ONLINE METHODS

Sample collection and histopathological assessment. ESCC tissue samples were collected from Cancer Institute/Hospital, Chinese Academy of Medical Sciences (CAMS) and Linxian Cancer Hospital. All the samples used in this study were residual specimens collected after diagnosis. No patient received treatment before sample collection. Matched normal tissues (germline controls) were collected from adjacent esophageal epithelial five centimeters away from the border of surgical area. All tumor/normal samples were subject to hematoxylin staining and histopathological review to assess the presence of tumor cells, normal esophageal epithelial cells, lymphocytic infiltration and necrotic cells. Tumor cellularity was scored visually in a semiquantitative fashion and only those tumors with >70% malignant cells were chosen for DNA/RNA sequencing. All patients signed independent informed consent forms for the sampling and molecular analyses. This study has been approved by the Ethics Committee/Institutional Review Board of Cancer Institute/Hospital, CAMS and Linxian Cancer Hospital.

Cell culture and related chemicals. All of the human ESCC cell lines and 293T cells were grown in Dulbecco modified Eagle medium (DMEM) with 10% fetal calf serum (FCS) and were maintained at 37 °C in a 5% CO2 air-humidified incubator. ESCC cell lines KYSE30, KYSE50, KYSE70, KYSE110, KYSE150, KYSE450, KYSE510 and YES2 were kindly shared by Y. Shimada (Kyoto University). KYSE140 and KYSE180 were generously provided by Xin-Yuan Guan (Department of Clinical Oncology, University of Hong Kong). All ESCC cell lines were regularly authenticated and tested for absence of mycoplasma recently[55]. TPA was purchased from Sigma-Aldrich. KPT-330 was provided by Karyopharm Therapeutics.

Short-term cell proliferation assays. Cells were placed into 96-well plates at 2×10^3 to 4×10^3 cells/well and incubated for an additional 4–5 days in DMEM with 1–10% FCS. The MTT (3-(4, 5-dimethylthiazol–2-yl)-2,5-diphenyl tetrazolium bromide) assay was performed as described previously[56].

Assessment of apoptosis. Annexin V assay (BD Pharmingen) was performed to assess apoptosis according to the manufacturer's instructions. Briefly, cells were harvested after exposure to KPT-330, washed twice with PBS, incubated with FITC/PE-conjugated Annexin V and propidium iodide/7-AAD for 15 min, and measured by flow cytometry (FCM) using FACScan (Becton Dickinson).

Xenografts in NOD/SCID mice. A total of six male NOD/SCID mice at 5–6 weeks of age were provided by Cancer Science Institute of Singapore. We mixed 2×10^6 of KYSE150 cells (Scramble and shFAT2 stable cells) with 100 µl of Matrigel solution (BD Biosciences) per injection, and the mixture was injected subcutaneously on the upper flanks of NOD/SCID mice. Mice were randomly allocated to either Scramble or shFAT2 groups. After 19 d, the mice were killed to weigh and analyze the dissected tumors. No blinding of investigators was performed. No statistical methods were used to determine the sample size of mice. Animal study was done in compliance with ethical regulations of Institutional Animal Care and Use Committee of National University of Singapore.

Reverse transcription and real-time PCR. Total RNA was reverse transcribed to cDNA with Superscript III (Invitrogen) according to manufacturer's protocols. Real-time PCR reactions were performed in triplicate for every sample using the 7500 Real-time PCR System (Applied Biosystems). Primers are listed in **Supplementary Table 13a**.

Lentiviral-based expression and shRNA vectors. The open reading frame (ORF) of human *FAT1* and *ZNF750* transcripts were generously provided by T. Chan's group from Memorial Sloan-Kettering Cancer Center[27] and P. Khavari's group from Veterans Affairs Palo Alto Healthcare System[24], respectively. Both ORFs were subcloned into lentiviral-based expression vector SHC003 (Sigma-Aldrich) using NheI and FseI cloning sites. SHC003-Turbo-GFP was used as control (Sigma-Aldrich). The lentiviral-based shRNA vectors (shXPO1 and shFAT2 sequences are listed in **Supplementary Table 13b**) were generated with PLKO.1 backbone (Sigma-Aldrich) using AgeI and EcoRI cloning sites. SHC002-Scramble shRNA was used as control (Sigma-Aldrich). The cloning primers are listed in **Supplementary Table 13c**.

Transfections, viral particle production and infection. ESCC cells and 293T cells were transfected with Lipofectamine 2000 according to the manufacturer's instructions (Invitrogen). siRNA sequences targeting *FAT1* or *ZNF750* are listed in **Supplementary Table 13d**. Lentiviral particles were produced and harvested using MISSION Lentiviral Packaging System (Sigma-Aldrich). The ESCC cells were infected with the lentiviral particles for 48 h in the presence of 8 µg/ml polybrene (Sigma-Aldrich).

Whole-exome sequencing (WES) and targeted deep sequencing (TDS). We sonicated 1.5 µg (WES) or 1 µg (TDS) of genomic DNA to generate a peak target size of 200 bp. DNA was captured using the SureSelect Human All Exon 50M (WES) or customized cRNA beads (**Supplementary Table 14**, TDS) according to the manufacturers' protocols. Captured DNAs were subjected to massively parallel sequencing using HiSeq2000 with 75-to 100-bp paired-end reads.

To detect somatic nucleotide variations from WES, we used previously described in-house algorithms[57–59]. Briefly, the sequencing reads were aligned to a human reference genome (hg19) using BWA version 0.5.8 with default parameters. PCR duplicate reads were removed using Picard (http://www.picard.sourceforge.net/). Before summarizing base-call data, we eliminated low-quality reads, including those reads that had either more than five mismatches to the reference sequences or whose mapping quality was less than 30. The significance of each candidate mutation was evaluated by Fisher's exact test by enumerating the number of the reference base and the candidate SNV in both tumor and germline control. Candidate mutations with *P* values less than 0.01 were adopted as provisional candidates for somatic mutations. In addition, the following nucleotide positions were eliminated from further analysis: those at which the depth was less than 10 in either tumor or control, or where the most frequent SNV or indel accounted for less than 7% of all reads in the tumor. Germline SNPs were eliminated using sequencing data from paired normal DNA. Finally, a list of candidate somatic mutations was generated by excluding synonymous SNVs and other variants registered in dbSNP131, the 1000 Genomes Project or our in-house SNP database constructed from 180 patients[57,58].

To make somatic mutation calling with the TDS approach of the frequency cohort more rigorous, the following nucleotide positions were further removed: those at which the supporting depth from both directions was less than 5 in either tumor or control, or where the most frequent SNV or indel accounted for less than 8% of all reads in the tumor.

To detect probable somatic mutations of ESCC cell lines that do not have paired germline controls, nucleotide positions were further removed if the supporting depth from both directions was less than 5, the most frequent SNV or indel accounted for less than 8% of all reads, or the frequency of the SNV or indel was between 45% and 55% without copy number abnormalities.

Tissue microarray (TMAs) and immunohistochemistry (IHC). Paraffin-embedded tissue microarrays (TMAs) containing 50 primary ESCC tumors and the corresponding normal epithelia were used for IHC. For each case, histologically normal tissues adjacent to tumors were examined as control. TMA slides were initially deparaffinized using xylene, rehydrated with xylene and ethanol, immersed in 3% hydrogen peroxide solution for 10 min, heated with citrate at 95 °C for 25 min and cooled at room temperature for 60 min. The slides were incubated overnight at 4 °C with the following antibodies: XPO1 (sc-5595, Santa Cruz Biotechnology; 1:50), ZNF750 (HPA023012, Sigma-Aldrich; 1:100), FAT1 (HPA023882, Sigma-Aldrich; 1:25), or FBXW7 (H00055294-M02, Abnova; 1:200), FGFR1 (9740, Cell Signaling Technology, 1:25) and visualized using PV-9000 Polymer Detection System following the manufacturer's instructions (Golden Bridge International). Counterstaining was carried out with hematoxylin. The results were separately evaluated by two pathologists. Protein expression was evaluated on the basis of staining intensity, graded on the following scale: 0 (negative), 1 (weak), 2 (moderate) and 3 (strong).

