Molecular Medicine

An Entity Evolving into a Community: Defining the Common Ancestor and Evolutionary Trajectory of Chronic Lymphocytic Leukemia Stereotyped Subset #4

Lesley-Ann Sutton,[1] Giorgos Papadopoulos,[2*] Anastasia Hadzidimitriou,[1,3] Stavros Papadopoulos,[2] Efterpi Kostareli,[4] Richard Rosenquist,[1] Dimitrios Tzovaras,[2] and Kostas Stamatopoulos[1,3,5]*

[1]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden; [2]Information Technologies Institute, Center for Research and Technology Hellas, Thessaloniki, Greece; [3] Institute of Applied Biosciences, Center for Research and Technology Hellas, Thessaloniki, Greece; [4] Division of Epigenomics and Cancer Risk Factors, German Cancer Research Center (DKFZ), Heidelberg, Germany; and [5] Hematology Department and HCT Unit, G. Papanicolaou Hospital, Thessaloniki, Greece

Patients with chronic lymphocytic leukemia (CLL) assigned to stereotyped subset #4 express highly homologous B-cell receptor immunoglobulin (BcR IG) sequences with intense intraclonal diversification (ID) in the context of ongoing somatic hypermutation (SHM). Their remarkable biological and clinical similarities strongly support derivation from a common ancestor. We here revisited ID in subset #4 CLL to reconstruct their evolutionary history as a community of related clones. To this end, using specialized bioinformatics tools we assessed both IGHV–IGHD–IGHJ rearrangements (n = 511) and IGKV–IGKJ rearrangements (n = 397) derived from eight subset #4 cases. Due to high sequence relatedness, a number of subclonal clusters from different cases lay very close to one another, forming a core from which clusters exhibiting greater variation stemmed. Minor subclones from individual cases were mutated to such an extent that they now resembled the sequences of another patient. Viewing the entire subset #4 data set as a single entity branching through diversification enabled inference of a common sequence representing the putative ancestral BcR IG expressed by their still elusive common progenitor. These results have implications for improved understanding of the ontogeny of CLL subset #4, as well as the design of studies concerning the antigenic specificity of the clonotypic BcR IGs.

Online address: http://www.molmed.org
doi: 10.2119/molmed.2014.00140

INTRODUCTION

Patients with chronic lymphocytic leukemia (CLL) assigned to stereotyped subset #4 are characterized clinically by an early age at diagnosis and an indolent disease course and molecularly by B-cell receptor immunoglobulins (BcR IGs) that exhibit a series of distinctive immunogenetic features (1,2). More specifically, they are IgG-switched (a rarity in CLL since the great majority of CLL clones, >90% of all cases, express IgM/IgD) and are composed of heavy chains encoded by the *IGHV4-34* gene and light chains encoded by the *IGKV2-30* gene (3–5). The antigen-binding sites of subset #4 are equally interesting, being composed of a variable heavy complementarity determining region 3 (VH CDR3) that is long and enriched in positively charged residues (reminiscent of pathogenic anti-DNA antibodies) (3,4). Anti-DNA is the most common specificity in autoreactivity, with DNA binding often acquired through surface-active basic amino acids; predominantly arginine (R) but also, to a lesser extent, lysine (K) (6–8). This point is worthy of note since the VH CDR3 of subset #4 is defined by a (R/K)RYY motif which is deemed to not only be "CLL-biased" but also exclusive to subset #4 as it has never been found outside this context (3,4). In addition, both the VH and variable kappa (VK) domains of subset #4 demonstrate a high impact of somatic hypermutation (SHM) and are remarkable for carrying shared ("stereotyped") SHM, that is, identical changes at the same codon position of the variable domain (3,9).

Subset #4 is also outstanding due to intense intraclonal diversification (ID)

**LAS and GP are both first authors.*
***Address correspondence to** Lesley-Ann Sutton, Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, SE-751-85 Uppsala, Sweden. Phone: +46-18-4714852; E-mail: lesley.sutton@igp.uu.se.*
Submitted July 16, 2014; Accepted for publication December 1, 2014; Epub (www.molmed.org) ahead of print December 2, 2014.

within their IG genes in the context of ongoing SHM, alluding to an active, ongoing interaction with antigen(s) (10,11). Indeed, by conducting a large-scale longitudinal study of subset #4 we previously established: (i) the existence in most cases of distinct "clusters" of subcloned sequences; (ii) a hierarchical pattern of subclonal evolution, thus revealing which SHMs were negatively or positively selected overtime; and, (iii) subclonal drift, that is, temporal changes in the relative size of different clusters of sequences (12).

Nevertheless, this study only investigated clonal evolution at an individual case level and hence could not shed light on the clonal ancestry of subset #4 as a whole, which is relevant since the remarkable biological and clinical similarities of subset #4 cases strongly support derivation from a common ancestor. In an attempt to trace the ontogeny of subset #4, we here sought to revisit ID in subset #4 and reconstruct their evolutionary history by determining the structure of a community of related clones profiled at different time points for both IG heavy and light chains.

MATERIALS AND METHODS

Patient Group

Peripheral blood samples were collected at multiple time points from eight CLL patients meeting the International Workshop on Chronic Lymphocytic Leukaemia (iwCLL) criteria; these eight patients, on the basis of both their IG gene sequence features and our previously established criteria, were assigned to subset #4 (1,3,4,13). Patients' demographics and clinical and molecular data are summarized in Supplemental Table 1. Cases were analyzed over a six-year period (range 7 to 72 months, median 20 months) and no patient received treatment during sampling (Supplemental Table 1). The diagnostic sample was available, and called time point 1, for 6 of the 8 patients analyzed. No diagnostic samples were available for the remaining two patients (P0103 and P2451) and therefore the initial sample (time point 1) analyzed for these patients were 81 and 63 months post diagnosis, respectively. Written informed consent was obtained in accordance with the Declaration of Helsinki and the study was approved by the local ethics review committee.

PCR Amplification, Subcloning and Sequence Analysis of IGHV–IGHD–IGHJ and IGKV–IGKJ Gene Rearrangements

PCR amplification using the high-fidelity Accuprime Pfx polymerase (Invitrogen [Thermo Fisher Scientific, Waltham, MA, USA]), subcloning and sequence analysis and interpretation were performed as described previously. The sequence data evaluated herein has been reported previously (1,3,9–12).

