

Gene Expression Profiling of B cell Chronic Lymphocytic Leukemia Reveals a Homogeneous Phenotype Related to Memory B Cells

Ulf Klein *, Yuhai Tu !, Gustavo A. Stolovitzky !, Michela Mattioli *, Giorgio Cattoretti*, Hervé Husson §, Arnold Freedman §, Giorgio Inghirami ||, Lilla Cro ¶, Luca Baldini ¶, Antonino Neri ¶, Andrea Califano **, and Riccardo Dalla-Favera *

* Institute for Cancer Genetics, the Departments of Pathology and Genetics & Development, Columbia University, New York, NY10032, U.S.A.

! IBM T.J. Watson Research Center, Yorktown Heights, New York, NY 10598, U.S.A.

§ Department of Adult Oncology Dana-Farber Cancer Institute, and Department of Medicine, Harvard Medical School, Boston, MA 02115, U.S.A.

|| Department of Pathology, New York University Medical Center, New York 10016, U.S.A.

¶ Ematologia 1, Department of Hematology, Ospedale Maggiore, I.R.C.C.S., Milan, Italy

** First Genetic Trust, Inc., Lyndhurst, New Jersey 07071, U.S.A.

Running foot: Gene expression profiling of B cell chronic lymphocytic leukemia

Word count: 9,134 Character count: 50,693

Corresponding author: Riccardo Dalla-Favera, Institute for Cancer Genetics, Columbia University, 1150 St. Nicholas Ave., New York, NY 10032, U.S.A.

Phone: +1 212 304-7381 Fax: +1 212 304-5537 e-mail: rd10@columbia.edu

Key words: somatic hypermutation, germinal center, CD5, DNA microarray, cluster analysis

Abbreviations: B-CLL: B cell chronic lymphocytic leukemia, M-CLL: IgV mutated CLL, UM-CLL: IgV unmutated CLL, GC: germinal center, CB: centroblast, CC: centrocyte.

Abstract

B cell-derived Chronic Lymphocytic Leukemia (CLL) represents a common malignancy whose cell derivation and pathogenesis are unknown. Recent studies have shown that >50% of CLLs display hypermutated immunoglobulin variable region (IgV) sequences and a more favorable prognosis, suggesting that they may represent a distinct subset of CLLs which have transited through germinal centers (GC), the physiologic site of IgV hypermutation. To further investigate the phenotype of CLLs, their cellular derivation and their relationship to normal B cells, we have analyzed their gene expression profiles using oligonucleotide based DNA chip microarrays representative of ~12,000 genes. The results show that CLLs display a common and characteristic gene expression profile that is largely independent of their IgV genotype. Nevertheless, a restricted number of genes (<30) have been identified whose differential expression can distinguish IgV mutated versus unmutated cases and identify them in independent panels of cases. Comparison of CLL profiles with those of purified normal B cell subpopulations indicates that the common CLL profile is relatively more related to that of memory B cells than to those derived from naïve B cells, CD5⁺ B cells, and GC centroblasts and centrocytes. Finally, this analysis has identified a subset of genes specifically expressed by CLL cells of potential pathogenetic and clinical relevance.

Introduction

B cell chronic lymphocytic leukemia (CLL) represents the most common leukemia in the Western countries with an estimated incidence of 1 per 100,000 per year (1). This disease is characterized by the monoclonal expansion of B lymphocytes in the peripheral blood, bone marrow and lymphoid organs, and by an indolent course which ultimately becomes aggressive and invariably lethal (1). Current knowledge of the pathogenesis of CLL is limited because no specific genetic alteration has yet been associated with this disease. In particular, CLL is not associated with reciprocal balanced chromosomal translocations, but rather with specific deletions (2) suggesting the loss of presently unidentified tumor suppressor genes. CLL cells have a low proliferative rate and a prolonged life span, suggesting that their primary alteration may be a defect in apoptosis (1).

The cellular origin of CLL is unknown and has been the object of recent controversy. CLL cells are characterized by the expression of the cell surface markers CD5, CD23, CD19, and low levels of sIgM/IgD, a pattern not shared by any known B cell subpopulation (1, 3). The expression of the CD5 antigen was originally taken to suggest that B-CLL originates from CD5⁺ B cells (4), which are usually characterized by unmutated immunoglobulin variable region (IgV) genes (5, 6). However, recent studies have shown that 50-70% CLL have undergone IgV hypermutation (7-9), a phenomenon that characterizes normal B cells undergoing a T cell-dependent germinal center (GC) reaction (10) and malignant B cells derived from GC or post-GC B cells (11, 12). This finding has led to the hypothesis that CLL cases displaying mutated IgV may derive from a cell that had transited through the GC, whereas those with germline IgV may derive from GC-independent cells (8, 9, 13). This hypothesis has both biological and clinical relevance since the two subgroups have different prognosis, with IgV mutated CLL displaying a more benign clinical course (13, 14).

To further investigate the phenotype of CLL subtypes, their cellular derivation and their relationship with normal B cells, we have analyzed their gene expression profiles using oligonucleotide-based DNA chip microarrays (Affymetrix) representative of ~12,000 genes. The gene expression profiles have been comparatively analyzed: i) between mutated (M-CLL) and unmutated (UM-CLL) cases, to determine

whether the two subgroups are different; ii) with those of normal B cell subpopulations, including GC-independent (CD5⁺), pre-GC (naï ve), GC [centroblasts (CB) and centrocytes (CC)], and post-GC (memory) B cells, to investigate their cellular derivation; and iii) with those of normal B cell subpopulations and of other malignancies derived from mature B cells to identify genes that are specifically expressed in CLL.

Methods

CLL cases. Peripheral blood from CLL patients who had not received treatment was taken after informed consent. CLL cells were enriched from peripheral blood mononuclear cells (PBMC) by magnetic cell separation using the MidiMACS® system (Miltenyi Biotech, Auburn, CA). Following Ficoll-Isopaque density centrifugation, PBMC were incubated with CD19-Microbeads (MB) (Miltenyi Biotech). Magnetically labeled cells were isolated by passing the cell suspension over an LS-column (Miltenyi Biotech). All isolation steps were performed on ice and using ice-cold solutions. Sequences of IgV_H genes of the 34 CLL cases were determined as described (15).