Protein blotting. Cells were lysed on ice with lysis buffer (50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 0.5% Nonidet P-40) containing complete protease and

phosphatase inhibitor cocktail (Roche). The rest of the procedures for protein blot (western blot) were performed as described previously[60]. Antibodies specific for cyclin B1 (4135), cleaved-PARP (9541), $p21^{WAF1}$(2947) and BIM (2933) were purchased from Cell Signaling Technology. Antibodies specific for c-Myc (SC-788), cyclin D1 (sc-8396), Bcl-xl (sc-8392), P53 (sc-6243), XPO1 (sc-5595) and PUMA (sc-28226) were purchased from Santa Cruz Biotechnology. Antibody against FAT1 (HPA023882), ZNF750 (HPA023012) and β-Actin (A5316) were purchased from Sigma-Aldrich.

RNA extraction and paired-end sequencing. We extracted 6 μg of total RNA from each sample according to the manufacturer's instruction of RNeasy Micro kit (QIAGEN), followed by mRNA purification using Oligotex mRNA Mini Kit (QIAGEN). RNA quality was assessed by Agilent 2100 Bioanalyzer, and the RNA integrity number was >7.0, with 28S/18S >0.7. The cDNA libraries were constructed following the TruSeq RNA Sample Preparation Guide (Illumina). Briefly, first-strand and second-strand cDNAs were synthesized from the purified mRNAs, and fragmented by ultrasonic waves. After performing end repair, cDNAs were modified by 3′-end adenylation and adaptor ligation. PCR was performed to enrich the cDNA templates in order to complete the cDNA library construction. The concentration of the final cDNA library was more than 1 ng/μl, and the size of the cDNA fragments was 350–400 bp. Sequencing reagents were prepared according to the Genome Analyzer IIx User Guide (Illumina). cDNA fragments were treated with cluster generation and loaded to the lanes of Illumina flow cells, and paired-end sequencing was performed using the 2 × 100 nt multiplex program. The raw sequenced reads yielded more than 50 million bases per sample.

Array-based CGH and SNP-array. Array-CGH experiments were performed using standard protocols as previously described (44K human genome CGH microarrays, Agilent Technologies)[9]. Human genomic DNA (PROMEGA) was used as reference. Briefly, 500 ng of reference genomic DNA and tumor DNA were digested with AluI and RsaI (PROMEGA). The resulting reference DNA was labeled with cyanine-5 dUTP and the tumor DNA with cyanine-3 dUTP, respectively (Agilent Technologies). After clean-up, labeled DNA probes were mixed and hybridized to CGH microarray for 24–40 h. Washing, scanning and data extraction were performed according to standard protocols. Array-CGH data were analyzed using Genomic Workbench (Agilent), BRB-CGH tools and MD-SeeGH (see URLs). Mean $\log_2$ tumor/reference ratios of all probes in a chromosome region greater than 1.0 were determined to be high-amplitude amplification, and ratios less than −1.0 were determined to be high-amplitude deletion. These values were converted based on the following presumptions: (i) for amplifications, copy number >6 was present in more than 70% cancer cell populations; (ii) for deletions, loss of both alleles was present in more than 70% cancer cell populations and (iii) normal cell contamination was less than 30% (ref. 60).

Human 250K arrays (Affymetrix) were used for SNP-array assays, and results were analyzed as previously described using CNAG/AsCNAR software[61,62].

Interphase cell nuclei preparations and fluorescence *in situ* hybridization (FISH). In order to examine interphase cell nuclei, we cut ESCC tissue samples into small pieces in phosphate-buffered saline (PBS), followed by treatment with a hypotonic solution (0.075 mol/L KCl) for 30 min and three successive changes of the fixative solution (methanol/acetic acid, 3:1). Interphase cell nuclei in suspension were kept at 4 °C overnight and then stored at −20 °C. Prior to the hybridization, the nuclear suspensions were dropped onto slides and dried at room temperature for 2–3 d.

FISH was performed on the interphase cell nuclei. By random priming using BioPrime DNA labeling system (Invitrogen), chromosome enumeration probes (CEP) and bacterial artificial chromosome (BAC)-DNA probes were directly labeled with Green-dUTP (Abbott Molecular) and Cy3-dUTP (GE Healthcare), respectively. BAC-DNA clones selected for *FGFR1* were NONSC11F1 (Chr8: 38,170,901–38,368,835).

The slides for FISH analyses were pretreated with RNase A (100 mg/ml in 2 × saline sodium citrate [SSC]) and pepsin (50 mg/ml in 0.01 mol/l HCl), subsequently denatured in 70% formamide/2 × SSC at 73–75 °C for 3 min, quickly cooled with two rinses of 2 × SSC at 4 °C, dehydrated in a gradient series of ethanol (75%, 85% and 100%) and air dried. The labeled probes were precipitated, re-dissolved in the hybridization solution (50% formamide, 10% dextran sulfate, 1% Tween-20, 2 × SSC), denatured at 75 °C for 8 min and quick-chilled on ice for 2 min. Hybridization was performed in a humid chamber at 37 °C for 24–48 h. Posthybridization washes were performed in 50% formamide/2 × SSC for 15 min at 43 °C and were performed twice for 3 min each in 2 × SSC. The slides were dehydrated in 75%, 85% and 100% ethanol, air dried, counterstained with 4,6-diamidino-2-phenylindole (DAPI) (1 mg/ml) and covered with coverslips.

Microscopy and digital image analysis. FISH images were captured using a Zeiss Axio fluorescence microscope equipped with a cooled charged-coupled device (CCD) camera (Princeton Instruments) or a JAI M4 Plus CCD camera (Metasystems International). All of the fluorescent images were captured with individual single-band-pass filters specific for visualizing DAPI, Green, Cy3 fluorochromes. Pseudo-color images were constructed and analyzed using MetaMorph (Universal Imaging Corporation) or Metacyte module of Metafer imaging systems (Metasystems International).

Amplifications were assessed as the difference value between BAC-DNA and CEP probes was at least 2 in more than 30% cells. Cluster signals and ratios (BAC-DNA/CEP) >2 were judged as the high-level amplifications.

Analysis of significantly altered pathways. To analyze significantly altered pathways, we performed unbiased Gene Ontology (GO) enrichment in the overall ranked MutSigCV list[63]. As a parallel approach, WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) was applied to identify the KEGG pathways that are enriched in the MutSigCV list. Multiple-test corrected *P* values were calculated using a hypergeometric test for pathways that are enriched compared to human genome reference gene set[64,65].

Statistical analysis. *Q* values for significantly mutated genes were calculated using default setting according to MutSigCV[13]. The *P* value of the two mutational events as mutually exclusive events is defined by calculating the joint probability. The statistical analyses of the following assays were conducted using the two-tailed Student's *t*-test upon verification of the assumptions (for example, normality); otherwise the nonparametric test was applied: short-term cell proliferation, xenografts proliferation, q-PCR, apoptosis induction, mRNA levels examined from GSE20347 (ref. 11), GSE23400 (ref. 54) and CCLE.