Visualization of Clonal Evolution in Subcloned IG Gene Sequences

Sequence data was processed using the Damerau-Levenshtein edit distance algorithm (14–16). The Damerau-Levenshtein distance, as defined in this paper, is a multivariate function of two parameters; in this study these two parameters are the amino acid (or nucleotide) sequences. The defined distance is used for the computation of the difference between two IG chains. It is a distance metric in the sense that, given the amino acid (or nucleotide) sequences s1, s2, s3, the following conditions apply:

- Nonnegativity: $d(s1, s2) \geq 0$;
- Nondegeneracy: $d(s1, s2) = 0$ if and only if s1 = s2;
- Symmetry: $d(s1, s2) = d(s2, s1)$;
- Triangle inequality: $d(s1, s2) + d(s2, s3) \geq d(s1, s3)$.

The analytical form of the Damerau-Levenshtein distance between two chains, a and b, having lengths M and N, respectively, is defined by the following:

$$\mathbf{lev}_{a,b}(i,j) = \begin{cases} \max(i,j), & \text{if } \min(i,j) = 0, \\ \min\begin{cases} \mathbf{lev}_{a,b}(i-1,j)+1 \\ \mathbf{lev}_{a,b}(i,j-1)+1 \\ \mathbf{lev}_{a,b}(i-1,j-1)+\begin{cases} 0, & \text{if } a_i = b_j \\ 1, & \text{if } a_i \neq b_j \end{cases} \end{cases} & \text{otherwise} \end{cases}$$

where **lev** is a two-dimension matrix with M rows and N columns and the (i,j) entry is $\mathbf{lev}_{a,b}(i,j)$, where $i = 0, 1, 2, \ldots, M\text{-}1$ and $j = 0, 1, 2, \ldots, N\text{-}1$. This matrix has one in the element when a match is found between the i-th letter of the chain a and the j-th element of chain b; otherwise this element is equal to 0. Application of this algorithm to the entire IG variable domain was used to illustrate the diversity/similarity between subcloned sequences obtained from different time points across all patients. Consequently, each clonal sequence was compared with the entire data set, i.e., all clonal sequences obtained irrespective of time point or case, and this strategy provided a more robust evolutionary model for CLL subset #4 than inferring clonal relations at an individual case level. Modeling the genesis of subset #4 in this manner facilitated the deconvolution of sequence changes that occurred during the life of the clone. The distance matrix (17,18) resulting from the comparison process was used to interconnect each clone to the remainder in a minimum spanning tree which was subsequently visualized using purpose-built tools (19–21). Clones were positioned within this tree according to their individual distances, thus forming clusters which illustrated clonal relatedness beyond the individual patient level.

To explore the functional similarities of observed sequence changes, we followed the ImMunoGeneTics information system (IMGT) classification of the 20 common amino acids for the properties of hydropathy and chemical characteristics (http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/) and performed the following comparisons: (i) amino acid sequence distance including only replacement mutations; (ii) amino acid sequence distance when considering amino acids with similar physicochemical properties as single equivalent entities; (iii) amino acid sequence distance when considering amino acids within the same hydropathy group as single equivalent entities; and (iv) nucleotide sequence distance.

Focusing on both the VH and VK CDR3, hierarchical visualization was performed and by determining which nu-

cleotide or amino acid had the highest probability of appearing at a certain position, a hypothetical VH and VK CDR3 sequence from which all subset #4 CDR3 sequences derive could be constructed. More specifically, a hierarchical tree structure comprised of nodes and branches was assembled. Within this structure, the root node corresponded to the derived (proposed) ancestral sequence, and the branches were determined based on the calculated optimal string distance of each node. The string distance of a node indicated its position from the root node and also its proximity to the other nodes.

RESULTS

Composite Clusters of Subset #4 IG Sequences: Convergent Patterns of Subclonal Evolution

Clustering at the amino acid level. Analysis of the IGHV–IGHD–IGHJ amino acid sequences produced six distinct clusters (Figure 1A). Four of these clusters were composed of subclonal sequences obtained from different patients (P0907, P1422, P3020 and P1939), with each individualized cluster exhibiting a distinctive dispersion of clones, thus reflecting the varying extent of ID among subset #4 cases (Supplemental Figure 1). The remaining two clusters largely consisted of sequences from two patients each (composite clusters). The first such cluster contained sequences from patients P3916 and P2920 that grouped closely together. The second multimember cluster primarily contained sequences from two patients, P0103 and P2451; however, seven subcloned sequences from patient P1422, stemming from two different time points also clustered within this group, while the majority of sequences from patient P1442 clustered separately and at some distance away (Figures 1A,B). Thus, subcloned sequences from individual subset #4 cases clustered close together and behaved as if they were clonally related, that is, as if they stemmed from a community of clones with common ancestry. The observed branching may be indicative of special selective pressures occurring in parallel in distinct subclones.

Similar analysis of the IGKV–IGKJ amino acid sequences produced five clusters (Figure 1C). Two clusters were located within a very close distance, forming a more central core from which a further two clusters emanated. More specifically, the first cluster was formed by two patients (P2920 and P0907), while the second closely neighboring cluster contained the subcloned sequences from P0103 and P2451. Subcloned sequences from P3916 bridged these two clusters (Figure 1D). Clonal sequences from P1422 formed one of the two more distant clusters, while the other cluster was composed of clonal sequences from P1939. Patient P3020, previously found to carry limited ID despite bearing the highest SHM load (within both the IG heavy and light chain), was distanced from all other clusters. As with the cluster analysis of IG heavy chain sequences, we noted that individual IG kappa sequences occasionally were separated from their respective clusters and, instead, attached to clusters generated by other patients and located some distance away. Hence, the pattern of clustering evidenced from the kappa light chain sequences is analogous to that of their partner heavy chains, thereby reenforcing the idea that subset #4 essentially constitutes a community of related clones that follow closely similar ontogenetic and evolutionary pathways.

