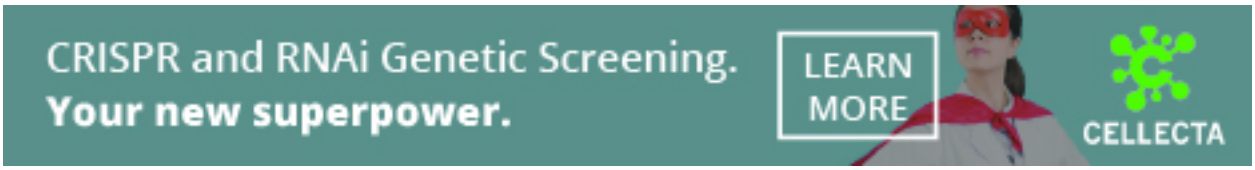
Network properties derived from deep sequencing of the human B-cell receptor repertoires delineates B-cell populations

Rachael Bashford-Rogers, Anne Palser, Brian Huntly, et al.

Genome Res. published online June 6, 2013
Access the most recent version at doi:10.1101/gr.154815.113



NETWORK PROPERTIES DERIVED FROM DEEP SEQUENCING OF THE HUMAN B-CELL RECEPTOR REPERTOIRES DELINEATES B-CELL POPULATIONS.

Rachael J. M. Bashford-Rogers[1], Anne L. Palser[1], Brian J. Huntly[2], Richard Rance[1], George S. Vassiliou[1], George A. Follows[3] and Paul Kellam[1,4]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, CB10 1SA;

[2]CIMR, Addenbrooke's Hospital, Hills Road, Cambridge, United Kingdom, CB2 0XY;

[3]Department of Hematology, Addenbrooke's Hospital, Hills Road, Cambridge, United Kingdom, United Kingdom, CB2 0QQ;

[4]Research Department of Infection, Division of Infection and Immunity, University College London, Gower Street, London, United Kingdom, WC1E 6BT

Correspondence should be addressed to Paul Kellam (pk5@sanger.ac.uk)

SCIENTIFIC SECTION DESIGNATION: METHODS AND RESOURCES

Abstract

The adaptive immune response selectively expands B- and T-cell clones following antigen recognition by B- and T-cell receptors (BCR and TCR) respectively. Next-generation sequencing is a powerful tool for dissecting the BCR and TCR populations at high-resolution, but robust computational analyses are required to interpret such sequencing. Here, we develop a novel computational approach for BCR repertoire analysis using established next-generation sequencing methods coupled with network construction and population analysis. BCR sequences organize into networks based on sequence diversity, with differences in network connectivity clearly distinguishing between diverse repertoires of healthy individuals and clonally expanded repertoires from individuals with chronic lymphocytic leukemia (CLL) and other clonal blood disorders. Network population measures defined by Gini Index and cluster sizes quantify the BCR clonality status and are robust to sampling and sequencing depths. We show, for the first time, B-cell CLL tumor clone evolution over the course of therapy. BCR network analysis therefore allows the direct and quantifiable comparison of BCR repertoires between samples and intra-individual population changes between temporal or spatially separated samples.

Introduction

Healthy humans have approximately $3x10^9$ B-cells in the peripheral blood and this population consists of a repertoire of distinct B-cells expressing different B-cell receptors (BCRs) necessary to bind diverse antigens and produce an effective humoral immune response. BCRs consist of two identical heavy chain (IgH) and two identical light-chain proteins, where the antigen-binding regions are highly diversified (Tonegawa 1983; Woof and Burton 2004). BCR diversity is generated in a number of ways. The IgH gene locus encodes for multiple distinct copies of the variable (V), diversity (D) and joining (J) gene segments (Jung et al. 2006), with functional IgH BCR genes generated by site-specific V-D-J recombination (Latchman 2005; Schatz and Swanson 2010). The imprecise joining of the V-D-J gene segments leads to random deletion and insertion of nucleotides during recombination events, resulting in sequence diversification at the junctional regions (Figure 1A). Rearranged BCR genes are further diversified by helper T-cell mediated somatic hypermutation (SHM) through the action of activation-induced cytosine deaminase. Through clonal affinity selection for enhanced antigen binding, non-germline SHM-mediated variation contributes significantly to the diversification of the mature B-cell repertoire (Brezinschek et al. 1995; Dorner et al. 1998; Weinstein et al. 2009; Batrak et al. 2011).

The diversification and selection dynamics of BCR repertoires in healthy individuals and those with infection, autoimmunity, immunodeficiency or B-cell malignancies remain poorly understood but can have important clinical implications. For example, the majority of B-cells in individuals with B-cell malignancies typically express a single dominant clonal BCR sequence (Arber 2000; Campbell et al. 2008) and continued intra-clonal tumor evolution by SHM in patients with B-cell lymphomas has been observed (Stamatopoulos et al. 1996; Bagnara et al. 2006; Volkheimer et al. 2007). Importantly, patients with chronic lymphocytic leukemia (CLL) with mutated BCR sequences in the tumor clone compared to the germline have a prognostically inferior survival rate and requirement of early treatment compared to those with unmutated malignant clones (Caligaris-Cappio and Ghia 2008). The BCR sequence repertoire of an individual therefore represents a surrogate of their B-cell clonality

status in health and disease, with the potential to give new insights into adaptive immune response as well as providing diagnostic and prognostic power when used clinically.

Previous studies have mainly produced descriptive analyses of the BCR populations. Isoelectric focusing (IEF) spectrotype methods (Williamson et al. 1973; Rieben et al. 1996; Satoh et al. 1996) preceded the advent of sequencing technologies (Arnaout et al. 2011) and is not quantitative. The potential size of the human repertoire is estimated to be 10^{11} unique BCR sequences therefore deep, high-throughput sequencing is necessary for sampling this repertoire robustly and to identify different subsets of BCRs (Dimitrov 2010; Benichou et al. 2012). There are several methods for isolation, amplification and sequencing of B-cell repertoires. Multiplex PCR amplification, using degenerate PCR primers complementary to germline V and J segments have been designed and validated previously (van Dongen et al. 2003; Lukowsky et al. 2006; Bruggemann et al. 2007; Evans et al. 2007; van Krieken et al. 2007; Vargas et al. 2008), used in numerous biological studies (Sanchez et al. 2003; Campbell et al. 2008; Boyd et al. 2009; Boyd et al. 2010; Krause et al. 2011; Jager et al. 2012; Lev et al. 2012; Maletzki et al. 2012) and optimized for clinical use (McClure et al. 2006; Harris et al. 2012; Sproul and Goodlad 2012) although the potential for biased PCR amplification remains. 5' rapid amplification of cDNA ends (5' RACE) has also been used (Bertioli 1997; Freeman et al. 2009; Varadarajan et al. 2011; Warren et al. 2011), but can suffer from low efficiency and high levels of non-specific amplification, contamination by short fragments from RNA degradation or incomplete cDNA synthesis. Both methods, employ PCR and therefore have a risk of systematic over/underrepresentation of immunoglobulin sequences either through different primer annealing or different amplification efficiencies of the distinct V families (Sandberg et al. 2005).