55. Zhu, Y.H. *et al.* Downregulation of the novel tumor suppressor DIRAS1 predicts poor prognosis in esophageal squamous cell carcinoma. *Cancer Res.* **73**, 2298–2309 (2013).
56. Lee, D.H. *et al.* Synergistic effect of low-dose cucurbitacin B and low-dose methotrexate for treatment of human osteosarcoma. *Cancer Lett.* **306**, 161–170 (2011).
57. Sato, Y. *et al.* Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.* **45**, 860–867 (2013).
58. Sakaguchi, H. *et al.* Exome sequencing identifies secondary mutations of SETBP1 and JAK3 in juvenile myelomonocytic leukemia. *Nat. Genet.* **45**, 937–941 (2013).
59. Yoshida, K. *et al.* Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64–69 (2011).
60. Lin, D.C. *et al.* Genomic and functional characterizations of phosphodiesterase subtype 4D in human cancers. *Proc. Natl. Acad. Sci. USA* **110**, 6109–6114 (2013).
61. Nannya, Y. *et al.* A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res.* **65**, 6071–6079 (2005).
62. Yamamoto, G. *et al.* Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of affymetrix single-nucleotide-polymorphism genotyping microarrays. *Am. J. Hum. Genet.* **81**, 114–126 (2007).
63. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
64. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* **33**, W741–W748 (2005).
65. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.* **41**, W77–W83 (2013).

Published OnlineFirst June 11, 2014; DOI: 10.1158/2159-8290.CD-14-0104

RESEARCH ARTICLE

Acquired Initiating Mutations in Early Hematopoietic Cells of CLL Patients

Frederik Damm[1,3], Elena Mylonas[1,3], Adrien Cosson[4,5], Kenichi Yoshida[12,15], Véronique Della Valle[1,3,9], Enguerran Mouly[1,3,9], M'boyba Diop[1,3], Laurianne Scourzic[1,3,9], Yuichi Shiraishi[13,14], Kenichi Chiba[13,14], Hiroko Tanaka[13,14], Satoru Miyano[13,14], Yoshikane Kikushige[16,17], Frederick Davi[4,5,6], Jérôme Lambert[7], Daniel Gautheret[3,10], Hélène Merle-Béral[4,5,6], Laurent Sutton[11], Philippe Dessen[1,3], Eric Solary[2,3,8,9], Koichi Akashi[16,17], William Vainchenker[2,3,9], Thomas Mercher[1,3,9], Nathalie Droin[2,3,9], Seishi Ogawa[12,15], Florence Nguyen-Khac[4,5,6], and Olivier A. Bernard[1,3,8,9]

Published OnlineFirst June 11, 2014; DOI: 10.1158/2159-8290.CD-14-0104

ABSTRACT Appropriate cancer care requires a thorough understanding of the natural history of the disease, including the cell of origin, the pattern of clonal evolution, and the functional consequences of the mutations. Using deep sequencing of flow-sorted cell populations from patients with chronic lymphocytic leukemia (CLL), we established the presence of acquired mutations in multipotent hematopoietic progenitors. Mutations affected known lymphoid oncogenes, including *BRAF*, *NOTCH1*, and *SF3B1*. *NFKBIE* and *EGR2* mutations were observed at unexpectedly high frequencies, 10.7% and 8.3% of 168 advanced-stage patients, respectively. *EGR2* mutations were associated with a shorter time to treatment and poor overall survival. Analyses of *BRAF* and *EGR2* mutations suggest that they result in deregulation of B-cell receptor (BCR) intracellular signaling. Our data propose disruption of hematopoietic and early B-cell differentiation through the deregulation of pre-BCR signaling as a phenotypic outcome of CLL mutations and show that CLL develops from a pre-leukemic phase.

SIGNIFICANCE: The origin and pathogenic mechanisms of CLL are not fully understood. The current work indicates that CLL develops from pre-leukemic multipotent hematopoietic progenitors carrying somatic mutations. It advocates for abnormalities in early B-cell differentiation as a phenotypic convergence of the diverse acquired mutations observed in CLL. *Cancer Discov; 4(9); 1088–1101. ©2014 AACR.*

See related commentary by Jiang and Elemento, p. 995.

INTRODUCTION

Cancer develops from an individual cell that accumulates acquired mutations. Appropriate medical care requires thorough understanding of the natural history of the disease, including the identification and order of occurrence of the mutations, the cell of origin, and the clonal organization of the tumor cells. In addition, because the transformation process can capture preexisting somatic mutations (1, 2), their driver nature needs to be fully established, based on their recurrence and their functional consequences. Such in-depth investigations identify initial driver mutations, which are relevant as targets for therapy.

Chronic lymphocytic leukemia (CLL), the most frequent adult leukemia in Western countries, is characterized by an accumulation of mature B lymphocytes (3). The CLL tumor cells are clonal, as assessed by rearrangement of the immunoglobulin heavy chain (*IGH*) gene, and express low levels of surface B-cell receptor (BCR). In a fraction of patients, the *IGH* variable gene segment (*IGHV*) rearrangement is mutated, reflecting normal somatic hypermutation triggered by antigen recognition. Patients with *IGHV* mutations have a better prognosis than those without *IGHV* mutations.

Investigation of CLL samples by massively parallel sequencing has identified a number of acquired somatic mutations (4, 5), but no individual gene is mutated in more than 20% of the patients. The products of these mutated genes are involved in RNA metabolism, genome stability and cell cycle, control of the Notch pathway, Wnt signaling, and inflammation (4). Transformation may also depend on specific *IGH* rearrangements and BCR intracellular signaling cascades (6, 7). The cell of origin of CLL is currently debated. Immunophenotype and expression profiling analyses pointed at mature $CD5^+$ B cells (8), but the involvement of early hematopoietic cells in CLL development has been emphasized in xenograft experiments, which showed that the hematopoietic stem/progenitor cells from patients with CLL show biased and abnormal differentiation toward the B-lymphoid lineage in immunodeficient mice (9).

To investigate the natural history of CLL, we embarked on a thorough analysis of CLL samples using massive parallel sequencing and cellular analyses.

Authors' Affiliations: [1]Institut National de la Santé et de la Recherche Medicale (INSERM) U985; [2]INSERM U1009; [3]Institut Gustave Roussy, Villejuif; [4]INSERM U1138; [5]Université Pierre et Marie Curie-Paris 6; [6]Service d'Hématologie Biologique, Hôpital Pitié-Salpêtrière, APHP; [7]Université Paris Diderot, Hôpital Saint-Louis, SBIM; [8]Ligue Nationale Contre le Cancer, Equipe labellisée, Paris; [9]Université Paris-Sud; [10]Université Paris-Sud, Institut de Génétique et Microbiologie, CNRS UMR 8621, Orsay; [11]Service d'Hématologie Clinique, Centre Hospitalier Victor Dupouy, Argenteuil, France; [12]Cancer Genomics Project, Graduate School of Medicine; [13]Laboratory of DNA Information Analysis and [14]Laboratory of Sequence Data Analysis, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo; [15]Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University, Kyoto; [16]Medicine and Biosystemic Science; and [17]Center for Cellular and Molecular Medicine, Kyushu University Graduate School of Medical Sciences, Fukuoka, Japan

Note: Supplementary data for this article are available at Cancer Discovery Online (http://cancerdiscovery.aacrjournals.org/).