Clustering based on shared amino acid properties. Further comparisons were performed at the amino acid level by permitting a degree of ambiguity through the use of amino acid equivalences, that is, following the IMGT grouping of amino acids into classes based on distinct physicochemical or hydropathic properties. Excluding amino acids with the same physicochemical properties from the sequence distance-defining algorithm, that is, considering such amino acids as equal and, hence, not resulting in an overall change, resulted in a slight alteration to cluster formation. At the IG heavy chain level, five distinct clusters were now observed as opposed to the six clusters initially observed when physicochemical properties were included; this change primarily resulted from the clustering of P3020 with P0907 (previously occurring as two distinct clusters) (Figure 2A). For subset #4 kappa light chains, the effect on cluster formation was very minor (Figure 2B).

When amino acids within the same hydropathy groups were considered as equal, it was noted that mutations introduced by SHM did not lead to striking changes in hydropathy for any patient. Nevertheless, one difference observed when clustering in this manner concerned the subcloned IG heavy chain sequences of P3020 which, although remaining as a distinct cluster, now emerged from the cluster produced by patient P0907; prior to taking physicochemical properties or hydropathy into consideration, this cluster hailed from the central core cluster (Supplemental Figures 2, 3). The observed clustering based on shared amino acid properties indicates strong functional constraints for preservation of critical physicochemical properties; the limited range of permissible amino acids potentially reflects selection events governed by structural constraints for optimal antigen recognition.

Clustering at the nucleotide level. Finally, clustering based on changes within IG heavy chain nucleotide sequences produced an individual and distinct cluster for six patients, while patients P0103 and P2451 remained clustered together (Figure 3A). Within the IG kappa chains, although the clusters generated shared similarities to cluster formation at the amino acid level, the two central cores were completely distanced from each other and, instead, a major cluster was formed by four patients (P0103, P2451, P3916 and P1939) (Figure 3B). Since these four patients all carry a 10-amino acid VK CDR3, the enhanced segregation of clusters observed at the nucleotide level is likely attributable to the additional three nucleotides that these sequences carry compared with cases carrying a 9-amino acid VK CDR3; thus accounting