Previous studies have qualitatively shown diverse IgH repertoires in healthy patients contrasting with clonal populations in malignancies (Sanchez et al. 2003; Campbell et al. 2008; Boyd et al. 2009; Carulli et al. 2011; Logan et al. 2011; Maletzki et al. 2012), and that distinct subsets of B-cells within the same individual have distinct repertoires (Wu et al. 2010). 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 (Stewart et al. 1997; Sanchez et al. 2003; Campbell et al. 2008; Boyd et al. 2009; Weinstein et al. 2009; Boyd et al. 2010; Wu et al. 2010; Jager et al. 2012; Lev et al. 2012; Maletzki et al. 2012). However, computational assignment of V-D-J sequences to reference databases results in many incompletely assigned IgHV, D and J genes even when the germline alleles are known (Weinstein et al. 2009). This is most likely due to SHM masking the identity of the germline genes present in the NGS, or the existence of allelic variation relative to the reference IgH genes. Further, investigation of V-D-J gene usage frequencies utilizes only part of the BCR sequence diversity with important information about the V-D-J joining regions and mutational relationships not considered.

Here we propose that analysis of the BCR sequence relationships using the full BCR V-D-J sequence is more informative for human BCR repertoire analysis than V-D-J gene classification. We show that human BCR repertoire diversity can be interpreted through full V-D-J genotype diversity using BCR networks, previously shown to be an intuitive way for understanding B-cell repertoires in zebrafish (Ben-Hamo and Efroni 2011). In such networks, the lowest level of organization in a population of B-cells, namely independent B-cells, is 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). Using degenerate PCR-based methods we focus on sequencing RNA populations to maximize analysis of functionally rearranged BCRs rather than any non-functional first BCR allele defective rearrangements present in the genomic DNA from B-cell populations, but with the disadvantage that unequal numbers of RNA molecules per cell have the potential to inflate or deflate detected B-cell populations in the repertoire. Through sequencing the BCRs from samples with clonally expanded B-cell populations (peripheral blood from patients with CLL and human lymphoblastoid cell lines (LCLs)) as well as diverse BCR populations from peripheral blood from healthy individuals, we show network analysis provides a robust framework to understand vast sequencing repertoires by sequence relationships that clearly distinguish between B-cell quantitatively

using network measures. This framework is complimentary to existing phylogenetic methods, and we show, for the first time, B-cell tumor clone evolution over the course of therapy. These methods are robust to sampling and sequencing depths as well as different sequencing technologies, thereby allowing the direct comparison of multiple tumor samples from the same and different patients.

Results

Next generation sequencing of IgH variable genes

We amplified by RT-PCR the expressed rearranged IgHV-D-J loci from mRNA from human B-cell populations using the consensus IgHJ primer and FR1 or FR2 IgHV family primers (Figure 1A and Table S1) (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 1). Samples yielded an average of 42,324 sequencing reads after filtering for quality and presence of IgH sequence (Table S2). Two additional samples from CLL patient A (pre and post treatment) were sequenced on the MiSeq platform. We also analyzed the BCR 454 sequence datasets from *Boyd et al.* (Boyd et al. 2009), which includes three healthy individuals and five patients with clonal blood disorders (Table S6).

The combined per-base error-rate for the RT-PCR and sequencing process for the 454 platform was similar to other studies (Wang et al. 2007; Boyd et al. 2009) (1.74×10^{-4}, 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). Similarly the combined per-base error-rate for MiSeq was 2.06×10^{-4}.

To initially assess the clonality of our samples, we determined the percentage of reads identical to the most abundant BCR sequence in each sample. The percentage of reads corresponding to the highest expressed 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 1B). There was no significant difference in the percentage of identical reads between the LCL and CLL patient samples (p-value=0.0594). Therefore, we confirm that 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) (Figure S2).

Validation of sequencing to represent the B-cell populations

To assess the reproducibility of the RT-PCR sequencing method to sample the BCR repertoire, we compared the number of overlapping sequences from two types of technical repeats on a range of samples: i) repeating the RT-PCR and re-sequencing (RT-PCR repeats) and ii) repeating the 454 sequencing from the same RT-PCR product (sequencing repeats). The percentage of the sequences shared (no more than 1bp different) between sequencing runs was calculated using all-against-all alignments. This showed over 98% and 30% reproducibility for LCL and healthy PB samples respectively (Figure 1C), probably due to the increased probability of resampling more abundant BCR types in LCLs. It is clear the sequencing overlaps between RT-PCR repeats (Figure 1D, green bars) are not significantly different to those between sequencing repeats (Figure 1D, purple bars) (p-value=0.738 by paired t-test), suggesting that our RT-PCR amplification and sequencing depth is sufficient to be representative of the major clonal BCR population in the sample.

Comparison between independent primer sets suggests limited primer bias.

To assess whether multiplex PCR methods cause significant PCR amplification bias, we determined the correlation between IgHV gene usages for samples independently amplified by the FR1 and/or FR2 primer sets. The IgHV gene usage frequency distributions for both healthy individuals and LCLs resemble those seen by *Arnaout et al.* (Arnaout et al. 2011) (Figure S3A and B). IgHV gene usage frequencies between FR1 and FR2 amplified samples from eight LCL samples and four healthy individual samples are highly correlated (R-value=0.984, Figure S3C) suggesting there is minimal primer amplification bias. Performing RT-PCR repeats using the FR1 primer set measures the stochastic effect of resampling the RNA populations and again we found a strong correlation between IgHV gene usage frequencies between replicates (R-value=0.972, Figure S3D). Importantly, for both of these comparisons, we see a strong linear correlation between IgHV gene usage frequencies when percentages are greater than 5% of the overall population, but BCR sequences at a

frequency of 0-5% show some effect of stochastic sampling, irrespective of primer set usage. Therefore, overall, we do not see significant primer amplification bias under conditions used here.

Limitations of V-D-J classification

We classified the V-D-J genes for each read by sequence similarity to germline sequences from the ImMunoGeneTics database (IMGT) (Lefranc et al. 2009) (Figure S4). The majority of sequences could be classified to their most closely related reference sequences for IgHV and IgHJ genes (average of 99.8% and 96.1% 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. 2010). We find no significant difference between the percentage of classified V, D and J genes of our dataset compared to that of *Boyd et al.* (Boyd et al. 2009) (Figure S4). To overcome limitations of IgH V-D-J gene classification, we propose that the use of complete V-D-J sequence information and mutational relationships would be more informative and robust framework for BCR repertoire analysis and B-cell population structure than simple V-D-J gene classification.

BCR sequences naturally organize into networks based on sequence diversity

For each sample, filtered and trimmed 454 or MiSeq sequences were used directly to generate a sequence network (Figure 2A). Each 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. Clusters are groups of interconnected vertices. Differences in network architectures are clearly seen by comparing B-cell populations from healthy individuals and LCLs. In LCLs, the majority of 454 sequences fall within a small

number of clusters, as these samples are predominantly comprised of a small number of large B-cell clone types (Figure 2Bi). In contrast, healthy individuals have sparsely connected networks where most sequences are unique, thus yielding small vertices indicative of high BCR sequence diversity in the sampled repertoire (Figure 2Bii). From healthy individuals 4.8-32.2% of BCR sequences fall within clusters of 3 or more reads, with the largest cluster representing 16.7% (4023 reads) of the total population in healthy individual 10. Sequences within a cluster are most likely related to a single V-D-J BCR progenitor. Alignment of sequences within the clusters shows the nucleotide differences are distributed along the length of the 454 sequences (Figure S5 A-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 S5D). As expected, the LCLs showed no significant difference between the mutational proportions of the CDR and FWR regions. Mismatches are not found primarily at the V-D or D-J boundary suggesting cluster sequences are derived from the same B-cell precursor.

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 2Ci) whereas others have more limited clonal expansion (Figure 2Cii) and expansion of two dominant clusters (Figure 2D). 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 largely sparsely connected age-matched healthy individuals.

