Integrative Genomic Analysis Implicates Gain of PIK3CA at 3q26 and MYC at 8q24 in Chronic Lymphocytic Leukemia

Jennifer R. Brown1,7, Megan Hanna3,8, Bethany Tesar1, Lillian Werner2, Nathalie Pochet8,10, John M. Asara5,7, Yaoyu E. Wang4, Paola dal Cin6, Stacey M. Fernandes1, Christina Thompson1, Laura MacConaill3,6, Catherine J. Wu1,7, Yves Van de Peer10, Mick Correll4, Aviv Regev8,9, Donna Neuberg2, and Arnold S. Freedman1,7

Abstract

Purpose: The disease course of chronic lymphocytic leukemia (CLL) varies significantly within cytogenetic groups. We hypothesized that high-resolution genomic analysis of CLL would identify additional recurrent abnormalities associated with short time-to-first therapy (TTFT).

Experimental Design: We undertook high-resolution genomic analysis of 161 prospectively enrolled CLLs using Affymetrix 6.0 SNP arrays, and integrated analysis of this data set with gene expression profiles.

Results: Copy number analysis (CNA) of nonprogressive CLL reveals a stable genotype, with a median of only 1 somatic CNA per sample. Progressive CLL with 13q deletion was associated with additional somatic CNAs, and a greater number of CNAs was predictive of TTFT. We identified other recurrent CNAs associated with short TTFT: 8q24 amplification focused on the cancer susceptibility locus near MYC in 3.7%; 3q26 amplifications focused on PIK3CA in 5.6%; and 8p deletions in 5% of patients. Sequencing of MYC further identified somatic mutations in two CLLs. We determined which catalytic subunits of phosphoinositide 3-kinase (PI3K) were in active complex with the p85 regulatory subunit and showed enrichment for the α subunit in three CLLs carrying PIK3CA amplification.

Conclusions: Our findings implicate amplifications of 3q26 focused on PIK3CA and 8q24 focused on MYC in CLL. Clin Cancer Res; 18(14); 3791–802. ©2012 AACR.

Introduction

Chronic lymphocytic leukemia (CLL) is the most common leukemia of adults but still incurable. Prognosis at diagnosis is widely variable, and the key cytogenetic abnormalities determined by FISH remain one of the best predictors of prognosis and treatment response (1). However, disease course still varies significantly within these cytogenetic groups, and our ability to predict prognosis remains limited. Once treated, CLL inevitably relapses and each subsequent remission gets shorter.

The advent of high-resolution array–based technologies lends itself to the detailed characterization of cancers on multiple levels including, but not limited to, copy number, gene expression, protein expression and modification, and methylation. Copy number analyses have been undertaken extensively in cancer, allowing the recent publication of a article that surveys the landscape of somatic copy number alterations (CNA) primarily in solid tumors but including acute lymphoblastic leukemia (ALL; ref. 2). A small high-resolution study in CLL described wide variability in the number of CNAs observed (3), though a larger study looking at newly diagnosed patients (4) found a generally stable genome in previously untreated patients. Recent efforts have dissected the structure of 13q and 11q deletions in detail (5, 6, 7) and associated the number of CNAs with overall survival (8). We hypothesized that high-risk CLLs would likely harbor additional recurrent CNAs, not part of the canonical CLL FISH panel, but which would likely reflect on disease pathogenesis. We therefore undertook a large integrative study of CLL, using both copy number analysis with direct comparison to cognate germline and gene expression profiling, to characterize the CLL genome at very high resolution.

Materials and Methods

CLL patients

One hundred and sixty-one patients with CLL enrolled on a prospective cohort natural history study at Dana-Farber...
Clinical Cancer Research
Clin Cancer Res; 18(14) July 15, 2012

Translational Relevance
CLL is a heterogeneous disease in which genetic markers are the most effective determinants of prognosis. In our study, we have identified and characterized 3 recurrent genomic abnormalities that are associated with progressive CLL. Our very high-resolution platform allowed us to show that the targets of two of these abnormalities are MYC and PIK3CA, both well-known oncogenes in solid tumors but lacking a well-described role in CLL. Our work provides the foundation for assessing the potential importance of these genetic abnormalities as prognostic markers in prospective clinical trials. Furthermore, drugs that target PIK3CA are already in clinical trials in CLL, so determining whether the PIK3CA amplifications identified in this work will predict sensitivity to treatment with these drugs will be critically important.

Gene expression profiling
All 146 CLLs, for which adequate quality RNA was obtained, were assessed for their mRNA expression profile by hybridization to the Affymetrix U133 Plus 2.0 array. RNA was isolated from viable frozen tumor cells using the Qiagen RNeasy Mini kit. RNA quality was assessed by A260:280 ratios and by RNA integrity number (RIN) analysis on a Bioanalyzer before submission for gene expression profiling. All expression profiles were processed using Robust Multi-Array Average (RMA), implemented by the PreprocessData-set module in GenePattern (http://www.broadinstitute.org/cancer/software/genepattern; refs. 12, 13). Probes were collapsed to unique genes by selecting the probe with the maximal average expression for each gene. Batch effects were further removed using the ComBat module in GenePattern (13).