Isolation of normal human B cell subsets and tumor cells. A detailed description of the gene expression profiles of normal B cell subpopulations, including the methods for purification and characterization, will be reported elsewhere (manuscript in preparation). Briefly, tonsils were obtained from routine tonsillectomies performed at the Children's Hospital of Columbia University, New York, NY. The specimens were kept on ice immediately after surgical removal, and all cell isolation steps were performed on ice. B cell subsets were isolated by magnetic cell separation (see above). The isolation strategies were based on recent characterizations of the human B cell subsets (16-19). For the isolation of naï ve B cells, tonsillar MC were first incubated with anti-CD27, anti-CD10 (both Pharmingen), and anti-CD3 (Coulter/Immunotech), then with anti-IgG₁-MB and anti-CD14-MB. Magnetically labeled cells were depleted by passing the cell suspension over a LD-column (Miltenyi Biotech). The flow-through was incubated first with anti-IgD-FITC (Pharmingen), then with anti-FITC-MB. IgD-positive B cells were isolated using LS-columns. CB were isolated in a single step by staining tonsillar MC with anti-CD77 (Coulter/Immunotech), followed by incubation with Mouse-Anti-Rat-IgM (MARM; Pharmingen), and finally anti-IgG₁-MB. The cells were passed over a LS-column. CC: tonsillar MC were first incubated with anti-CD77, then with anti-CD39 (Coulter/Immunotech), anti-CD3, and MARM, followed by anti-IgG₁-MB staining. The cells were passed subsequently over a LS-column with G21 needle and a LD-column. The resulting flow-through was first incubated with anti-CD10, then with anti-IgG₁-MB. CD10⁺ CC were isolated using LS-columns. Memory B cells: tonsillar MC were incubated first with anti-CD10, anti-CD3, and anti-CD38 (Pharmingen), the latter at a titer of 1:1000 to facilitate a selective depletion of CD38^{high} plasma cells and GC B cells, then with anti-IgG₁-MB and anti-CD14-MB. Labeled cells were depleted using a LD-column. The flow-through was first stained with anti-CD27-FITC, then with anti-FITC-MB. CD27⁺ cells were isolated using a LS-column.

Cord blood was obtained after informed consent. Cord blood B cells almost exclusively consist of CD5⁺ B cells. To enrich for CD5⁺ B cells from cord blood, mononuclear cells obtained by Ficoll-Isopaque density centrifugation were incubated with CD19-MB (Miltenyi Biotech). The stained cells were passed over a LS-column.

Follicular lymphoma (FL) cells were isolated from single cell suspensions from digested lymph nodes or spleen which were histologically involved with FL. These cell suspensions were depleted of T cells, monocytes, NK cells using anti-CD4, -CD8, -CD14, -CD56, followed by immunomagnetic beads as previously described (20). FL cells were isolated from 6 individuals with relapsed FL who had not received any therapy for at least 6 months prior to biopsy. Diffuse large cell lymphoma (DLCL) and Burkitt lymphoma (BL) tumor cells were enriched by magnetically depleting CD3-, CD4-, CD8-, CD14-, CD15-, CD16-, CD56-, and glycophorin-expressing non-tumor cells. All samples were obtained according to appropriate Human Protection Committee validation and informed patient consent.

***Generation of cRNA and Microarray Hybridization.* Total RNA was isolated in two steps using Trizol (Life Technologies), followed by RNeasy (Qiagen) purification. Double strand cDNA was generated from 5 mg of total RNA using a poly dT oligonucleotide that contains a T7 RNA polymerase initiation site and the SuperScript Choice System Kit (Life Technologies). cDNA was phenol/chloroform extracted. Biotinylated cRNA was generated by *in vitro* transcription using the Bio Array" High Yield" RNA Transcript Labeling Kit (ENZO Diagnostics, Inc., Farmingdale, NY). The cRNA was purified using RNeasy. cRNA was fragmented according to the Affymetrix protocol, and 15 mg of biotinylated cRNA were hybridized to U95A microarrays (Affymetrix). Following scanning (scanner from Affymetrix), the expression values for the genes were determined using Affymetrix GENECHIP software, using the Global Scaling option that allows a number of experiments to be normalized to one target intensity, thus facilitating comparison between multiple experiments.**

***Data Processing.* The Affymetrix expression data (average differences) was processed as follows. The small and negative expression levels were clipped-off to be equal to a cutoff value arbitrarily chosen as 20. The logarithm of this clipped-off data was subsequently used throughout analyses.**

***Dendrogram.* The hierarchical clustering algorithm used to generate the dendrogram is based on the average-linkage method (21, 22). To construct the dendrogram, a subset of genes was used out of the total of 12588 gene segments present on the microarray, whose expression levels vary the most among the 40 samples, and which are thus most informative. For the hierarchical clustering shown in Fig.1, only genes were chosen whose average change in expression level from the mean across the whole panel was at least 2-fold (2337 genes selected). Independent analyses were performed using all genes and only gene segments whose average change in expression level was at least 3-fold and 4-fold, respectively. The expression values of each selected gene is normalized to have zero mean and unit standard deviation. The distance between two individual samples is calculated by Euclidean distance with the normalized expression values.**

***Analysis of gene expression profiles.* We used the [Genes@Work](#) software platform which is a gene expression analysis tool based on the pattern discovery algorithm SPLASH (for: structural pattern localization analysis by sequential histograms) (23, 24). Genes@Work is used here to perform both supervised and unsupervised gene expression analysis.**

Rather than inferring clusters by comparing the gene expression values of each possible pair of experiments, as it is the case for the average-linkage method (21, 22), Genes@Work proceeds by discovering one or more global gene expression signatures that are common to an entire set of at least n experiments (the support experiments), where n is a user-selectable parameter called the minimum support. A Genes@Work pattern represents (a subset of) the genes that are differentially expressed in (a subset of) the phenotype set with respect to the control set. A Genes@Work pattern can be represented as

a matrix composed of columns for the experiments of the phenotype and control set and rows for the differentially expressed genes (see description of matrices). The number of experiments in the Genes@Work pattern, $n' e n$, is called the pattern support. Differential expression is determined as follows: independently for each gene, an expected gene expression probability density $p(e)$ is computed empirically from the experimental set [the method differs in the case of supervised vs. unsupervised clustering (see below)]. Then, given the group of n cells, the algorithm builds a cluster with all of the genes such that the integral of their $p(e)$ over the expression range of the group is less than a predefined threshold d . The statistical significance of a Genes@Work pattern is computed based on the probability of observing a similar pattern (i.e., a pattern with the same number of support experiments and support genes) in a set of random cells whose genes are distributed according to the empirical probability density $p(e)$, the null hypothesis (24). Genes@Work patterns are then ranked according to their statistical significance.

Supervised vs. unsupervised clustering. In supervised clustering, a phenotype set and a control set is defined. The aim is to identify a set of genes that optimally discriminate between phenotype and control sets. The expression probability density $p(e)$ is determined independently for each gene as the one most likely to produce the expression values observed in the control set, given the expected error in the measure. This is accomplished by convolving the expression values in the control set with a gaussian kernel, with a standard deviation equal to the error determined from repeatability experiments as a function of the expression level. For details, see Ref. 24. The algorithm is run twice, with either set chosen in turn as phenotype, and the resulting gene expression patterns are then fused into a single one.

In unsupervised clustering, a control set is not available as the algorithm is applied to determine the most likely subsets without any *a-priori* knowledge. In this case the expression probability density is computed differently (24): either as a uniform probability density over the expected expression range of the full set, or as a normal-distribution. The latter is used with a mean equal to that of the gene expression over the full experimental set and a standard deviation equal to 1/4 of the experimental one.