Population measures capture network and sample diversity

We next aimed to quantify BCR population measures to allow comparison and interpretation of B-cell repertoire dynamics and biology. We investigated several parameters to describe different aspects of the B-cell populations. Gini Index is an unevenness measure.

When applied to the vertex size distribution for a given sample, these measures indicate the overall clonal nature of a sample, and when applied to the cluster size distributions, these measures indicate the overall SHM of a sample. The maximum cluster size 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, we include a measure of the second maximum cluster size as a percentage of reads in a sample.

The LCL samples, due to the more restricted 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 3A) suggesting 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 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), indicating 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 3A point (a)). This was also verified by independent sequencing using the FR2 primer set (Figure S6). 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.

We generated networks from the sequences derived from *Boyd et al.* (Boyd et al. 2009) to validate these population measures on independent BCR sequence data. We show 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 (Figures S7A), occupying equivalent regions of the cluster and vertex Gini Index graphs to samples within this study. Therefore, we conclude that Gini Index population

268 measure robustly separates distinct B-cell populations into different regions based on the
269 clonal nature of the sample.

270 We then determined whether we could separate monoclonal expansions, biclonal
271 expansions and diverse B-cell populations using the maximum cluster sizes and second
272 maximum cluster sizes (Figure 3B). We show that the CLL and LCL samples have maximum
273 cluster sizes >30% of the total reads compared to maximum cluster sizes of healthy individual
274 samples of <20%. However, the LCLs and CLLs collectively occupy two distinct regions in
275 this space. One group exhibits a single dominant clonal sequence, where the remainder of the
276 clusters are <5% of the total reads (Figure 3B surrounded by the dashed line). The second
277 group of samples has two dominant clusters above 40% and 20% of the total reads
278 respectively. CLL patient 5 repertoire comprises two dominant clonal groups each utilizing
279 different V-D-J genes ([IGHV3-66*03/IGHD6-19*01/IGHJ3*02] and [IGHV6-1*01/IGHD3-
280 3*01/IGHJ4*02] respectively), where the two clones are unlinked and represented by
281 completely different BCR sequences (Figure 2D and S8). Limited polyclonal expansions were
282 observed also in 5/8 of the LCL samples reflecting that EBV transformation of peripheral B-
283 cells frequently results in polyclonal LCLs. Using the dataset from *Boyd et al.* (Boyd et al.
284 2009), we show the same phenomenon of polyclonal expansions in a subset of samples
285 (patients with CLL/SLL and FL/SLL, Figure 2Eiii) where the maximum cluster sizes are
286 >35% and second maximum cluster sizes are >19% of the total reads (Figure S7B). Therefore
287 the polyclonal status of the tumor samples can be determined using B-cell network
288 reconstruction and analysis.

289 An important requirement of this approach is that the network diversity measures
290 must not be highly dependent on the depth of sequencing (scale invariant) and volume of PB
291 sample. If a given diversity measure is scale invariant for B-cell networks then the network
292 diversity measure should be the same regardless of the depth of sampled sequences, i.e. a
293 subset of 454 sequences should yield the same network diversity measure as the full set of
294 sequences. We tested all the proposed population measures as a function of sequencing depth
295 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 data size (Figure S9). Below 20%, small deviations in the Gini Index measures are seen. As these network measures had minimal standard deviation over all sub-sampling ranges, they are therefore robust parameters for inter-sample comparison.

Minimal effect of sequencing errors on network properties

We determined whether 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, we estimated the expected number of vertices comprising a cluster that could be due to sequencing error, given our experimentally derived PCR and sequencing error-rates. We find that all the samples have cluster sizes greater than that expected due to error alone, even at twice the measured error-rate (Figure S10). Therefore, the connectivity patterns of networks predominantly reveal differences in clonal expansions of B-cell populations rather than total sequencing errors. We propose that clusters identified in BCR networks are therefore derived from B-cells that share a common pro-B-cell progenitor with rearranged V-D-J that have subsequently expanded and diversified.

Directly comparing the Gini Index measures of V-D-J sequences from samples amplified independently by distinct primer sets (FR1 or FR2 primer sets) showed a strong positive linear correlation between the two primer sets, with R-values of 0.999 and 0.996 respectively for the vertex and cluster size diversities (Figure S11A). This supports a lack of PCR or sampling bias or an effect of sequencing errors for independent RT-PCRs with FR1 compared to FR2 primer sets. Further, we find that networks generated to include edge lengths of up to 5 base changes faithfully retain the network architecture for both the LCL and healthy individual samples (Figure S11B).

BCR repertoire network properties relates to CLL development

To assess the sensitivity of this analysis method, we use 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 performed. We find 90.9% of all reads in the undiluted sample fall within the leukemic cluster (Figure 3C and D). Using our methods, we can detect the leukemic clonal sequences at dilutions as low as 1:100,000. 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. We see significant increases in maximum cluster size above that of the healthy individual in CLL dilutions of 1:100 or less. Therefore, 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 types by comparing to healthy control BCR clonality (Sayala et al. 2007).

We therefore sought to understand the relationship between the BCR population measures and the CLL clinical information for each patient. Interestingly, there was a strong correlation between the length of time since CLL diagnosis and the vertex Gini Index (Figure 3E). This suggests longer disease times lead to larger vertices representing larger tumor clonal populations, in agreement with previous studies (Kelly et al. 2002; Hayes et al. 2010).

By BCR deep sequencing and network analysis, we hypothesize that we can follow the dynamics of dominant clonal populations at multiple time points. To test this, we sampled a stage B CLL patient (patient A) during the course of therapy. A pre-treatment sample was taken immediately before entering second cycle of Chlorambucil treatment, and second sample taken 28 days later. The BCRs were sequenced on the MiSeq platform yielding 40,414 and 36,197 high-quality reads respectively. The most abundant BCR sequences in the two samples were identical. Network construction shows a single dominant cluster at both time points, which decreases from 86% to 53% of total reads, reflecting the reduction in WBC (Figure 4A). This corresponds to the vertex Gini Index reducing from 0.892 to 0.713 and the cluster Gini Index reducing from 0.244 to 0.138. A composite network was generated of all

the sequences from the dominant clusters in both time-points, where the vertex sizes correspond to frequencies of each BCR at either the pre or post treatment samples (Figure 4B). The proportions of each unique BCR sequence between the pre and post-treatment samples give a strong linear relationship (R-value=0.999995, Figure 4C), where the post-treatment proportional frequencies are 62% of those in the pre-treatment sample, indicating that all BCR clones within the leukemic cluster are equally affected by this treatment.

To understand the evolution of the leukemic cluster over time, a maximum parsimony tree was fitted (Schliep 2011) encompassing sequences that were observed at least 6 times for the two samples (Figure 4D). Bootstrapping was performed to evaluate the reproducibility of the trees, showing strong tree support (>95% certainty for all branches). 122/231 of the BCR sequences are unique to the pre-treatment sample (e.g. Figure 4, clone 1) compared to 13/231 unique to the post-treatment sample (e.g. Figure 4, clone 2). However, the sequences that are primarily observed in the post-treatment sample show divergence from the dominant leukemic clone, suggesting leukemic cluster BCR evolution during treatment (e.g. clone 2). A similar analysis was performed on an independent Boyd et al. (Boyd et al. 2009) for a patient with chronic lymphocytic leukemia and small lymphocytic lymphoma samples separated by 3 months (Figure S12) showing distinct clonal diversification patterns.