F. Damm and E. Mylonas contributed equally to this article.

S. Ogawa, F. Nguyen-Khac, and O.A. Bernard share senior authorship of this article.

Corresponding Authors: Olivier A. Bernard, INSERM U985, 39 rue Camille Desmoulins, Villejuif 94805, France. E-mail: olivier.bernard@inserm.fr; Florence Nguyen-Khac, Unité de Cytogénétique Hématologique, Service d'Hématologie Biologique, GH Pitié-Salpêtrière/Charles Foix, 83 Bd de l'Hôpital, Paris 75013, France. E-mail: florence.nguyen-khac@psl.aphp.fr; and Seishi Ogawa, Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan. E-mail: sogawa-tky@umin.ac.jp

doi: 10.1158/2159-8290.CD-14-0104

Published OnlineFirst June 11, 2014; DOI: 10.1158/2159-8290.CD-14-0104

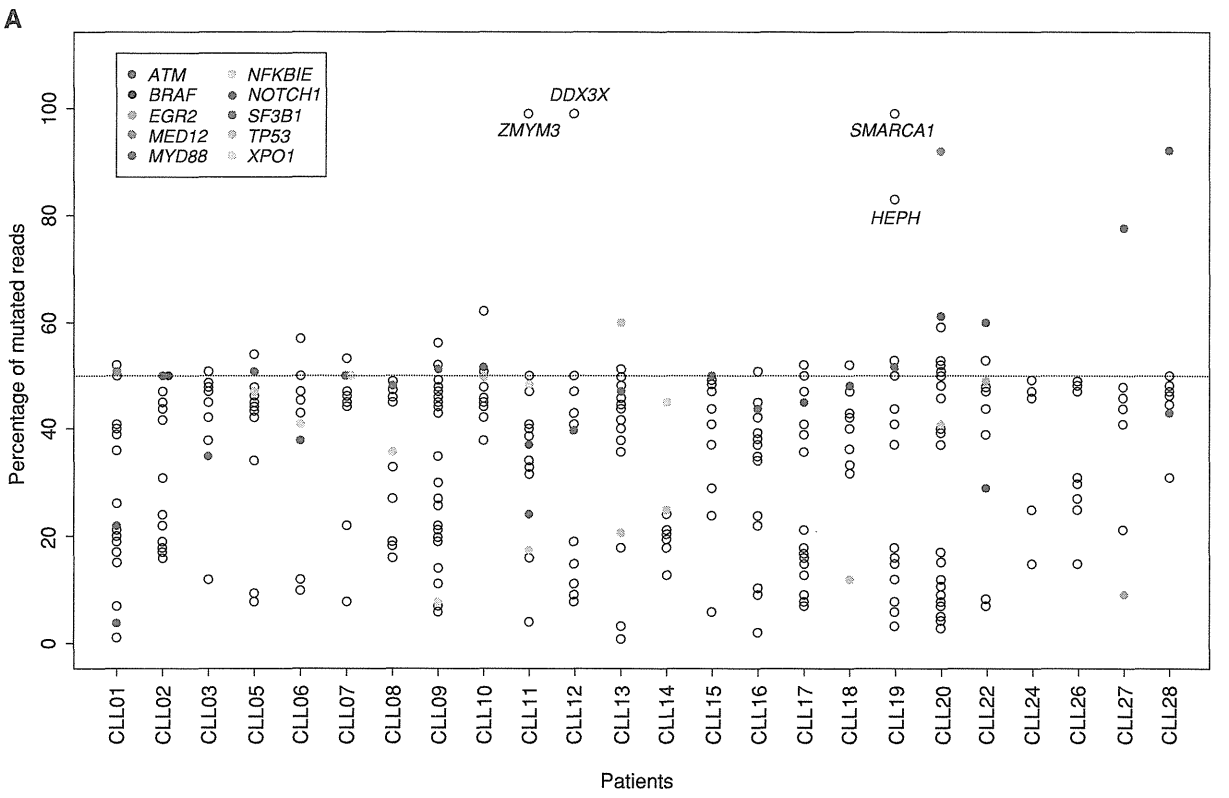

Figure 1. Somatic mutations identified in 24 patients with CLL. **A,** variant allele frequencies in the tumor fraction. Genes analyzed in the extension cohort are colored. *ZMYM3*, *DDX3X*, *SMARCA1*, *MED12*, and *HEPH* genes are on chromosome X. See Supplementary Table S1 for details. The patient numbers are in abscissa. (*continued on following page*)

RESULTS

SF3B1 Mutations Are Detected in Nonlymphocytic Cellular Fractions of CLL Patients

To search for CLL mutations in the hematopoietic progenitor cell fraction, we first investigated the distribution of *SF3B1* mutations in the hematopoietic tree of patients with *SF3B1*-mutated CLL, because this gene is frequently mutated in both CLL and myelodysplastic syndrome (MDS), a chronic stem cell–derived myeloid tumor (10, 11). Sanger sequencing of the mutational hotspots in the *SF3B1* gene using DNA from 50 patients with CLL identified 7 patients carrying an *SF3B1* mutation. We next flow-sorted cells according to the expression of mutually exclusive cell-surface markers: CD34 (which marks the immature progenitor cells compartment at the apex of hematopoietic differentiation), and markers of mature cells, including CD3 (T cells), CD14 (monocytes), and CD19 (both normal and tumor B cells). Sequencing analyses of DNA from these cellular fractions showed wild-type *SF3B1* sequences in the $CD3^+$ cells and the mutated *SF3B1* sequence in the CD19 fraction in all seven cases. The mutation was also observed in the $CD34^+$ and/or $CD14^+$ fractions in two patients (Supplementary Fig. S1), suggesting it was acquired in these patients in an early progenitor cell that was able to participate in both lymphoid and myeloid differentiation.

Acquired Mutations Are Detected in Multipotent Progenitors in the Majority of CLL Patients

We next used whole-exome sequencing of DNA from flow-sorted cell populations derived from 24 patients with CLL (17 *IGHV*-unmutated and 7 *IGHV*-mutated; Supplementary Table S1). Results of the *IGH* gene rearrangements were always compatible with monoclonal proliferation. Viable cells were flow-sorted to purities greater than 96% (see flow chart description; Supplementary Fig. S2A and S2B). Comparison of exome sequences from tumor cells and T lymphocytes (essentially spared by CLL mutations as shown for *SF3B1* mutation) identified a total of 415 somatic mutations predicted to result in protein-coding changes in 361 different genes with a median of 17 mutations/patient (range, 7–34; Supplementary Table S1; Fig. 1A). Some mutations were present in virtually all $CD19^+$ cells, whereas the allelic ratio of other mutations showed that they were present only

