MolecularMedicine

Partial versus Productive *Immunoglobulin Heavy* Locus Rearrangements in Chronic Lymphocytic Leukemia: Implications for B-Cell Receptor Stereotypy

Eugenia Tsakou,[1,2] Andreas Agathagelidis,[3,4] Myriam Boudjoghra,[5] Thorsten Raff,[6] Antonis Dagklis,[3] Maria Chatzouli,[2] Tatjana Smilevska,[4] George Bourikas,[1] Helene Merle-Beral,[5] Eleni Manioudaki-Kavallieratou,[2] Achilles Anagnostopoulos,[3] Monika Brüggemann,[6] Frederic Davi,[5] Kostas Stamatopoulos,[3,4] and Chrysoula Belessi[2]

[1]Hematology Department, Democritus University of Thrace, Alexandroupolis, Greece; [2]Hematology Department, Nikea General Hospital, Piraeus, Greece; [3]Institute of Agrobiotechnology, Center for Research and Technology Hellas, Thessaloniki, Greece; [4]Hematology Department and Hematopoietic Cell Transplantation Unit, G. Papanicolaou Hospital, Thessaloniki, Greece; [5]Laboratory of Hematology and Université Pierre et Marie Curie, Hôpital Pitié-Salpètrière, Paris, France; and [6]II. Medizinische Klinik und Poliklinik, Universitätsklinikum Schleswig-Holstein, Kiel, Germany

The frequent occurrence of stereotyped heavy complementarity-determining region 3 (VH CDR3) sequences among unrelated cases with chronic lymphocytic leukemia (CLL) is widely taken as evidence for antigen selection. Stereotyped VH CDR3 sequences are often defined by the selective association of certain immunoglobulin heavy diversity (*IGHD*) genes in specific reading frames with certain *immunoglobulin heavy joining* (*IGHJ*) genes. To gain insight into the mechanisms underlying VH CDR3 restrictions and also determine the developmental stage when restrictions in VH CDR3 are imposed, we analyzed partial *IGHD-IGHJ* rearrangements (D-J) in 829 CLL cases and compared the productively rearranged D-J joints (that is, in-frame junctions without junctional stop codons) to (a) the productive *immunoglobulin heavy variable* (*IGHV*)*-IGHD-IGHJ* rearrangements (V-D-J) from the same cases and (b) 174 D-J rearrangements from 160 precursor B-cell acute lymphoblastic leukemia cases (pre-B acute lymphoblastic leukemia (ALL)). Partial D-J rearrangements were detected in 272/829 CLL cases (32.8%). Sequence analysis was feasible in 238 of 272 D-J rearrangements; 198 of 238 (83.2%) were productively rearranged. The D-J joints in CLL did not differ significantly from those in pre-B ALL, except for higher frequency of the *IGHD7-27* and *IGHJ6* genes in the latter. Among CLL carrying productively rearranged D-J, comparison of the *IGHD* gene repertoire in productive V-D-J versus D-J revealed the following: (a) overuse of IGHD reading frames encoding hydrophilic peptides among V-D-J and (b) selection of the *IGHD3-3* and *IGHD6-19* genes in V-D-J junctions. These results document that the *IGHD* and *IGHJ* gene biases in the CLL expressed VH CDR3 repertoire are not stochastic but are directed by selection operating at the immunoglobulin protein level.

Online address: http://www.molmed.org
doi: 10.2119/molmed.2011.00216

INTRODUCTION

A major mechanism for the generation of a broad repertoire of human antibody specificities is the rearrangement of one each from a cluster of multiple and distinct immunoglobulin (IG) heavy and light variable (V), diversity (D: for heavy chains only) and joining (J) genes (1). This combinatorial diversity is further increased by the junctional diversity that results from differential trimming of nucleotides from the ends of the rearranged genes (exonuclease activity) along with the addition of random nucleotides between them (terminal deoxynucleotidyl transferase [TdT] activity) (2,3). This step ensures that the somatically created complementarity-determining region 3 (CDR3) is the most diverse part of the variable domain.

The CDR3 of the V domain of the heavy chain (VH) plays a critical role in antigen recognition to the extent that the more similar the primary VH CDR3 sequences of two IGs, the more similar their folding and, likely, their specificities (4). Thanks to the accumulated effects of both combinatorial and junctional diversity, normally, the chances

Address correspondence to *Kostas Stamatopoulos, Hematology Department and HCT Unit, G. Papanicolaou Hospital, 57010 Exokhi, Thessaloniki, Greece. Phone: +30-2313-307992; Fax: +30-2313-307579. E-mail: kostas.stamatopoulos@gmail.com.*
Submitted June 23, 2011; Accepted for publication September 26, 2011; Epub (www.molmed.org) ahead of print September 27, 2011.

The Feinstein Institute for Medical Research North Shore LIJ

that two independent B-cell clones will carry identical VH CDR3 are considered as negligible.

Against this background, chronic lymphocytic leukemia (CLL) is uniquely characterized by the existence of subsets of cases with quasi-identical (stereotyped) VH CDR3 sequences within their B-cell receptors (BcRs) that collectively account for up to 30% of the cohort (5,6). Such striking BcR similarity in unrelated CLL cases has been widely interpreted as evidence for a role for a limited set of antigens or structurally related epitopes in leukemogenesis (5,7–10).