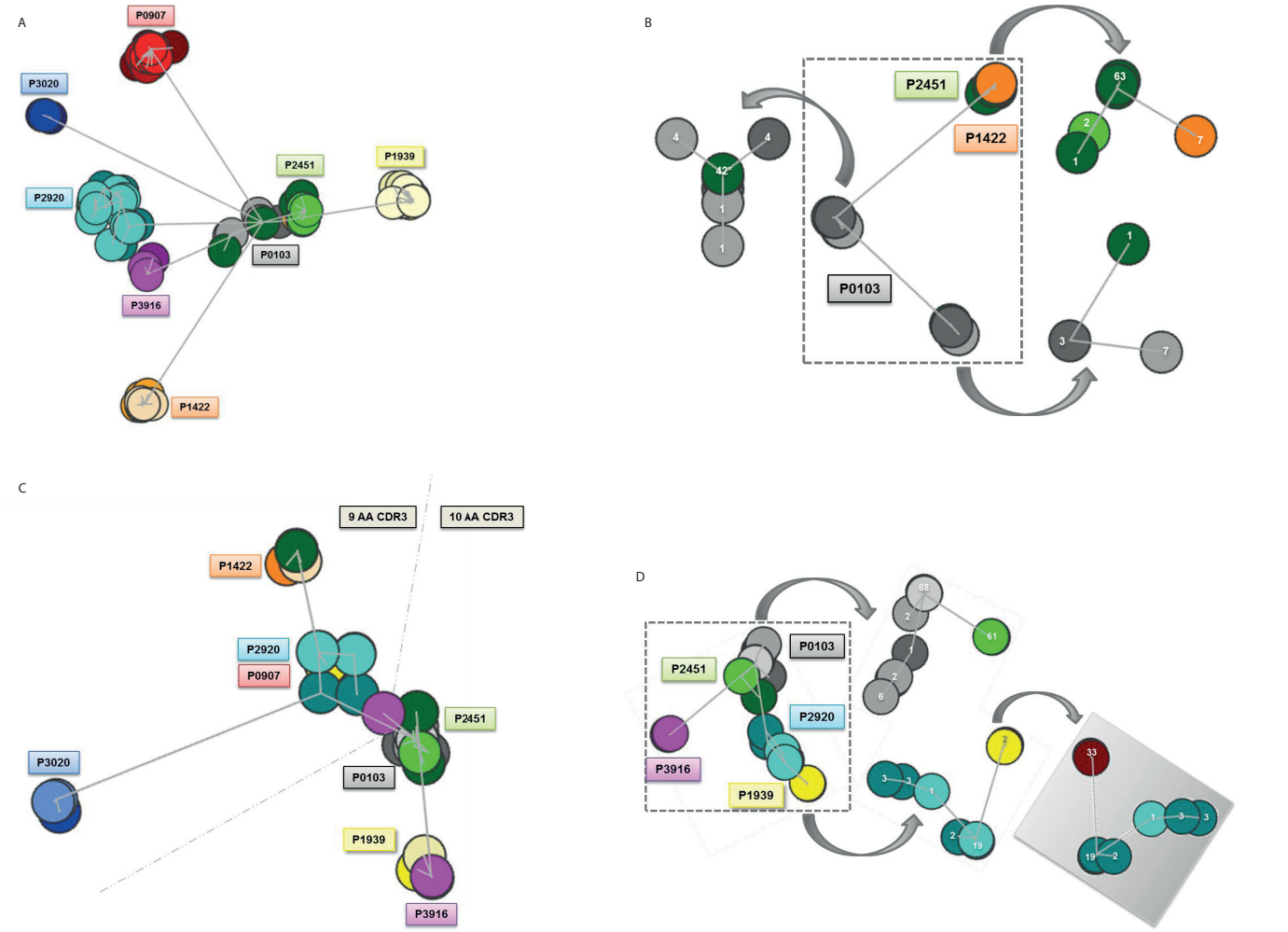

Figure 1. Composite clusters of subset #4 IG sequences at the amino acid level. Figure 1A illustrates cluster formation following analysis of the IGHV–IGHD–IGHJ amino acid sequences (n = 511). Six distinct clusters were observed: a central core was created by clonal sequences from two patients, P0103 and P2451, and from this core radiated a further five clusters. Figure 1B provides a more detailed view of the composition of this central core. The central core is framed by dotted lines and each cluster is then dissected further. The seven sequences from P1422 segregated from the parent cluster were observed initially at diagnosis as a minor subclone, were represented by only a single subcloned sequence at the second time point (1/33 subcloned sequences; 3%) and were undetectable at the third time point. Figure 1C details cluster formation following analysis of the IGKV–IGKJ amino acid sequences (n = 397). Figure 1D provides a more complete view of the major cluster resulting from analysis of the IGKV–IGKJ subclonal sequences. The major cluster is surrounded by dotted lines, and a comprehensive breakdown of each cluster is provided. As observed with the IG heavy chain sequences, we noted that individual IG kappa sequences occasionally were separated from their respective clusters, and instead attached to distant clusters. This was particularly noted for three clonal sequences, one from P1939 and two from P2451, which carried 9-amino acid VK CDR3s while their remaining clonal sequences all carried a 10-amino acid VK CDR3; the longer VK CDR3 is created by an additional proline at codon 115 and an equal proportion of subset #4 cases in this study carried either a 9-amino acid VK CDR3 or a 10-amino acid VK CDR3. During cluster formation, patients with identical sequences become hidden by the last patient to be analyzed and found to harbor the exact same sequence. Thus, while it may initially appear that P0907 is absent from the clustering analysis in Figure 1C, it is merely obscured by another patient. This is illustrated in the reverse image of the P1939 (yellow)/P2920 (blue) cluster provided in Figure 1D, with the subclonal sequences of P0907 indicated by the red circle. Each circle represents subcloned sequences from one of the eight subset #4 patients included in the study. Identical sequences overlap and are thus represented by a single circle. Circles are color coded to match the patient tag and different shades of the same color indicate subclonal sequences from the same patient but from a different time point. The number within each circle indicates how many sequences carried that specific rearrangement. In Figure 1C, subcloned sequences with a 9-amino acid VK CDR3 lie above the dashed gray line while subclonal sequences from patients with a 10-amino acid VK CDR3 lie below the line. The number of circles appearing for each case is related to the level of intraclonal diversification observed. The asterisk beside the number 42 in Figure 1B indicates that this circle represents sequences from more than one patient.

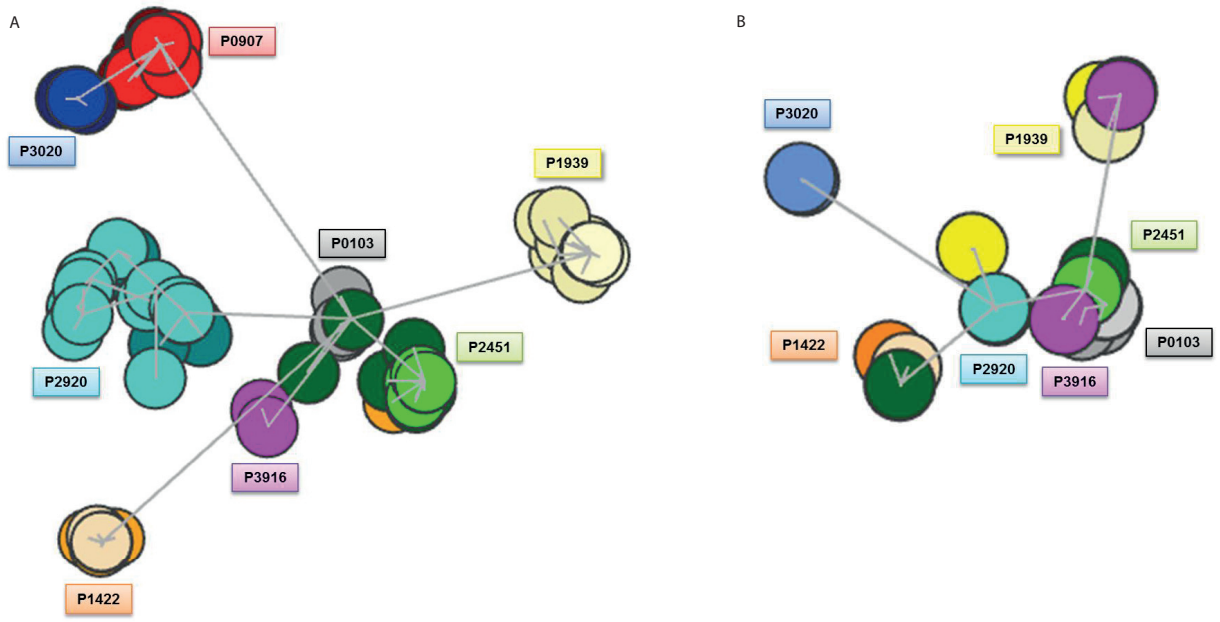

Figure 2. Cluster formation when considering amino acids within the same physicochemical groups as equals. Figure 2A illustrates clustering of the IG heavy chains (n = 511) while figure 2B concerns clustering of the IG kappa light chains (n = 397). When considering amino acids within the same physicochemical groups (as defined by IMGT) as equals, a new cluster was formed at the heavy chain level between P3020 and P0907 (previously represented by two distinct clusters) while the effect on IG kappa light chains was minor and predominantly related to the separation of P2920 and P0907 from the central cluster. Circles are color coded to match the patient tag and different shades of the same color indicate subclonal sequences from the same patient but from a different time point. The number of circles appearing for each case is related to the level of ID observed.

for three additional sequence changes as opposed to one at the amino acid level.

Taken collectively, this detailed computational reconstruction of CLL subset #4 clonal evolution based on merged IG sequence data for all eight cases (at either an amino acid or nucleotide level) reveals a convergent and unified tumorigenic evolutionary process. Thus, this framework is indicative of a "consensus path" of evolution for subset #4 cases with the branched evolutionary growth perhaps reflecting selective pressures honing their BcR affinities.

Tracing the Origins of CLL Subset #4: Molecular Phylogeny of CDR3 Sequences

Both the VH and VK CDR3s were visualized hierarchically with the aim of constructing a CDR3 sequence at both the nucleotide and amino acid level, which then could be considered as the root from which all subset #4 CDR3 sequences derive. Comparison of each CDR3 sequence to the derived root sequence, using the same algorithmic process applied throughout the entire variable domain, enabled us to identify the mutational path followed by each individual patient.