Published OnlineFirst June 11, 2014; DOI: 10.1158/2159-8290.CD-14-0104

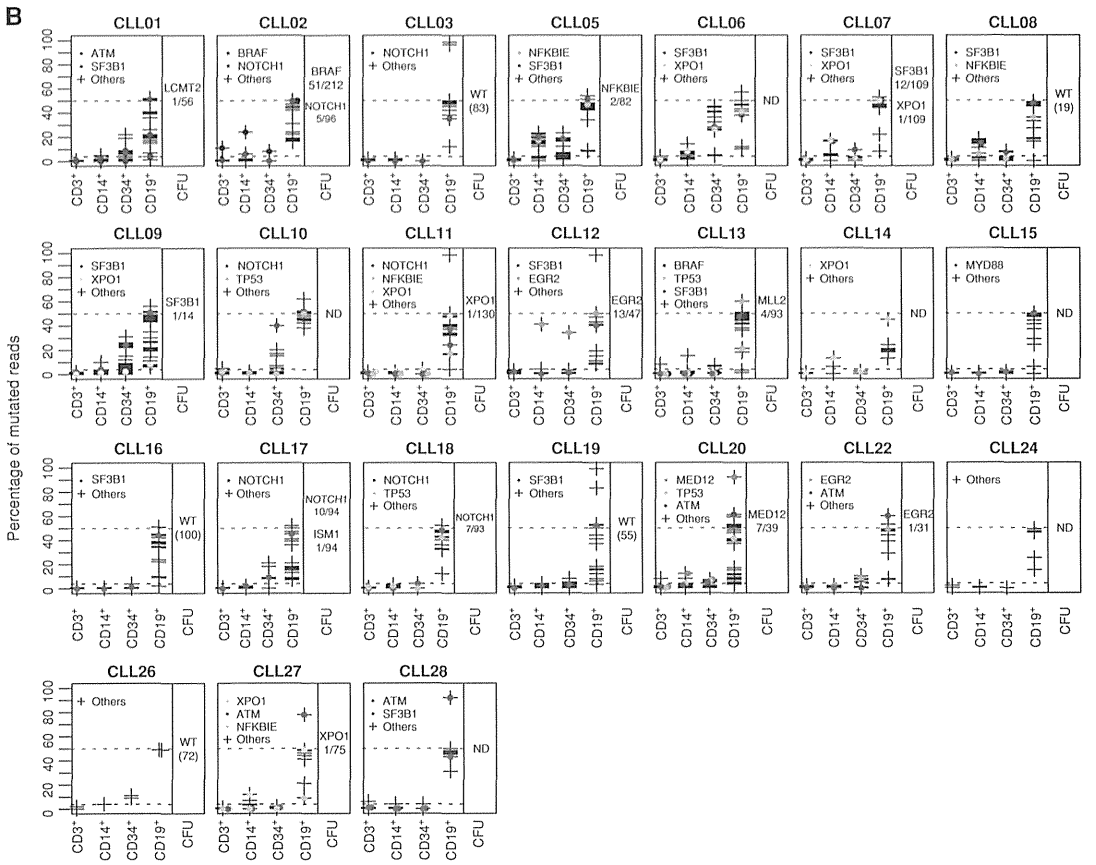

Figure 1. ***(Continued)*** **B,** variant allele frequencies in the hematopoietic fractions, sorted on the basis of the cell-surface expression of the following antigens (in abscissa): CD34 (progenitors), CD3 (T cells), CD19 and CD5 (tumor cells), and CD14 (monocytes). Column on the right, the genotyping of colony forming units (CFU) from single CD34$^+$ cells. If only wild-type (WT) colonies were observed, the number of analyzed colonies is indicated in parentheses. The ratio of 50% and 4% are indicated by dotted lines, orange and black, respectively. ND, not done.

in a fraction of these cells, indicating that they were secondarily acquired.

We used targeted deep resequencing to simultaneously validate and quantify the mutation burden in DNA from the sorted fractions (Fig. 1B and Supplementary Table S1). Sorting impurity and aberrant antigen expression should be taken into account when analyzing cell-sorted fractions. Mutation burdens below 4% that potentially result from sorting contamination were regarded as negative. Among the 24 patients analyzed, only 3 (CLL03, 15, and 24) were devoid of mutations in the CD34$^+$ progenitor or the CD14$^+$ monocyte fractions. All 3 patients carry mutated *IGHV* rearrangements in their CLL cells. In the other 21 patients, at least one mutation was detectable in the CD14$^+$ or in the CD34$^+$ fractions. Two patients (CLL14 and 27) showed mutations in the CD14$^+$ and not in the CD34$^+$ fraction and, conversely, 6 patients (CLL10, 11, 16, 17, 18, and 22) showed mutation in the CD34$^+$ but not in the CD14$^+$ fraction. In 13 samples (CLL01, 02, 05, 06, 07, 08, 09, 12, 13, 19, 20, 26, and 28), at least one mutation was detected in both fractions. The presence of a CLL mutation in immunophenotypical progenitor (CD34$^+$) or myeloid (CD14$^+$) primary cells confirmed the involvement of immature cells in CLL pathogenesis. The burden of mutations detected in the immature hematopoietic cells (referred to hereafter as early mutations) was always among the highest mutation burdens in CLL cells, consistent with their occurrence at the initial steps of CLL development. However, in all patients only a subset of the CLL mutations was observed in the progenitors or myeloid fractions.

Because cell-surface expression of myeloid antigens is not sufficient to attest to the myeloid nature of a progenitor cell, we next tested the myeloid differentiation capacities of the mutated progenitors. We sorted single CD34$^+$CD19$^-$ progenitor cells and grew them *in vitro* in myeloid conditions. Viable cells were available for 18 patients. The cloning efficiency was close to 60% for each patient (exceptions were CLL08 and CLL09) and colonies were confirmed as myeloid (erythroid, megakaryocytic, or/and granulo-monocytes) by FACS immunophenotyping of randomly chosen colonies. Colony genotyping confirmed the presence of CLL mutations in myeloid cells in 13 patients, whereas 5 patients (CLL03, 08, 16, 19, and 26) did not show mutated colonies

Published OnlineFirst June 11, 2014; DOI: 10.1158/2159-8290.CD-14-0104

(Fig. 1B and Supplementary Table S1). Although the absence of mutation may be due to the low number of colonies in some patients (as for patient CLL08), the other 4 patients clearly did not show mutated cells in over 50 colonies analyzed. In addition, the frequency of mutated colonies differed from the estimated mutation burden in the sorted $CD34^+$ fractions, supporting the idea that not all mutated progenitors could grow in these myeloid culture conditions. We also investigated the myeloid colonies from 17 patients for the presence of the CLL *IGH* rearrangement using rearrangement-specific PCR. No *IGH*-rearranged colonies were detected for 11 patients (CLL01, 02, 03, 08, 09, 11, 12, 13, 16, 20, and 26). A low number of IGH-rearranged colonies was observed for 6 patients: CLL05 (8/96), 07 (2/109), 17 (1/96), 27 (2/96), 18 (3/96), and 19 (1/59). Nucleotide sequence analyses showed that the variable joining gene segments (VJ) junction amplified from the colonies matched the tumor cell rearrangement in patients 05 and 07. Colonies from patients 17 and 19 showed a rearrangement differing from those of their CLL counterpart. Half of the colonies from patients 18 and 27 showed the same rearrangement as the corresponding tumor cells, whereas the other half carried other VJ junctions. Of note, every VJ-positive colony also carried an early mutation.

Together, these data demonstrate the presence of CLL mutations in a multipotent hematopoietic progenitor fraction in the majority of patients with CLL. Reasoning that the mutations had been originally acquired in a single cell, the high proportion of mutated cells in the $CD34^+$ or $CD14^+$ fractions demonstrates that the cell carrying the identified mutation had some clonal advantage and accumulated over time. The mutations seem to variably affect hematopoietic differentiation, as judged from the mutation burden detected in the hematopoietic fractions (see Supplementary Fig. S3A).

Some patients showed an overall normal balance between myeloid and B-lymphoid differentiation. They showed multilineage involvement indicative of an unbiased differentiation of the mutated stem/progenitor cells (for example, CLL02, 07, 12, and 20 in Fig. 1B and Supplementary Table S1). In our settings, a mutation would be detected only if it induces the accumulation of the mutated cell in the given fraction. If a mutation induces accumulation at a late step and not at early steps of differentiation, accumulation will occur in the mature cells ($CD14^+$) and not the immature cells ($CD34^+$). For example, patients CLL14 and CLL27 would belong to this first group of patients.

A second type of patient (for example, CLL10, 11, and 22) shows an unbalanced involvement of myeloid cells (a lower mutational burden than in $CD34^+$ progenitors), suggesting that the early mutations bias the mutated stem/progenitor cells toward the lymphoid lineage or specifically allow the accumulation of lymphoid-primed progenitors.

A third type of patient (CLL03, 15, and 24 in this series) lacks detectable mutation in either the myeloid or the progenitor compartments, suggesting either a strict commitment toward lymphoid differentiation or the involvement of a lymphoid-primed progenitor. Alternatively, these patients may follow a different transformation pathway. The numbers and burden of mutations did not differ statistically between these 3 patients and the others (Supplementary Table S1).

