Leukemia (2010) 24, 125–132

www.nature.com/leu

ORIGINAL ARTICLE

A different ontogenesis for chronic lymphocytic leukemia cases carrying stereotyped antigen receptors: molecular and computational evidence

N Darzentas[1], A Hadzidimitriou[1,2], F Murray[3], K Hatzi[4], P Josefsson[5], N Laoutaris[6], C Moreno[7,8], A Anagnostopoulos[2], J Jurlander[5], A Tsaftaris[1], N Chiorazzi[4], C Belessi[6], P Ghia[9,10], R Rosenquist[3], F Davi[11] and K Stamatopoulos[2]

[1]Centre for Research and Technology Hellas, Institute of Agrobiotechnology, Thessaloniki, Greece; [2]Hematology Department and HCT Unit, G Papanicolaou Hospital, Thessaloniki, Greece; [3]Rudbeck Laboratory, Department of Genetics and Pathology, Uppsala University, Uppsala, Sweden; [4]The Feinstein Institute for Medical Research, North Shore–LIJ Health System, Manhasset, NY, USA; [5]Department of Hematology, Rigshospitalet, Copenhagen, Denmark; [6]Department of Hematology, Nikea General Hospital, Athens, Greece; [7]Institute of Hematology and Oncology, Hospital Clinic, University of Barcelona, Barcelona, Spain; [8]Institut d'Investigacions Biomediques August Pi i Sunyer, Hospital Clinic, University of Barcelona, Barcelona, Spain; [9]Laboratory of B-Cell Neoplasia, Division of Molecular Oncology and Clinical Unit of Lymphoid Malignancies, Department of Oncology, Università Vita-Salute San Raffaele, Milan, Italy; [10]Istituto Scientifico San Raffaele, Milan, Italy and [11]Laboratory of Hematology, Hôpital Pitié-Salpètrière, Université Pierre et Marie Curie, Paris, France

Chronic lymphocytic leukemia (CLL) is uniquely characterized by the existence of subsets of cases with quasi-identical, 'stereotyped' B-cell receptors (BCRs). Herein we investigate this stereotypy in 2662 patients with CLL, the largest series yet, using purpose-built bioinformatics methods based on sequence pattern discovery. Besides improving the identification of 'stereotyped' cases, we demonstrate that CLL actually consists of two different categories, based on the BCR repertoire, with important biological and ontogenetic differences. The first (~30% of cases) shows a very restricted repertoire and is characterized by BCR stereotypy (clustered cases), whereas the second includes cases with heterogeneous BCRs (nonclustered cases). Eleven major CLL clusters were identified with antigen-binding sites defined by just a few critically positioned residues, regardless of the actual immunoglobulin (IG) variable gene used. This situation is closely reminiscent of the receptors expressed by cells participating in innate immune responses. On these grounds, we argue that whereas CLL cases with heterogeneous BCRs likely derive from the conventional B-cell pool, cases with stereotyped BCRs could derive from progenitor cells evolutionarily adapted to particular antigenic challenges, perhaps intermediate between a true innate immune system and the conventional adaptive B-cell immune system, functionally similar to what has been suggested previously for mouse B1 cells.
Leukemia (2010) **24,** 125–132; doi:10.1038/leu.2009.186;
published online 17 September 2009
Keywords: CLL; B-cell receptor; stereotypy; pattern; repertoire

Introduction

The study of immunoglobulin heavy variable region (*IGHV*) genes has revolutionized our concepts about chronic lymphocytic leukemia (CLL).[1] It was long known that the *IGHV* gene repertoire in CLL is biased and distinct from that of normal peripheral blood B cells.[2–4] That was considered as an indirect indication of a restriction also in terms of antigen specificity. In principle, however, a wide IG heavy and light variable gene repertoire is not required for the production of specific antibodies to most antigens, as it depends mainly on the molecular features of the heavy-chain complementarity-determining region 3 (HCDR3). This is the region of the heavy chain that is very critical in the makeup of the antigen-binding site[5] to such an extent that the more similar the primary HCDR3 sequences of two IGs, the more similar their folding and, likely, their specificities.[6]

Along these lines, it was noteworthy that CLL is uniquely characterized by the existence of subsets of cases with remarkably restricted, 'stereotyped' HCDR3 sequences within their B-cell receptors (BCRs), strongly implying the recognition of structurally similar epitopes, likely selecting the leukemic clones.[7–14] As we and others have shown, these HCDR3 stereotypes may also share unique molecular and clinical features.[12–17] For instance, the *IGHV3-21*/*IGLV3-21* subset has been associated with a poor prognosis irrespective of *IGHV* mutational status.[12,15–17] In contrast, cases assigned to the *IGHV4-34*/*IGKV2-30* subset are relatively young,[13] uniformly express IgG-switched IGs[10,13] and tend to follow a very indolent course of the disease.[13] These findings suggest that a certain BCR stereotype can be critical in determining the prognosis and clinical outcome of subsets of CLL patients.

Therefore, the identification of shared amino-acid patterns leading to BCR stereotypy may offer useful hints that will assist in investigating the nature of the selecting antigens and their interactions with CLL progenitors or the malignant cells themselves and also have implications for the assignment of patients into different categories with distinctive biological and clinical features. Though theoretically easy, this task has never been fully accomplished, mainly due to several shortcomings of sequence-alignment-based tools that cannot properly address large-scale analysis of the complex and rather short HCDR3 sequences,[18] hampering the possibility to fully understand the distinctive features of stereotypy within the CLL BCRs.

In this study we introduce purpose-built, sophisticated computational tools specifically developed for sequence pattern discovery in HCDR3 amino-acid sequences to allow clustering of 2662 geographically distant CLL patients, the largest series of CLL sequences reported to date. Our analysis demonstrates that, based on shared sequence patterns, almost 30% of CLL HCDR3 sequences can be assigned to different clusters with unique characteristics. The use of these novel tools allowed identification of more distant sequence relationships among clustered cases; these relationships offer for the first time a comprehensive

Correspondence: Dr P Ghia, Department of Oncology, Università Vita-Salute San Raffaele, c/o DIBIT 4A3, Via Olgettina 58, Milano 20132, Italy.
E-mail: ghia.paolo@hsr.it
Received 30 March 2009; revised 4 July 2009; accepted 23 July 2009; published online 17 September 2009

overview of the HCDR3 'landscape' in CLL with the creation of a tree-like hierarchy, based on the identified clusters.

Of great importance, our analysis of the IG features of the identified clusters indicates that the IGHV gene repertoire restrictions typical of CLL are essentially a property of the clustered cases, clearly segregating them from the nonclustered cases. This suggests that the biological mechanisms that shape the IG repertoire of the CLL precursor cell population(s) may differ between CLL cases with clustered vs nonclustered sequences. The biased repertoire of the former subgroup could reflect an origin from a B-cell population intermediate between a true innate immune system and the conventional adaptive B-cell immune system, functionally similar to what has been suggested previously for mouse B1 cells.