VH CDR3 formation starts at the early stages of B-cell development by the rearrangement of an *immunoglobulin heavy constant delta* (*IGHD*) to an i*mmunoglobulin heavy joining group* (*IGHJ*) gene, followed by the rearrangement of an *immunoglobulin heavy variable group* (*IGHV*) gene to the partially rearranged *IGHD-IGHJ* genes. Partial *IGHD-IGHJ* rearrangements (D-J) hallmark the differentiation of common lymphoid precursors into the B-cell lineage and are already present in CD34$^+$/CD19$^-$/CD10$^+$ precursor cells (11,12). Partial *IGHD-IGHJ* rearrangements (even if productively rearranged) do not always recombine to an *IGHV* gene. In this case, the *IGHD-IGHJ* remains as a partial D-J rearrangement in the *immunoglobulin heavy* (*IGH)* locus. The biological mechanism, which determines whether a partial *IGHD-IGHJ* rearrangement will develop into an *IGHV-IGHD-IGHJ* rearrangement is still unknown.

Studies in the mouse have shown a biased usage of *IGHD* genes in reading frames (RFs) encoding for hydrophilic amino acids in *IGHV-IGHD-IGHJ* rearrangements (V-D-J) along with counter-selection of *IGHD-IGHJ* rearrangements with *IGHD* genes in RF encoding for hydrophobic amino acids or stop codons (13–18) . Relevant studies have also shown that *IGHD-IGHJ* with *IGHD* genes rearranged in RF encoding for hydrophobic amino acids can be translated into a truncated μ protein, called Dμ (13,19). The Dμ protein was found to be expressed on the cell surface of murine pre-B cells assembled with surrogate light chains and the coreceptors CD79A (Igα) and CD79B (Igβ). This defective pre-B receptor acts as a signal transducing receptor and may interfere with B-cell development by inhibiting further *IGHD-IGHJ* replacements as well as *IGHV-IGHD-IGHJ* rearrangements (20).

Partial *IGHD-IGHJ* rearrangements are readily found in pre-B cells. They have also been reported in a proportion of both precursor B-cell acute lymphoblastic leukemia cases (21) and mature B-cell malignancies (22–24). Accumulating evidence from previous reports suggests that these rearrangements may offer a glimpse into the ontogenetic history of normal and malignant B cells as well as the mechanisms underlying the diverse IG repertoires observed at different developmental stages or different B-cell neoplasms.

With this in mind, we here analyzed the molecular features of D-J rearrangements in a series of 829 patients with CLL, which we subsequently compared with (a) the productive V-D-J from the same CLL clones and (b) the D-J obtained from a series of 160 patients with acute lymphoblastic leukemia (ALL), selected for comparison as representative of a malignancy of early B cells, thus contrasting CLL, a malignancy of mature B cells. Our aim was to gain insight into the mechanisms that shape the IG repertoire and determine the developmental stage when VH CDR3 restrictions are imposed in CLL.

MATERIALS AND METHODS

Patient Group

Blood samples were collected from 829 consecutive typical CLL cases from collaborating institutions in France and Greece who met the International Workshop on CLL/National Cancer Institute (iwCLL/NCI) diagnostic criteria (25). Blood and/or bone marrow samples were also collected from 160 cases with B-cell lineage ALL from Germany, diagnosed according to World Health Organization criteria and immunological marker analysis and selected on the basis of carrying partial *IGHD-IGHJ* rearrangements, amplified with BIOMED-2 protocols (26). The study was approved by the local ethics review committee of each institution. In all CLL cases, the tumor load was at least 80%.

Polymerase Chain Reaction Amplification and Sequence Analysis of *IGHV-IGHD-IGHJ* Gene Rearrangements

Mononuclear cells were separated from blood samples (and/or bone marrow) by centrifugation on a Ficoll-Hypaque gradient. Polymerase chain reaction (PCR) amplification of *IGHV-IGHD-IGHJ* gene rearrangements was performed on either genomic DNA (gDNA) or complementary DNA (cDNA), as previously described (27). Purified PCR amplicons were subjected to direct sequencing on both strands. Sequence data were analyzed using the IMGT® databases and the IMGT/V-QUEST tool (http://www.imgt.org) (28,29).

PCR Amplification and Sequence Analysis of Partial *IGHD-IGHJ* Gene Rearrangements

PCR amplification of partial *IGHD-IGHJ* gene rearrangements was performed on gDNA using seven subgroup-specific IGHD primers in combination with a consensus IGHJ primer, following the BIOMED-2 experimental protocol (26). Primers were designed to anneal to the intron upstream of the *IGHD* gene in distinct positions relative to the recombination signal sequence (RSS) elements for each *IGHD* subgroup, resulting in PCR product sizes specific for each IGHD subgroup, ranging from 100 to 420 bp. Primer annealing to intron sequences precludes the possibility of amplification of the *IGHD-IGHJ* joint of the coexpressed *IGHV-IGHD-IGHJ* rearrangement (26). In all cases, the clonal amplification PCR product was easily discriminated from the polyclonal "background” (if present) obtained of residual normal B cells that was faint and did not hinder the detection of a predominant PCR band.

Purified PCR products were subjected to direct sequencing on both strands, and sequence analysis was performed by a multistep procedure using the Basic Local Alignment Search Tool (BLAST, http://blast.ncbi.nlm.nih.gov), the Expert Protein Analysis System (ExPASy, http://au.expasy.org) and databases, algorithms and tools from IMGT®, the international ImMunoGeneTics information system (http://www.imgt.org).

As a first step, we identified the rearranged *IGHD* and *IGHJ* genes by aligning each sequence to the entire *IGHD-IGHJ* germline sequence (GenBank accession number EMB/X97051) using the “align2sequences” tool of “specialized BLAST.” The imprint of exonuclease and TdT activity in the D-J joint was determined by comparing each sequence to the corresponding germline *IGHD* and *IGHJ* gene according to the IMGT rules for junction analysis of the immunoglobulin genes (http://www.imgt.org/IMGT_jcta/share/textes/userGuide.html#VDJends). The patterns “m,” “m-“ and “‘mm—” were trimmed from the 3′ end of the D-REGION, the patterns “m,” “-m” and “—mm” were trimmed from the 5′ end of the J-REGION, where “m” indicates a mutation and “-” indicates an identical nucleotide by comparison with the corresponding germline sequences, ensuring that at least two nucleotides at each end of the *IGHD* and *IGHJ* genes are identical to germline. Following this rule, we then determined all nucleotide changes in these sequences that met the criteria for being considered as mutations.