Early Mutations Affect Genes Recurrently Mutated in CLL and Other Malignancies

Mutations detected in the progenitors of patients with CLL affected genes already known to be mutated in CLL, in other hematologic malignancies, or even in other cancers, supporting their active role in transformation (Supplementary Table S2). Early mutations were observed in the *NOTCH1*, *SF3B1*, *TP53*, and *XPO1* genes are among the most frequently mutated genes in CLL (4, 5, 12). Genes such as *BRAF* and *MLL2* are mutated in CLL and in other B-cell malignancies (13–15). A few *EGR2*- and *NFKBIE*-mutated patients have been reported in CLL (4, 5, 16). To further establish the importance of the early mutations identified in our patients, we investigated the recurrence of some of them by direct Sanger sequencing of the mutational hotspots of *BRAF*, *EGR2*, *MED12*, *MYD88*, *NFKBIE*, *NOTCH1*, *SF3B1*, *TP53*, and *XPO1* in the 168 untreated patients with stage B and C CLL who were sampled at inclusion in a clinical trial (www.clinicaltrials.gov, NCT00931645; Supplementary Table S3; ref. 17). A total of 113 mutations in 84 patients were identified, and 84 of 168 (50.0%) patients presented with at least one mutation of this nine-gene panel (Fig. 2A and B and Supplementary Table S4). Inactivating mutations of *NFKBIE* were found in 10.7% (18 of 168; Fig. 2C) of the patients. Missense mutations of *EGR2* were observed in 8.3% (14 of 168; Fig. 2C) of the patients and associated with higher $CD38^+$ expression (median, 70% vs. 17%; $P = 0.009$), a known poor prognosis marker, a shorter time to treatment (median, 15.4 vs. 1.2 months; $P = 0.0006$), and a shorter 5-year overall survival (56.2 vs. 80.4 months; $P = 0.04$; Fig. 2D).

Deregulation of BCR Signaling as a Phenotypic Convergence of Early Mutations in CLL

Normal BCR and pre-BCR signaling occurs through BRAF, which activates ERK proteins (18), which in turn phosphorylate and activate the ternary complex factor–serum response factor (SRF) dimer, resulting in the upregulation of a set of immediate early genes, including *EGR2* (19, 20). *BRAF* and *EGR2* mutations may therefore affect the BCR signaling, which is abnormal in CLL (7). *BRAF* mutations, most frequently V600E, have been described in a variety of human malignancies, including hairy cell leukemia (13, 21), another malignant B-cell disease. In CLL, *BRAF* mutations target amino acids located in the P-loop of the kinase (Supplementary Table S4; ref. 22), leading to weaker activation than the canonical V600E mutations (18). Ectopic expression of the CLL-mutant BRAF-G469R in Ba/F3 cells showed a constitutive ERK phosphorylation and *Egr2* transcription (Fig. 3A and B). To analyze the impact of BRAF-G469R in B-cell differentiation, we transduced hematopoietic progenitors with BRAF–wild-type (WT), BRAF-G469R, or empty murine stem cell virus (MSCV) vector and engrafted the cells in irradiated syngeneic recipients. Animals were analyzed after 5 weeks, before the onset of any gross hematologic disorders. Careful analyses of the immunoglobulin M (IgM)–positive B-cell compartment showed a decrease in the proportion of B cells in the BRAF-G469R mice, as compared with MSCV or BRAF-WT mice (Fig. 3C and D). In addition, the mean fluorescence of IgM was significantly lower in B220-IgM–positive BRAF-G469R–expressing cells than in their WT or MSCV counterparts (Fig. 3E). A similar abnormal (IgM low, IgD^-) B-cell population was present in the spleen of the BRAF-G469R mice (Fig. 3C and E).

Published OnlineFirst June 11, 2014; DOI: 10.1158/2159-8290.CD-14-0104

A

Gene	Mutated patients	Overall frequency (%)
SF3B1	27	16.1
NFKBIE	18	10.7
NOTCH1	15	8.9
TP53	14	8.3
EGR2	14	8.3
XPO1	14	8.3
BRAF	6	3.6
MED12	3	1.8
MYD88	2	1.2
Patients with at least one mutation	84/168	50%

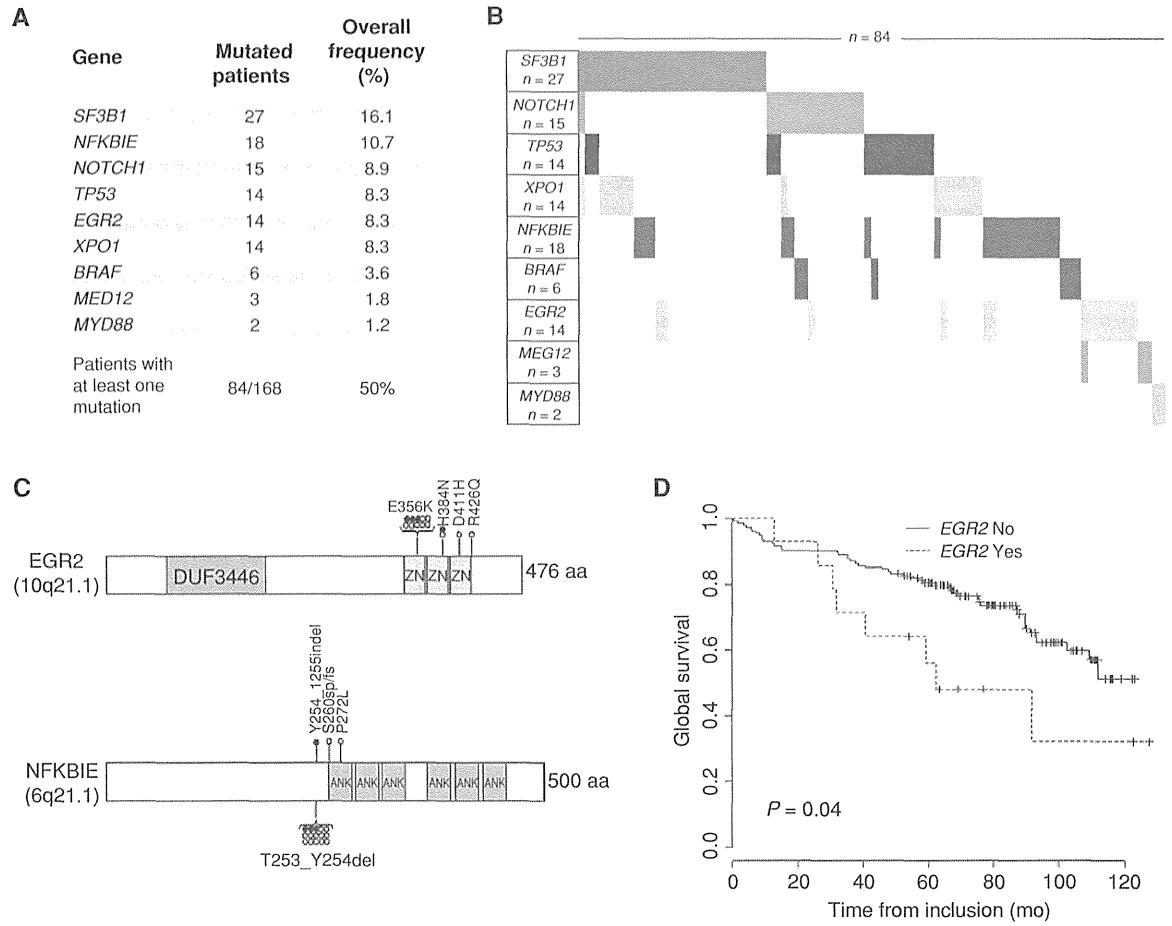

Figure 2. Gene mutation profile in 168 patients with CLL. Frequencies (**A**) and distributions (**B**) of identified mutations. **C,** localization of identified mutations in EGR2 and NFKBIE proteins. Circles, mutations; filled circle, proven somatic mutation. **D,** overall survival of patients with CLL according to *EGR2* mutation status (log-rank test).