Patients and methods

Patient group

A total of 2662 patients with CLL from collaborating institutions in Denmark ($n=189$), Finland ($n=34$), France ($n=767$), Greece ($n=481$), Italy ($n=238$), Spain ($n=92$), Sweden ($n=504$) and New York ($n=357$) were studied for IGHV gene repertoire and mutational status; this set is herein defined as 'internal'. All cases displayed the typical CLL immunophenotype as described earlier[7,11–14] and met the diagnostic criteria of the National Cancer Institute Working Group. Written informed consent was obtained according to the Declaration of Helsinki Principles and the study was approved by the local ethics review committee of each institution.

PCR amplification of CLL IGH rearrangements

PCR amplification and sequence analysis of IGHV-D-J rearrangements were performed as previously described.[7,11–14] Sequence data were analyzed using the IMGT database and tools (http://imgt.cines.fr). Only in-frame rearrangements were evaluated.

Collection of sequence data from public databases

IGHV-D-J sequences were retrieved from the IMGT/LIGM-DB database. Redundant, poorly annotated, out-of-frame, incomplete and identical sequences were excluded from the analysis. The final collection was comprised of 5528 unique IGHV-D-J sequences from B-cell lymphoproliferative disorders (including 182 sequences from CLL, here defined as 'external'), normal B cells, 'immune dysregulation' disorders (allergy, asthma, immunodeficiency) and autoreactive B cells (Supplementary Tables I-II).

Discovery of sequence patterns within HCDR3 amino-acid sequences

To comprehensively and efficiently identify amino-acid sequence patterns within the collected HCDR3 amino-acid sequences, we used the TEIRESIAS algorithm, a computational tool developed by the Bioinformatics and Pattern Discovery group at the IBM Computational Biology Center[19] and available from http://cbcsrv.watson.ibm.com/download.phtml.html. TEIRESIAS is an established pattern discovery method, used in diverse and challenging biological studies.[20] It requires a set of user-defined parameters, the complete set of which, as used in this analysis, is described in detail in Supplementary Materials and methods File. The adopted parameters ensured an extremely high level of sensitivity of the procedure, enabling the connection of all pairs of sequences that shared at least 50% amino-acid identity and 70% similarity. This higher similarity threshold was requested based on the equivalence in terms of 'hydropathy', 'volume', 'chemical characteristics' among the different residues ((AVLI), (CM), (ST), (KRH), (DE), (NQ), (F), (W), (P), (G), (Y)), as outlined in the IMGT classification of amino acids.[21] This distinction between identity and similarity thresholds was based on established criteria for the identification of HCDR3 stereotypy in CLL[7–14,17] and on published concepts such as amino-acid substitution matrices.[22] The identified patterns were subjected to a subsequent and stricter filtering process, using the following criteria: (1) sequence relatedness (as defined above), (2) HCDR3 length and (3) location within HCDR3 (Figure 1).

Clustering of HCDR3 sequences

Sequences were clustered based on the patterns they shared following an in-house clustering approach, implemented in the PERL programming language. The algorithm takes the highest scoring connections between sequences, and starts building clusters on a first (ground) level, called level 0. Sequences in each level 0 cluster are guaranteed to exhibit at least 50% amino-acid identity and 70% amino-acid similarity against all other members of that cluster, to have sequence lengths no more than two amino acids different, and to display the identified sequence patterns in a position no more than two amino acids apart. The common sequences between level 0 clusters led to their grouping in clusters at progressively higher levels of hierarchy, based on the existence of common amino-acid patterns. For this reason, sequences belonging to higher-level clusters, though largely retaining very similar—yet not always identical—sequence lengths and pattern location, do not necessarily exhibit 50% amino-acid identity and 70% amino-acid similarity against all other members of the same cluster.

Data mining and methods availability

Data were processed with the PERL programming language and clusters visualized with Biolayout 3D,[23] a freely available

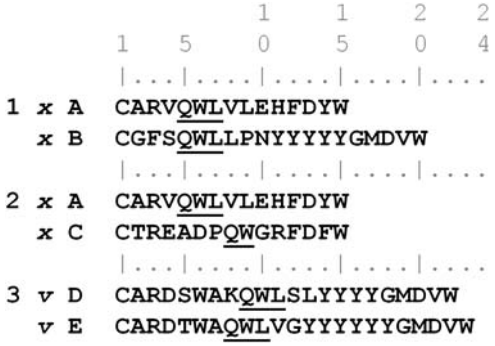

Figure 1 Pattern location and HCDR3 length filters. Three different scenarios tackled by the pattern filters, based on the location (offset) of patterns within HCDR3 amino-acid sequences, and the lengths of those sequences. Five sequences (A–E) of different lengths and a marker pattern, QWL, are used. In scenario 1, the link between sequences A and B is rejected because although the pattern's locations are identical, the sequence lengths are more than two amino acids different. In scenario 2, the lengths of the sequences are similar enough, however the pattern is in locations that differ by three amino acids. In scenario 3, the sequences are connected because both the locations and the lengths are within the limits. The length and offset criteria were put in place taking into consideration IG three-dimensional structural constraints.[6]

(http://www.biolayout.org/) interactive network (such as the clusters and their members) viewer. We make all PERL code and instructions available on request (as already stated, the TEIRESIAS algorithm can be downloaded from IBM Research).

Investigating the phylogeny of the IGHV *germ-line genes*

We compared the deduced amino-acid sequences of all known human vs mouse *IGHV* genes available in the IMGT/V-QUEST reference directory sets. Furthermore, we constructed sequence distance trees of functional human *IGHV* genes based on their amino-acid sequences using the neighbor-joining method of the freely available (http://www.megasoftware.net/) Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.[24]

Results

Pattern discovery in HCDR3 sequences

A total of 2844 in-frame IGHV-D-J sequences from patients with CLL were included in the analysis; 2662 sequences were 'internal', whereas the remaining 182 sequences were retrieved from public databases and, therefore, called 'external'. Detailed information on IG gene repertoires, *IGHV* gene mutational status and HCDR3 features is provided in Supplementary Tables III-VI and Supplementary Figures 1–4. Combinatorial pattern discovery using the TEIRESIAS algorithm (see Patients and methods) was performed in all the 2844 CLL sequences. The output was a list of 483 243 patterns, which were subjected to a subsequent multiple and strict filtering process (see Patients and methods), based on sequence relatedness (in terms of identity and similarity), HCDR3 amino-acid length and location (offset) within HCDR3. This resulted in a final list of 2182 patterns (a reduction of 99.5% compared to the initial set of patterns) along with the sequences each pattern appeared in, and a score describing the quality of the pattern-based connections between those sequences. An equivalent analysis was also performed in the non-CLL data set from the public databases, and in the complete data set combining both the CLL and non-CLL sequence data (Supplementary Figures File).