Subsequently, each nucleotide sequence was translated in all RFs using the “translate tool” from ExPASy. Given that, in a productive rearrangement, the *IGHJ* gene must be read in the RF that encodes the motif WGXG (where “W” is tryptophan, “G” is glycine, “X” is any amino acid and “G” is glycine, respectively), each amino acid sequence carrying the WGXG motif was further compared with the amino acid sequences of the corresponding *IGHD* gene in all three RFs, to define the proper IGHD RF in each rearrangement. According to IMGT®, IGHD RF1/RF2/RF3 was counted from the first/second/third nucleotide of the D-REGION (that follows the 5′D-HEPTAMER), respectively.

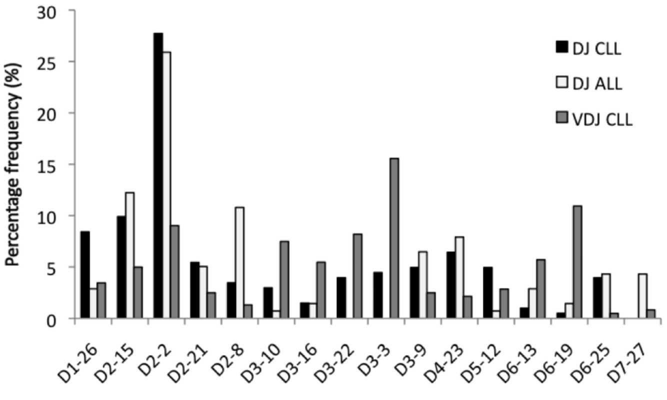

Figure 1. Differential usage of selected *IGHD* genes in D-J in CLL, D-J in ALL and productive V-D-J in CLL.

Rearrangements carrying a stop codon in the *IGHD-IGHJ* joint were considered as unproductive (following the IMGT definition of the functionality), since they could not lead to a productive *IGHV-IGHD-IGHJ* rearrangement. Productively rearranged D-J joints with a stop codon within the *IGHD* gene sequence were retained in the analysis, since they could eventually lead to a productive *IGHV-IGHD-IGHJ* rearrangement after removal of the stop codon during *IGHV* to *IGHD-IGHJ* recombination (for example, by exonuclease trimming of the *IGHD* gene, a feature often seen in productive rearrangements from normal, autoreactive or malignant B-cell clones, including CLL).

Statistical Analysis

Descriptive statistics for discrete parameters included counts and frequency distributions. For quantitative variables, statistical measures included means, medians, standard deviation and min–max values. Significance of bivariate relationships between variables was assessed using χ^2 and *t* tests. For all comparisons, a significance level of *P* at 0.05 was set, and all statistical analyses were performed with the statistical package SPSS version 17.0 (SPSS, Chicago, IL, USA).

All supplementary materials are available online at www.molmed.org.

RESULTS

Molecular Features of Productive IGHV-IGHD-IGHJ Rearrangements in CLL

Productive *IGHV-IGHD-IGHJ* rearrangements (V-D-J) were successfully sequenced in all 829 CLL cases included in the study. A total of 12 of 829 CLL cases (1.4%) carried double productive *IGHV-IGHD-IGHJ* rearrangements. Using the 98% cutoff value for identity to the closest germline gene, 455 of 842 (54%) *IGHV* sequences were considered as “mutated,” whereas the remaining 387 of 842 (46%) were considered as “unmutated” (283 of 387 unmutated sequences had 100% identity to germline). In keeping with previous reports (30,31), the most frequent *IGHV* genes in the mutated subgroup were *IGHV4-34*, *IGHV3-23* and *IGHV3-7*, whereas the *IGHV1-69*, *IGHV1-2* and *IGHV4-39* genes predominated in the unmutated subgroup (Supplementary Table 1). Following previously described criteria (6), overall, 215 of 829 cases (25.9%) were assigned to different subsets with stereotyped BcR.

Table 1. *IGHD* gene RF usage and hydropathicity in productively rearranged *IGHD-IGHJ* (D-J) versus productive *IGHV-IGHD-IGHJ* (V-D-J) in CLL and ALL.

	D-J in CLL	V-D-J in CLL[a]	D-J in ALL	V-D-J in ALL[a]
IGHD RF				
RF1	73 (36.1)	40 (20.8)	48 (34.5)	14 (28.6)
RF2	64 (31.7)	100 (52.1)[b]	43 (31.0)	17 (34.7)
RF3	65 (32.2)	52 (27.1)	48 (34.5)	18 (36.7)
IGHD RF				
Hydrophilic RF	70 (34.6)	118 (61.5)[b]	51 (36.7)	18 (36.7)
Hydrophobic RF	69 (34.2)	51 (26.6)	45 (32.3)	22 (44.9)
RF with stop codon(s)	63 (31.2)	23 (11.9)	43 (31.0)	9 (18.4)

Data are n (%).

[a] In 6 of 198 and 4 of 53 V-D-J in CLL and ALL, respectively, the *IGHD* gene was not determined.

[b] Statistically significant higher frequency for RF2 and hydrophilic RF was observed in productive V-D-J junctions versus productively rearranged D-J joints in CLL ($P < 0.001$ for both comparisons).