To examine gene expression in relation to each CNA, differential gene expression was determined between samples with and without the CNA. Genes were selected according to a t test P value <0.05, using the Comparative Marker Selection and Extract Comparative Marker Results modules in GenePattern (13, 14). The significance (nominal P value) of each marker gene was computed using a permutation test based on 1,000 replications. Samples were ordered in heat maps on the basis of the correlation of their expression phenotype to that of the samples with the given CNA. Those with a positive correlation are displayed to the right of the gap in the heat maps, and were analyzed in gene set enrichment analysis (GSEA), as detailed in the Supplementary Methods.

Immunoprecipitation of p85 and shotgun liquid chromatography/tandem mass spectrometry
Approximately 10^8 previously frozen PBMCs from individuals with and without the 3q26 amplification were thawed, ficolled, and lysed, and a minimum of 0.5 mg cell lysate was subjected to immunoprecipitation with anti-p85 (N-SH2 antibody; Millipore) followed by mass spectrometry analyses to identify p85 interacting proteins. Further details of this procedure and the quantification technique can be found in the Supplementary Methods.

Results
Characteristics of patient cohort
The clinical characteristics of the patients are shown in Supplementary Table S2. Seventy-eight percent were untreated at the time of sampling, whereas 22% had been previously treated. An additional 21% of patients were treated in the subsequent follow-up period, for a total...

Cancer Institute (DFCI; Boston, MA) were studied. The diagnosis of CLL, according to the World Health Organization (WHO) criteria was confirmed in all cases. Any individual 18 years or older seen at DFCI with CLL/small lymphocytic lymphoma (SLL) was eligible. The study was approved by the Dana-Farber Institutional Review Board, and all subjects signed written informed consent. CLL cells were collected from peripheral blood as detailed in the Supplementary Methods. The median follow-up from diagnosis is 81 months.

Genome-wide DNA profiling
Genome-wide DNA profiles were obtained using the Genome-wide Human SNP Array 6.0 (Affymetrix), run on the Genetic Analysis Platform at the Broad Institute of Harvard and MIT (Cambridge, MA), according to the manufacturer’s protocol. Tumor DNA [peripheral blood mononuclear cells (PBMC) or isolated B cells] and at least one matched germline control (saliva, granulocytes, or both) were run for every sample. The quality of all DNAs was verified in 3 independent PCR reactions before use. The data were initially analyzed by the Genomic Identification of Significant Targets in Cancer (GISTIC) method, which identifies significant deletions and amplifications based on the analysis of the frequency of occurrence and the amplitude of each aberration in the tumor samples alone, as previously described (2, 9). The GISTIC analysis used the 5 nearest neighbor normalization method and removed all catalogued germ line copy number variants (CNV; ref. 10). Statistical significance was assessed using a permutation test on the basis of the overall pattern of aberrations across the genome and accounted for multiple hypothesis testing using a false discovery rate (FDR) framework, with q values <0.05 considered significant. To confirm that all abnormalities identified by GISTIC were somatic, the paired somatic and germline samples were manually reviewed using the Integrative Genomics Viewer (IGV) from the Broad Institute.
of 43.6% treated, with an overall median time from diagnosis to treatment of 41 months (0.4–161.2 months). Unmutated IGHV and ZAP-70 positivity were both highly predictive of reduced TTFT and overall survival (OS), as previously reported, although data were missing for a subset of patients (Supplementary Tables S3 and S4).

Eighty-two percent of the 161 CLLs showed at least one CNA by high-resolution single-nucleotide polymorphism (SNP) array (Fig. 1 and Table 1). GISTIC analysis on the entire population identified the known common CLL abnormalities at frequencies that would be expected in a largely untreated cohort: 57% deletion 13q, 6.2% deletion 11q, 5.6% deletion 17p, and 12% trisomy 12 (Fig. 1A and B and Table 1). Deletions 17p and 11q are known to have an adverse prognosis with short TTFT and OS (1) and this finding was confirmed in our study for TTFT, whereas OS was adversely impacted at this point only for 17p deletion (Supplementary Tables S3 and S4).

Number of CNAs predicts TTFT

GISTIC analysis revealed a paucity of CNAs in comparison with most solid tumors (Fig. 1A and B and Table 1), as previously reported (8), and some of the identified regions were actually rare germline CNVs when compared with cognate germline. Our analysis revealed that 1 of 4 significant amplifications and 14 of 29 significant deletions identified by GISTIC were also present in the corresponding germline. Therefore, we assessed total somatic CNAs per sample using not only the GISTIC analysis but also direct tumor–normal comparison and manual review. The median number of acquired somatic CNAs in the overall population was 1, but in treated patients the median was 2 (Supplementary Tables S1 and S5; see Supplementary Methods). Choosing a cut-off at the median or one CNA greater served to dichotomize groups with markedly different TTFTs (Fig. 1D). In CLLs with deletions of 11q or 17p, a median of 3 CNAs per patient was observed; 41% of patients with 3 CNAs or more and 70% of patients with 4 CNAs or...
more also had 11q or 17p deletions (Supplementary Table S5). Even in patients without 11q or 17p deletions, however, increasing number of CNAs remained predictive of short TTFT, suggesting that number of CNAs is an independent adverse predictor (Fig. 1E), as previously reported (8).