Clustering of sequences based on HCDR3 patterns

Based on shared patterns, 783 of 2844 (27.5%) CLL sequences were placed in 339 clusters at the ground level (level 0) that included 2–21 sequences each. HCDR3 sequences can appear in more than one level 0 cluster highlighting complex relationships, which were then used for further clustering. The common sequences between level 0 clusters led to their grouping in clusters at four progressively higher levels of hierarchy (levels 1–4) (Figure 2; Supplementary Table VII; Supplementary Figure 5). As they reflected more distant sequence relationships in the form of more widely shared sequence patterns, higher-level clusters (level 2 and above) were considerably larger in size (up to 86 cases each) but, interestingly, were characterized by notable homogeneity with regard to *IGHV* gene usage (Table 1;

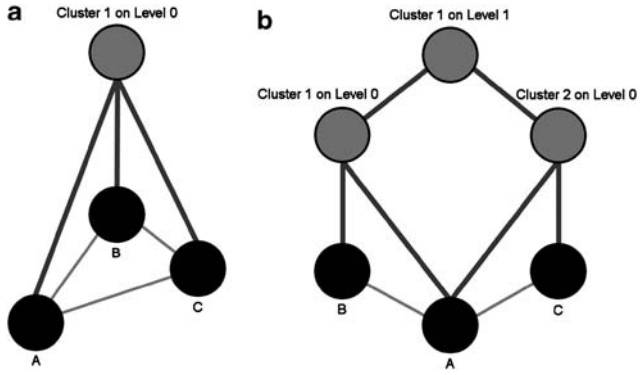

Figure 2 Two simplified clustering scenarios. (**a**) Sequence A is connected to B, forming cluster 1 on level 0; sequence C is then added to the same cluster because it is connected to both A and B. (**b**) Sequence A is connected to B, forming cluster 1 on level 0; sequence C is connected to A but not to B, so it takes a copy of A and forms cluster 2 on level 0; because the two clusters have A in common, they are connected on the next level of hierarchy to form cluster 1 on level 1.

Table 1 Major high-level clusters (archetypes) in the present study

Cluster	*Size*	*Predominant* IGHV *gene(s)*	IGHD *gene(s)*	IGHJ *gene*	IGK(L)V *gene*	*Consensus HCDR3 pattern*[a]
2-0000	27	*IGHV1-69* (all)	*IGHD3-16* (all)	*IGHJ3*	*IGKV3-20*	ARGG.YDY[AVLI]WGSYR..DAFD
2-0005	24	*IGHV4-39* (all)	*IGHD6-13* (all)	*IGHJ5*	*IGKV1-39/1D-39*	A…SSSW……WFDP
2-0013	14	*IGHV1-2* (all)	*IGHD1-26* (all)	*IGHJ6*	*IGKV4-1*	A…YYYYGMDV
2-0021	15	*IGHV1-58* (6/15) *IGHV1-69* (6/15)	*IGHD3-3* (all)	*IGHJ4*	Not available	A……DFWSG…
2-0023	7	*IGHV4-34* (all)	*IGHD2-15* (all)	*IGHJ6*	*IGKV3-20*	A..FYC.G..C…Y G.D[AVLI]
3-0000	47	*IGHV1-69* (23/47) *IGHV3-48* (9/47)	*IGHD3-3* (all)	*IGHJ6*		……D..SG……Y.Y.MDV
3-0001	82	*IGHV3-21* (78/82)	ND	*IGHJ6*	*IGLV3-21*	..[DE]
3-0002	86	*IGHV1-3* (31/86) *IGHV1-2* (24/86) *IGHV5-a* (10/86)	*IGHD619* (all)	*IGHJ4*	*IGKV1-39/1D-39*	A[KRH]…..DY
3-0003	27	*IGHV1-69* (all)	*IGHD3-10* (all)	*IGHJ6*	Variable	A….GV[AVLI]…….YY.MDV
3-0004	32	*IGHV1-69* (22/32) *IGHV3-48* (4/32)	*IGHD2-2* (all)	*IGHJ6*	Variable	A……..[AVLI]……YGMDV
3-0005	39	*IGHV4-34* (34/39) *IGHV3-72* (5/39)	*IGHD5-5* (15/39) *IGHD4-17* (10/39) *IGHD3-10* (7/39)	*IGHJ6*	*IGKV2-30*	[AVLI]…….[KRH]……[DE][AVLI]
4-0000	59	*IGHV1-69* (49/59)	*IGHD22* (28/59) *IGHD310* (23/59)	*IGHJ6*	Variable	A……..Y.MDV

[a]The dot is considered a wildcard and represents any amino acid at that position, whereas each pair of square brackets represents one position and means that any of the enclosed amino acids can be found at that position.

Supplementary Figures 6 and 7). Furthermore, they exhibited a notable restriction in HCDR3 lengths (Table 1; Supplementary Figure 7), despite the fact that, by design, the length criterion was more relaxed in higher- compared to ground-level clusters. Higher-level clusters also exhibited restricted IGKV/IGLV usage in most cases for which evaluation of paired IGH-IGK/L sequences was possible (Table 1).

In the non-CLL data set, level 0 clusters were significantly smaller in size (most included 2–3 members) than the corresponding CLL clusters. Furthermore, in stark contrast to high-level clusters in CLL, high-level clusters in the non-CLL group were characterized by marked *IGHV* gene heterogeneity. Finally, for the combined data set, 2493 sequences were placed in clusters. Of interest, CLL sequences formed significantly larger clusters compared to non-CLL sequences, which formed mainly two- or three-member-only level 0 clusters (Supplementary Figure 8).

Effects of clustering on the CLL IG repertoire

The IGHV repertoire of CLL sequences in clusters differed significantly from nonclustered cases (Figure 3; Supplementary Figure 9; Supplementary Table VIII). In detail, the six most frequent *IGHV* genes in high-level clusters (level 2 and above)—*IGHV1-69, IGHV3-21, IGHV4-34, IGHV1-2, IGHV1-3* and *IGHV4-39*—accounted for 73.9% of the *IGHV* gene repertoire at level 2 and 79.2% at level 3 but only 45.8% at the cohort level (χ^2-test: $P<0.0001$ for either comparison). Important differences could be also observed between 'low-level' (that is, levels 0–1) vs 'high-level' (that is, levels 2–4) clusters regarding the ranking of *IGHV* genes. Thus, certain genes were represented with an increasing frequency at each successive level of clustering (the foremost example being *IGHV3-21*); in contrast, other genes (for example, *IGHV3-7*, *IGHV3-23*, *IGHV3-30*) frequently represented at cohort level and usually somatically mutated, were progressively suppressed when their frequency among clustered sequences was compared to the cohort (Figure 4).