With regards to the VH CDR3, the derived root sequences for both nucleotide and amino acids, were GCG AGA GGC TAC GCG GAT ACA GCT GTG GTT AGG AGG TAC TAC TAT TAC GGT ATG GAC GTC and ARGYADTAVVRRYYYYGMDV, respectively. These sequences would have been created through the association of the *IGHV4-34* and *IGHJ6* genes with the *IGHD5-18* gene in reading frame 1. Within these sequence strings, GGC TAC GCG (translation: GYA) and AGG AGG (translation: RR) cannot be assigned to the germline sequence of any IGHD and/or IGHJ gene, and thus would correspond to nontemplated regions (N1 and N2, respectively). Regarding the VK CDR3, the derived root nucleotide and amino acid sequences were ATG CAA GGC ACA CAC TGG CCC CCG TAC ACT and MQGTHWPPYT, respectively, and would have been created by the association of the *IGKV2-30* and *IGKJ2* genes.

Comparison of the complete VH CDR3 amino acid sequence data set to the root revealed that no patient's sequence exactly matched the root. That said, clonal sequences exhibiting the least differences were from patients P1422, P1939, P2451 and P3916 while P0907, P2920 and P3020 were those located furthest away (Figure 4A). Within the VK CDR3 data set, 41% (162/397) of all kappa light chain sequences were identical to the derived root. These sequences were from patients P0103 (n = 75), P2451 (n = 60) and P3916 (n = 27), while their few remaining sequences together with the subclonal sequences obtained from all other patients, contained only one or two differences, thus explaining the limited branching observed from the root (Figure 4B). Overall, by adopting this strategy we could for the first time propose the

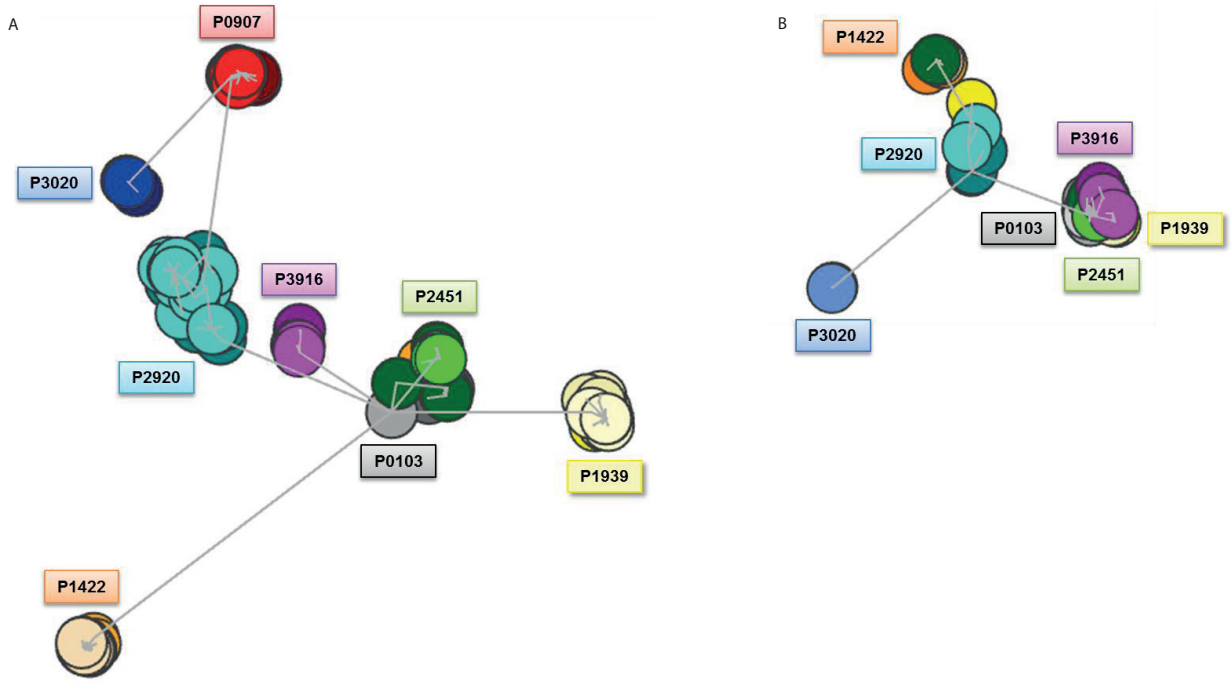

Figure 3. Composite clusters of subset #4 IG sequences at the nucleotide level. Figure 3A illustrates cluster formation following analysis of the IGHV–IGHD–IGHJ nucleotide sequences (n = 511). Seven distinct clusters were observed; six clusters represented a single patient each, while P0103 and P2451 remained clustered together, thus accounting for the seventh cluster. Figure 3B highlights cluster formation following analysis of the IGKV–IGKJ nucleotide sequences (n = 397) and highlights the distancing of the two central cores and instead the formation of a major cluster containing the subcloned sequences of four patients (P0103, P2451, P3916 and P1939). Circles are color coded to match the patient tag and different shades of the same color indicate subclonal sequences from the same patient but from a different time point. The number of circles appearing for each case is related to the level of intraclonal diversification observed.

preimmune VH and VK CDR3 which forms the subset #4 BcR IG.

DISCUSSION

CLL subset #4 lies at the intersection between autoimmunity and malignancy. The expression of *IGHV4-34* endows B cells with the capacity to recognize the N-acetyllactosamine (NAL) antigenic epitope present in both self and exogenous antigens via a germline-encoded motif located within the heavy variable framework region 1 of the *IGHV4-34* gene (22,23). This motif remains intact in all CLL subset #4 IG heavy chain sequences despite a heavy SHM load and intense ID (3,4,10). Notably, recombinant monoclonal antibodies from CLL subset #4 patients have been found to bind viable B cells, recognizing the NAL epitope present on B-cell CD45 (24,25). Additional features encoded in the subset #4 IG BcR sequence that hint at autoreactivity include: (i) the predicted high electropositivity of their long arginine-rich VH CDR3s, reminiscent of pathogenic anti-DNA antibodies; and (ii) the presence of recurrent SHMs typified by the frequent introduction of acidic residues, similar to edited anti-DNA antibodies (3,9).