IGHD3 subgroup genes predominated among V-D-J, followed by *IGHD2* subgroup genes. At the individual gene level, *IGHD3-3* was the most frequent gene (131/842 V-D-J, 15.6%), followed by the *IGHD6-19* (92/842 V-D-J, 10.9%) and *IGHD2-2* genes (76/842 V-D-J, 9%) (Supplementary Table 2) (Figure 1). Among the *IGHJ* genes, *IGHJ4* and *IGHJ6* were the most frequently used genes in V-D-J (365/842, 43.3%, and 244/842, 29%, respectively) (Supplementary Table 3).

Molecular Features of Partial *IGHD-IGHJ* Rearrangements in CLL

Partial *IGHD-IGHJ* rearrangements (D-J) were detected in 272 of 829 patients with CLL (32.8%). No significant associations were identified regarding the frequency of D-J in relation to *IGHV* gene usage or mutational status or BcR stereotypy observed in the productive rearrangement from the other allele, thereby indicating no intrinsic differences between CLL cases with different BcR.

Sequence analysis was feasible in 238 of 272 D-J, of which 198 (83.2%) were productively rearranged (no stop codon at the *IGHD-IGHJ* joint). Five D-J carried readily identifiable *IGHD-IGHD* gene fusions (Supplementary Table 4), four of which were productively rearranged. Overall, 21 different *IGHD* genes were identified in 198 productively rearranged D-J.

IGHD2 subgroup genes predominated among productively rearranged D-J, with *IGHD2-2* being the far most frequent gene (56 of 198 D-J, 28.3%), followed by *IGHD2-15* (19 of 198 D-J, 9.6%) (Figure 1; Supplementary Table 5). The *IGHJ4* and *IGHJ6* genes were used at similar frequencies, collectively accounting for 65.6% of the series (130 of 198 cases; Supplemental Table 6). All three RFs (RF 1, 2 and 3) of the *IGHD* genes were equally distributed among productively rearranged D-J (Table 1).

Additionally, no difference was identified regarding the usage of hydrophobic RF versus hydrophilic RF versus RF with stop codon(s) (Table 1). A total of 10 of 63 cases with the *IGHD* gene read in an RF with stop codon(s) had lost the germline-encoded stop codon during *IGHD-IGHJ* rearrangement. Evidence for TdT, 5′- and 3′-exonuclease activities was identified in most cases (Table 2).

Sequence changes in the form of point mutations were detected in 24 of 198 D-J (12.1%): in particular, 22 cases carried a single mutation, whereas the remaining two cases carried two mutations. Of 24 cases with mutations in D-J, 20 carried somatically hypermutated *IGHV* genes in the corresponding V-D-J, whereas the remaining four cases carried coding *IGHV* genes with 100% identity to germline.

Comparative Assessment of Partial *IGHD-IGHJ* versus Productive *IGHV-IGHD-IGHJ* Rearrangements in CLL

For all subsequent analyses, we focused on the subgroup of 198 cases carrying a productively rearranged D-J on one allele along with a productive V-D-J on the other allele and performed a detailed comparative assessment of the D-J joints and V-D-J junctions. Regarding the

Table 2. TdT and exonuclease activity in productively rearranged *IGHD-IGHJ* (D-J) versus *IGHV-IGHD-IGHJ* (V-D-J) in CLL and ALL.

	D-J in CLL	V-D-J in CLL	D-J in ALL	V-D-J in ALL	D-J versus V-D-J in CLL (*P*)	D-J versus V-D-J in ALL (*P*)	D-J in CLL versus ALL (*P*)
TdT activity							
Positivity (%)	96.9	92.0	90.0	89.8	0.03	0.92	0.01
Median number of nucleotides (range)	9 (1–76)	7 (1–39)	8 (1–34)	6 (1–81)	0.01	0.96	0.13
3′ *IGHD* exonuclease activity							
Positivity (%)	87.6	81.8	88.5	85.7	0.06	0.98	0.52
Median number of nucleotides (range)	4 (1–24)	4 (1–19)	4 (1–25)	5 (1–16)	0.74	0.12	0.26
5′ *IGHJ* exonuclease activity							
Positivity (%)	86.6	86.6	85.4	77.6	0.99	0.08	0.84
Median number of nucleotides (range)	5 (1–27)	5 (1–29)	6 (1–25)	5 (1–16)	0.39	0.06	0.06