In addition, the *IGHJ* gene usage of CLL sequences in clusters also differed significantly from nonclustered sequences. Specifically, the *IGHJ6* gene was represented with an increasing frequency in each successive level of clustering, whereas the *IGHJ4* gene was characterized by a decreasing frequency (Supplementary Table IX; Supplementary Figure 10). Comparison of sequences in different levels of clustering with regard to HCDR3 lengths showed that sequences of 9, 13, 20 and 22 amino-acid HCDR3 lengths were represented increasingly in each successive level (Figure 5; Supplementary Table X), reflecting the increasing homogeneity of clustered cases. Although high-level clusters showed such great homogeneity in terms of *IGHV* gene usage and HCDR3 length, the sequence relatedness in all level 2–3 CLL clusters could be explained by just a few critically positioned residues within the HCRD3 region, distinctive for each cluster (Table 1).

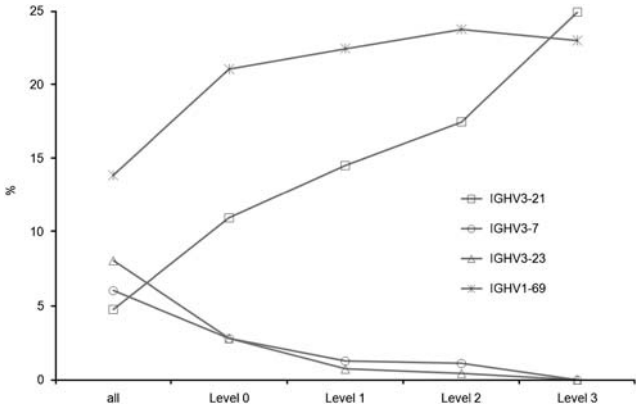

Figure 4 Effects of clustering on the immunoglobulin heavy variable (*IGHV*) gene repertoire. The percentage of chronic lymphocytic leukemia (CLL) sequences with *IGHV1-69*, *IGHV3-21*, *IGHV3-23* and *IGHV3-7* along the process of clustering, from the whole cohort (all) to sequences in level 3 clusters. As the graph clearly shows, the percentage of sequences with *IGHV1-69* and *IGHV3-21* is increased by ~10 and ~20%, respectively, whereas the other two genes are almost nonexistent starting from level 1 clusters.

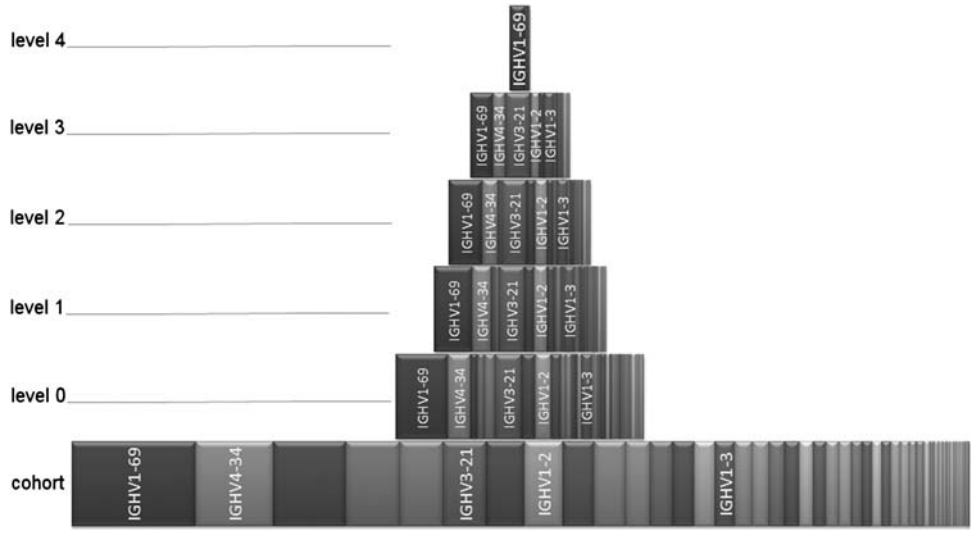

Figure 3 The effect of clustering on the numbers (quantity) and the IG repertoire (quality) of chronic lymphocytic leukemia (CLL) cases. Captured as a pyramid, the base of the pyramid represents the whole cohort in absolute numbers, with each colored 'stone' representing a different immunoglobulin heavy variable region (*IGHV*) gene. Consecutive layers of 'stones' represent the consecutive levels of clustering, all the way to the single cluster of level 4. In each layer, the most significant *IGHV* genes are marked.

Phylogeny of human IGHV *genes: a CLL perspective*

To dissect the biological and genetic reasons behind the *IGHV* gene bias shown among clustered cases, we constructed sequence distance trees of functional human *IGHV* genes based on their amino-acid sequences using the neighbor-joining method. Examination of the *IGHV3* subgroup tree revealed a branching that was reflected in the repertoires of clustered vs nonclustered CLL sequences. In particular, *IGHV3-21*, the foremost example of a gene found clustered in CLL rearrangements (Figures 3,4), belongs to a branch clearly distinct from other branches that include, for instance, the *IGHV3-23*, *IGHV3-30*, *IGHV3-33* and *IGHV3-7* genes (Figure 6a).[25] In accordance to this observation, the latter four genes were essentially absent from the *IGHV* gene repertoire of CLL sequences in level 2 clusters and above. In addition, though the *IGHV3-21* gene predominates in the major level 3 CLL cluster 0001, a few non-*IGHV3-21* cases were present in this cluster, utilizing the *IGHV3-48* and *IGHV3-11* genes that interestingly belong to the same branch of the *IGHV3* subgroup tree as *IGHV3-21*.