Progression in 13q deletion is associated with presence of other CNAs

Ninety-one patients (57%) had 13q deletions, and GISTIC analysis categorized them into 3 groups on the basis of decreasing size labeled R1, R2, and R3 (Table 1 and Fig. 2A). As shown in Fig. 2A, all 91 patients were at least partially deleted in the R3 region, which includes the previously described minimally deleted region extending from DLEU2 and miR-15a/16-1 to DLEU7 (6, 15). Interestingly, we were unable to define a universally present minimally deleted region (MDR) in our data set, as 3 patients had partial deletion of DLEU2 but lacked deletion of miR-15a/16-1 (Supplementary Fig. S1, section 2), and one patient completely lacked deletion of the DLEU2 and miR-15a/16-1 region, instead showing only deletion of DLEU7 (Supplementary Fig. S1, section 3). A second patient showed biallelic deletion of DLEU7 with only monoallelic deletion of the DLEU2 miR-15a/16-1 region (Supplementary Fig. S1, section 3). These findings are consistent with recent results suggesting that multiple genes in this region contribute to the CLL phenotype (16–19).

Longer 13q deletions were seen in 30 patients and were subdivided into 2 groups by GISTIC, with a subset having deletions extending just past the RB gene (and thereby deleting regions R2 and R3 in Fig. 2A and Table 1), and the remainder having very long deletions including RB and extending up to 40 Mb (thereby deleting regions R1, R2, and R3 in Fig. 2A and Table 1). Longer 13q deletions that delete RB have been labeled type II deletions and reported to carry a poor prognosis (5–7), but in our cohort, no significant difference in TTFT was observed between patients with short 13q deletions confined to R3 and those with longer 13q deletions (Fig. 2B), defined as either those extending to R2 only, or those extending through R1 and R2.