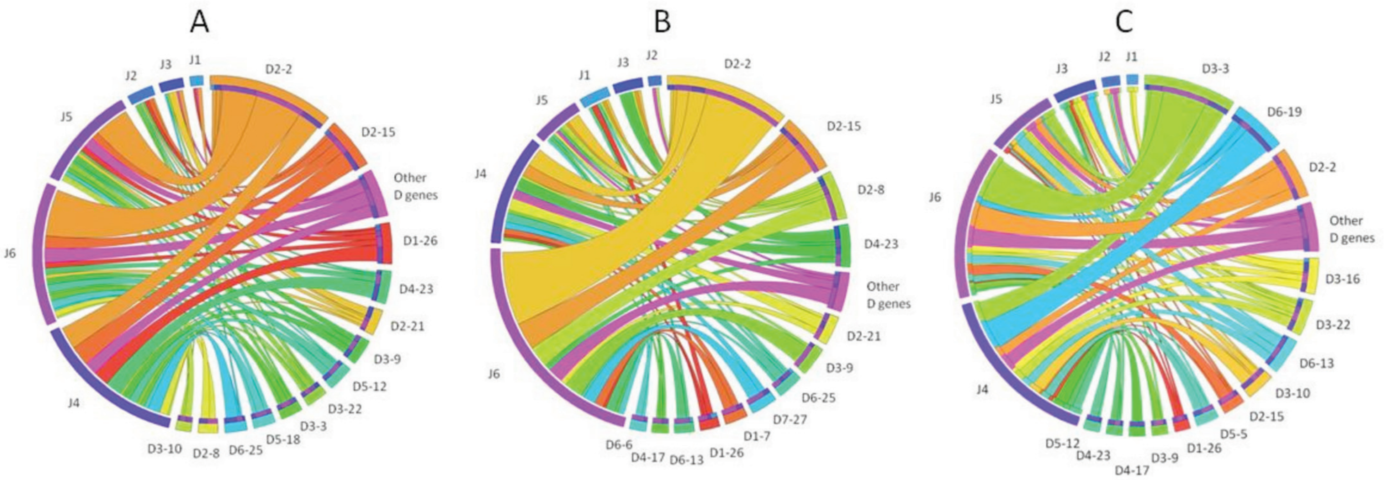

Figure 2. Graphical representation of *IGHD* and *IGHJ* gene associations in CLL and ALL. (A) Partial *IGHD-IGHJ* (D-J) rearrangements in CLL. (B) Partial *IGHD-IGHJ* (D-J) rearrangements in ALL. (C) Productive *IGHV-IGHD-IGHJ* (V-D-J) rearrangements in CLL. The graphs were prepared by using the Circos software package (http://mkweb.bcgsc.ca/circos). Strong biases are evident in the case of (a) the *IGHD2-2* gene, which is very frequently recombined to the *IGHJ6* gene in D-J; (b) the *IGHD3-3* gene, which is recurrently recombined to the *IGHJ6* gene in V-D-J; and (c) the increased usage of the *IGHD7-27* and *IGHJ6* genes in D-J in ALL.

molecular features of rearranged *IGHD* genes in V-D-J, a preferential usage of *IGHD* genes in RF2 versus RF1 or RF3 and hydrophilic RF versus hydrophobic RF or RF with stop codon(s) was recorded. Evidence for TdT, 5′- and 3′-exonuclease activities was found in most cases (Table 2). Comparison of the productively rearranged D-J joints versus productive V-D-J junctions revealed that, except for exonuclease activities, which were similar in both D-J and V-D-J, the D-J and V-D-J differed significantly.

In particular, TdT activity was recorded in a significantly higher proportion of D-J versus V-D-J ($P = 0.03$); additionally, a higher number of inserted nucleotides was found in D-J versus V-D-J ($P = 0.01$) (Table 2). Furthermore, certain genes were overrepresented (*IGHD3-3*, *IGHD6-19*, *IGHD3-10*) or underrepresented (*IGHD2-2*, *IGHD2-15*, *IGHD1-26*, *IGHD4-23*) in V-D-J versus D-J ($P < 0.01$ for all comparisons) (Figure 1). All genes overrepresented in V-D-J versus D-J exhibited biased usage of a single RF (*IGHD3-3*: RF2; *IGHD3-10*: RF2; *IGHD6-19*: RF3), whereas no such bias was observed in the (few) D-J using the same genes ($P < 0.001$ for all comparisons). Additionally, a significantly higher frequency ($P < 0.001$) of hydrophilic RF was recorded among V-D-J (Table 1).

The evident selection for the *IGHD3-3* gene in productive V-D-J in CLL prompts a closer examination of its properties. As previously shown for productive rearrangements in CLL (32,33), this gene is preferentially used in RF2, thus encoding for a relatively long peptide sequence (10 amino acids: YYDFWSGYYT) with a predicted acidic isoelectric point value (3.8). Additionally, rearrangements carrying the *IGHD3-3* gene are biased to the usage of the *IGHJ6* gene (YYYYYGMDV) (32,33), leading to the formation of long VH CDR3 enriched in tyrosine residues (Figure 2). In principle, increased usage of any single amino acid restricts the overall VH CDR3 variability (34). However, tyrosine may be considered as an exception in that it allows for many forms of inter- and intramolecular bonding that enable aromatic stacking interactions and, thus, increase VH CDR3 flexibility and, likely, the range of antigenic epitopes that can be "accommodated" in the antigen-binding loop (34). Interestingly, usage of *IGHD3-3* in RF2 among V-D-J, especially those using the *IGHV1-69* gene, has been reported as an adverse prognostic indicator in CLL (32).

Molecular Configuration of Partial *IGHD-IGHJ* Rearrangements in ALL and Comparison to CLL

A total of 160 cases with B-cell lineage ALL were selected for analysis on the basis of carrying a partial *IGHD-IGHJ* rearrangement. Overall, within this cohort, 174 D-J and 74 V-D-J were successfully analyzed. Fourteen of 160 cases carried double D-J, whereas 9 cases carried double V-D-J. Fifteen D-J bore multiple *IGHD* genes in the form of *IGHD-IGHD* gene fusions. In 13 of 15 such junctions, two *IGHD* genes were identified, whereas the remaining two junctions bore three *IGHD* genes (Supplementary Table 7).

In 10 ALL cases, the same *IGHD-IGHJ* joint was identified in both their D-J and V-D-J, indicative of clonal evolution (Supplementary Table 8); four of these cases carried *IGHD-IGHD* gene fusions. Cases with evidence of clonal evolution were excluded from further repertoire analysis.