Pattern discovery. Unsupervised clustering: following computation of the expression probability of each gene, pattern discovery is performed starting with a support value $n = n_0 - m$, where n_0 is equal to the number of experiments in the set and m is the size of the smallest subset which is expected to be characterized by a discernible sub-phenotype (typically, $m = n_0 / 4$). For a given support n , the pattern z score $z_p(p_i, n)$ of each discovered pattern, p_i , is computed. (The pattern z score z_p , is inversely related to the pattern p -value; when z_p is sufficiently large, this relation is approximately $p = 1/z_p^2$.) The pattern $P(n)$, corresponding to the maximum pattern z score $Z_p(n)$, is selected. $Z_p(n) = 1$ if no patterns are found for the given support value. This process is repeated iteratively for decreasing values of n , until a sudden increase in the pattern z score is detected. This corresponds to a maximum of the first derivative of the pattern z score function for a support value equal to n . Such extrema can be determined, as a first approximation, when the following condition is satisfied: $Z_p(n - 1)/Z_p(n) \approx Z_p(n)/Z_p(n + 1)$. The pattern $P(n)$ for the corresponding value of n is called the interest pattern. The experiments that support the interest pattern become a subset and the genes that support the interest pattern become the subset-signature. For a discussion of the pattern z score, see Ref. 24.

Supervised clustering: the gene expression probability densities are computed and pattern discovery is performed iteratively with a decreasing value of the support, starting at N (the size of the phenotype

group), until either a significant gene cluster size is obtained, or discovered patterns appear to be no longer statistically significant.

For the unsupervised clustering analysis, an optimal value for $d = 0.1$, or 10% of the total probability, is chosen. For supervised clustering, due to the significantly higher sensitivity, a smaller value of $d = 0.01$, or 1% of the total probability, is used.

Graphic representations of gene expression patterns (matrices). Columns represent individual experiments, and rows represent individual genes present on the expression microarray. To generate a pseudo-color map, first a gene- and experiment-specific change of variables, from the original measurement v into z_{ge} , is computed using the formula:

$$v - (m_P + m_C) / 2$$

$$z_{ge} = \frac{v - (m_P + m_C) / 2}{(s_P + s_C) / 2},$$

$$(s_P + s_C) / 2$$

where m_P and s_P are respectively the mean and standard deviation computed from the gene expression values for that gene in the phenotype group, and m_C and s_C are their corresponding values computed from the control group. The value of this function is then plotted using a pseudo-color map that represents $z_{ge} = 0$ as black, $z_{ge} > 0$ as progressively brighter hues of red, and $z_{ge} < 0$ as progressively brighter levels of green. $z_{ge} = 4$ and $z_{ge} = -4$ correspond to complete saturation of the red and the green, respectively. The resulting pseudo-color map associates the same colors to measurements that are off by the same number of standard deviations from their expected value.

For each gene, the statistical significance of the differential expression across the phenotype and control sets (gene z score, z_g), is computed using the formula:

$$m_P - m_C$$

$$\frac{m_P - m_C}{(s_P + s_C) / 2} = z_g$$

$$(s_P + s_C) / 2$$

In contrast to the generally used fold-ratio (m_P / m_C) to describe gene expression changes among cell subsets, z_g represents the differential expression between phenotype and control samples relative to the variability of their expression levels. Rows are divided into two groups. First, the genes that are overexpressed in the first experiment set are reported by decreasing value of the z_g score (most significant first). Then, the genes that are underexpressed are reported, ordered by increasing value of the z_g score (most significant first).

Classifier and classification method. The classifier is a scoring function based on the values of a set of genes (gene cluster) which are differentially expressed in two sets of cell types and can thus be used for cell type classification. The higher the score, the more likely it is that a cell type is related to the

phenotype set. Pattern discovery methods are used here to identify gene clusters. The union of the genes of all statistically significant patterns is used to define the classifier (see above for pattern discovery method). Given a set of genes, the scoring function is defined quantitatively as:

$$R_1 = \frac{\sum_{i=1}^{n_c} v_i (m_i^{(1)} + m_i^{(2)}) / 2}{\sum_{i=1}^{n_c} (m_i^{(1)} - m_i^{(2)}) / 2}$$

where $f(x) = \max(-1, \min(1, x))$, n_c is the number of genes in the classifier, and $i=1,2,\dots,n_c$ is the label for each individual gene in the classifier, m_i is the expression level of gene i for the new cell, $m_i^{(1)}$ and $m_i^{(2)}$ represent the mean expression levels of gene i for cells in the phenotype and control set, respectively.

Online supplemental material. Genbank accession number, Affymetrix entry numbers, and the normalized primary data of the genes shown in Figures 2, 4, and 5 are available at the journal's web site (www.jem.org).

Results

A panel of 34 CLLs, characterized for their typical cell surface phenotype and presence or absence of IgV mutations, was used for this study. From 20 of these samples, tumor cells were purified by magnetic cell separation of CD19⁺ cells, while 14 cases that showed a representation of >80% of malignant cells in the peripheral blood were used as unpurified cell populations (Table I). As in previous reports (13, 14), CLLs with $\leq 2\%$ basepair difference to the corresponding germline IgV gene were considered as UM-CLLs. Eighteen of 34 cases, comparably distributed between purified and non-purified samples (11 and 7, respectively), were shown to be M-CLL carrying 6 to 33 mutations per case.

RNA extracted from these cases was converted into labeled cRNA and hybridized to U95A Affymetrix Gene Chips representative of $\sim 12,000$ genes, including mostly known genes (>80%). Gene expression profiles were analyzed using two main approaches: i) unsupervised clustering, which can identify distinct cell types (e.g. CLL cases) which have not been classified *a-priori*; and ii) supervised clustering, which allows the identification of differentially expressed genes between cell types (CLL cases) defined *a-priori* according to a given criterium (e.g. presence of IgV mutations). Unsupervised and supervised clustering were obtained using two algorithms, the average-linkage method (21, 22) or the pattern discovery algorithm SPLASH used by the Genes@Work software platform (see Methods) (23, 24). The latter is capable of capturing subtle differences in gene expression by combining an optimal non-linear transformation of the gene expression values coupled with an analytical method to compute the statistical significance of the identified clusters (see Methods and below).

CLLs display a common gene expression profile independent of IgV mutations. To determine whether M- and UM-CLLs are phenotypically different, we first analyzed their gene expression profiles by two independent unsupervised methods. When analyzed using Genes@Work (not shown) or clustering by the average-linkage method (Fig.1), CLLs displayed a common profile that is clearly distinguishable from that of FL (see first branching in the dendrogram; Fig. 1). Purified and non-purified cases were readily recognized as different (second branching; Fig. 1), most likely reflecting the contribution of normal cells contaminating the non-purified cases. However, in both subgroups, M- and UM-CLLs were not distinguishable and appear intermingled (Fig.1). Lowering the stringency of the analysis to the lowest limits of statistical significance for pattern discovery, or varying the selection criteria for the genes used in clustering (see Methods), did not lead to further separation of M- and UM-cases (not shown). These results indicate that M- and UM-CLL have a common pattern of expression for most of the 12,000 genes analyzed, suggesting that they have a largely common phenotype.

A small subset of differentially expressed genes allows the classification of M- and UM-CLLs. Subtle differences in gene expression among closely related cell populations may escape detection using unsupervised clustering analyses. Thus, to determine whether M- and UM-CLLs have subtle differences in gene expression, the two subgroups were compared by supervised clustering using Genes@Work.