Table 1. Somatic MDRs identified by GISTIC

<table>
<thead>
<tr>
<th>Genomic location of CNA</th>
<th>N (%)</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>GISTIC gains</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2p16.3</td>
<td>3 (1.9)</td>
<td>32,764,999</td>
<td>70,027,926</td>
</tr>
<tr>
<td>3q26</td>
<td>9 (5.6)</td>
<td>180,434,350</td>
<td>180,434,928</td>
</tr>
<tr>
<td>8q24</td>
<td>6 (3.7)</td>
<td>128,267,747</td>
<td>128,426,358</td>
</tr>
<tr>
<td>Trisomy 12</td>
<td>19 (11.8)</td>
<td>50,065,017</td>
<td>76,116,030</td>
</tr>
<tr>
<td>18q</td>
<td>3 (1.9)</td>
<td>18,902,673</td>
<td>19,673,569</td>
</tr>
<tr>
<td>Trisomy 19</td>
<td>1 (0.6)</td>
<td>2,467,076</td>
<td>23,835,347</td>
</tr>
<tr>
<td>Trisomy 7</td>
<td>1 (0.6)</td>
<td>77,753,344</td>
<td>77,909,776</td>
</tr>
<tr>
<td>13q</td>
<td>1 (0.6)</td>
<td>21,618,001</td>
<td>21,819,902</td>
</tr>
<tr>
<td>16p</td>
<td>1 (0.6)</td>
<td>187,977,019</td>
<td>190,299,620</td>
</tr>
<tr>
<td>16q</td>
<td>1 (0.6)</td>
<td>86,702,690</td>
<td>132,661,780</td>
</tr>
<tr>
<td>17p</td>
<td>1 (0.6)</td>
<td>141,945,240</td>
<td>142,213,197</td>
</tr>
<tr>
<td>GISTIC losses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1q44 loss</td>
<td>4 (2.5)</td>
<td>243,904,769</td>
<td>244,959,448</td>
</tr>
<tr>
<td>3p14.2 loss</td>
<td>1 (0.6)</td>
<td>34,513</td>
<td>85,197,968</td>
</tr>
<tr>
<td>3p26.2 loss</td>
<td>1 (0.6)</td>
<td>34,513</td>
<td>7,303,192</td>
</tr>
<tr>
<td>4q35.2 loss</td>
<td>1 (0.6)</td>
<td>187,977,019</td>
<td>190,299,620</td>
</tr>
<tr>
<td>5q14.3 loss</td>
<td>1 (0.6)</td>
<td>86,702,690</td>
<td>132,661,780</td>
</tr>
<tr>
<td>7q34 loss</td>
<td>7 (4.3)</td>
<td>102,877,806</td>
<td>113,811,014</td>
</tr>
<tr>
<td>8p23.1 loss</td>
<td>8 (5.0)</td>
<td>40,438,503</td>
<td>40,478,258</td>
</tr>
<tr>
<td>11q22.3 loss</td>
<td>10 (6.2)</td>
<td>47,405,627</td>
<td>47,786,494</td>
</tr>
<tr>
<td>13q14.11 R1</td>
<td>11 (6.8)</td>
<td>49,538,817</td>
<td>50,081,506</td>
</tr>
<tr>
<td>13q14.2 R2</td>
<td>19 (11.8)</td>
<td>91,704,917</td>
<td>93,186,479</td>
</tr>
<tr>
<td>13q14.3 R3</td>
<td>61 (37.9)</td>
<td>91,704,917</td>
<td>93,186,479</td>
</tr>
<tr>
<td>Biallelic</td>
<td>22 (13.7)</td>
<td>91,704,917</td>
<td>93,186,479</td>
</tr>
<tr>
<td>14q31.1</td>
<td>2 (1.2)</td>
<td>21,507,073</td>
<td>21,555,877</td>
</tr>
<tr>
<td>17p12</td>
<td>9 (5.6)</td>
<td>7,258,530</td>
<td>7,990,010</td>
</tr>
<tr>
<td>20p arm</td>
<td>2 (1.2)</td>
<td>9,756</td>
<td>26,223,105</td>
</tr>
<tr>
<td>22q11.22</td>
<td>53 (32.9)</td>
<td>21,507,073</td>
<td>21,555,877</td>
</tr>
</tbody>
</table>
Our analysis further showed that 24% of the del 13q patients carried biallelic 13q deletions. Seventeen of 61 patients (28%) with short deletions confined to the R3 region had biallelic loss, as compared with 5 of 30 patients (16.7%) with long deletions extending to R1 or R2 ($P = 0.3$). Of the latter 5 patients, the regions of biallelic deletion were confined to the much smaller R3 region around miR-15a/16-1, whereas the longer deletion regions were monoallelic. In previous work, we and others have reported that biallelic 13q deletions are associated with a longer TTFT (4, 20). However, in this study, TTFT was similar for both groups (Fig. 2C). This finding was true for the entire cohort as well as for those patients untreated at time of sampling (Fig. 2D). Thus, additional CNAs identified by high-density SNP array can most effectively identify patients with 13q deletion who will progress rapidly to treatment. Furthermore, no association of number of CNAs with long versus short 13q deletions was observed ($P = 0.24$).

Gene expression profiling analysis was used to assess genes located in the 13q deletion region, focusing in particular on DLEU2 and TRIM13 (deleted in R3) and RB1 (deleted in R2). DLEU2 and TRIM13 at 13q14 showed significant downregulation only in samples with biallelic deletion (Fig. 2E). Samples with monoallelic deletion at GISTIC regions R1 and R2, resulting in deletion of one copy of the RB1 gene, did show significant downregulation of RB1 compared with those samples intact at 13q or deleted only at GISTIC R3, which does not include RB1 (Fig. 2E).

**CNAs associated with disease progression**

To identify recurrent CNAs associated with high-risk disease, we compared the molecular profiles of patients...
who remained untreated to the overall cohort (Fig. 1B and C). The genomes of untreated patients were significantly more stable, showing predominantly 13q deletion or trisomy 12 (Fig. 1C). Patients who had undergone therapy either before or after sampling showed 3 highly significant additional abnormalities: amplification at 3q26.32 and 8q24.21 and deletion at 8p (Fig. 1A and B). The amplification events include focal amplifications at the known oncogenes PIK3CA and MYC, respectively.

**Focal amplification of 3q26.32 affects PIK3CA**

Nine patients (9 of 161 or 5.6%) had amplification of 3q26.32. Three of these patients showed focal somatic amplification of a small region (<350 bp) corresponding to exon 21 of PIK3CA (Fig. 3A), whereas the remainder had large gains (Fig. 3A). PIK3CA encodes the p110-α isoform of the phosphoinositide 3-kinase catalytic subunit (PI3K), 1 of 4 class I isoforms, which also include β, γ, and δ, and PIK3CA is mutated by amplification or activating mutation in many solid tumors (21, 22). A subset of the amplifications we observed were confirmed by FISH, with a probe to the MECOM gene at 168.8 Mb (near PIK3CA at 180 Mb). FISH showed that 3 patients had 3 copies, and one patient had 3 to 6 copies (data not shown). These amplifications also seemed to be associated with a mildly reduced TTFT (P < 0.0001; Fig. 3B); the numbers are too small to assess focal and broad