We next investigated the consequences of *EGR2* mutations. The *EGR2* gene encodes a versatile transcription factor that participates in the control of cellular differentiation, including myeloid (23), B-lymphoid, and T-lymphoid differentiation (24, 25). All *EGR2* mutations identified in CLL were heterozygous missense mutations, and, with the exception of R426Q, were located within the zinc-finger domains (Fig. 2C). In addition, *EGR2* mutations were detected as early molecular events in 2 patients (CLL12 and CLL22; Fig. 1B and Supplementary Table S1).

To investigate the functional consequences of *EGR2* mutations, we first expressed GST fusion proteins, including the zinc-finger region of WT or two EGR2 mutants (E356K and H384N). Electrophoretic mobility shift assays (EMSA) using a biotinylated probe corresponding to a high-affinity EGR2 site (26) showed specific binding of the WT and the H384N proteins (Fig. 4A), although H384N binding seemed weaker than WT despite comparable protein amounts (Fig. 4B). The interaction of the E356K protein with the probe was not observed in this assay. To investigate their ability to regulate transcription, we expressed the WT and the mutant forms of EGR2 in the murine multipotent hematopoietic cell line EML. Expression levels of all EGR2 isoforms were comparable (Fig. 4C) and were associated with slower growth (Fig. 4E). Cells expressing WT EGR2 showed a progressive reduction in the expression of the cell-surface markers B220 (B lymphoid) and Gr1 (myeloid; Fig. 4D). Growth slowdown and loss of $B220^+$ and $Gr1^+$ cells occurred even faster in cells expressing mutant EGR2 (Fig. 4D and E), indicating that the mutations had a functional impact. We investigated the expression of known EGR2 target genes using RNA obtained from sorted GFP^+ cells, 3 days after transduction, to detect primary transcriptional changes induced by EGR2 expression. As shown in Fig. 4F, WT and mutated EGR2 proteins interfered with the expression of EGR2 targets. The effects of the three EGR2 proteins were similar on *Csf1* transcription, whereas WT-EGR2 was stronger than E356K, which was stronger than H384N, in the transactivation of *Gadd45b*. Taken together, these results indicate that the EGR2 mutations in CLL do not functionally inactivate the protein but rather affect the

Published OnlineFirst June 11, 2014; DOI: 10.1158/2159-8290.CD-14-0104

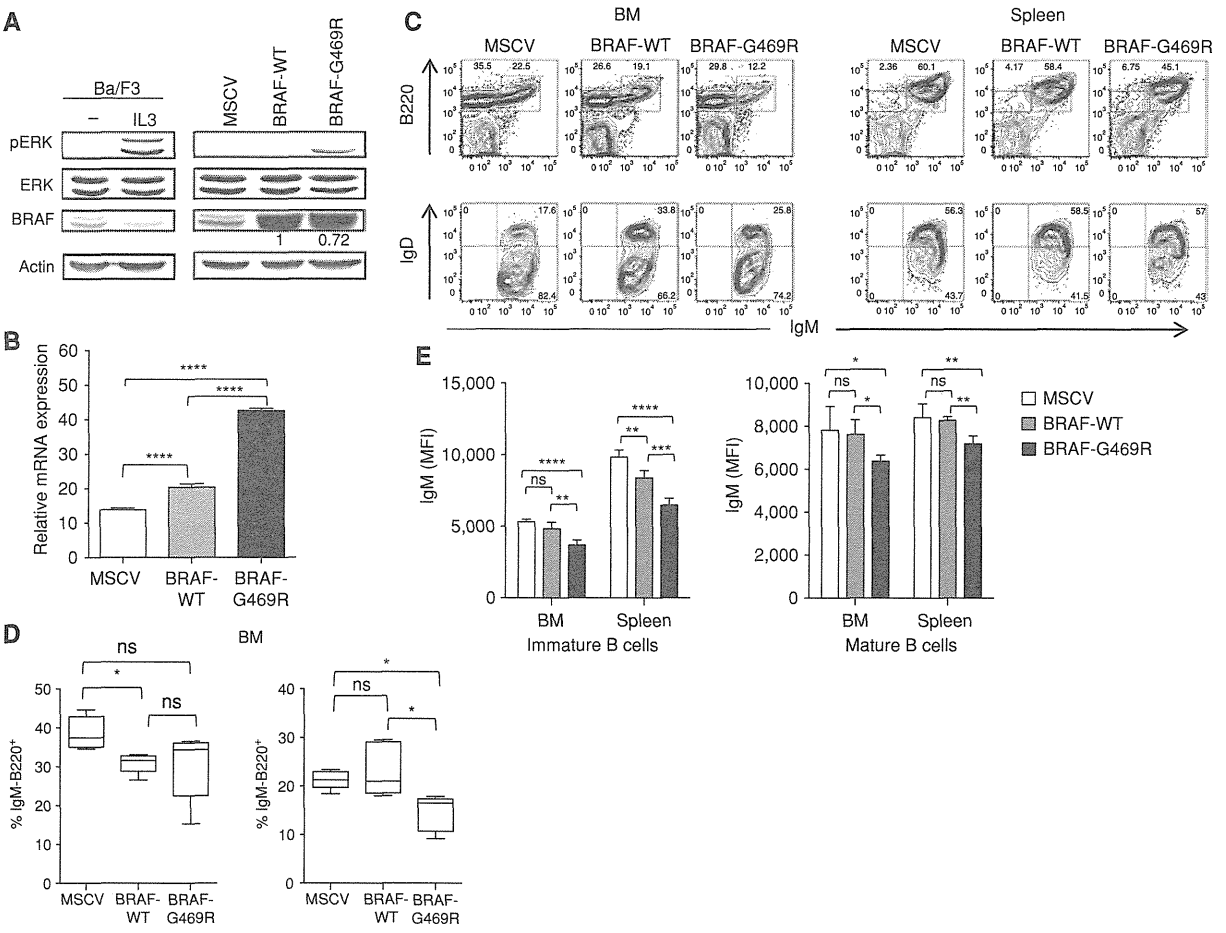

Figure 3. Functional analyses of BRAF-G469R. **A,** expression of BRAF-G469R from MSCV results in constitutive ERK activation. Right, ERK phosphorylation was detected in IL3-starved Ba/F3 cells expressing mutated BRAF, but not BRAF-WT or empty control. Antibodies are indicated and BRAF expression normalized to β-actin and BRAF-WT expression. **B,** real-time quantitative polymerase chain reaction (RQ-PCR) evaluation of *Egr2* expression in transduced Ba/F3 cells, normalized with respect to *Gapdh*. **C,** representative FACS analysis of bone marrow (BM; left) and spleen (right) cells from MSCV, BRAF-WT, and BRAF-G469R mice analyzed 5 weeks after engraftment. Plots are gated on donor (CD45.2$^+$) GFP$^+$ cells and the percentages of gated cells are shown. **D,** mean percentage (±SD; $n = 5$ mice) of bone marrow GFP$^+$ donor cells expressing B220 and negative (left) or positive (right) for membrane IgM. **E,** mean fluorescence intensity (MFI) of IgM in GFP$^+$ donor B cells, immature (IgM$^+$ IgD$^-$; left) and mature (IgM$^+$ IgD$^+$; right). BM or spleen origin is indicated (****, $P < 0.0001$; ***, $P < 0.001$; **, $P < 0.01$; *, $P < 0.05$; ns, $P > 0.05$).

transcription of EGR2 target genes. Whether the differential activities of the mutants are due to differences in DNA binding or interaction with transcription cofactors at target genes will require additional investigation.