We first compared a set of 20 cases including 9 UM- and 11 M-CLL. In order to avoid the influence of contaminating cells and to obtain the maximum specificity for the CLL phenotype, these cases were selected among the purified ones. Fig.2a shows that a set of 23 genes is differentially expressed in M- versus UM-CLLs, the majority of them (20 of 23) being upregulated in UM-CLLs versus M-CLLs. Several of these genes are of unknown function, while the remaining ones encode products of heterogeneous nature.

To validate the specificity of this differential gene expression profile, we tested whether it could be used to classify an independent panel of 14 cases into M- and UM-CLLs. To test the sensitivity of the classifier and its potential clinical use, unpurified cases were used for this analysis. Fig.2b shows that the profile shown in Fig.2a could correctly classify all 7 M-CLLs, and 5 of 7 UM-CLLs (p -value of < 0.025). Thus, the 23 genes differentially expressed in M- versus UM-CLL represent a consistent phenotypic difference between the two subgroups (see Discussion).

The gene expression profile of CLL is related to that of memory B cells. Phenotypic as well as IgV gene analyses could not conclusively identify the cell of origin of CLL. To address this issue, we compared the gene expression profiles of CLL to those of the major human B cell subsets, namely GC CD77⁺ (CB) and CD77⁻ (CC) B cells, pre-GC (naï ve) and post-GC (memory) B cells (16-19), as well as GC-independent CD5-positive B cells. GC, naï ve (IgD⁺CD27⁻), and memory (CD27⁺) B cells were purified by magnetic cell separation from tonsillar mononuclear cells, while CD5⁺ B cells were isolated from umbilical cord blood (see Methods). Each B cell subset was isolated from five individuals, and cRNA generated from these fractions was hybridized to the U95A array as described above.

To determine whether CLLs are more related to GC (CB and CC) or non-GC (naï ve and memory) B cells, we first compared the CLL profiles to those differentially expressed by these two B cell subgroups (Fig.3a). Genes that distinguish CB and CC cells from naï ve and memory B cells as identified by supervised clustering are shown in the left panel of Fig.3a, while the CLL samples are aligned to the right to visualize the expression of the respective genes in CLL cells. Genes known to be differentially expressed among GC and non-GC B cells (CD10, CD38, CD39, CD44, CD69, A-myb, Ki67, BCL-6,

bcl-2) (25-29) are indicated as internal controls. Fig.3a (right panel) shows that the gene expression profile of CLL is significantly more related to that of the non-GC naïve and memory B cells (see Fig.3d for statistical analysis). The same approach showed that the gene expression profile of CLL is more related to naive and memory cells than to CD5⁺ B cells (Fig. 3b,d). Analogous results were obtained using the less sensitive unsupervised clustering by the average-linkage method (data not shown). Finally, when CLLs were compared to naive versus memory B cells (Fig.3c), the results showed that 14 of the 20 CLLs were significantly more related to memory than to naive B cells (*p*-value of < 0.025). Note that, although less clear-cut than in the previous comparative analysis (Fig.3a,b), the relatedness to memory B cells is significant considering that memory and naïve B cells differ in the expression of only ~140 of 12,000 genes (data not shown). While the difference was not statistically significant for the remaining six cases, none of the CLL cases was more related to naïve than to memory B cells. The degree of relatedness to memory B cells was not significantly different for M- and UM-CLLs. Taken together, these results indicate that, independently of their IgV mutational status, CLLs are more related to memory cells than to naïve, GC, or CD5⁺ B cells.

This observation prompted a direct analysis of the differences in gene expression profiles between CLL and memory B cells by supervised clustering. Fig.4 shows the results of this analysis with the differentially expressed genes organized according to putative functional categories, including proliferation, apoptosis, cytokines/chemokines and receptors, adhesion, and cytokinesis. A number of proliferation-associated genes were downregulated in CLL cells (e.g. *c-Myc*, average fold difference in expression levels (CLL vs. memory cells), 9 fold; *cyclin B*, 13 fold; and *E2-C*, 80 fold). Apoptotic functions appear to be suppressed in CLL with the anti-apoptotic gene *bcl-2* upregulated (4 fold) as expected (1), and various genes encoding pro-apoptotic molecules uniformly downregulated [*BID*, 33 fold; *Rad9* (30), 9 fold; *DRAK1* (31), 3 fold; and *DRAK2* (31), 4 fold]. The IL-4 pathway appears to be activated in CLL, based on the previously observed (32) upregulated expression of the gene encoding the IL4-receptor (6 fold), on the observed downregulation of *SOCS-1* (Fig. 4), an inhibitor of the IL4-signaling pathway (33) (19 fold), and consistent with the fact that CLL cells are responsive to IL4 *in vitro* (1) (see Discussion). Finally, the expression of various genes encoding adhesion-associated molecules appears differentially regulated in CLL compared to memory B cells. Overall, these results indicate that CLL differs from memory B cells in the expression of numerous genes that suggest a more quiescent, anti-apoptotic phenotype, with distinct cytokine and chemokine response and adhesion properties.

Identification of genes specifically expressed in CLL. To identify genes specifically up-or down-regulated in CLL cells, we used supervised clustering to compare the gene expression profiles of CLL cases to those of normal B cell subsets (naïve, CB, CC, memory) and to those derived from various non-Hodgkin lymphoma (NHL) subtypes, including FL, DLCL, and BL. Fig.5 indicates that 32 genes are specifically expressed (or overexpressed), while >50 genes appear downregulated in CLL. Several of the genes upregulated in CLL are involved in signal transduction pathways: *CDC25* [average difference in expression level (CLL vs. all others): 14 fold over background] is a Ras guanine nucleotide exchange factor (GEF) (34, 35). *EPAC* (35 fold over background) plays a role in the cAMP signal transduction pathway via *Rap1* (36), a Ras-related GEF involved in B cell receptor signaling and oncogenesis (see Discussion). Significantly overexpressed genes include also: i) the cell surface receptors *Ror1* (19 fold), an orphan tyrosine kinase receptor, and the thromboxane A2 receptor (4 fold); and ii) several genes related to TGF β signaling, such as fibromodulin (>250 fold over background), associated with modulation of TGF β signaling and cell adhesion (37), the TGF β -inducible *TIEG2*, a

SP1-like transcription factor (38) (10 fold over background), and BIGH3/TGFB1 (39) (38 fold). Genes specifically downregulated in CLL include genes involved in cell cycle progression and DNA replication and metabolism (E2-C, CIP2/KAP, CDC2, cyclin B, ribonucleotide reductase, thymidine kinase, dihydrofolate reductase, topoisomerase IIa), suggesting a markedly quiescent phenotype.

Discussion

Gene expression profiling allows a more comprehensive examination of cell phenotypes than the ones based on the analysis of individual or small numbers of genes, proteins, or signaling pathways. This technology has been used in this study to address three open questions regarding CLL: does CLL include one or two biological phenotypes since it displays a discordant IgV mutational status and clinical behavior? What is the cellular origin of CLL within the B cell lineage? Can a CLL-specific gene expression profile be identified? The results obtained directly address these questions and have implications relevant for the pathogenesis and, possibly, for the clinical management of this disease.