---

**Figure 3.** A, amplification of 3q26.32, focused on PIK3CA exon 21. Bottom, segmented (top) and raw (bottom) data for tumor and normal as noted. B, TTFT for 3q26 patients with or without deletions of 11q or 17p. Neither, median 87 months (n = 136); 3q26 amplification alone, 62 months (n = 7); deletions 11q or 17p alone, 27 months (n = 15); 3q26 amplification with deletions 11q or 17p, 3.5 months (n = 2; P = 0.0001). C, PIK3CA expression determined by gene expression profiling. D, immunoprecipitation of PIK3CA from 2 CLLs with broad and 2 with focal 3q26 amplifications, and 5 controls with quantitation of bands below. E, percentage of p110-α, β, and γ catalytic subunits in complex with p85, in samples with broad 3q26 amplification (n = 3) and controls without amplification (n = 6), as measured by mass spectrometry (P < 0.05). F, ratio of α5 to p110 catalytic subunit, in complex with p85 regulatory subunit (P = 0.02). Samples (n = 3) with broad amplifications and 6 control samples, one of which had lower spectral counts (50). G, heat map showing the CLL cohort ordered by correlation of gene expression with the 3q26 gain samples. Samples with 3q26 gain are labeled in red at right, and samples to the right of the gap show positive correlation with the 3q26 expression pattern. Tick marks at right indicate the locations of genes present on chromosome 3q26 (see Supplementary Table S20 for names).
amplification patients separately (Supplementary Fig. S2). Five of these 9 patients are deceased, as compared with 18 of 152 without this amplification, suggesting a possible effect on overall survival, although again the numbers are small.

The significance of the focal PIK3CA amplifications compared with the broad amplifications is unclear. The focal amplifications affect the kinase domain of the protein, which is a hotspot for somatic mutation in solid tumors, but would not be expected to increase RNA or protein expression. In fact, when we analyze PIK3CA RNA expression by gene expression profiling (GEP), we see increased expression in the broad but not the focal amplification patients (Fig. 3C). Immunoprecipitation of the α-isof orm of PI3K from CLLs with focal and broad 3q26 amplifications also shows increased protein in the broad amplification samples but not in the focal amplification samples (Fig. 3D). To assess the functional significance of increased expression in the samples with broad gains, we determined which catalytic subunits of PI3K were in complex with the p85 regulatory subunit in CLL cells. To accomplish this, we carried out immunoprecipitation experiments with an antibody to the p85 regulatory subunit and then used mass spectrometry to identify the proteins in complex with p85, in 3 patients with broad amplifications and 6 control patients with no copy number gain. We found that in the control CLL patients, the δ subunit of PI3K was the predominant p110 catalytic subunit associated with p85 (48%). In contrast, in the 3 CLLs carrying broad amplifications of PIK3CA that we were able to assess, δ represented only 33% of associated catalytic subunits (P = 0.04; Fig. 3E), and the α subunit was enriched in complex relative to δ (α:δ ratio for gain samples 1.24 vs. 0.43 for controls, P = 0.02; Fig. 3F). These results suggest that at least the broad 3q26 amplifications result in altered PI3K subunit composition in these CLLs.

Given that mutations in the PI3K pathway (PIK3CA and PTEN in particular) are well described in solid tumors, we sequenced the entire coding regions of PIK3CA, PIK3CD, PIK3CG, PTEN, and PIK3R1 and genotyped AKT E17K in 188 CLLs. No somatic mutations were identified, suggesting that point mutation is not a mechanism of activation of these genes in CLL, even though the PI3K pathway has been shown to be constitutively activated (23, 24).

We evaluated whether the gene expression profiling data showed any pattern associated with 3q26 gain (Fig. 3G). Supervised analysis identified 2,981 genes that were differentially expressed between CLLs with and without 3q26 gain (Supplementary Table S6). The CLL samples were then ordered on the basis of the correlation of their gene expression with that of the 3q26 gain samples (Fig. 3G). In the 3q26 gain samples themselves, as well as those samples with gene expression that positively correlated with the 3q26 gain samples, an exploratory analysis using GSEA identified increased expression of gene sets composed of genes repressed by Polycomb complexes in embryonic stem (ES) cells (Supplementary Table S7). This finding is discussed further below.

**Amplification of 8q24 affects MYC**

Amplification at 8q24 was present in 6 of 161 CLLs (3.7%; Fig. 4A). Amplification of MYC was confirmed by FISH, with 2 patients harboring 3 intact copies of MYC, 1 patient 4 copies, and 1 patient had one rearranged copy with 2 intact copies (data not shown). Two patients had focal amplification of the "gene desert" regulatory region approximately 360 kb centromeric to MYC, indicated by the GISTIC plot at the bottom of Fig. 4A. This 8q24 "gene desert" region contains multiple SNPs that have been implicated by genome-wide association studies (GWAS) in susceptibility to multiple solid tumors as well as CLL (25–27). 8q24 amplification was associated with short TITF, which seemed to be independent of co-occurrence with high-risk deletions of 11q and 17p, although the numbers in each group are very small (P = 0.0001; Fig. 4B). Western blot showed increased MYC expression in samples with gain compared with several controls without gain (Fig. 4C).