Discussion

Deep-sequencing of B-cell and T-cell repertoires offers the potential for quantitative understanding of the adaptive immune system in health and disease. Here we use deep-sequencing of B-cell receptor V-D-J population frequencies and novel analyses of BCR repertoires at the level of clonal populations. The observation of frequent multiple identical BCR sequences in tumors and much lower frequency identical BCR sequences in PB from healthy individuals suggests that we rarely sequence multiple identical RNA molecules from a single B-cell. 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 effects of RT-PCR or sequencing error and

amplification bias on our analysis, often of concern for deep-sequencing, are minimal. We show a strong linear correlation between the network parameters of samples that have been RT-PCR amplified using different primer sets to distinct regions of the IgH variable RNA transcript suggesting that the PCR methods here have limited primer or amplification bias. We confirm the dominant clonal sequences for the CLL patients by Sanger sequencing, and show that in all cases the samples have cluster sizes notably greater than that expected due to the measured total process error-rate. Therefore, these observed V-D-J clusters are likely to have undergone mutational processes greater than process errors.

We define for the first time B-cell V-D-J sequence population measures that describe the clonality of the sequences and quantify both the effect of B-cell sequence diversification (cluster size) and clonal proliferation (vertex size) using the Gini Index as an unevenness measure. The maximum and second maximum cluster size is used to assess dual clonal expansions. 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. We show 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.

We observed variation between the BCR repertoires in healthy individuals and in CLL. One healthy individual showed a more developed network, defined by an increase in connectivity, with corresponding higher Gini Index values and larger maximum cluster sizes compared to the other healthy individuals. This clone was not germline in sequence and could be a result of antigen specific memory B-cell expansion or an undiagnosed malignant

transformation. Variation in network structures between individual healthy zebrafish BCR repertoires was also observed in the study by *Ben-Hamo et al.* (Ben-Hamo and Efroni 2011), where higher connectivity suggested an immune response within the individual. Similarly in CLL, assessing the maximum and second maximum cluster sizes we identify patients with more than one BCR clonal expansion where the two dominant clones have different V-D-J gene usages. This 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 selection in CDRs compared to malignant-driven expansion (Figure S5). Multiple separate clonal B-cell populations have been observed in previously published data in a subset of patients identified by different V and J chain usages (Hsi et al. 2000; Boyd et al. 2009), but the clinical significance of these findings are not known.

Time-dependent evolution of BCR networks may however provide a powerful means of assessing B-cell tumor evolution and response to therapy as well as the dynamics of a healthy B-cell repertoire. CLL vertex Gini Index is correlated with the time an individual has been living with CLL (Figure 3E). This coupled with the observation of *in vivo* evolution of BCR clones in CLL (Figure 4) suggests BCR sequencing in CLL may provide an additional prognostic value for the disease. Divergent evolution from a common leukemic ancestor has previously been observed in CLL, possibly through the accumulation of driver mutations with selective advantages in growth over other sub-clones (Campbell et al. 2008). Hypermutations within the IgH locus may also play a driver role in clonal expansions (Ghia and Caligaris-Cappio 2006). Therefore BCR sequencing and subsequent network and evolutionary analysis may play an important role in identifying population changes. However, an evolutionary model for BCR diversity in health and disease, similar to the models used in infectious disease phylogenetics is needed to fully explore these possibilities. Nevertheless, for the first time we show the short-term effect of therapy on the B-cell repertoire in CLL, and demonstrate how networks lend themselves to phylogenetic approaches. These methods are

435 sensitive and informative for characterizing of B-cell populations in health and B-cell
436 malignancies.
437
438

Methods

Samples

Peripheral blood mononuclear cells (PBMCs) were isolated from 10ml of whole blood from healthy volunteers and CLL patients using Ficoll gradients (GE Healthcare). 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). Research was approved by relevant institutional review boards and ethics committees (07/MRE05/44)

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 S1. 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. 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. 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.

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) or by 250bp paired-ended MiSeq (Illumina). Raw 454 or MiSeq reads were filtered for base quality (median >32) using the QUASR program (http://sourceforge.net/projects/quasr/). MiSeq forward and reverse reads were merged together if they contained identical overlapping region of >65bp, or otherwise discarded. 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.

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.

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).

Network assembly and analysis

The network generation algorithm is summarized in Figures 2A and S1. 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 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.

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).

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. Here we draw 1000 independent samples 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, $1 \leq i \leq \infty$.

Temporal evolution of dominant clones:

Sequences from the dominant clusters that were observed at least 6 times for the two samples underwent a multiple alignment using the ClustalW2 algorithm (www.ebi.ac.uk/Tools/clustalw2/index.html) with default parameters. A phylogenetic tree was fitted using unrooted parsimony methods implemented in R (http://cran.r-project.org/web/packages/phangorn/). Model tests were performed on different substitution models, for which JC+G+I substitution model was found to be optimum and thus used here. 1000 bootstrap samples of individual nucleotides in the multiple alignment was used to assess the reproducibility of the phylogenetic trees. The proportional difference in expression, *diff(i)*, of a given sequence i between month 0 and month 3 was calculated by difference in expression between the two time points and divided by the sum of the expression of the sequence over both times:

$$Diff(i) = \frac{E_i(t=3) - E_i(t=0)}{E_i(t=3) + E_i(t=0)}, \qquad -1 \leq Diff(i) \leq 1$$

where $E_i(t=0)$ and $E_i(t=3)$ is the expression of sequence i at 0 months and 3 months respectively.

DATA ACCESS

The IgH sequences discussed can be found as accession number ERP002120 in the European Nucleotide Archive (ENA).

ACKNOWLEDGEMENTS

This work was supported by the Wellcome Trust. We would like to thank the Cambridge Cancer Trials Centre and nurse specialists Gwyn Stafford, Rosie Tween, Lisa Walbridge and Joanna Baxter, and the patients and staff of Addenbrooke's Haematology Translational Research Laboratory and Cambridge Blood and Stem Cell Biobank which is supported by the BRC.

AUTHORSHIP

Contribution: R.J.M.B-R., A.L.P. and P.K. designed the study; R.J.M.B-R. performed experiments and analyzed the data; G.A.F. and G.S.V. provided patient samples; R.R. performed the 454 sequencing; B.J.H, G.A.F. and G.S.V. provided advice for the project; R.J.M.B-R., A.L.P. and P.K. wrote the paper, and all authors reviewed and approved the manuscript.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**(3): 403-410.

Arber DA. 2000. Molecular diagnostic approach to non-Hodgkin's lymphoma. *J Mol Diagn* **2**(4): 178-190.

Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, Nusbaum C, Rajewsky K, Koralov SB. 2011. High-resolution description of antibody heavy-chain repertoires in humans. *PloS one* **6**(8): e22365.

Bagnara D, Callea V, Stelitano C, Morabito F, Fabris S, Neri A, Zanardi S, Ghiotto F, Ciccone E, Grossi CE et al. 2006. IgV gene intraclonal diversification and clonal evolution in B-cell chronic lymphocytic leukaemia. *British journal of haematology* **133**(1): 50-58.

Batrak V, Blagodatski A, Buerstedde JM. 2011. Understanding the immunoglobulin locus specificity of hypermutation. *Methods Mol Biol* **745**: 311-326.

Ben-Hamo R, Efroni S. 2011. The whole-organism heavy chain B cell repertoire from Zebrafish self-organizes into distinct network features. *BMC Syst Biol* **5**: 27.

Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. 2012. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* **135**(3): 183-191.

Bertioli D. 1997. Rapid amplification of cDNA ends. *Methods Mol Biol* **67**: 233-238.

Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD et al. 2010. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *Journal of immunology* **184**(12): 6986-6992.

Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD et al. 2009. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci Transl Med* **1**(12): 12ra23.