CLLs display a common phenotype independent of IgV mutations. The gene expression profiles shown in Fig. 1 demonstrate that CLL has a characteristic gene expression profile that is clearly distinguishable from FL, a malignancy derived from mature B cells and displaying an indolent clinical course similar to CLL. This observation confirms the results obtained by Alizadeh et al. using a different DNA microarray technology (oligonucleotide versus cDNA-based arrays) and a only partially overlapping set of genes examined (12,000 non-lymphocyte-biased versus 18,000 lymphocyte-biased) (40).

These results also show that CLLs have a homogeneous phenotype independent of the presence of IgV mutations. This phenotype is defined by the common pattern of expression of 12,000 genes, with only 23 being differentially expressed in the M- versus UM-subtypes. Although significant differences in protein expression or modification cannot be excluded by gene expression profiling, this result does not support the hypothesis that CLL may include distinct biological entities (8, 13), as it is the case, for instance, for DLCL (40). Rather, these findings strongly suggest that all CLLs may derive from a common cell precursor through a common pathogenetic mechanism (see below).

CLLs are related to memory B cells. Based on various phenotypic similarities, mainly the presence of the CD5 marker and on initial reports that CLL displayed unmutated IgV, it has long been proposed that CLL may derive from the malignant transformation of CD5⁺ B cells (4), the human equivalent of mouse B1 cells. Subsequently, the identification of M-CLLs led to the hypothesis that a fraction of CLL cases may derive from a GC-experienced B cell, possibly a memory B cell (M-CLL) (8, 9, 13). The gene expression profiles shown here suggest that CLL do not derive from CD5⁺ B cells or from GC B cells, while they indicate that both M- and UM-CLLs are more related to memory than to naïve B cells. This conclusion is based on a comparative evaluation and cannot formally exclude the existence of a presently unrecognized B cell subpopulation that is more similar to CLL than memory B cells. However, this conclusion is in agreement with other well established traits shared by CLL and memory B cells, namely the frequent presence of IgV mutations and, most notable, the expression of the CD27 marker (14, 17-19, 41). The fact that CLL cells express the CD5 and CD23 markers, which are not typical of memory B cells, is not inconsistent with their derivation from memory B cells since the expression of these markers could represent an abnormal trait of the transformed phenotype rather than a characteristic of the normal precursor of CLL. In the case of CD23, whose expression is induced by IL-4 (42), this hypothesis is

supported by the abnormal expression of several genes involved in the IL-4 pathway (upregulation of IL-4 receptor, downregulation of SOCS1; see Results) and by the abnormal sensitivity of CLL to IL-4 *in vitro* (1). Finally, a derivation of CLL cells from memory B cells is also consistent with the fact that CLLs lack chromosomal translocations which are thought to occur in developing B cells rearranging their antigen receptors or in GC B cells (see below).

The notion that CLL may all derive from memory B cells leaves open the question of why a sizable fraction of them expresses unmutated IgV genes, which are typical for pre-GC, naïve B cells or GC-independent cells involved in T-independent humoral responses. One possible explanation is provided by the observation that a minor fraction of CD27⁺ B cells carry unmutated IgV (17, 18) and may have entered the memory cell pool without acquiring IgV mutations, possibly due to already high affinity for the antigen. Such CD27⁺ unmutated memory B cells may therefore represent the normal counterpart of UM-CLL. Alternatively, the precursor of UM-CLL may be a cell that has acquired a memory phenotype via encounter with a T-independent (and thus GC-independent and IgV mutation-independent) pathway (9). This hypothesis implies that the antigen recognized by M-CLL and UM-CLL may be of different nature, consistent also with the observation that CDRIII length and IgV region usage differs among the two subgroups (9, 43). A difference in the nature and/or time of exposure to antigen may also explain the different clinical behavior of the two subgroups.

Overall, the finding that CLLs, including the UM subtype, display a gene expression profile similar to memory B cells warrants further investigations on the possible heterogeneity of the memory B cell pool (44, 45). The few genes that are differentially expressed in M- versus UM-CLL (Fig.2a) may turn useful in dissecting this heterogeneity, although further characterization is needed because they represent a heterogeneous group of genes, some of unknown function.

Implications for CLL pathogenesis. The observation that all CLLs share a common gene expression profile suggests that they derive from a common pathogenetic pathway. This notion is consistent with the relatively homogeneous cytogenetic profile of CLL, characterized by few common chromosomal abnormalities and, particularly, by the strong association with 13q14 deletions, present in the majority of cases (up to 70%) (46-48). These deletions are often present as a single chromosomal abnormality and are thought to reflect the loss or inactivation of a still unknown tumor suppressor gene (46-48). The homogeneity of the gene expression profile of CLL suggests that this alteration or some functional equivalent (e.g. inactivating point mutations) may be present also in those cases lacking cytogenetically or molecularly detectable 13q14 deletions.

The observation that CLL is more related to memory B cells than to any other known normal B cell subset suggests that the multistep process leading to CLL may actually initiate in memory B cells. This hypothesis is consistent with the notion that, unique among lymphoid malignancies, CLL lack reciprocal balanced chromosomal translocations (2). These aberrations are thought to occur either during Ig VDJ recombination in maturing B cells, or during Ig hypermutation and isotype switch in mature B cells within the GC (11), and, accordingly, are common in B cell lymphoma, most of which derive from mature GC B cells (12, 49). Thus, if the transformation process leading to CLL initiates in memory B cells, it cannot involve chromosomal translocations since the mechanisms involved in these aberrations have been inactivated in these cells. This notion is consistent with the predominant presence in CLL of genetic alterations, such as deletions and amplifications, common in tumors deriving from tissues not physiologically subjected to antigen receptor gene rearrangements or hypermutation (50).

The comparative analysis of the gene expression profiles of CLL versus memory B cells (Fig.4) or other normal and neoplastic cells (Fig.5) provides a significant body of new information to dissect the CLL phenotype. First, the downregulation of a number of pro-apoptotic genes, together with the already known upregulation of the anti-apoptotic molecule bcl-2, is consistent with the documented long-lived apoptosis-resistant phenotype of CLL (1). Second, unique among normal and neoplastic B cells, CLL cells consistently display a significant overexpression of the EPAC and CDC25 genes, both encoding guanine nucleotide exchange factors that activate, respectively, Rap1 and Ras, the small GTPases that control pleiotropic transcriptional responses via the Raf/ERK pathway (51). This observation has possible pathogenetic relevance and warrants further studies since deregulation of the Ras pathway represents one of the most common alterations in human tumors (52), while recent evidence suggests that Rap1 can be deregulated by chromosomal translocations in lymphoid malignancy (53). Finally, the upregulation of mRNA for several cytokine or chemokine receptors compared to memory cells (IL-4R, TGF β type III receptor, CCR7; Fig.4) or to other B cells (Ror1 and thromboxane A2 receptor; Fig.5) suggests that CLL cells may be abnormally responsive to certain stimuli. While each of the genes specifically expressed in CLL requires validation for differential expression at the protein level, these observations are potentially important for CLL pathogenesis and are amenable to experimental testing.