Because we observed focal gains affecting a region previously implicated as a MYC regulatory region, the likely target of 8q24 amplification seemed to be the MYC locus. Because known mutations in exon 1 of MYC lead to Burkitt lymphoma, we sequenced exon 1 of MYC in 188 CLL samples. We found one sample with a MYC Thr58Ala mutation, which has been previously described in Burkitt lymphoma and shown to abrogate a regulatory phosphorylation site, leading to activation of MYC (28). Interestingly this mutation impairs FBXW7-mediated degradation of MYC by the proteasome (29), and we have recently identified recurrent mutations in FBXW7 in CLL (30), suggesting that MYC may be a target of FBXW7 in CLL. We also identified a second somatic mutation in MYC, a heterozygous insertion mutation that duplicates 9 amino acids of the N-terminal interaction and transactivation domain (Supplementary Fig. S3); the patient carrying this mutation had a very short TITF and died 49 months from diagnosis. Although infrequent, these somatic mutations suggest another possible mechanism of MYC involvement in CLL.

Analysis of GEP data comparing patients with 8q24 amplification to patients without identified 5,307 genes, whose expression was significantly different, and the samples were again ordered on the basis of the correlation of their gene expression pattern with that of the 8q24 amplified samples (Fig. 4D, Supplementary Table S8). 8q24 samples and the samples with gene expression similar to 8q24 samples again showed enrichment for gene sets repressed by Polycomb complexes in ES cells, similar to the findings in the 3q26 gain samples. Interestingly, many of the genes differentially regulated in the 3q26 and 8q24 gain samples were shared between them (1,290 genes, FDR-corrected P = 3.4 x 10⁻¹¹⁵; Supplementary Table S10).

**Deletion of 8p predicts short TITF**

Deletion at the 8p locus was observed in 8 of 161 samples (5.0%). The common region of deletion was broad, spanning 11.0 to 29.6 Mb (Fig. 5A). A previous report found that 28% of a small number of 17p-deleted patients also harbored deletion at the 8p locus (31). In our study, 3 of 8
patients with 8p deletion had a coexistent 17p deletion and a fourth had a coexistent 11q deletion. Six of the 8 patients had not been treated at the time of sampling, indicating that the deletion occurred \textit{de novo}. Deletion at 8p was associated with short TTFT with 7 of 8 patients subsequently undergoing therapy and rapidly ($P < 0.0001$; Fig. 5B). TTFT was short independent of deletions 17p or 11q, although the numbers are small (Fig. 5B). OS also seemed poor, with 4 of 8 of these patients deceased, compared with 19 of 153 in the overall cohort, although again the numbers are small.

Coanalysis with GEP available for 7 of the 8 CLLs with 8p loss identified 807 genes that were differentially regulated, including 63 located on chromosome 8p itself (Fig. 5C, Supplementary Table S11). Using GSEA on samples with gene expression correlated with the 8p deletion samples again identified upregulation of genes that are repressed by Polycomb in ES cells, although a significant overlap with the differentially regulated genes in the 3q26 and 8q24 samples was not observed (Supplementary Table S12).

**Single-sample GSEA (SSGSEA) identifies common expression signature previously associated with hematopoietic stem cells**

Exploratory GSEA analysis of all 3 CNA groups identified increased expression of gene sets composed of targets of Polycomb-based silencing in ES cells (32). Previous work in ES cells has identified several core components of their gene expression signature: a Polycomb cluster of genes bound by Polycomb complex factors; a Core cluster of genes bound by the pluripotency factors Oct4, Sox2, and Nanog; and a Myc cluster of genes targeted by Myc (32). Differential expression of these clusters has been described in hematopoietic
stem cells (HSC) and a variety of cancers (33). Therefore, to further characterize the finding of Polycomb cluster over-expression in our CNA groups, we used SSGSEA to test these gene sets previously reported as ES cell gene sets (32, 34, 35). SSGSEA showed that the CNA group in each case, which included samples with the CNA itself as well as those samples with a positively correlated gene expression pattern, was enriched for gene modules previously associated with self-renewing long-term HSCs, specifically showing induction of ES Polycomb (SUZ12, EED, and H3K27ME3) and ‘core ES’ gene sets (WEINBERG_ES_CORE_NINE and WEINBERG_ES_2), together with repression of ES cell gene sets reflecting proliferation, such as MYC and proliferation gene sets (Supplementary Tables S13–S15; refs. 32, 34). Control samples in contrast showed the opposite pattern of enrichment, with induction of ES Myc and proliferation modules and repression of ES Polycomb modules, a pattern previously associated with short-term HSCs (32, 34). Interestingly, the same SSGSEA analysis conducted on samples with 17p deletion, 11q deletion, high-number of CNAs (≥2), and unmutated IGHV failed to identify any consistent pattern between the high- and low-risk groups (Supplementary Tables S16–S19). Therefore, given the similar results among the 3 CNA groups, we assessed for overlap among the samples with GEP that correlated with each CNA. Here, we found substantial overlap among the CLL samples related to each CNA (FDR-corrected P values: $6.4 \times 10^{-12}$ for the overlap between 3q26 and 8p, $3.1 \times 10^{-21}$ for the overlap between 3q26 and 8q24, and $3.2 \times 10^{-12}$ for the overlap between 8p and 8q24). We therefore investigated whether these samples sharing gene expression patterns correlated with all 3 CNAs showed any common clinical features or shorter TTFT. We found that the distribution of clinical features in this group was similar to the overall group (Supplementary Fig. S4A). However, the samples sharing the CNA gene expression pattern showed a significantly shorter TTFT compared with control samples, with median 61 months compared with 161 months for the control group ($P = 0.03$; Supplementary Fig. S4B). These results suggest that patterns of gene expression previously associated with HSC biology may play a heretofore uninvestigated role in CLL. Future work will need to test this hypothesis in other CLL cohorts.