Brezinschek HP, Brezinschek RI, Lipsky PE. 1995. Analysis of the heavy chain repertoire of human peripheral B cells using single-cell polymerase chain reaction. *Journal of immunology* **155**(1): 190-202.

Bruggemann M, White H, Gaulard P, Garcia-Sanz R, Gameiro P, Oeschger S, Jasani B, Ott M, Delsol G, Orfao A et al. 2007. Powerful strategy for polymerase chain reaction-based clonality assessment in T-cell malignancies Report of the BIOMED-2 Concerted Action BHM4 CT98-3936. *Leukemia* **21**(2): 215-221.

Caligaris-Cappio F, Ghia P. 2008. Novel insights in chronic lymphocytic leukemia: are we getting closer to understanding the pathogenesis of the disease? *J Clin Oncol* **26**(27): 4497-4503.

Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, Follows GA, Green AR, Futreal PA, Stratton MR. 2008. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci U S A* **105**(35): 13081-13086.

Carulli G, Marini A, Ciancia EM, Bruno J, Vignati S, Lambelet P, Cannizzo E, Ottaviano V, Galimberti S, Caracciolo F et al. 2011. Discordant lymphoma consisting of splenic mantle cell lymphoma and marginal zone lymphoma involving the bone marrow and peripheral blood: a case report. *J Med Case Rep* **5**: 476.

Dimitrov DS. 2010. Therapeutic antibodies, vaccines and antibodyomes. *mAbs* **2**(3): 347-356.

Dorner T, Brezinschek HP, Foster SJ, Brezinschek RI, Farner NL, Lipsky PE. 1998. Delineation of selective influences shaping the mutated expressed human Ig heavy chain repertoire. *Journal of immunology* **160**(6): 2831-2841.

Evans PAS, Pott C, Groenen PJTA, Salles G, Davi F, Berger F, Garcia JF, van Krieken JHJM, Pals S, Kluin P et al. 2007. Significantly improved PCR-based clonality testing in B-cell malignancies by use of multiple immunoglobulin gene targets. Report of the BIOMED-2 Concerted Action BHM4-CT98-3936. *Leukemia* **21**(2): 207-214.

Frazer KA Ballinger DG Cox DR Hinds DA Stuve LL Gibbs RA Belmont JW Boudreau A Hardenbol P Leal SM et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164): 851-861.

Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. 2009. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res* **19**(10): 1817-1824.

Ghia P, Caligaris-Cappio F. 2006. The origin of B-cell chronic lymphocytic leukemia. *Semin Oncol* **33**(2): 150-156.

Harris S, Bruggemann M, Groenen PJTA, Schuuring A, Langerak AW, Hodges E. 2012. Clonality analysis in lymphoproliferative disease using the

626 BIOMED-2 multiplex PCR protocols: experience from the EuroClonality
627 group EQA scheme. *Journal of Hematopathology*: 91-98.
628 Hayes GM, Busch R, Voogt J, Siah IM, Gee TA, Hellerstein MK, Chiorazzi N, Rai KR, Murphy EJ.
629 2010. Isolation of malignant B cells from patients with chronic lymphocytic leukemia (CLL)
630 for analysis of cell proliferation: Validation of a simplified method suitable for multi-center
631 clinical studies. *Leukemia Res* **34**(6): 809-815.
632 Hsi ED, Hoeltge G, Tubbs RR. 2000. Biclonal chronic lymphocytic leukemia. *Am J Clin Pathol*
633 **113**(6): 798-804.
634 Jager U, Fridrik M, Zeitlinger M, Heintel D, Hopfinger G, Burgstaller S, Mannhalter C, Oberaigner W,
635 Porpaczy E, Skrabs C et al. 2012. Rituximab serum concentrations during immuno-
636 chemotherapy of follicular lymphoma correlate with patient gender, bone marrow infiltration
637 and clinical response. *Haematologica* **97**(9): 1431-1438.
638 Jung D, Giallourakis C, Mostoslavsky R, Alt FW. 2006. Mechanism and control of V(D)J
639 recombination at the immunoglobulin heavy chain locus. *Annu Rev Immunol* **24**: 541-570.
640 Kelly LM, Yu JC, Boulton CL, Apatira M, Li J, Sullivan CM, Williams I, Amaral SM, Curley DP,
641 Duclos N et al. 2002. CT53518, a novel selective FLT3 antagonist for the treatment of acute
642 myelogenous leukemia (AML). *Cancer Cell* **1**(5): 421-432.
643 Krause JC, Tsibane T, Tumpey TM, Huffman CJ, Briney BS, Smith SA, Basler CF, Crowe JE, Jr.
644 2011. Epitope-specific human influenza antibody repertoires diversify by B cell intraclonal
645 sequence divergence and interclonal convergence. *Journal of immunology* **187**(7): 3704-3711.
646 Latchman D. 2005. Gene Regulation (Advanced Texts)
647 Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E,
648 Brochet X, Lane J et al. 2009. IMGT, the international ImMunoGeneTics information system.
649 *Nucleic Acids Res* **37**(Database issue): D1006-1012.
650 Lev A, Simon AJ, Bareket M, Bielorai B, Hutt D, Amariglio N, Rechavi G, Somech R. 2012. The
651 kinetics of early T and B cell immune recovery after bone marrow transplantation in RAG-2-
652 deficient SCID patients. *PloS one* **7**(1): e30494.
653 Lin MM, Zhu M, Scharff MD. 1997. Sequence dependent hypermutation of the immunoglobulin heavy
654 chain in cultured B cells. *Proc Natl Acad Sci U S A* **94**(10): 5284-5289.
655 Logan AC, Gao H, Wang C, Sahaf B, Jones CD, Marshall EL, Buno I, Armstrong R, Fire AZ,
656 Weinberg KI et al. 2011. High-throughput VDJ sequencing for quantification of minimal
657 residual disease in chronic lymphocytic leukemia and immune reconstitution assessment.
658 *Proceedings of the National Academy of Sciences of the United States of America* **108**(52):
659 21194-21199.
660 Lukowsky A, Marchwat M, Sterry W, Gellrich S. 2006. Evaluation of B-cell clonality in archival skin
661 biopsy samples of cutaneous B-cell lymphoma by immunoglobulin heavy chain gene
662 polymerase chain reaction. *Leuk Lymphoma* **47**(3): 487-493.
663 Maletzki C, Jahnke A, Ostwald C, Klar E, Prall F, Linnebacher M. 2012. Ex-vivo clonally expanded B
664 lymphocytes infiltrating colorectal carcinoma are of mature immunophenotype and produce
665 functional IgG. *PloS one* **7**(2): e32639.
666 McClure RF, Kaur P, Pagel E, Ouillette PD, Holtegaard CE, Treptow CL, Kurtin PJ. 2006. Validation
667 of immunoglobulin gene rearrangement detection by PCR using commercially available
668 BIOMED-2 primers. *Leukemia* **20**(1): 176-179.
669 Morrow JS. 1977. Toward a more normative assessment of maldistribution: the Gini Index. *Inquiry*
670 **14**(3): 278-292.
671 Rieben R, Frauenfelder A, Nydegger UE. 1996. Spectrotype analysis of human ABO antibodies:
672 evidence for different clonal heterogeneity of IgM, IgG, and IgA antibody populations. *Vox*
673 *Sang* **70**(2): 104-111.
674 Sanchez ML, Almeida J, Gonzalez D, Gonzalez M, Garcia-Marcos MA, Balanzategui A, Lopez-Berges
675 MC, Nomdedeu J, Vallespi T, Barbon M et al. 2003. Incidence and clinicobiologic
676 characteristics of leukemic B-cell chronic lymphoproliferative disorders with more than one
677 B-cell clone. *Blood* **102**(8): 2994-3002.
678 Sandberg Y, van Gastel-Mol EJ, Verhaaf B, Lam KH, van Dongen JJ, Langerak AW. 2005. BIOMED-
679 2 multiplex immunoglobulin/T-cell receptor polymerase chain reaction protocols can reliably
680 replace Southern blot analysis in routine clonality diagnostics. *J Mol Diagn* **7**(4): 495-503.
681 Satoh M, Akizuki M, Yamagata H, Nakayama S, Homma M. 1996. Restricted heterogeneity and
682 changing spectrotypes in autoantibodies to La/SS-B. *Autoimmunity* **24**(4): 229-236.
683 Sayala HA, Rawstron AC, Hillmen P. 2007. Minimal residual disease assessment in chronic
684 lymphocytic leukaemia. *Best Pract Res Clin Haematol* **20**(3): 499-512.
685 Schatz DG, Swanson PC. 2010. V(D)J Recombination: Mechanisms of Initiation. *Annu Rev Genet*.