Similar phylogenetic relatedness, reflected in the *IGHV* gene repertoire of major CLL archetypes, was evident among *IGHV1* subgroup genes.[25] The *IGHV1-69* gene belongs to a branch of the *IGHV1* subgroup tree that includes the *IGHV1-58* gene and excludes other *IGHV1* subgroup genes (for example, *IGHV1-2*/*IGHV1-3*/*IGHV1-8*) (Figure 6b). The *IGHV1-69* gene formed a major high-level CLL cluster (2-0021) along with its closest 'relative', that is, *IGHV1-58*, but not with the other *IGHV1* subgroup genes. Conversely, the major level 3 CLL cluster 3-0002 included rearrangements of the latter *IGHV1* subgroup genes (especially, *IGHV1-2* and *IGHV1-3*) but contained no *IGHV1-69* sequences, even though this gene predominated in high-level clusters. Furthermore, cases using *IGHV1-2*/*IGHV1-3*/*IGHV1-8*/*IGHV1-18* as well as the *IGHV5-a* and *IGHV7-4-1* genes belonged to the same cluster (3-0002), reflecting their membership in the IGHV1-5-7 phylogenetic clan.[25,26]

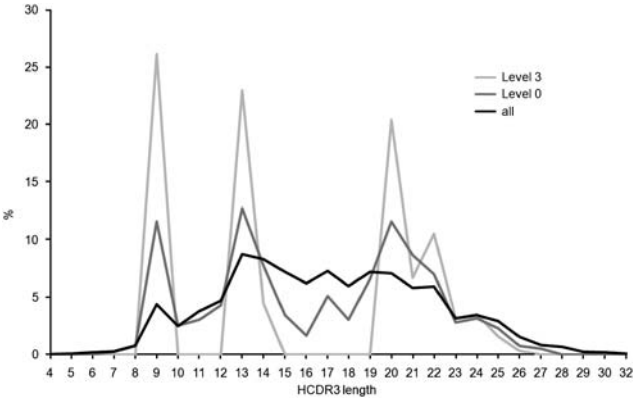

Figure 5 Effects of clustering on HCDR3 length. The distribution of chronic lymphocytic leukemia (CLL) HCDR3 sequence lengths of all (that is, nonclustered) sequences (black line), HCDR3 sequences in level 0 clusters (dark gray line) and HCDR3 sequences in level 3 clusters (light gray line). Levels 1 and 2 were omitted from the Figure for clarity. The 'enrichment' of clusters in sequences of four different lengths is striking. Specifically, sequences of lengths of 9, 13, 20 and 22 amino acids add up to the vast majority (~80%) of level 3 clusters.

Evolutionary considerations

As it is a central notion that sequence conservation very likely reflects antigen-binding activity,[24–26] we compared human *IGHV* gene sequences with those from different species. For such an analysis to be meaningful, the evolutionary distance between the species should be neither too proximate nor too distant, therefore we chose the mouse as the best reference species and system. Using the IMGT V-QUEST tool, we compared 124 human and 186 mouse germ-line *IGHV* sequences, including all alleles of all *IGHV* genes and noted only the best match of each sequence. This approach produced a pair-wise nucleotide identity matrix that included 67 human and 157 mouse *IGHV* sequences.

Our analysis showed that the most conserved genes between the human and the mouse are *IGHV3-21* along with *IGHV3-48* and *IGHV3-11*, which together form the major level 3 CLL

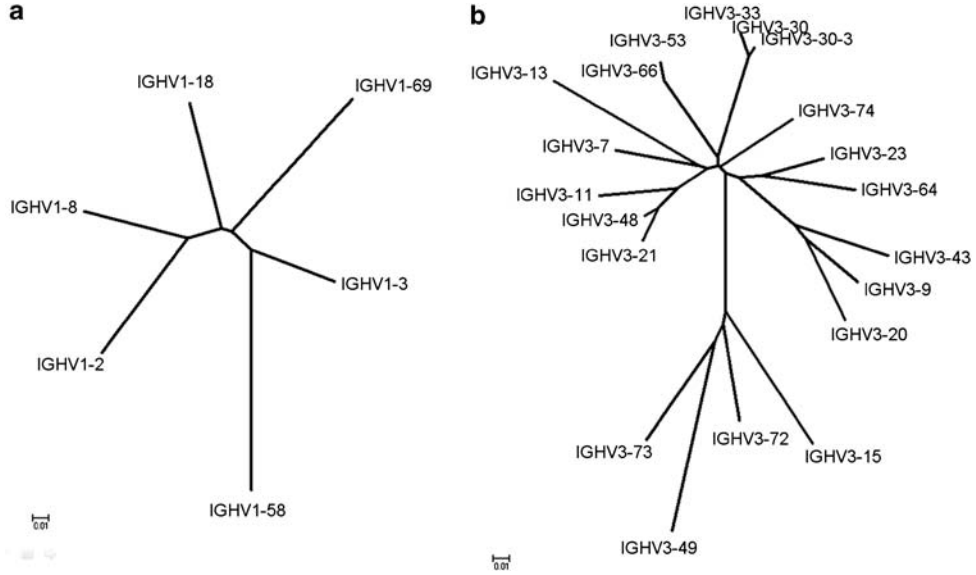

Figure 6 Sequence distance trees of human immunoglobulin heavy variable (*IGHV*) genes. Evolutionary relationships among human (**a**) *IGHV3* and (**b**) *IGHV1* subgroup genes. These unrooted trees were built using the neighbor-joining method.

cluster 0001. These three 'leaves' of the same branch in the human *IGHV* gene tree exhibit 85–89% nucleotide identity to the mouse *IGHV5-17* gene, strongly suggesting conservation by some functional constraint.

Discussion

Over the past decade, the concept of stereotyped BCRs with highly similar if not identical IG sequence patterns revolutionized our thinking about CLL biology.[27] Considering the extremely low probability (10^{-12}) of co-expression of identical BCRs, this unique feature of the CLL IG repertoire supports the notion that CLL development and evolution is not a simple stochastic event and indicates a role for antigen in driving the cell of origin for at least subsets of CLL cases.[3,7–14,17] From a clinical perspective, evidence that BCR stereotypy may assist in the biologically and clinically relevant categorization of patients with CLL is progressively emerging[7–14,17,28] and suggests that CLL patient outcome could be a reflection of ongoing BCR signaling in the context of other co-signals.

For these reasons, the systematic recognition of IG sequence patterns becomes an issue of prime importance. However, it remains unresolved, because the analytical methods used in all previous studies,[7–14,17] including ours, have relatively limited potential to effectively investigate data sets ever-increasing in size and complexity. Therefore, we developed a fast and yet transparent and manageable bioinformatics platform that encapsulates available experience and knowledge around a core of cutting-edge combinatorial pattern discovery in the form of the TEIRESIAS algorithm.[19] This choice was pivotal because, by overcoming the shortcomings of previously considered alignment-based methods, the algorithm offered an unconstrained, accurate and complete picture of HCDR3 stereotypy and, subsequently, of the intricate and often subtle sequence relationships in CLL, unprecedented by previous studies. Therefore, besides improving the quality and the reproducibility of the analysis, it helped highlight interesting features that bear relevant biological implications.