The route to malignancy for CLL subset #4 clones may thus be a multifactorial phenomenon, beginning with autoreactive precursors that undergo positive selection by DNA, nucleosomes and/or surface structures of apoptotic cells (26,27). Thereafter, modifications introduced by SHM may curtail this autoreactivity, thus rendering these clones anergic (28–30), though still capable of reactivation either through their BcRs and/or other immune receptors, namely toll-like receptors (TLRs) (31–35). While this scenario bodes well for our understanding of the evolutionary pathway followed by subset #4 clones, despite much ingenuity and effort, our knowledge about the specific eliciting antigen(s) for subset #4 remains limited. Along these lines, it is relevant to mention that recombinant monoclonal antibodies derived from subset #4 patients lacked detectable reactivity with DNA, however, upon removal of SHMs (reversion to germline configuration), these antibodies regained the ability to strongly bind DNA (24). Nevertheless, owing to difficulties in defining the unmutated progenitor rearrangement, mainly due to the extensive SHM present within subset #4 clones, the contribution made by the somatically generated CDR3s to auto-antibody specificity (24,25,36–39) may have been underestimated, thus obscuring the actual antibody-antigen interactions (40–42).

In an attempt to clarify and enhance our understanding of the ontogeny of CLL subset #4 B cells, we sought not

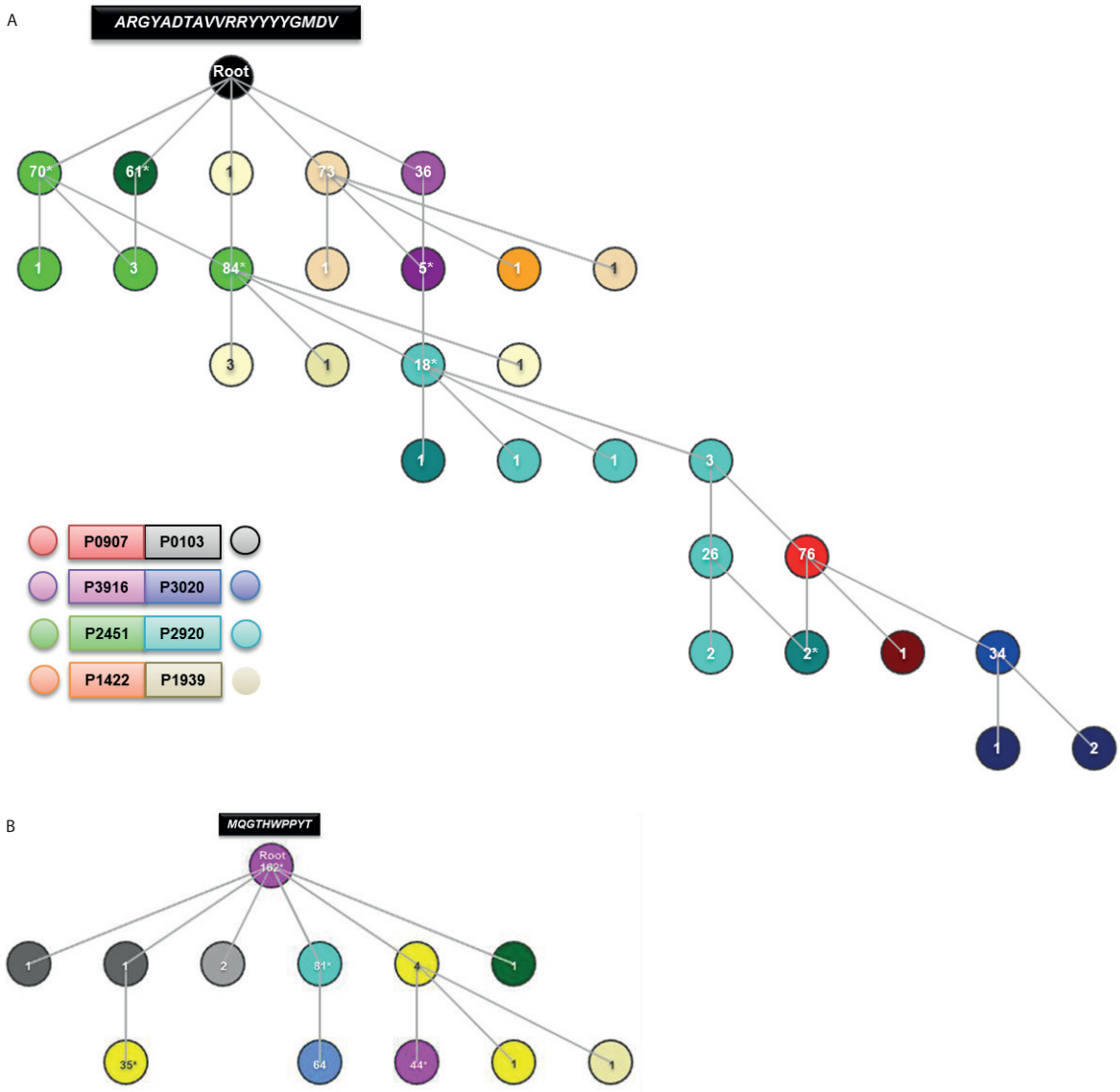

Figure 4. Molecular phylogeny of the VH and VK CDR3 sequences of subset #4. Figure 4A illustrates how hierarchical visualization of the VH CDR3 amino acid sequence from all patients facilitated the construction of a VH CDR3 sequence that can now be considered as the root from which all subset #4 VH CDR3 sequences are derived. Since no patient's VH CDR3 sequence exactly matched the derived root, they are visually placed as branches. Sequences that have an equal number of amino acid changes from the root are placed at the same level (within the same row), since they branch from the root in a similar manner. Figure 4B illustrates the above phenomena for the VK CDR3 amino acid sequences. The limited branching evidenced is indicative of sequence relatedness, with only one or two differences between any patient sequence and the derived root. Identical sequences overlap and are thus represented by a single circle. The asterisk indicates that this circle represents sequences from more than one case. The number within each circle indicates how many sequences carried that specific rearrangement.

only to reconstruct the evolutionary history of subset #4 clones viewed as a single antibody lineage, that is, the sequence of changes introduced into the lineage during the development of the clone, but also to identify the common ancestral sequence from which all subset #4 cases are derived—a task hitherto unattainable due to the heavy SHM load within the antigen-binding sites. One means to obtain insight into the trajectory of subset #4 clones would be through characterization of their genetic sequence, with the greatest insight obtained from longitudinal sampling. Consequently, for this purpose, we drew on a community of related clones profiled at different time points, for both heavy and light chains, derived from 8 subset #4 cases (12). The Damerau–Levenshtein distance algorithm, in the form of a

purpose-built computational tool, was applied and enabled us to infer both the unmutated ancestral rearrangement and the maturation intermediates, and hence gain further insight into the interplay between mutational constraints and selection on antigen-binding affinity.