Clinical implications. The results of these studies have potential clinical application in several areas. First, the ability to distinguish M- versus UM-CLL by gene expression profiling (Fig. 2) represents a potential prognostic test for CLL since these two groups have a distinct clinical course. It is unlikely that this analysis will find clinical application using expensive microarrays with 12,000 genes such as the ones used in this study since IgV sequencing is more rapid and economical. Rather, these results should lead to the development of simple and inexpensive cytochemical assays recognizing the products of few of the genes differentially expressed in M- versus UM-CLL. Second, the products of genes specifically expressed in CLL (Fig. 5) represent potential markers for the diagnosis of CLL, its differential diagnosis from related B cell malignancies, or the detection of small numbers of CLL cells in minimal residual disease contexts. Finally, the products of these same genes represent potential therapeutic targets based on their specific expression in CLL cells versus normal cells. The successful use of monoclonal antibody therapy targeting the CD20 molecule in B cell lymphoma (54) suggests that cell surface receptors that are abnormally expressed in CLL (e.g Ror1 and thromboxane A2 receptors) may represent good candidates for initial testing of this approach.

Further DNA microarray studies may be aimed at determining the expression profiles of cytogenetically distinct CLL cases or cases displaying different clinical behavior (55).

Literature

1. Caligaris-Cappio, F., and T.J. Hamblin. 1999. B-cell chronic lymphocytic leukemia: a bird of a different feather. *J. Clin. Oncol.* 17:399-408.
2. Döhner, H., S. Stilgenbauer, K. Döhner, M. Bentz, and P. Lichter. 1999. Chromosome aberrations in B-cell chronic lymphocytic leukemia: reassessment based on molecular cytogenetic analysis. *J. Mol. Med.* 77:266-281.
3. Kipps, T.J. 1998. Chronic lymphocytic leukemia. *Curr. Opin. Hematol.* 5:244-53.

4. Dighiero, G., T. Kipps, H.W. Schroeder, N. Chiorazzi, F. Stevenson, L.E. Silberstein, F. Caligaris-Cappio, and M. Ferrarini. 1996. What is the CLL B-lymphocyte? *Leuk. Lymphoma* 22 Suppl 2:13-39.
5. Brezinschek, H.P., S.J. Foster, R.I. Brezinschek, T. Dörner, R. Domiati-Saad, and P.E. Lipsky. 1997. Analysis of the human VH gene repertoire. Differential effects of selection and somatic hypermutation on human peripheral CD5(+)/IgM+ and CD5(-)/IgM+ B cells. *J. Clin. Invest.* 99:2488-2501.
6. Fischer, M., U. Klein, and R. Küppers. 1997. Molecular single-cell analysis reveals that CD5-positive peripheral blood B cells in healthy humans are characterized by rearranged V κ genes lacking somatic mutation. *J. Clin. Invest.* 100:1667-1676.
7. Schroeder, H.W., Jr., and G. Dighiero. 1994. The pathogenesis of chronic lymphocytic leukemia: analysis of the antibody repertoire. *Immunol. Today* 15:288-294.
8. Oscier, D.G., A. Thompsett, D. Zhu, and F.K. Stevenson. 1997. Differential rates of somatic hypermutation in V(H) genes among subsets of chronic lymphocytic leukemia defined by chromosomal abnormalities. *Blood* 89:4153-4160.
9. Fais, F., F. Ghiotto, S. Hashimoto, B. Sellars, A. Valetto, S.L. Allen, P. Schulman, V.P. Vinciguerra, K. Rai, L.Z. Rassenti, T.J. Kipps, G. Dighiero, H.W. Schroeder, Jr., M. Ferrarini, and N. Chiorazzi. 1998. Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. *J. Clin. Invest.* 102:1515-1525.
10. Rajewsky, K. 1996. Clonal selection and learning in the antibody system. *Nature* 381:751-758.
11. Küppers, R., U. Klein, M.L. Hansmann, and K. Rajewsky. 1999. Cellular origin of human B-cell lymphomas. *N. Engl. J. Med.* 341:1520-1529.
12. Stevenson, F., S. Sahota, D. Zhu, C. Ottensmeier, C. Chapman, D. Oscier, and T. Hamblin. 1998. Insight into the origin and clonal history of B-cell tumors as revealed by analysis of immunoglobulin variable region genes. *Immunol. Rev.* 162:247-259.
13. Hamblin, T.J., Z. Davis, A. Gardiner, D.G. Oscier, and F.K. Stevenson. 1999. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic. *Blood* 94:1848-1854.
14. Damle, R.N., T. Wasil, F. Fais, F. Ghiotto, A. Valetto, S.L. Allen, A. Buchbinder, D. Budman, K. Dittmar, J. Kolitz, S.M. Lichtman, P. Schulman, V.P. Vinciguerra, K.R. Rai, M. Ferrarini, and N. Chiorazzi. 1999. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* 94:1840-1847.
15. Pasqualucci, L., A. Neri, L. Baldini, R. Dalla-Favera, and A. Migliazza. 2000. BCL-6 mutations are associated with immunoglobulin variable heavy chain mutations in B-cell chronic lymphocytic leukemia. *Cancer Res.* 60:5644-5648.
16. Pascual, V., Y.J. Liu, A. Magalski, O. de Bouteiller, J. Banchereau, and J.D. Capra. 1994. Analysis of somatic mutation in five B cell subsets of human tonsil. *J. Exp. Med.* 180:329-339.

17. Klein, U., K. Rajewsky, and R. Küppers. 1998. Human immunoglobulin (Ig)M+IgD+ peripheral blood B cells expressing the CD27 cell surface antigen carry somatically mutated variable region genes: CD27 as a general marker for somatically mutated (memory) B cells. *J. Exp. Med.* 188:1679-1689.
18. Tangye, S.G., Y.J. Liu, G. Aversa, J.H. Phillips, and J.E. de Vries. 1998. Identification of functional human splenic memory B cells by expression of CD148 and CD27. *J. Exp. Med.* 188:1691-1703.
19. Agematsu, K., S. Hokibara, H. Nagumo, and A. Komiyama. 2000. CD27: a memory B-cell marker. *Immunol. Today* 21:204-206.
20. Ghia, P., V.A. Boussiotis, J.L. Schultze, A.A. Cardoso, D.M. Dorfman, J.G. Gribben, A.S. Freedman, and L.M. Nadler. 1998. Unbalanced expression of bcl-2 family proteins in follicular lymphoma: contribution of CD40 signaling in promoting survival. *Blood* 91:244-251.
21. Hartigan, J.A. 1975. *Clustering Algorithms*. Wiley, New York.
22. Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95:14863-14868.
23. Califano, A. 2000. SPLASH: structural pattern localization analysis by sequential histograms. *Bioinformatics* 16:341-357.
24. Califano, A., G. Stolovitzky, and Y. Tu. 2000. Analysis of gene expression microarrays for phenotype classification. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 8:75-85.
25. Liu, Y.J., C. Barthelemy, O. de Bouteiller, C. Arpin, I. Durand, and J. Banchereau. 1995. Memory B cells from human tonsils colonize mucosal epithelium and directly present antigen to T cells by rapid up-regulation of B7-1 and B7-2. *Immunity* 2:239-248.
26. Golay, J., V. Broccoli, G. Lamorte, C. Bifulco, C. Parravicini, A. Pizzey, N.S. Thomas, D. Delia, P. Ferrauti, D. Vitolo, and M. Introna. 1998. The A-Myb transcription factor is a marker of centroblasts in vivo. *J. Immunol.* 160:2786-2793.
27. Cattoretti, G., C.C. Chang, K. Cechova, J. Zhang, B.H. Ye, B. Falini, D.C. Louie, K. Offit, R.S. Chaganti, and R. Dalla-Favera. 1995. BCL-6 protein is expressed in germinal-center B cells. *Blood* 86:45-53.
28. Onizuka, T., M. Moriyama, T. Yamochi, T. Kuroda, A. Kazama, N. Kanazawa, K. Sato, T. Kato, H. Ota, and S. Mori. 1995. BCL-6 gene product, a 92- to 98-kD nuclear phosphoprotein, is highly expressed in germinal center B cells and their neoplastic counterparts. *Blood* 86:28-37.
29. van Der Vuurst De Vries, A.R., and T. Logtenberg. 1999. A phage antibody identifying an 80-kDa membrane glycoprotein exclusively expressed on a subpopulation of activated B cells and hairy cell leukemia B cells. *Eur. J. Immunol.* 29:3898-3907.
30. Komatsu, K., T. Miyashita, H. Hang, K.M. Hopkins, W. Zheng, S. Cuddeback, M. Yamada, H.B. Lieberman, and H.G. Wang. 2000. Human homologue of *S. pombe* Rad9 interacts with