A total of 30 D-J carried a stop codon in the *IGHD-IGHJ* joint (including seven of the remaining 11 cases with *IGHD-IGHD* gene fusions). Among the remaining 134 productively rearranged D-J, *IGHD2* subgroup genes predominated, with *IGHD2-2* being the most frequent gene (36 of 134 D-J, 26.9%) followed by *IGHD2-15* (16 of

134 D-J, 11.9%) (Figure 1; Supplementary Table 9). *IGHD7-27*, the most *IGHJ*-proximal gene (that is, of all *IGHD* genes the most proximal to the *IGHJ* gene cluster), was identified in 6 of 134 (4.5%) cases. *IGHJ6* was by far the predominant gene (73 of 134 D-J, 54.5%); collectively, *IGHJ4*, *IGHJ5* and *IGHJ6* were found in 118 of 134 D-J (88.1%) (Supplementary Table 10). All three RFs of the *IGHD* genes were equally distributed among the productively rearranged D-J, and there was no difference in the usage of hydrophilic RF versus hydrophobic RF versus RF with stop codon(s) (Table 1). Five of 43 cases with the *IGHD* gene read in an RF with stop codon(s) had lost the germline-encoded stop codon during *IGHD-IGHJ* rearrangement. Evidence for TdT, 5′- and 3′-exonuclease activities was identified in most cases (Table 2). Single sequence changes in the form of point mutations were detected in 9 of 134 D-J (6.7%); in all cases, the mutation was located within the *IGHD* gene sequence.

A total of 53 of 134 cases (39.5%) carrying D-J on one IGH allele also carried a V-D-J on the second IGH allele; only 19 of 53 V-D-J (36%) were in-frame. In keeping with the literature (21), *IGHV6-1*, the most *IGHJ*-proximal *IGHV* gene (that is, of all *IGHV* genes the most proximal to the *IGHJ* gene cluster), was overrepresented, being used in 11 of 53 V-D-J (20.7%). Detailed information on *IGHV*, *IGHD* and *IGHJ* gene usage is given in Supplementary Tables 11–13. Mutations (1–3/case) were present within the *IGHV* gene in 8 of 53 V-D-J (15.1%), of which four carried out-of-frame junctions.

Comparison of V-D-J versus D-J in regard to the molecular characteristics of the junctions (preferred IGHD RF, TdT or exonuclease activities) revealed no significant differences. In contrast, *IGHD3* subgroup genes predominated in V-D-J over D-J (18 versus 3 of 53 cases, $P = 0.02$, and 7 versus 2 of 19 cases with productive V-D-J, $P = 0.05$), whereas the opposite was observed for *IGHD2* subgroup genes (18 versus 29 of 53 cases, $P = 0.03$, and 5 versus 12 of 19 cases with productive V-D-J, $P = 0.03$).

Comparison of the productively rearranged D-J in B-cell lineage ALL versus CLL demonstrated few significant differences: (a) preferential usage of the *IGHD7-27* and *IGHJ6* genes in ALL ($P < 0.001$) and (b) lower frequency of cases with TdT activity in D-J in ALL ($P = 0.01$) (Table 2).

DISCUSSION

CLL is distinctive for the existence of subsets of cases with identical or highly homologous (stereotyped) VH CDR3 sequences that account for up to 30% of the cohort (5,6). This finding strongly indicates a role for selection by antigen in shaping the expressed CLL IG repertoire (5,9). Although this is a reasonable conclusion, also supported by several lines of recent research (7–9), the possibility that CLL progenitors might exhibit inherently biased *IGHD-IGHJ* rearrangements, thus eventually accounting for restricted VH CDR3 motifs, has not been formally tested.

We addressed this issue from a novel perspective by assessing in parallel the molecular configuration of partial D-J versus productive V-D-J rearrangements in a cohort of 829 CLL cases. To gain better insight into the mechanisms that shape the IG repertoire and determine more accurately the developmental stage when VH CDR3 restrictions are imposed in CLL, we compared the CLL profiles to those observed in ALL, taking advantage of a large series of ALL cases carrying D-J rearrangements. This approach uniquely enabled to document significant differences between productive V-D-J and productively rearranged D-J in CLL, which, together with the few observed differences among D-J in CLL versus ALL, indicate that the expressed IG repertoire in CLL is not stochastically shaped, but rather seems to be directed by selection operating at the IG protein level.

Previous reports have demonstrated the existence of D-J in various B-cell malignancies (namely, ALL, multiple myeloma, hairy cell leukemia) at frequencies ranging from 22% to 66% (21–24). In all reported series, the number of D-J rearrangements eventually analyzed was relatively small (36–50 cases), precluding definitive conclusions, especially with regard to gene usage among different entities. Here, we document that partial *IGHD-IGHJ* rearrangements can be detected in roughly one-third of CLL cases. In addition, we provide a more comprehensive view on the molecular profile and gene repertoires of such rearrangements in B-cell malignancies by reporting results from the analysis of a combined dataset of 412 such rearrangements from CLL and ALL.

As in previous reports (21–24), we found that *IGHD2* subgroup genes predominated in D-J in both CLL and ALL. At the individual *IGHD* gene level, we observed increased rearrangement frequency between 5′ genes of the *IGHD* cluster, for example, *IGHD2-2*, and 3′ genes of the *IGHJ* cluster, for example, *IGHJ4*, *IGHJ5* and *IGHJ6*, suggestive of secondary rearrangements on the same allele (Figure 2). Comparative assessment of productively rearranged D-J in CLL versus ALL revealed an overall similar *IGHD* and *IGHJ* gene usage profile (with very few exceptions), thus indicating no significant disease-related biases regarding molecular events early in B-cell ontogeny. The presence of mutations in a minority of D-J could be plausibly attributed to a “bystander mutagenesis” effect, where nonexpressed rearrangements are mutated without selection for expression of a functional antigen receptor, as previously reported for either normal or neoplastic B cells (35–38).