Discussion

We report the results of a large-integrated analysis of SNP array screening and gene expression profiling of the CLL genome. Significant advantages of our data set include the use of an extremely high-resolution platform and comparison with matched germline DNA, allowing clear determination of somatic events and filtering of previously undescribed germline CNVs. We find that CLL is quite
genomically stable compared with most solid tumors, with a median of only 1 CNA per genome in stably untreated patients, often 13q deletion or trisomy 12. This estimate is lower than earlier studies without matched germline controls (4, 36), but similar to more recent studies that included matched germline analysis (5). The genomic stability of indolent CLL is unsurprising given that many cases display a benign course with minimal progression for years.

We also found that an increasing number of somatic CNAs were predictive of short TTFT, as previously reported (8, 36). This finding was true in the entire cohort and in those lacking 17p and 11q deletions, suggesting that increasing CNAs is independently associated with short TTFT. Similarly, additional CNAs were the major predictor of short TTFT in the context of 13q deletion. The size of 13q deletion and whether it was mono- or biallelic have both been reported to have prognostic significance, but in this cohort neither feature was predictive. The most important predictor of TTFT was the presence of any additional somatic CNA defined by SNP array. Our data are therefore an extension of Döhner and colleagues’ original observation using FISH (1), but our conclusions are based on a much higher resolution platform, and thus may allow more definitive prognostic prediction.

Although the number of CNAs has prognostic significance, it remains likely that specific recurrent CNAs may target genes important in CLL pathogenesis. We were therefore interested in identifying other genomic regions targeted recurrently in CLL and found 3 that were significantly associated with requiring therapy in our data set: amplification of 3q26 focused on PIK3CA, amplification of 8q24 focused on the known GWAS cancer risk region near MYC, and 8p loss. All 3 broad chromosomal regions have been reported previously, but the targets of these broad events in CLL were not previously hypothesized. The very high-resolution platform used here allowed us to identify very focal CNAs that suggest likely targets in 2 of these cases, PIK3CA and MYC. In fact, a recent study by Beroukhim and colleagues that catalogued the CNAs observed in more than 3,000 cancers found that amplifications do most commonly involve either the whole chromosome arm or are focal (2), similar to what we observe in CLL. In the Beroukhim study, both PIK3CA and MYC were targets of recurrent gains in cancer.

To date, genomic alterations in PI3K have not been reported in CLL, although both amplifications and activating point mutations occur in solid tumors (21, 22). Here, we report amplifications of the PIK3CA locus in CLL, but we did not observe activating point mutations. A similar pattern of PIK3CA amplification without mutation has been described in mantle cell lymphomas (37), and amplifications in endometrial cancer show a distinct phenotype compared with somatic mutation, suggesting that amplification and somatic mutation may have distinct consequences (38). The PI3K pathway is constitutively (23, 24) and inducibly activated by multiple cell surface signals in CLL, including the B-cell receptor pathway (39, 40). Some interest has focused on the question of which PI3K catalytic isoform is most important in CLL, given that the δ isoform is highly expressed and the δ knockout mouse shows impairment in the B-cell compartment (39, 41, 42). Our data suggest that the δ isoform of p110 is most common in active complex with the p85 regulatory subunit in CLL, but that this balance can shift toward α, at least in several samples with α amplification. Our data suggest that PIK3CA amplification may be one of many mechanisms contributing to PI3K activation in CLL. Currently a δ-specific PI3K inhibitor is showing marked clinical activity in CLL (43); whether pan-PI3K or PI3K α inhibitors will have similar potency remains to be determined, as does the effect of PIK3CA amplification on the activity of the δ inhibitor. Ultimately, prospective validation of the frequency of PIK3CA amplification in CLL will be required to determine its importance in the disease.