BCL-2/BCL-xL and promotes apoptosis. *Nat. Cell Biol.* 2:1-6.

31. Sanjo, H., T. Kawai, and S. Akira. 1998. DRAKs, novel serine/threonine kinases related to death-associated protein kinase that trigger apoptosis. *J. Biol. Chem.* 273:29066-71.
32. Douglas, R.S., R.J. Capocasale, R.J. Lamb, P.C. Nowell, and J.S. Moore. 1997. Chronic lymphocytic leukemia B cells are resistant to the apoptotic effects of transforming growth factor-beta. *Blood* 89:941-947.
33. Losman, J.A., X.P. Chen, D. Hilton, and P. Rothman. 1999. Cutting edge: SOCS-1 is a potent inhibitor of IL-4 signal transduction. *J. Immunol.* 162:3770-3774.
34. Martegani, E., M. Vanoni, R. Zippel, P. Coccetti, R. Brambilla, C. Ferrari, E. Sturani, and L. Alberghina. 1992. Cloning by functional complementation of a mouse cDNA encoding a homologue of CDC25, a *Saccharomyces cerevisiae* RAS activator. *EMBO J.* 11:2151-2157.
35. Wei, W., R.D. Mosteller, P. Sanyal, E. Gonzales, D. McKinney, C. Dasgupta, P. Li, B.X. Liu, and D. Broek. 1992. Identification of a mammalian gene structurally and functionally related to the CDC25 gene of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* 89:7100-7104.
36. de Rooij, J., F.J. Zwartkruis, M.H. Verheijen, R.H. Cool, S.M. Nijman, A. Wittinghofer, and J.L. Bos. 1998. Epac is a Rap1 guanine-nucleotide-exchange factor directly activated by cyclic AMP. *Nature* 396:474-477.
37. Soo, C., F.Y. Hu, X. Zhang, Y. Wang, S.R. Beanes, H.P. Lorenz, M.H. Hedrick, R.J. Mackool, A. Plaas, S.J. Kim, M.T. Longaker, E. Freymiller, and K. Ting. 2000. Differential expression of fibromodulin, a transforming growth factor-beta modulator, in fetal skin development and scarless repair. *Am. J. Pathol.* 157:423-433.
38. Cook, T., B. Gebelein, K. Mesa, A. Mladek, and R. Urrutia. 1998. Molecular cloning and characterization of TIEG2 reveals a new subfamily of transforming growth factor-beta-inducible Sp1-like zinc finger-encoding genes involved in the regulation of cell growth. *J. Biol. Chem.* 273:25929-25936.
39. Skonier, J., M. Neubauer, L. Madisen, K. Bennett, G.D. Plowman, and A.F. Purchio. 1992. cDNA cloning and sequence analysis of beta ig-h3, a novel gene induced in a human adenocarcinoma cell line after treatment with transforming growth factor-beta. *DNA Cell. Biol.* 11:511-522.
40. Alizadeh, A.A., M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, L.M. Staudt, and et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403:503-511.
41. Ranheim, E.A., M.J. Cantwell, and T.J. Kipps. 1995. Expression of CD27 and its ligand, CD70, on chronic lymphocytic leukemia B cells. *Blood* 85:3556-3565.
42. Punnonen, J., G. Aversa, B.G. Cocks, and J.E. de Vries. 1994. Role of interleukin-4 and interleukin-13 in synthesis of IgE and expression of CD23 by human B cells. *Allergy* 49:576-586.

43. **Widhopf, G.F., 2nd, and T.J. Kipps. 2001. Normal B cells express 51p1-encoded Ig heavy chains that are distinct from those expressed by chronic lymphocytic leukemia B cells. *J. Immunol.* 166:95-102.**
44. **Dono, M., S. Zupo, N. Leanza, G. Melioli, M. Fogli, A. Melagrana, N. Chiorazzi, and M. Ferrarini. 2000. Heterogeneity of tonsillar subepithelial B lymphocytes, the splenic marginal zone equivalents. *J. Immunol.* 164:5596-5604.**
45. **Weller, S., A. Faili, C. Garcia, M.C. Braun, F.F. Le Deist, G.G. de Saint Basile, O. Hermine, A. Fischer, C. Reynaud, and J. Weill. 2001. CD40-CD40L independent Ig gene hypermutation suggests a second B cell diversification pathway in humans. *Proc. Natl. Acad. Sci. U. S. A.* 98:1166-1170.**
46. **Migliazza, A., F. Bosch, H. Komatsu, E. Cayanis, S. Martinotti, E. Toniato, E. Guccione, X. Qu, M. Chien, V.V. Murty, G. Gaidano, G. Inghirami, P. Zhang, S. Fischer, S.M. Kalachikov, J. Russo, I. Edelman, A. Efstratiadis, and R. Dalla-Favera. 2001. Nucleotide sequence, transcription map, and mutation analysis of the 13q14 chromosomal region deleted in B-cell chronic lymphocytic leukemia. *Blood* 97:2098-2104.**
47. **Mabuchi, H., H. Fujii, G. Calin, H. Alder, M. Negrini, L. Rassenti, T.J. Kipps, F. Bullrich, and C.M. Croce. 2001. Cloning and characterization of CLLD6, CLLD7, and CLLD8, novel candidate genes for leukemogenesis at chromosome 13q14, a region commonly deleted in B-cell chronic lymphocytic leukemia. *Cancer Res.* 61:2870-2877.**
48. **Corcoran, M.M., O. Rasool, Y. Liu, A. Iyengar, D. Grander, R.E. Ibbotson, M. Merup, X. Wu, V. Brodyansky, A.C. Gardiner, G. Juliusson, R.M. Chapman, G. Ivanova, M. Tiller, G. Gahrton, N. Yankovsky, E. Zabarovsky, D.G. Oscier, and S. Einhorn. 1998. Detailed molecular delineation of 13q14.3 loss in B-cell chronic lymphocytic leukemia. *Blood* 91:1382-1390.**
49. **Klein, U., T. Goossens, M. Fischer, H. Kanzler, A. Braeuninger, K. Rajewsky, and R. Küppers. 1998. Somatic hypermutation in normal and transformed human B cells. *Immunol. Rev.* 162:261-280.**
50. **Lengauer, C., K.W. Kinzler, and B. Vogelstein. 1998. Genetic instabilities in human cancers. *Nature* 396:643-649.**
51. **Zwartkruis, F.J., and J.L. Bos. 1999. Ras and Rap1: two highly related small GTPases with distinct function. *Exp. Cell. Res.* 253:157-165.**
52. **Bos, J.L. 1989. ras oncogenes in human cancer: a review. *Cancer Res.* 49:4682-9.**
53. **Hussey, D.J., M. Nicola, S. Moore, G.B. Peters, and A. Dobrovic. 1999. The (4;11)(q21;p15) translocation fuses the NUP98 and RAP1GDS1 genes and is recurrent in T-cell acute lymphocytic leukemia. *Blood* 94:2072-2079.**
54. **White, C.A., R.L. Weaver, and A.J. Grillo-Lopez. 2001. Antibody-targeted immunotherapy for treatment of malignancy. *Annu. Rev. Med.* 52:125-145.**
55. **Stratowa, C., G. Löffler, P. Lichter, S. Stilgenbauer, P. Haberl, N. Schweifer, H. Döhner, and K.K. Wilgenbus. 2001. cDNA microarray gene expression analysis of B cell chronic lymphocytic**