686 Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**(4): 592-593.
687 Sproul AM, Goodlad JR. 2012. Clonality testing of cutaneous lymphoid
688 infiltrates: practicalities, pitfalls and potential uses. *Journal of*
689 *Hematopathology*: 69-82.
690 Stamatopoulos K, Kosmas C, Stavroyianni N, Loukopoulos D. 1996. Evidence for immunoglobulin
691 heavy chain variable region gene replacement in a patient with B cell chronic lymphocytic
692 leukemia. *Leukemia* **10**(9): 1551-1556.
693 Stewart JJ, Lee CY, Ibrahim S, Watts P, Shlomchik M, Weigert M, Litwin S. 1997. A Shannon entropy
694 analysis of immunoglobulin and T cell receptor. *Molecular immunology* **34**(15): 1067-1082.
695 Tonegawa S. 1983. Somatic generation of antibody diversity. *Nature* **302**(5909): 575-581.
696 van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL, Delabesse E,
697 Davi F, Schuuring E, Garcia-Sanz R et al. 2003. Design and standardization of PCR primers
698 and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations
699 in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-
700 3936. *Leukemia* **17**(12): 2257-2317.
701 van Krieken JH, Langerak AW, Macintyre EA, Kneba M, Hodges E, Sanz RG, Morgan GJ, Parreira A,
702 Molina TJ, Cabecadas J et al. 2007. Improved reliability of lymphoma diagnostics via PCR-
703 based clonality testing: report of the BIOMED-2 Concerted Action BHM4-CT98-3936.
704 *Leukemia* **21**(2): 201-206.
705 Varadarajan N, Julg B, Yamanaka YJ, Chen H, Ogunniyi AO, McAndrew E, Porter LC, Piechocka-
706 Trocha A, Hill BJ, Douek DC et al. 2011. A high-throughput single-cell analysis of human
707 CD8(+) T cell functions reveals discordance for cytokine secretion and cytolysis. *J Clin Invest*
708 **121**(11): 4322-4331.
709 Vargas RL, Felgar RE, Rothberg PG. 2008. Detection of clonality in lymphoproliferations using PCR
710 of the antigen receptor genes: Does size matter? *Leukemia Res* **32**(2): 335-338.
711 Volkheimer AD, Weinberg JB, Beasley BE, Whitesides JF, Gockerman JP, Moore JO, Kelsoe G,
712 Goodman BK, Levesque MC. 2007. Progressive immunoglobulin gene mutations in chronic
713 lymphocytic leukemia: evidence for antigen-driven intraclonal diversification. *Blood* **109**(4):
714 1559-1567.
715 Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW. 2007. Characterization of mutation
716 spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res*
717 **17**(8): 1195-1201.
718 Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, Webb JR, Holt RA. 2011. Exhaustive
719 T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen
720 selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res*
721 **21**(5): 790-797.
722 Weinstein JA, Jiang N, White RA, 3rd, Fisher DS, Quake SR. 2009. High-throughput sequencing of
723 the zebrafish antibody repertoire. *Science* **324**(5928): 807-810.
724 Williamson AR, Salaman MR, Kreth HW. 1973. Microheterogeneity and allomorphism of proteins.
725 *Ann N Y Acad Sci* **209**: 210-224.
726 Woof JM, Burton DR. 2004. Human antibody-Fc receptor interactions illuminated by crystal
727 structures. *Nature reviews Immunology* **4**(2): 89-99.
728 Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. 2010. High-throughput
729 immunoglobulin repertoire analysis distinguishes between human IgM memory and switched
730 memory B-cell populations. *Blood* **116**(7): 1070-1078.
731
732
733

734 **FIGURE LEGENDS:**

735 **Figure 1. Sequencing of B-cell receptor repertoires. A)** Representation of the genomic
736 rearrangement process during V-D-J recombination to generate the heavy chain B-cell
737 receptor. B-cell receptor amplification was performed by reverse transcription on total RNA
738 by single J region primer, and subsequent multiplex PCR amplification. **B)** The percentage of
739 reads corresponding to the highest expressed B-cell receptor sequence for each sample,
740 separated into sample type: healthy individuals, chronic lymphocytic leukemia patients (CLL)
741 and human lymphoblastoid cell lines (LCL). Two-sided t-tests was performed between the
742 sample subsets, with the p-values indicated above. **C)** Percentage of sequences shared
743 between runs for technical repeats for a) the RT-PCR and re-sequencing (RT-PCR repeats,
744 green bars) and b) the 454 sequencing from the same RT-PCR product (sequencing repeats,
745 purple bars). For each sample, two repeats were performed and the percentage of reads shared
746 between the repeats is shown (each repeat is compared to the other so two bars are shown per
747 sample).

748

749 **Figure 2. B-cell receptor repertoires from different samples. A)** Schematic diagram
750 showing the method by which the sequencing networks are generated: each vertex represents
751 a unique sequence, where the relative size of the vertex is proportional to the number of 454
752 sequencing reads that were identical to the vertex sequence. Edges are created between
753 vertices that differ by one base (indel or substitution). The vertex colors correspond to the
754 relative abundance of the corresponding sequences, where red, orange and yellow indicates
755 observation of a sequence in >90%, between 40-90% and <40% of the reads in the sample
756 respectively. **B)** Comparison of BCR sequence networks between **i)** a typical LCL sample and
757 **ii)** a typical healthy individual. **C)** BCR sequence networks of CLL patients with **i)** extensive
758 clonal enlargement and **ii)** limited clonal expansion. **D)** BCR sequence networks of CLL
759 patient 5 showing expansion of two dominant clusters. **E)** Networks generated from
760 sequencing dataset from *Boyd et al.* (Boyd et al. 2009) of **i)** healthy donor 1, **ii)** patient 2 with
761 follicular lymphoma (FL) and **iii)** patient 3 with FL and small lymphocytic lymphoma (SLL).

Figure 3. Measures differentiating between B-cell receptor populations. A) Cluster Gini Index plotted against vertex Gini Index for thirteen healthy individual samples, eleven chronic lymphocytic leukemia (CLL), and eight human lymphoblastoid cell line (LCL) samples. Point (a) corresponds with healthy individual 10. The red box and grey dashed box distinguish between the regions occupied between diverse and clonal populations respectively. **B)** The second maximum cluster sizes plotted against the maximum cluster sizes. The red, grey dashed and black solid boxes distinguish between the regions occupied between unexpanded populations, monoclonal expanded populations and biclonally expanded populations respectively. **C)** 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.* (Boyd et al. 2009) and **D)** 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 +/- standard deviation of maximum cluster size for healthy individuals for this dataset. **E)** Correlation between the Gini Index and the length of time since chronic lymphocytic leukemia (CLL) diagnosis for each patient in our dataset, with corresponding R^2-value.