Through this approach the focused evolution of subset #4, the evolution of single entities into a community of related clones, was clearly evidenced with most patient clusters found lying very close to each other due to a high degree of sequence relatedness. The branching observed within such clusters could perhaps reflect specific selective pressures that occurred in parallel in distinct subclones, as a means to fine-tune their BcR affinities. Importantly, exploring the evolutionary trajectory of subset #4 enabled us to suggest for the first time the common ancestral sequence from which all subset #4 cases likely descend. By determining the most probable sequence of mutations, the mutationally preferred pathway, the unmutated common ancestor (including the predicted VH CDR3) could be inferred, which could now serve as a template for antigen reactivity studies (which should better predict antigen specificities compared with previous studies). Defining the antigens bound by the CLL cells should aid in unraveling the path to malignancy in subset #4. We thus reason that knowledge of the subset #4 ancestral rearrangement could provide a blueprint for the resolution of crystal structures, which would not only further define structural characteristics of the #4 antibody, but also provide detailed molecular insights into the nature of contact sites between the antibody and antigen.

CONCLUSION

The tale of CLL subset #4 is truly intriguing; bestowed with autoreactive properties at birth, they fortuitously escape immunological tolerance and exist in an anergic state in the periphery, only to reemerge as immunocompetent cells (potentially due to dual engagement of the BcRs and TLRs). That said, the story is far from complete and unresolved issues relate to where and under what influence SHM (and also switching to the IgG isotype) occurs, and whether specific modalities of BcR/TLR collaboration and/or regulation may eventually impact on the biological behavior of the clones. Nevertheless, results from this study unveil new leads in the ontogeny of CLL subset #4 clones and bring fresh insights, which may directly impact the design of studies concerning the antigenic specificity of the clonotypic BcR IGs. Although it is difficult to predict how revelations in biological understanding may translate into improved immunological interventions, it seems reasonable to think that once a detailed understanding of the B-cell ontogeny of CLL subset #4 is achieved, doors for therapeutic strategies may open, for example, the design of peptides that would inhibit or alter the consequences of antigen-antibody interactions.

ACKNOWLEDGMENTS

This work was supported in part by the Swedish Cancer Society, the Swedish Research Council, and the Lion's Cancer Research Foundation, Uppsala; the ENosAI project (code 09SYN-13-880) cofunded by the EU and the Hellenic General Secretariat for Research and Technology; the KRIPIS action, funded by the Hellenic General Secretariat for Research and Technology and the European Regional Development Fund of the EU under the O.P. Competitiveness and Entrepreneurship, NSRF 2007-2013.

DISCLOSURE

The authors declare that they have no competing interests as defined by *Molecular Medicine*, or other interests that might be perceived to influence the results and discussion reported in this paper.

REFERENCES

1. Stamatopoulos K, *et al*. (2007) Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: Pathogenetic implications and clinical correlations. *Blood*. 109:259–70.
2. Baliakas P, *et al*. (2014) Clinical effect of stereotyped B-cell receptor immunoglobulins in chronic lymphocytic leukaemia: a retrospective multicentre study. *Lancet Haematol*. 1:e74–84.
3. Murray F, *et al*. (2008) Stereotyped patterns of somatic hypermutation in subsets of patients with chronic lymphocytic leukemia: implications for the role of antigen selection in leukemogenesis. *Blood*. 111:1524–33.
4. Agathangelidis A, *et al*. (2012) Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. *Blood*. 119:4467–75.
5. Vardi A, *et al*. (2014) IgG-switched CLL has a distinct immunogenetic signature from the common MD variant: ontogenetic implications. *Clin*. *Cancer Res*. 20:323–30.
6. Krishnan MR, Jou NT, Marion TN. (1996) Correlation between the amino acid position of arginine in VH-CDR3 and specificity for native DNA among autoimmune antibodies. *J. Immunol*. 157:2430–9.
7. Li Z, Schettino EW, Padlan EA, Ikematsu H, Casali P. (2000) Structure-function analysis of a lupus anti-DNA autoantibody: central role of the heavy chain complementarity-determining region 3 Arg in binding of double- and single-stranded DNA. *Eur*. *J*. *Immunol*. 30:2015–26.
8. Jang YJ, Stollar BD. (2003) Anti-DNA antibodies: aspects of structure and pathogenicity. *Cell*. *Mol*. *Life Sci*. 60:309–20.
9. Hadzidimitriou A, *et al*. (2009) Evidence for the significant role of immunoglobulin light chains in antigen recognition and selection in chronic lymphocytic leukemia. *Blood*. 113:403–11.
10. Sutton LA, *et al*. (2009) Extensive intraclonal diversification in a subgroup of chronic lymphocytic leukemia patients with stereotyped IGHV4–34 receptors: implications for ongoing interactions with antigen. *Blood*. 114:4460–8.
11. Kostareli E, *et al*. (2010) Intraclonal diversification of immunoglobulin light chains in a subset of chronic lymphocytic leukemia alludes to antigen-driven clonal evolution. *Leukemia*. 24:1317–24.
12. Sutton LA, *et al*. (2013) Temporal dynamics of clonal evolution in chronic lymphocytic leukemia with stereotyped IGHV4–34/IGKV2–30 antigen receptors: longitudinal immunogenetic evidence. *Mol*. *Med*. 19:230–6.
13. Hallek M, *et al*. (2008) Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines. *Blood*. 111:5446–56.
14. Dang QT, Phan TH. (2010) Determining Restricted Damerau-Levenshtein Edit-Distance of Two Languages by Extended Automata [Internet]. In: *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2010 IEEE RIVF International Conference on*; 2010 Nov 1–4; Hanoi. [cited 2015 Mar 31]. Available from: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=5632914&queryText%3DDetermining+Restricted+Damerau-Levenshtein
15. Gomez-Alonso C, Valls A. (2008) A Similarity Measure for Sequences of Categorical Data Based

