to determine inter-group edges. Due to the reported prevalence of 454 derived sequencing errors in homopolymeric regions (Albers et al., 2011), homopolymeric indels were not included in the total number of mismatches in pairwise comparisons. The network analyses were performed using igraph implemented in R (http://igraph.sourceforge.net/index.html).

2.2.8. Diversity measure calculations

The Gini index was calculated by ordering the cluster sizes from largest to smallest and creating a cumulative frequency distribution, where $R = \{r_1, r_2, \dots, r_n\}$, r_i is the cumulative size of the all the largest clusters until the ith largest cluster and normalized such that $r_n = 1$. The Gini index is $Gini\ index\ (g) = \sum_{i=1}^{N} \frac{(r_i - (i/N))}{N}$, where N is the number of clusters (Morrow, 1977).

2.2.9. Estimation of cluster sizes due to sequencing error

The Poisson distribution can estimate the expected number of reads containing i errors from the (central) vertex of size n reads, given an estimated error rate. The expected number of sequences with i errors is $n.p_i$, where $p_i = P(X = i) = \frac{\lambda\, e^{-\lambda}}{i!}$, and λ is the expected number of mutations per read. A cluster is defined as a set of interconnected vertices, in which edges are generated between vertices that differ by a single base. A vertex v is only included in a cluster when the minimum distance from v to any of the sequences in the cluster containing the central vertex is one. Thus, all the sequencing errors at $i=1$ generate vertices that have edges connecting to the central vertex. At $i > 1$, a vertex with set of mutations M_x will be connected to the cluster only if there exists a vertex in the cluster with a set of mutations M_y such that $\left|\frac{M_x}{M_y}\right| = |\{x \in M_x | x \notin M_y\}| = 1$ (i.e. there is only one mutation in M_x that is not in M_y). Therefore the probability of vertices due to i sequencing errors is estimated by drawing $S[n, i]$ samples from a multinomial distribution, for which the probability of the possible vertices that could connect to the cluster is given by $S[n, i] = \prod_{j=1}^{i-1} \frac{E[n, j-1]}{l} . p_i$, where l is the length of the sequence and $E[n, j]$ is the estimated number of vertices that are in the cluster which are at distance of j from the central node. 1000 independent samples were drawn from the multinomial distribution to estimate the average number of vertices at distances i from the central vertex, and therefore the cluster size due to sequencing error can be estimated by summing over the expected number of vertices at all i, $1 \leq i \leq \infty$.

2.2.10. Phylogenetic analysis of BCR sequences

BCR sequences related to the largest cluster 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) as determined by *phangorn* in R (Schliep, 2011).

2.2.11. Linear discriminant analysis of BCR repertoire parameters

Each numerical B-cell repertoire feature was normalised to sum to one over all samples. The *lda* function in R was performed to find a linear combination of features that best separates sample types (Rindskopf, 1997), projected over the first and second LDA dimensions. Hierarchical clustering of samples was performed using *hclust* in R (Murtagh and Contreras, 2012), where the distance measures between any two samples *i* and *j* was determined by:

$$d = \sqrt{\left(LDA_1^i - LDA_1^j\right)^2 + \left(LDA_2^i - LDA_2^j\right)^2}$$

Where LDA_1^i and LDA_2^i are the first and second LDA dimension values for sample *i* respectively.