A role for MYC in the initiation or progression of CLL has been much less clear, although transgenic mice expressing MYC together with BAFF have recently been reported to develop a CLL-like disease (44). This study also found that higher MYC expression in CLL patient samples was associated with shorter TTFT (44). Genomic analyses of Richter transformation, namely CLL that has transformed to a higher grade lymphoma, have identified MYC amplification as a common event thought to be acquired at the time of transformation (45). Here, we report MYC amplifications in CLL without transformation, through whole chromosome arm amplification or focal amplification of the 8q24 risk region near MYC. We also identify rare somatic mutations in MYC. These findings suggest multiple mechanisms of MYC activation in CLL, albeit at low frequency. The 8q24 gains identified here, while uncommon, are associated with short TTFT, and therefore likely a poor prognosis.

We identify 2 CLLs with focal gains of the 8q24 risk region, in which a SNP (rs2456449) has been associated with germline risk of CLL (25). Multiple studies have shown that this region can act as an enhancer for MYC (46–48). These focal amplifications may therefore represent somatic amplification of a germline risk allele, as described previously in neuroblastoma (49). If alleles identified by GWAS truly promote the risk of malignancy, additional similar instances will likely be identified over time.

Interestingly, gene expression analyses of our 3 CNA groups identified an overlapping set of CNAs with a shared expression pattern associated with induction of ES Polycomb gene sets, repression of ES Myc and proliferation gene sets, and induction of small, specific modules of ES "core" factors and targets, all of which have been previously associated with long-term self-renewing HSCs (32). These findings raise the possibility of a role for histone methylation in CLL pathogenesis and an association with HSC biology, but further work will be required to validate this finding and determine its significance to CLL.

In summary, our comprehensive integrated analysis of CLL has characterized 3 recurrent CNAs associated with reduced TTFT. Two of these CNAs affect PIK3CA and MYC foci. These CNAs will require validation in uniformly treated prospective cohorts to better define their incidence.
and prognostic significance. These studies together with emerging sequencing data will hopefully define key molecular subgroups of CLL that will clarify prognosis and inform novel therapeutic avenues in the coming years.

Disclosure of Potential Conflicts of Interest
J.R. Brown reports that he has served as a consultant for Calistoga Pharmaceuticals. No potential conflicts of interest were disclosed by the other authors.

Authors’ Contributions
Conception and design: J.R. Brown, A.S. Freedman
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J.R. Brown
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): J.R. Brown, M. Hanna, B. Tesar, N. Pochet, J.M. Asara, Y.E. Wang, Y.V. de Peer, M. Correll, A. Regev, D. Neuberg, L. Werner
Writing, review, and/or revision of the manuscript: J.R. Brown, M. Hanna, B. Tesar, N. Pochet, L. MacConaill, C.J. Wu, D. Neuberg, A.S. Freedman
Fernandes, C. Thompson and L. MacConaill.
Freedman.
The authors thank the Genetic Analysis Platform of the Broad Institute of Harvard and MIT for running the SNP arrays, the DFCI Microarray core facility for gene expression profiling, and the CLL Research Consortium tissue bank for IGHD and ZAP-70 results.

Grant Support
M. Hanna and L. MacConaill are supported through the CCGD and the Dana-Farber Strategic Plan Initiative. Y.E. Wang and M. Correll are supported through the CCBB and the Dana-Farber Strategic Plan Initiative. C.J. Wu acknowledges support from the Leukemia and Lymphoma Translational Research Program and is a Damon-Runyon Clinical Investigator supported in part by the Damon-Runyon Cancer Research Foundation (CI-38-07). N. Pochet is a postdoctoral research fellow of the Fund for Scientific Research-Flanders (FWO Vlaanderen) and a Broad Fellow of the Broad Institute. A. Regev is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, an NIH Pioneer award, HHML, and the Merkin Foundation for Stem Cell Research at the Broad Institute. J.M. Asara is supported by SPO1-CA120964 and SP30-CA066516 from the NIH. A.S. Freedman is supported in part by NIH S P01 CA092655. J.R. Brown was supported by K23 CA115682 from the NIH and is a Scholar of the American Society of Hematology as well as a Scholar in Clinical Research of the Leukemia and Lymphoma Society. These studies were supported by the Okinawa Lipton Fund, the Melon Fund, and the Rosenbush Fund. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received September 9, 2011; revised April 2, 2012; accepted April 13, 2012; published OnlineFirst May 23, 2012.

Acknowledgments
The authors thank the patients who participated in this study as well as the clinic and research staff who assisted with sample collection. The authors also thank the Genetic Analysis Platform of the Broad Institute of Harvard and MIT for running the SNP arrays, the DFCI Microarray core facility for gene expression profiling, and the CLL Research Consortium tissue bank for IGHD and ZAP-70 results.

References


41. Asara JM, Christofk HR, FreiMARK LM, Cantley LC. A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen. Proteomics 2008;8:994–8.
Integrative Genomic Analysis Implicates Gain of \textit{PIK3CA} at 3q26 and \textit{MYC} at 8q24 in Chronic Lymphocytic Leukemia

Jennifer R. Brown, Megan Hanna, Bethany Tesar, et al.