We herein demonstrate for the first time that the well-known IG gene repertoire restriction does not extend to all CLL cases but it is essentially a feature of those cases expressing stereotyped BCRs, which can be assigned to clusters defined by shared HCDR3 motifs. These major clusters could therefore be considered as corresponding to 'archetypical' BCRs with widely shared sequence features. Along the same line, although certain *IGHV* genes (especially, *IGHV3-21* and *IGHV1-69*) were overrepresented among clustered cases, others (for example, *IGHV3-7*, *IGHV3-23*, *IGHV3-30*) were, on the contrary, under-represented compared to the cohort. These results could be interpreted as evidence for a role of certain germ-line specificities in forming BCR archetypes in the progenitors of at least some CLL cases. Alternatively, they could be considered as indicative of a distinct ontogenetic origin for clustered vs nonclustered cases. In this scenario, CLL cases with stereotyped BCRs could derive from different progenitor cell populations evolutionarily adapted to particular antigenic challenges by (1) inherent differences in the capacity to rearrange and express particular *IGHV* genes and, (2) predetermined preferences for expressing certain stereotyped BCRs.[29–31]

The highly diverse CDR3 sequences are the principal determinants of specificity in antigen recognition, at least in the primary repertoire.[32] Indeed, in mice constrained to use a single *IGHV* gene to create their B-cell repertoire *in vivo*, HCDR3 diversity can be sufficient to create specific binders *in vivo* against essentially any large protein.[5] Of note, however, such restricted repertoires fail to respond to some antigens like polysaccharides.[5] Therefore, HCDR3 diversity alone is not enough to realize the full potential of antibody diversity.[33] There are well-documented cases in which the *IGHV* gene seems to be the specificity-defining sequence;[34] in other instances, both the *IGHV* gene and HCDR3 length restrictions are important. HCDR3 length restrictions may be related to a proper positioning of key structural determinants.[35] Alternatively, the rearranged *IGHV* gene requires a specific HCDR3 length for functionality.[36] Finally, one should not overlook the fact that there are well-documented cases for a particular light-chain sequence restriction among antibodies of a certain specificity, while still allowing for extensive diversity in the HCDR3.[37]

In the present study, remarkable HCDR3 length restrictions were identified for the CLL BCR archetypes corresponding to high-level clusters. In particular, although level 3 clusters could include sequences with a range of HCDR3 lengths, collectively, they were 'enriched' in sequences of just four different HCDR3 lengths (Figure 5). For instance, CLL cluster 3-0001 includes stereotyped rearrangements mainly of the *IGHV3-21* gene (78 of 82 cases) with a short HCDR3,[7,11–14,16,17] despite usage of *IGHJ6*, the longest *IGHJ* gene in the human genome. This subset of CLL patients has been repeatedly reported to exhibit distinctive biological and clinical features, such as frequent expression of CD38 and aggressive disease course.[7,12,13,17] Besides the well-known restricted usage of the *IGHV3-21* gene, we demonstrate here that this cluster is characterized (and may be defined) by a simple ('degenerate') HCDR3 consensus pattern, describing a 9 amino-acid long HCDR3 with only one 'landmark' acidic residue in the third position ('9/Acidic-3'). This phenomenon is not unique to *IGHV3-21* gene but holds for all levels 2–3 CLL clusters (Table 1). Indeed, CLL BCR archetypes using certain *IGHV* genes or groups of related *IGHV* genes could be defined by consensus HCDR3 patterns with just a few critically positioned residues, independent of the actual *IGHV* gene used.

These findings again highlight the fact that the sequence relatedness between CLL cases goes beyond the simple 'name' identity of the gene used but depends on structural and, likely, functional features. The systematic analysis of the deduced amino-acid sequences of all known human *IGHV* genes and the construction of sequence distance trees performed in this paper, following well-established principles,[25,26,38] helped us to further clarify this issue. The same BCR archetype may be shared by cases using different *IGHV* genes, especially those sharing common ancestry[13,14,25,26,38] (for example, the *IGHV1-2*/*IGHV1-3*/*IGHV1-8*/*IGHV1-18*, *IGHV5-a*, *IGHV7-4-1* genes, all members of the IGHV1-5-7 phylogenetic clan,[25,26] in CLL level 3 cluster 3-0002).

The branching of the human *IGHV* gene tree[25,26] was reflected in the ability of certain distinct, albeit related, genes to form CLL BCR archetypes. An illustrative example is provided by *IGHV1* subgroup genes. In particular, the *IGHV1-69* and *IGHV1-58* genes, two leaves of a distinct branch, formed a certain high-level archetype (that is, cluster 2-0021) but were absent from another major archetype (that is, cluster 3-0002), which includes cases using several *IGHV1* subgroup genes in other branches of the tree. The branching of the *IGHV* gene tree was also reflected in the repertoires of clustered vs nonclustered CLL sequences. For example, among *IGHV3* subgroup genes, branches including certain genes frequent at the cohort level and present among the nonclustered cases (for example,

IGHV3-23/*IGHV3-30*/*IGHV3-33*) seem to be excluded from CLL sequences in high-level clusters.

In addition, it is impressive to realize that just six *IGHV* genes (*IGHV1-69*, *IGHV1-3*, *IGHV1-2*, *IGHV3-21*, *IGHV4-34* and *IGHV4-39*) accounted for almost 80% of cases belonging to high-level clusters with stereotyped BCR structures. If we consider that this limited set of BCRs is mainly germ-line encoded (with a few notable exceptions carrying somatic mutations, such as the IGHV3-21 and IGHV4-34 BCRs), the situation is reminiscent of receptors of the innate immune system.[29–31,39] Along these lines, we have previously found stereotyped IG rearrangements in $CD5^+$ B-cell lymphoproliferations resembling CLL that develop in TCL1 transgenic mice.[40] On these grounds, we suggested that the TCL1 clones likely derived from the B-1a subset,[40] consistent with finding the initial, preleukemic clonal expansions in the peritoneal cavity.[41]

In the mouse, only two combinations of genes account for 5–15% of all mouse peritoneal B1 cells. These rearrangements react with phosphatidylcholine (PtC)/bromelainized red blood cells,[42,43] and one uses the *VH12* gene.[43] Almost all PtC-specific VH12-expressing B1 cells[43,44] are enriched for HCDR3 sequences of 10 amino acids, including many with a glycine in the fourth position ('10/G4' rearrangements).[43,44] In addition, VH12 B1 cells also exhibit a very restricted light-chain repertoire due to the inability of VH12H chains to associate with most L chains.[38,44] Based on these features, it is very tempting to consider analogies with CLL level 3 cluster 3-0001, which corresponds to the subset of CLL cases with stereotyped IGHV3-21 BCRs carrying '9/Acidic-3' HCDR3s and characterized by expression of λ-light chains using the *IGLV3-21* gene. We have previously demonstrated that the light-chain rearrangements in cases of this cluster have followed the hierarchical pattern of L chain recombination (IGK→IGK→IGL) and have gone through several rearranging attempts before producing a functional light chain.[16] On these grounds, we previously suggested that negative selection may have acted on BCRs with an IGHV3-21/IGKV rearrangement.[16] However, in analogy to the case of VH12 B-1 cells, one could also envisage that CLL precursors expressing certain stereotyped H chains require a restricted set of corresponding IGKV/IGLV light chains (that is, IGHV3-21 with IGLV3-21) to be properly expressed.