Figure 4. B-cell leukemic clonal evolution. A) The B-cell sequence networks for patient A with chronic lymphocytic leukemia for samples i) prior to and **ii)** after second cycle of Chlorambucil treatment separated by 1 month with corresponding white blood cell counts. **B)** All sequences from the dominant clusters from both temporal samples were used to generate a composite network, and the differential frequencies at each time point is indicated by the relative vertex sizes. **C)** Correlation between the proportional frequencies of each unique BCR within the dominant clones of patient A with corresponding R-value and linear regression equation. **D)** An unrooted maximum parsimony tree was generated showing the relationships between sequences that were observed at least 6 times between the pre and post-

790 treatment samples, where the branch lengths are proportional to the number of varying bases
791 (evolutionary distance). The same tree with the tip colors showing the relative difference in
792 sequence abundance between the different time points, where green indicates observation of
793 sequence primarily in the pre-treatment sample, blue indicates predominant observations in
794 the post-treatment sample, and white indicates no change in frequency. Clones 1 and 2 refer
795 to examples of BCRs observed only in the pre or post-treatment samples respectively.
796

797 **Tables and figures:**
798
799 **Table 1. 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	-

800 * Abbreviations: LCL=Human lymphoblastoid cell line, CLL=chronic lymphocytic
801 leukemia, HCC=Hepatocellular carcinoma.
802