To investigate the functional consequences of *EGR2* mutations in patient samples, we analyzed RNA-seq data obtained from 16 CLL samples. Fifteen genes were downregulated, whereas 224 genes were specifically upregulated in EGR2-E356K samples ($n = 4$) as compared with EGR2 WT or unanalyzed patients ($n = 10$; $P < 0{,}01$; Supplementary Table S5). Hierarchical clustering using the 224 upregulated genes showed clustering of all 5 *EGR2*-mutated CLL samples, including the EGR2-H384R sample (Fig. 5A). An additional sample, which was not analyzed by exome sequencing and lacked acquired mutations in *EGR2*, clustered together with the EGR2-mutated samples, suggesting that other alterations mimic the effect of *EGR2* mutations. To investigate whether the differentially expressed genes might be direct EGR2 targets, we used ChIP-seq data obtained from primary human monocyte extracts via chromatin immunoprecipitation with anti-EGR2 antibodies (27). Peaks were observed close to 168 of the 224 upregulated genes, indicating that these genes were likely directly regulated by EGR2 ($P < 0.001$; see Supplementary Table S5 and Methods). To further investigate this point, we used publicly available CLL expression data (28) to identify 24 predicted EGR2 target genes using the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACne; see Methods and Supplementary Table S6). When used as a surrogate marker of EGR2 transcriptional activity, this signature showed transcriptional activity in *EGR2*-mutated samples (Fig. 5B). Together, these data confirm that expression of mutated EGR2 proteins interferes with the expression of EGR2 target genes *in vivo*.

Because EGR2 is downstream of normal BCR signaling, we next determined a BCR signaling signature. For this purpose,

Published OnlineFirst June 11, 2014; DOI: 10.1158/2159-8290.CD-14-0104

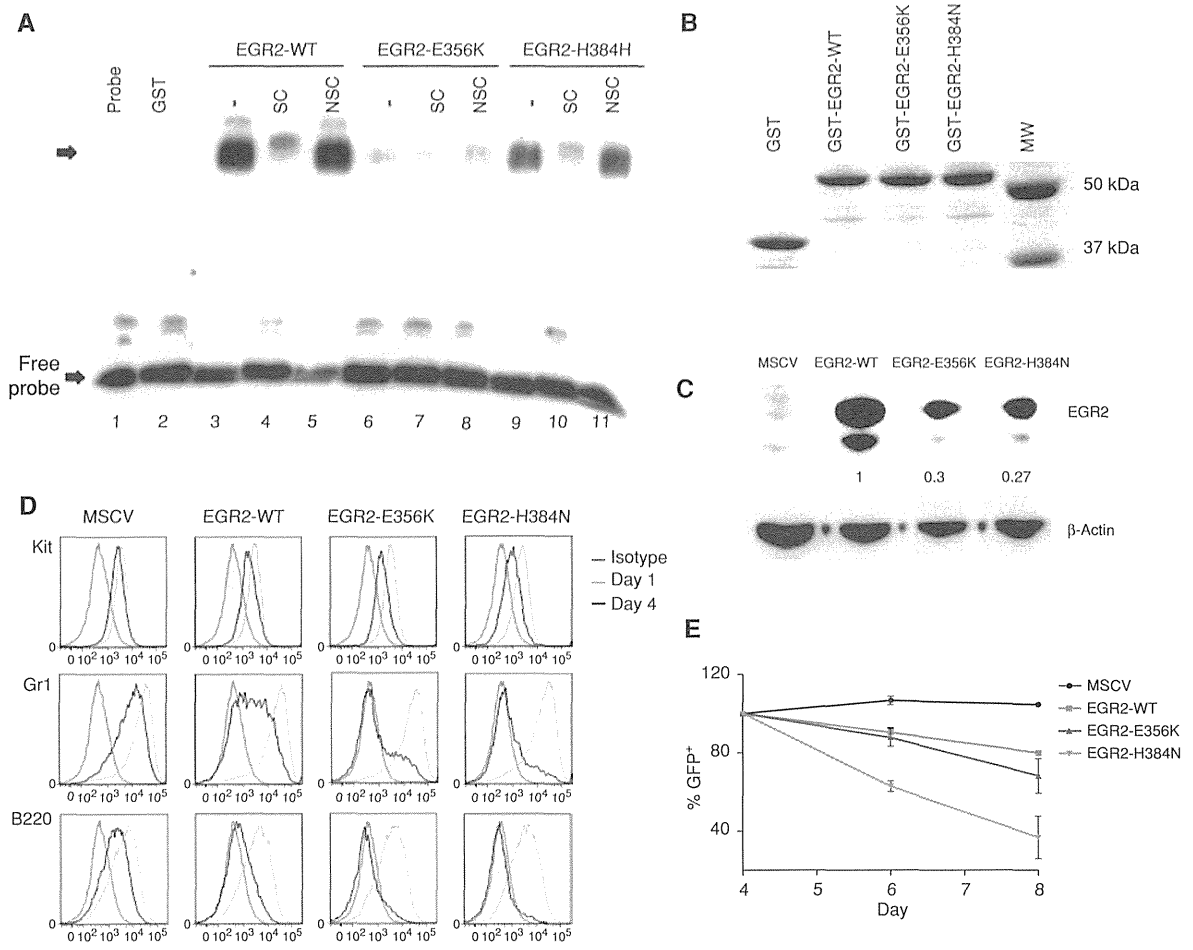

Figure 4. Functional analyses of EGR2 mutants. **A,** the zinc-finger regions of mutant EGR2 are less efficient than WT at binding EGR2 high-affinity sequences (CTCTGTACGCGGGGGCGGTTA) in EMSA. Shifted complexes (indicated by an arrow) are observed in lanes containing GST-EGR2-WT protein (lanes 3 and 5). Formation of this complex is inhibited in the presence of a 200-time molar excess of unlabeled probe used as specific competitor (SC; lanes 4, 7, and 10), but not in the presence of the same excess of a nonspecific competitor (NSC; lanes 5, 8, and 11). Shifted complexes are also detected in lanes corresponding to GST-EGR2-H384N (lanes 9 and 11), but are virtually absent in lanes corresponding to GST-EGR2-E356K (lanes 6 and 8). Lane 1, probe only; lane 2, native GST protein. **B,** qualitative and quantitative assessment of native and fusion GST proteins. MW, molecular weight markers. **C,** Western blot analyses using EGR2-specific antibodies show comparable levels of expression in transduced cells. Ratio was normalized to β-actin and EGR2-WT expression. **D,** expression of EGR2 proteins is associated with a decrease of Gr1 and B220 membrane expression. Analyses were gated on GFP+ cells. FACS analysis at day 1 (blue) and day 4 (black) posttransduction are shown. **E,** growth curve showing EML cells transduced by the indicated vector. Monitoring the proportions of GFP+ cells shows a decrease in the proportion of cells expressing WT and mutant EGR2 proteins, with a stronger decrease for the mutants. (*continued on following page*)

we defined a set of genes upregulated upon BCR stimulation of normal B cells, using available data (29, 30), and used Gene Set Enrichment Analyses (GSEA) to show that this signature is enriched in *EGR2*-mutated samples, with respect to non-mutated samples (Fig. 5C and D). Reciprocally, the EGR2-E356K signature was markedly enriched in BCR-stimulated samples, when compared with unstimulated B cells (Fig. 5E and F), further establishing the deregulation of intracellular BCR signaling in *EGR2*-mutated samples.

DISCUSSION

Here, we identified acquired mutations in the hematopoietic progenitors of patients with CLL and provided proof-of-principle for the role of these mutations during the natural history of the disease.

Our data identified early-mutated genes in patients with CLL. The high mutation burden observed in some patients, in the progenitor and/or mature myeloid fractions, underscores the notion that the identified mutations are functionally relevant and lead to the accumulation of mutated cells, in the progenitor and/or mature fractions. Some of those early drivers are well-known CLL oncogenes (i.e., *NOTCH1*, *XPO1*, and *SF3B1*). We also identified recurrent inactivating mutations of the *NFKBIE* gene in 10% of patients and as an early event in 1 patient. *NFKBIE* encodes an inhibitor of NF-kB activity with a specific role in B-lymphocyte biology (31–33). In addition, we showed that acquired missense