In both mouse and humans, the nature of the selective forces driving the evolution of the IGH loci is still a matter of controversy.[25,26,38] The different directions taken by the human and mouse genes will be partly due to chance.[25,26,38] However, they will also be affected by selection exerted by the antigens produced in the different environments that humans and mice have encountered over the past 70 million years of evolution. Therefore, it is remarkable that among all human *IGHV* genes, the *IGHV3-21* gene, along with *IGHV3-48* and *IGHV3-11*, all members of the same distinct branch of the *IGHV* tree, exhibit the highest similarity scores (85–89%) to a certain mouse *IGHV* gene (*IGHV5-17*). This homology may be considered as strong evidence for conservation by some functional constraint.

In conclusion, we have developed a powerful tool for pattern discovery in large cohorts of HCDR3 sequences. This approach becomes important if one takes into account the technical difficulties of crystallographic analysis (even *in silico*), which rule out its widespread application for determining the structure of the IGs expressed by the malignant B cells and their interactions with their cognate (mostly unknown) antigens. Our analysis reveals that the CLL BCR repertoire can be distinguished in two broad though different categories: the first (almost 30% of cases) is characterized by remarkable *IGHV* gene restrictions and BCR stereotypy (clustered cases), whereas the second includes cases with heterogeneous BCRs (non-clustered cases). Based on the evidence presented herein, the malignant clones belonging to the first category may derive from a B-cell population intermediate between a true innate immune system and the conventional adaptive B-cell immune system, functionally similar to what has been suggested previously for mouse B1 cells.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We thank Prof Marie-Paule Lefranc and Dr Veronique Giudicelli, Laboratoire d'Immunogenetique Moleculaire, LIGM, Universite Montpellier II, Montpellier, France, for their enormous support and help with the large-scale immunoglobulin sequence analysis throughout this project. We also thank Prof Göran Roos, Department of Medical Biosciences, Umeå University, Sweden; Prof Christer Sundström, Department of Genetics and Pathology, Uppsala University, Sweden; Dr Mats Merup, Department of Medicine, Karolinska University Hospital, Huddinge, Sweden; Dr Lyda Osorio, Department of Microbiology, Tumour and Cell Biology, Karolinska Institutet, Stockholm, Sweden; Dr Karin Karlsson, Department of Hematology, Lund University Hospital, Lund, Sweden and Prof Juhani Vilpo, Laboratory Centre, Tampere University Hospital, Tampere, Finland for providing samples and clinical data concerning Swedish and Finnish CLL patients. We also acknowledge the contribution of Dr Tatjana Smilevska, Dr Gerard Tobin, Dr Ulf Thunberg, Maria Norberg, Arifin Kaderi, Ingrid Thörn and Kerstin Willander to the sequence analysis. This work was supported by the General Secretariat for Research and Technology of Greece (INA-GENOME and ENTER programs); the BioSapiens Network of Excellence (contract number LSHG-CT-2003–503265); the Swedish Cancer Society; the Swedish Research Council, Medical Faculty of Uppsala University, Uppsala University Hospital; the Lion's Cancer Research Foundation, Uppsala, Sweden; the Associazione Italiana per la Ricerca sul Cancro—AIRC, Milano, Italy; Progetto Integrato Oncologia, Italian Ministry of Health, Rome, Italy; Fondazione Anna Villa e Felice Rusconi, Varese, Italy; Project C03/10 from 'Redes Temáticas de Investigación Cooperativa', Ministerio de Sanidad y Consumo (2003), Spain and the José Carreras International Foundation Against Leukemia (CR/07 and EM/07).

References

1 Chiorazzi N, Ferrarini M. B-cell chronic lymphocytic leukemia: lessons learned from studies of the B cell antigen receptor. *Annu Rev Immunol* 2003; **21**: 841–894.
2 Schroeder Jr HW, Dighiero G. The pathogenesis of chronic lymphocytic leukemia: analysis of the antibody repertoire. *Immunol Today* 1994; **15**: 288–294.
3 Johnson TA, Rassenti LZ, Kipps TJ. Ig VH1 genes expressed in B-cell chronic lymphocytic leukemia exhibit distinctive molecular features. *J Immunol* 1997; **158**: 235–246.
4 Fais F, Ghiotto F, Hashimoto S, Sellars B, Valetto A, Allen SL *et al.* Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. *J Clin Invest* 1998; **102**: 1515–1525.
5 Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* 2000; **13**: 37–45.
6 Zemlin M, Klinger M, Link J, Zemlin C, Bauer K, Engler GA *et al.* Expressed murine and human CDR-H3 intervals of equal length

exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *J Mol Biol* 2005; **334**: 733–749.
7 Tobin G, Thunberg U, Johnson A, Eriksson I, Söderberg O, Karlsson K *et al.* Chronic lymphocytic leukemias utilizing the VH3-21 gene display highly restricted Vlambda2-14 gene use and homologous CDR3s: implicating recognition of a common antigen epitope. *Blood* 2003; **101**: 4952–4957.
8 Ghiotto F, Fais F, Valetto A, Albesiano E, Hashimoto S, Dono M *et al.* Remarkably similar antigen receptors among a subset of patients with chronic lymphocytic leukemia. *J Clin Invest* 2004; **113**: 1008–1016.
9 Widhopf 2nd GF, Rassenti LZ, Toy TL, Gribben JG, Wierda WG, Kipps TJ. Chronic lymphocytic leukemia B cells of more than 1% of patients express virtually identical immunoglobulins. *Blood* 2004; **104**: 2499–2504.
10 Messmer BT, Albesiano E, Efremov DG, Ghiotto F, Allens SL, Kolitz J *et al.* Multiple distinct sets of stereotyped antigen receptors indicate a role for antigen in promoting chronic lymphocytic leukemia. *J Exp Med* 2004; **200**: 519–525.
11 Tobin G, Thunberg U, Karlsson K, Murray F, Laurell A, Willander K *et al.* Subsets with restricted immunoglobulin gene rearrangement features indicate a role for antigen selection in the development of chronic lymphocytic leukemia. *Blood* 2004; **104**: 2879–2885.
12 Ghia P, Stamatopoulos K, Belessi C, Moreno C, Stella S, Guida G *et al.* Geographic patterns and pathogenetic implications of IGHV gene usage in chronic lymphocytic leukemia: the lesson of the IGHV3-21 gene. *Blood* 2005; **105**: 1678–1685.
13 Stamatopoulos K, Belessi C, Moreno C, Boudjograh M, Guida G, Smilevska T *et al.* Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: pathogenetic implications and clinical correlations. *Blood* 2007; **109**: 259–270.
14 Murray F, Darzentas N, Hadzidimitriou A, Tobin G, Boudjogra M, Scielzo C *et al.* Stereotyped patterns of somatic hypermutation in subsets of patients with chronic lymphocytic leukemia: implications for the role of antigen selection in leukemogenesis. *Blood* 2008; **111**: 1524–1533.
15 Tobin G, Thunberg U, Johnson A, Thörn I, Söderberg O, Hultdin M *et al.* Somatically mutated Ig V(H)3-21 genes characterize a new subset of chronic lymphocytic leukemia. *Blood* 2002; **99**: 2262–2264.
16 Thorselius M, Krober A, Murray F, Thunberg U, Tobin G, Bühler A *et al.* Strikingly homologous immunoglobulin gene rearrangements and poor outcome in VH3-21-utilizing chronic lymphocytic leukemia independent of geographical origin and mutational status. *Blood* 2006; **107**: 2889–2894.
17 Bomben R, Dal Bo M, Capello D, Forconi F, Maffei R, Laurenti L *et al.* Molecular and clinical features of chronic lymphocytic leukaemia with stereotyped B cell receptors: results from an Italian multicentre study. *Br J Haematol* 2009; **144**: 492–506.
18 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997; **25**: 3389–3402.
19 Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics* 1998; **14**: 55–67.
20 Darzentas N, Rigoutsos I, Ouzounis CA. Sensitive detection of sequence similarity using combinatorial pattern discovery: a challenging study of two distantly related protein families. *Proteins* 2005; **61**: 926–937.
21 Pommie C, Levadoux S, Sabatier R, Lefranc G, Lefranc MP. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J Mol Recognit* 2004; **17**: 17–32.
22 Henikoff S, Henikoff JG. Performance evaluation of amino acid substitution matrices. *Proteins* 1993; **17**: 49–61.
23 Enright AJ, Ouzounis CA. BioLayout—an automatic graph layout algorithm for similarity visualization. *Bioinformatics* 2001; **17**: 853–854.
24 Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 2007; **24**: 1596–1599.
25 Kirkham PM, Schroeder Jr HW. Antibody structure and the evolution of immunoglobulin V gene segments. *Semin Immunol* 1994; **6**: 347–360.
26 Vargas-Madrazo E, Lara-Ochoa F, Ramirez-Benites MC, Almagro JC. Evolution of the structural repertoire of the human V(H) and Vkappa germline genes. *Int Immunol* 1997; **9**: 1801–1815.
27 Ghia P, Chiorazzi N, Stamatopoulos K. Microenvironmental influences in chronic lymphocytic leukaemia: the role of antigen stimulation. *J Intern Med* 2008; **264**: 549–562.
28 Kostareli E, Hadzidimitriou A, Stavroyianni N, Darzentas N, Athanasiadou A, Gounari M *et al.* Molecular evidence for EBV and CMV persistence in a subset of patients with chronic lymphocytic leukemia expressing stereotyped IGHV4-34 B-cell receptors. *Leukemia* 2009; **23**: 919–924.
29 Seidl KJ, MacKenzie JD, Wang D, Kantor AB, Kabat EA, Herzenberg LA. Frequent occurrence of identical heavy and light chain Ig rearrangements. *Int Immunol* 1997; **9**: 689–702.
30 Kantor AB, Merrill CE, Herzenberg LA, Hillson JL. An unbiased analysis of V(H)-D-J(H) sequences from B-1a, B-1b, and conventional B cells. *J Immunol* 1997; **158**: 1175–1186.
31 Tornberg UC, Holmberg D. B-1a, B-1b and B-2 B cells display unique VHDJH repertoires formed at different stages of ontogeny and under different selection pressures. *EMBO J* 1995; **14**: 1680–1689.
32 Davies DR, Cohen GH. Interactions of protein antigens with antibodies. *Proc Natl Acad Sci USA* 1996; **93**: 7–12.
33 Ohlin M, Zouali M. The human antibody repertoire to infectious agents: implications for disease pathogenesis. *Mol Immunol* 2003; **40**: 1–11.
34 Ohlin M, Borrebaeck CAK. Characteristics of human antibody repertoires following active immune responses *in vivo*. *Mol Immunol* 1996; **33**: 583–592.
35 Barrios Y, Jirholt P, Ohlin M. Length of the antibody heavy chain complementarity determining region 3 as a specificity-determining factor. *J Mol Recognit* 2004; **17**: 332–338.
36 Martin DA, Bradl H, Collins TJ, Roth E, Jack H-M, Wu GE. Selection of Ig μ heavy chains by complementarity determining region 3 length and amino acid composition. *J Immunol* 2003; **171**: 4663–4671.
37 Lamminmäki U, Westerlund-Karlsson A, Toivola M, Saviranta P. Modulating the binding properties of an anti-17beta-estradiol antibody by systematic mutation combinations. *Protein Sci* 2003; **12**: 2549–2558.
38 de Bono B, Madera M, Chothia C. *VH* gene segments in the mouse and human genomes. *J Mol Biol* 2004; **342**: 131–143.
39 Hardy RR. B-1 B cell development. *J Immunol* 2006; **177**: 2749–2754.
40 Yan XJ, Albesiano E, Zanesi N, Yancopoulos S, Sawyer A, Romano E *et al.* B cell receptors in TCL1 transgenic mice resemble those of aggressive, treatment-resistant human chronic lymphocytic leukemia. *Proc Natl Acad Sci USA* 2006; **103**: 11713–11718.
41 Bichi R, Shinton SA, Martin ES, Koval A, Calin GA, Cesari R *et al.* Human chronic lymphocytic leukemia modeled in mouse by targeted TCL1 expression. *Proc Natl Acad Sci USA* 2002; **99**: 6955–6960.
42 Mercolino TJ, Locke AL, Afshari A, Sasser D, Travis WW, Arnold LW *et al.* Restricted immunoglobulin variable region gene usage by normal Ly-1 ($CD5^+$) B cells that recognize phosphatidyl choline. *J Exp Med* 1989; **169**: 1869–1877.
43 Clarke S, McCray S. VH CDR3-dependent positive selection of murine VH12-expressing B cells in the neonate. *Eur J Immunol* 1993; **23**: 3327–3334.
44 Wang H, Clarke SH. Positive selection focuses the VH12 B-cell repertoire towards a single B1 specificity with survival function. *Immunol Rev* 2004; **197**: 51–59.

Supplementary Information accompanies the paper on the Leukemia website (http://www.nature.com/leu)