Figure 1

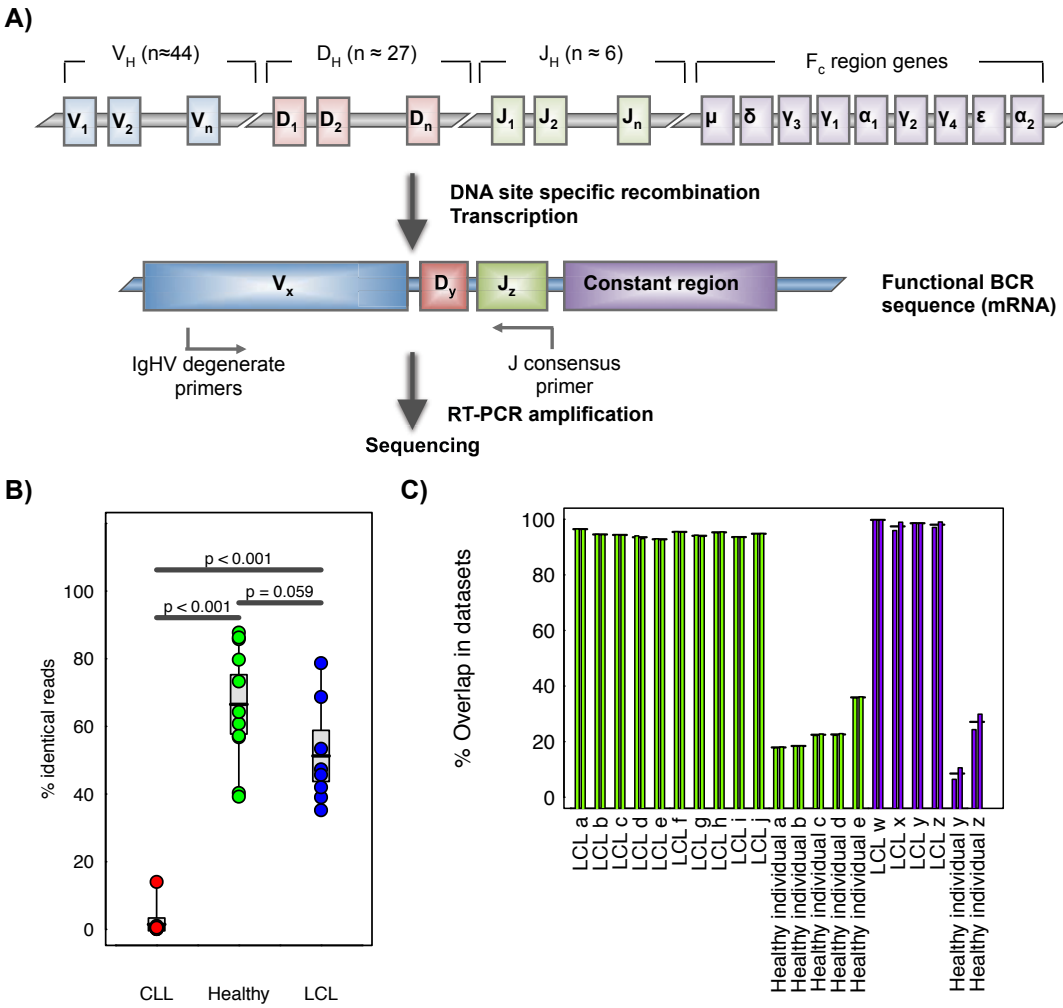


Figure 2

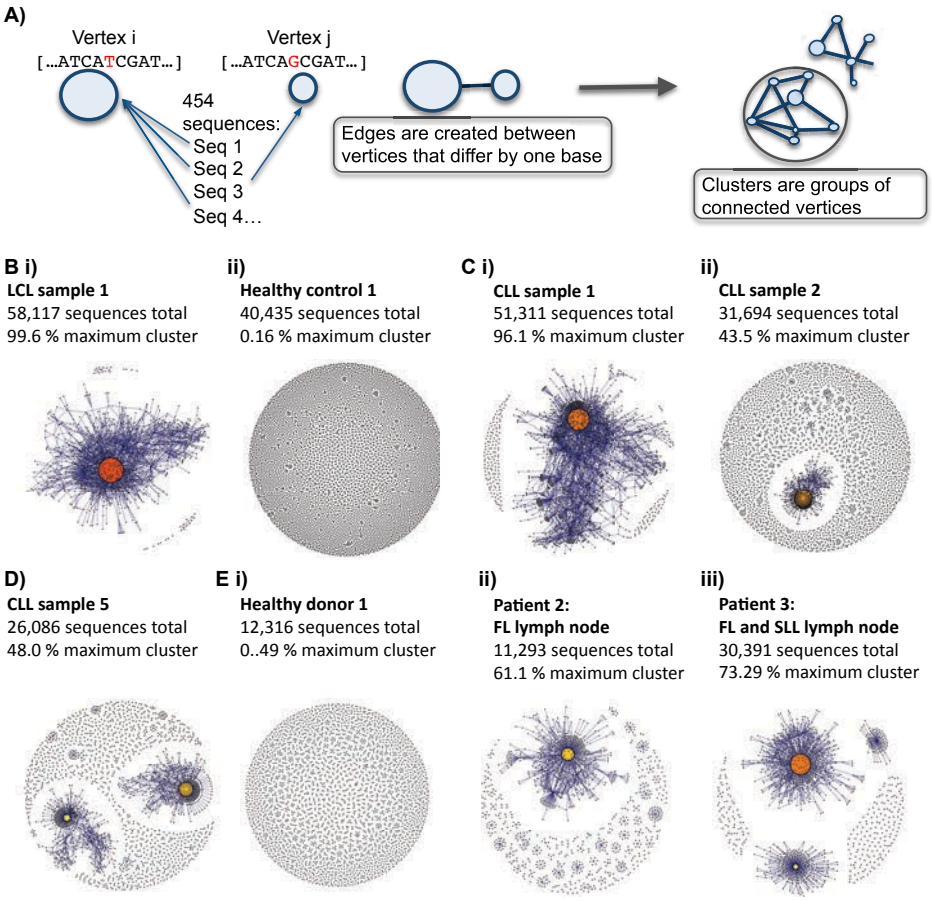
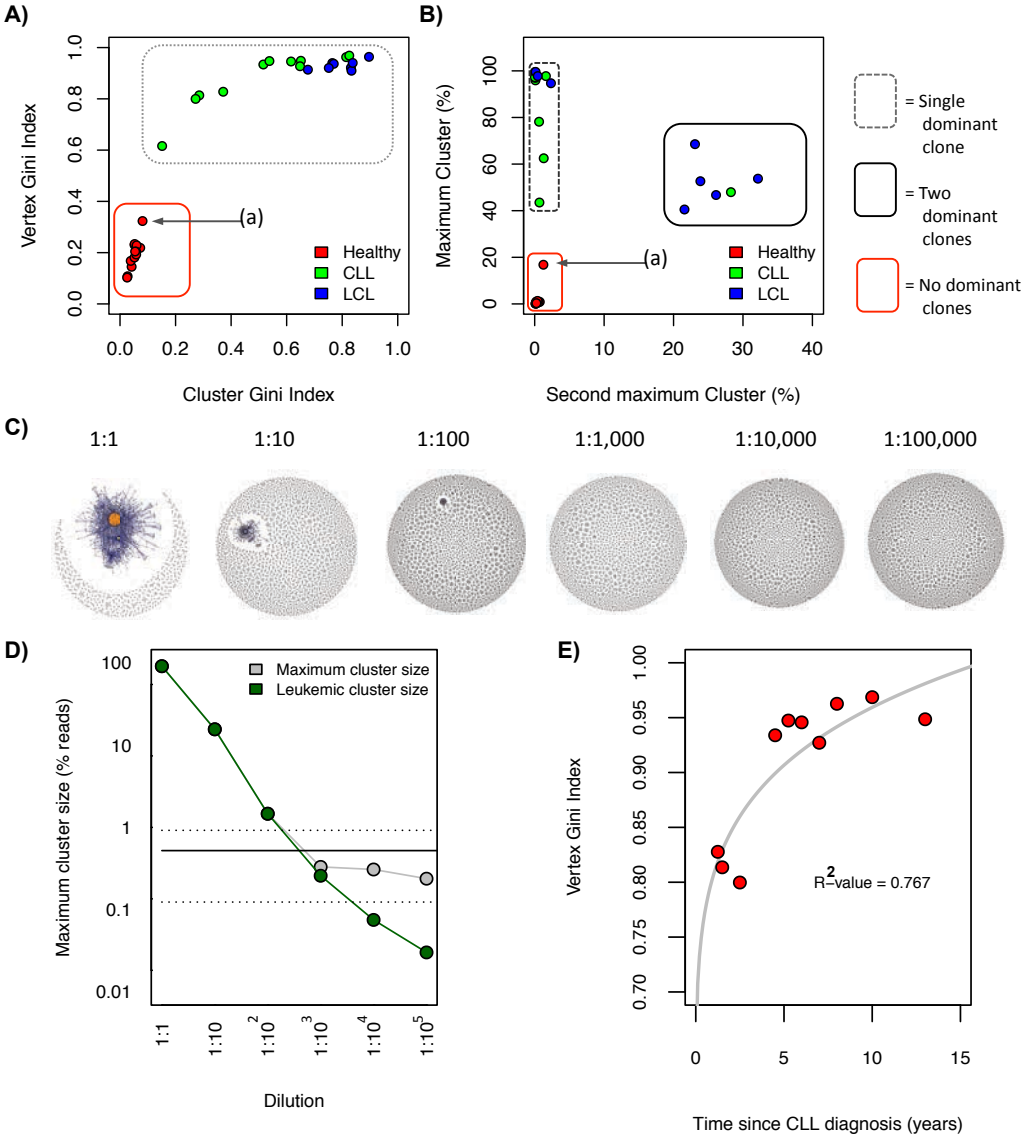
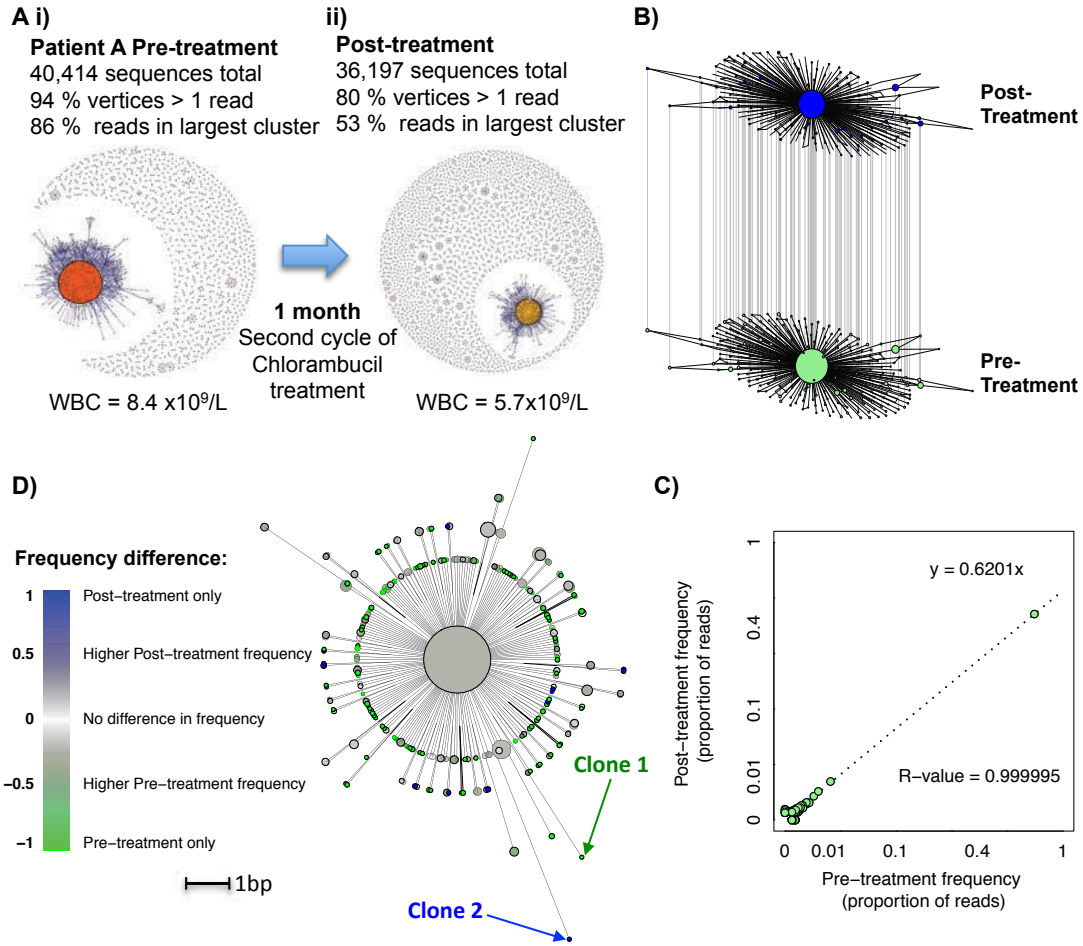


Figure 3



819 **Figure 4**
820
821



822
823