leukemia proposes potential new prognostic markers involved in lymphocyte trafficking. *Int. J. Cancer.* 91:474-480

56. Matsuda, F., E.K. Shin, H. Nagaoka, R. Matsumura, M. Haino, Y. Fukita, S. Taka-ishi, T. Imai, J.H. Riley, R. Anand, et al. 1993. Structure and physical map of 64 variable segments in the 3 0.8-megabase region of the human immunoglobulin heavy-chain locus. *Nat. Genet.* 3:88-94

Acknowledgements

We are grateful to Vladan Miljkovic for technical assistance. We also thank Laura Pasqualucci for discussions and Richard Baer for comments on the manuscript. U.K. was recipient of fellowships granted by the European Molecular Biology Organization and Human Frontiers Science Program. G.C. is recipient of an Aboodi Associate Professor Fellowship. H.H. was supported by the Cure for Lymphoma Foundation, M.M. by the Università degli Studi di Milano, A.N. by the Associazione Italiana Ricerca sul Cancro (AIRC). A.F. received support from the United States National Institute of Health (grant CA66996), the Leukemia and Lymphoma Society, and the Norman Hirschfield Foundation.

Figure Legends

Figure 1: M- and UM-CLLs share a common gene expression profile. Dendrogram and matrix showing the hierarchical clustering of 2337 selected genes (see Methods) of gene expression data generated from 34 CLL (P indicates purified cases) and 6 purified FL samples. The hierarchical clustering algorithm used is based on the average-linkage method (21, 22). FL, M- and UM-CLL samples are coded by red, blue, and green respectively. The matrix below the dendrogram depicts the gene expression values of the individual samples, with columns representing individual tumor samples and rows representing individual genes ordered according to hierarchical clustering. The color scale identifies relative gene expression changes normalized by the standard deviation, with 0 representing the mean expression level of a given gene across the panel.

Figure 2: Identification of genes differentially expressed in M- and UM-CLL. (a) Supervised cluster analysis of M- and UM-CLLs. Eight purified M- and eight purified UM-CLL samples were examined by supervised clustering using Genes@Work (23, 24). Columns represent individual CLL samples, rows correspond to genes. Color changes within a row indicate expression levels relative to the average of the sample population. Values are quantified by the scale bar that visualizes the difference in the z_{ge} score relative to the mean (0). Genes are ranked as described in the matrix description section. The support value for supervised analysis was chosen as $n = n_0 - 2$, where n_0 is the number of cells in the phenotype set, allowing for up to two unclustered cells per pattern in the phenotype set. Gene names and fold change are indicated; for GenBank accession numbers, see Supplementary Table I. Note that one of the discriminating genes is a IgV region (V4-31) that is not expressed in the CLL cases; this is due to the fact that: errors are expected for mutated IgV genes since single base pair variations affect the hybridization to oligonucleotides in the chip leading to misidentification of the IgV family member. (b) The gene expression profiles distinguishing M- from UM-CLL (classifier) can predict the M- versus UM- status of

an independent panel of unpurified CLL. The classifier (23 genes; Fig.2a) was applied to score 14 unpurified CLL cases (Table I), each identified by an open circle. The number of IgV gene mutations of each case is indicated within the open circle. The relatedness of each of the test samples to the two CLL subgroups is indicated by their proximity to either subgroup on the vertical axis (p -values are shown). The gray area marks the 95% confidence region, i.e. any sample with a score beyond this range can be assigned to one of the CLL subgroups with more than 95% confidence (see Methods).

Figure 3: The gene expression profile of CLL is related to that of memory B cells. Gene expression data sets generated from 20 purified CLL cases are compared to the genes differentially expressed between CB/CC and memory/naïve B cells (a), memory/naïve B cells vs. CD5⁺ cells (b), and memory vs. naïve B cells (c). Genes differentially expressed between the various B cell subpopulations were identified by supervised clustering using Genes@Work. Matrices (a-c) and gene ranking are as in Fig.2. Genes known to be differentially expressed among GC and non-GC B-cells and among naïve and memory B-cells, or on CLL cells (IL4R), are indicated (19, 25-29). Panel (d) shows the quantitative relatedness of the CLLs to the normal B cell populations as derived from panels a-c as described for Fig.2b. The gray area marks the 95% confidence region; the lower and upper margins of the gray area each correspond to a p -value of 0.025 (p -values decrease with increasing distance from the x-axis). Open and closed circles represent M- and UM-CLL cases, respectively.

Figure 4: Identification of genes specifically expressed in CLLs vs. memory B cells. Gene expression profiles of 10 randomly selected purified CLL cases (5 UM- and 5 M-CLL, respectively) were compared to those of 5 memory cell preparations by supervised clustering using Genes@Work. Matrices and gene ranking are as in Fig.2; the fold change is indicated. For GenBank accession numbers, see Supplementary Table II. Genes are grouped according to putative functional categories and ranked within each category. CD5 and CD23 are indicated as internal controls.

Figure 5: Identification of genes specifically expressed in CLL. Gene expression profiles of 10 randomly selected purified CLL cases (5 UM- and 5 M-CLL, respectively) were compared to those generated from normal (CB, CC, naïve and memory) B cell subpopulations, purified non-Hodgkin lymphoma cells (DLCL, BL, and FL), and DLCL and BL cell lines by supervised clustering using Genes@Work. Matrices and gene ranking are as in Fig.2. Genes and fold change are indicated at the right; for GenBank accession numbers, see Supplementary Table III.