Of great importance, the comparison of productively rearranged D-J joints versus productive V-D-J junctions in CLL revealed significantly different patterns regarding gene repertoires, IGHD RF use and hydropathicity profiles. Indeed, we documented a significant overrepresentation of *IGHD3* subgroup genes in general and of the *IGHD3-3* gene in particular among V-D-J versus D-J, alluding to selection at the amino acid level in the paratope. In contrast, *IGHD2* subgroup genes, although frequent in both V-D-J and D-J, regardless their origin (that is, from either CLL or ALL), were still overrepresented in D-J. Given that the recombination signal (RS) sequences of *IGHD2* and *IGHD3* subgroup genes do not differ to a significant extent such that might account for their observed frequencies in D-J, alternative explanations must be sought for, for example, differential accessibility of different parts of the IGHD cluster to the VDJ recombination machinery, similar to what has been reported in the mouse (39).

Biased usage of particular *IGHD* genes in certain RF underlies the creation of highly restricted antigen-binding motifs in many stereotyped V-D-J in CLL (27,40). This bias evidently reflects selection, as also attested by our finding of significantly increased usage of RF encoding for hydrophilic motifs in V-D-J junctions versus D-J joints in CLL, alluding to solvent-exposed interactions with antigenic epitopes.

In both mice and men, the usage of hydrophilic VH CDR3 sequences seems to be promoted by positive selection, whereas a consecutive stretch of hydrophobic amino acids may be negatively selected on the basis of structural limitations imposed by the antigen-binding site (41,42). Therefore, not unexpectedly, RF usage in V-D-J of both species reflects this general principle.

A confounding factor when attempting to draw further analogies as well as when assigning RF within a locus arises from the different approaches adopted for defining RF in mice or men. In most previous mouse studies, the preferred (that is, hydrophilic) RF is designated as RF1, whereas a far more sensible procedure is followed in the human, first proposed by Kabat's group, and adopted by IMGT, with RF1 counting from the first nucleotide of the D-REGION in germline configuration (that follows the 5′-HEPTAMER), thus defining three consecutively numbered "logical" RF (1,16,43). Hence, every *IGHD* gene has only one preferred RF, and this RF is not the same for all *IGHD* genes. Concerning hydrophilicity, the human *IGHD* genes encode for hydrophilic amino acids in different RFs, that is, RF2 for *IGHD2*, *IGHD3* and *IGHD4* subgroup genes; RF3 for *IGHD1*, *IGHD5* and *IGHD7* genes; and RF3 for *IGHD6* subgroup genes. Our findings among V-D-J from CLL cases evaluated in this study are in keeping with this general trend; however, this is clearly not the case for D-J.

Taken together, the results presented here strongly indicate that sequence restrictions leading to VH CDR3 stereotypy in CLL reflect selective pressures exerted on the BcR rather than intrinsic features of the *IGHD-IGHJ* gene rearrangement process during the early ontogenetic stages in the life history of CLL progenitor cells.

ACKNOWLEDGMENTS

The authors wish to thank Prof. Marie-Paule Lefranc and Dr. Veronique Giudicelli (Laboratoire d'Immunogenetique Moleculaire, Universite Montpellier II, Centre National de la Recherche Scientifique) for their invaluable support with immunoglobulin gene analysis and Dr. Paolo Ghia, Division of Molecular Oncology and Department of Oncology, Università Vita- Salute San Raffaele and Istituto Scientifico San Raffaele, Milan, Italy, and Dr. David Gonzalez, Section of Haemato-Oncology, Institute of Cancer Research, London, UK, for critically reading the manuscript and helpful suggestions. This work was supported by the ENosAI project (code 09SYN-13-880) and was co-funded by the European Union and the Hellenic General Secretariat for Research and Technology, the Greek Ministry of Health and Cariplo Foundation (Milan, Italy).

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. Lefranc MP, Lefranc G. (2001) *The Immunoglobulin FactsBook*. London: Academic Press. 457 pp.
2. Schissel MS. (2003) Regulating antigen-receptor gene assembly. *Nat. Rev. Immunol.* 3:890–9.
3. Schatz DG, Spanopoulou E. (2005) Biochemistry of V(D)J recombination. *Curr. Top. Microbiol. Immunol.* 290:49–85.
4. Zemlin M, *et al.* (2005) 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.* 334:733–49.
5. Chiorazzi N, Ferrarini M. (2011) Cellular origin(s) of chronic lymphocytic leukemia: cautionary notes and additional considerations and possibilities. *Blood*. 117:1781–91.
6. Darzentas N, *et al.* (2010) A different ontogenesis for chronic lymphocytic leukemia cases carrying stereotyped antigen receptors: molecular and computational evidence. *Leukemia.* 241:125–32.
7. Lanemo Myhrinder A, *et al.* (2008) A new perspective: molecular motifs on oxidized LDL, apoptotic cells, and bacteria are targets for chronic lymphocytic leukemia antibodies. *Blood*. 111:3838–48.
8. Chu CC, *et al.* (2008) Chronic lymphocytic leukemia antibodies with a common stereotypic rearrangement recognize nonmuscle myosin heavy chain IIA. *Blood*. 112:5122–9.
9. Rosen A, Murray F, Evaldsson C, Rosenquist R. (2010) Antigens in chronic lymphocytic leukemia: implications for cell origin and leukemogenesis. *Semin. Cancer Biol.* 20:400–6.
10. Johnson TA, Rassenti LZ, Kipps TJ. (1997) Ig VH1 genes expressed in B cell chronic lymphocytic leukemia exhibit distinctive molecular features. *J. Immunol.* 158:235–46.
11. Bertrad FE III, *et al.* (1997) Ig D(H) gene segment transcription and rearrangement before surface expression of the pan-B-cell marker CD19 in normal human bone marrow. *Blood*. 90:736–44.
12. Davi F, *et al.* (1997) Early onset of immunoglobulin heavy chain gene rearrangements in normal human bone marrow CD34+ cells. *Blood*. 90:4014–21.
13. Gu H, Kitamura D, Rajewsky K. (1991) B cell development regulated by gene rearrangement: arrest of maturation by membrane-bound D mu protein and selection of DH element reading frames. *Cell*. 65:47–54.
14. Tsubata T, Tsubata R, Reth M. (1991) Cell surface expression of the short immunoglobulin mu chain (D mu protein) in murine pre-B cells is differently regulated from that of the intact mu chain. *Eur. J. Immunol.* 21:1359–63.
15. Shimizu T, Yamagishi H. (1992) Biased reading frames of pre-existing DH—JH coding joints and preferential nucleotide insertions at VH—DJH signal joints of excision products of immunoglobulin heavy chain gene rearrangements. *EMBO J.* 11:4869–75.
16. Cohn M. (2008) A hypothesis accounting for the