on the Ordering of Common Elements. In: *Modeling Decisions for Artificial Intelligence: 5th International Conference, MDAI 2008, Sabadell, Spain, October 30-31, 2008, Proceedings*. Torra V, Narukawa Y (eds.) Springer, Berlin, pp. 134–45.
16. Schulz MA, Schmalbach B, Brugger P, Witt K. (2012) Analysing humanly generated random number sequences: a pattern-based approach. *PloS One*. 7:e41531.
17. Farris JS. (1972) Estimating phylogenetic trees from distance matrices. *Am. Nat*. 106:645–68.
18. Helmkamp LJ, Jewett EM, Rosenberg NA. (2012) Improvements to a class of distance matrix methods for inferring species trees from gene trees. *J. Comput. Biol*. 19:632–49.
19. Meyer F, Najman L. (2013) Segmentation, minimum spanning tree and hierarchies. In: *Mathematical morphology: From theory to applications*. L N, H T (eds.) John Wiley & Sons, Hoboken, New Jersey, USA, pp. 229–61.
20. Marsh JW, *et al*. (2010) Multilocus variable-number tandem-repeat analysis and multilocus sequence typing reveal genetic relationships among Clostridium difficile isolates genotyped by restriction endonuclease analysis. *J. Clin. Microbiol*. 48:412–8.
21. Erciyes K. (2013) Minimum spanning trees. In: *Distributed graph algorithms for computer networks*. Sammes AJ (ed.) Springer, pp. 69–82.
22. Silberstein LE, George A, Durdik JM, Kipps TJ. (1996) The V4–34 encoded anti-i autoantibodies recognize a large subset of human and mouse B-cells. *Blood Cells Mol. Dis*. 22:126–38.
23. Potter KN, Hobby P, Klijn S, Stevenson FK, Sutton BJ. (2002) Evidence for involvement of a hydrophobic patch in framework region 1 of human V4–34-encoded Igs in recognition of the red blood cell I antigen. *J. Immunol*. 169:3777–82.
24. Catera R, *et al*. (2008) Chronic lymphocytic leukemia cells recognize conserved epitopes associated with apoptosis and oxidation. *Mol. Med*. 14:665–74.
25. Catera R, *et al*. (2006) Polyreactive monoclonal antibodies synthesized by some B-CLL cells recognize specific antigens on viable and apoptotic T cells. *Blood*. 108:2813.
26. Pugh-Bernard AE, *et al*. (2001) Regulation of inherently autoreactive VH4–34 B cells in the maintenance of human B cell tolerance. *J. Clin. Invest*. 108:1061–70.
27. Cappione AJ, Pugh-Bernard AE, Anolik JH, Sanz I. (2004) Lupus IgG VH4.34 antibodies bind to a 220-kDa glycoform of CD45/B220 on the surface of human B lymphocytes. *J. Immunol*. 172:4298–307.
28. Cocca BA, *et al*. (2001) Structural basis for autoantibody recognition of phosphatidylserine-beta 2 glycoprotein I and apoptotic cells. *Proc. Natl. Acad. Sci. U. S. A*. 98:13826–31.
29. Li Y, Li H, Ni D, Weigert M. (2002) Anti-DNA B cells in MRL/lpr mice show altered differentiation and editing pattern. *J. Exp. Med*. 196:1543–52.
30. Behrendt M, *et al*. (2003) The role of somatic mutation in determining the affinity of anti-DNA antibodies. *Clin. Exp. Immunol*. 131:182–9.
31. Kostareli E, *et al*. (2009) Molecular evidence for EBV and CMV persistence in a subset of patients with chronic lymphocytic leukemia expressing stereotyped IGHV4–34 B-cell receptors. *Leukemia*. 23:919–24.
32. Ntoufa S, *et al*. (2012) Distinct innate immunity pathways to activation and tolerance in subgroups of chronic lymphocytic leukemia with distinct immunoglobulin receptors. *Mol. Med*. 18:1281–91.
33. Duty JA, *et al*. (2009) Functional anergy in a subpopulation of naive B cells from healthy humans that express autoreactive immunoglobulin receptors. *J. Exp. Med*. 206:139–51.
34. Andrews SF, Wilson PC. (2010) The anergic B cell. *Blood*. 115:4976–8.
35. Chatzouli M, *et al*. (2014) Heterogeneous functional effects of concomitant B cell receptor and TLR stimulation in chronic lymphocytic leukemia with mutated versus unmutated Ig genes. *J. Immunol*. 192:4518–24.
36. Herve M, *et al*. (2005) Unmutated and mutated chronic lymphocytic leukemias derive from self-reactive B cell precursors despite expressing different antibody reactivity. *J. Clin. Invest*. 115:1636–43.
37. Chu CC, *et al*. (2008) Chronic lymphocytic leukemia antibodies with a common stereotypic rearrangement recognize nonmuscle myosin heavy chain IIA. *Blood*. 112:5122–9.
38. Chu CC, *et al*. (2010) Many chronic lymphocytic leukemia antibodies recognize apoptotic cells with exposed nonmuscle myosin heavy chain IIA: implications for patient outcome and cell of origin. *Blood*. 115:3907–15.
39. Zwick C, *et al*. (2013) Autoantigenic targets of -cell receptors derived from chronic lymphocytic leukemias bind to and induce proliferation of leukemic cells. *Blood*. 121:4708–17.
40. Wellmann U, *et al*. (2005) The evolution of human anti-double-stranded DNA autoantibodies. *Proc. Natl. Acad. Sci. U. S. A*. 102:9258–63.
41. Lambrianides A, *et al*. (2007) Arginine mutation alters binding of a human monoclonal antibody to antigens linked to systemic lupus erythematosus and the antiphospholipid syndrome. *Arthritis Rheum*. 56:2392–401.
42. Zhang J, Jacobi AM, Wang T, Diamond B. (2008) Pathogenic autoantibodies in systemic lupus erythematosus are derived from both self-reactive and non-self-reactive B cells. *Mol. Med*. 14:675–81.