paradoxical expression of the D gene segment in the BCR and the TCR. *Eur. J. Immunol.* 38:1779–87.

17. Ivanov II, Link J, Ippolito GC, Schroeder HW Jr. (2002) Constraints on the Hydropathicity and Sequence Composition of HCDR3 are Conserved Across Evolution. In: *The Antibodies*. Volume 7. Capra JD, Zanetti M (eds.) Taylor & Francis Inc, New York, pp. 43–67.
18. Zemlin M, *et al.* (2008) Regulation of repertoire development through genetic control of DH reading frame preference. *J. Immunol.* 181:8416–24.
19. Reth MG, Alt FW. (1984) Novel immunoglobulin heavy chains are produced from DJH gene segment rearrangements in lymphoid cells. *Nature*. 312:418–23.
20. Horne MC, Roth PE, DeFranco AL. (1996) Assembly of the truncated immunoglobulin heavy chain D mu into antigen receptor-like complexes in pre-B cells but not in B cells. *Immunity*. 4:145–58.
21. Szczepa_ski T, *et al.* (2001) Precursor-B-ALL with D(H)-J(H) gene rearrangements have an immature immunogenotype with a high frequency of oligoclonality and hyperdiploidy of chromosome 14. *Leukemia.* 15:1415–23.
22. González D, *et al.* (2003) Incomplete DJH rearrangements of the IgH gene are frequent in multiple myeloma patients: immunobiological characteristics and clinical implications. *Leukemia.* 17:1398–403.
23. Martín-Jiménez P, *et al.* (2007) Molecular characterization of complete and incomplete immunoglobulin heavy chain gene rearrangements in hairy cell leukemia. *Clin. Lymphoma Myeloma.* 7:573–9.
24. González D, *et al.* (2005) Molecular characteristics and gene segment usage in IGH gene rearrangements in multiple myeloma. *Haematologica.* 90:906–13.
25. 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.
26. van Dongen JJ, *et al.* (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98–3936. *Leukemia.* 17:2257–317.
27. 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.
28. Brochet X, Lefranc MP, Giudicelli V. (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* 36:W503–8.
29. Lefranc MP, *et al.* (2009) IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* 37:D1006–12.
30. 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.
31. Fais F, *et al.* (1998) Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. *J. Clin. Invest.* 102:1515–25.
32. Tschumper RC, *et al.* (2008) Immunoglobulin diversity gene usage predicts unfavorable outcome in a subset of chronic lymphocytic leukemia patients. *J. Clin. Invest.* 118:306–15.
33. Agathaggelidis A, *et al.* (2010) The composition of the B cell receptor repertoire in 7428 cases of chronic lymphocytic leukemia: one third stereotyped, two third heterogeneous. What does this mean? *Blood.* 116:43.
34. Zemlin M, *et al.* (2003) 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.* 334:733–49.
35. Klein U, *et al.* (1998) Somatic hypermutation in normal and transformed human B cells. *Immunol. Rev.* 162:261–80.
36. Dörner T, Foster SJ, Farner NL, Lipsky PE. (1998) Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands. *Eur. J. Immunol.* 28:3384–96.
37. Goossens T, Klein U, Küppers R. (1998) Frequent occurrence of deletions and duplications during somatic hypermutation: implications for oncogene translocations and heavy chain disease. *Proc. Natl. Acad. Sci. U. S. A.* 95:2463–8.
38. Belessi C, *et al.* (2005) Analysis of expressed and non-expressed IGK locus rearrangements in chronic lymphocytic leukemia. *Mol. Med.* 11:52–8.
39. Sen R, Oltz E. (2006) Genetic and epigenetic regulation of IgH gene assembly. *Curr. Opin. Immunol.* 18:237–42.
40. Forconi F, *et al.* (2010) The normal IGHV1–69-derived B-cell repertoire contains stereotypic patterns characteristic of unmutated CLL. *Blood*. 115:71–7.
41. Raaphorst FM, Raman CS, Tami J, Fischbach M, Sanz I. (1997) Human Ig heavy chain CDR3 regions in adult bone marrow pre-B cells display an adult phenotype of diversity: evidence for structural selection of DH amino acid sequences. *Int. Immunol.* 9:1503–15.
42. Ippolito GC, *et al.* (2006) Forced usage of positively charged amino acids in immunoglobulin CDR-H3 impairs B cell development and antibody production. *J. Exp. Med.* 203:1567–78.
43. Kabat EA, Wu TT, Perry HM, Gottesman KS, Foeller C. (1991) *Sequences of Proteins of Immunological Interest.* 5th Edition. Bethesda (MD): U.S. Dept. of Health and Human Services, Public Health Service, National Institutes of Health. NIH publ. no. 91-3242.