

# Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia

Dan A. Landau,<sup>1,2,4,5,6,10</sup> Scott L. Carter,<sup>4,10</sup> Petar Stojanov,<sup>2,4,10</sup> Aaron McKenna,<sup>4</sup> Kristen Stevenson,<sup>3</sup> Michael S. Lawrence,<sup>4</sup> Carrie Sougnez,<sup>4</sup> Chip Stewart,<sup>4</sup> Andrey Sivachenko,<sup>4</sup> Lili Wang,<sup>1,2</sup> Youzhong Wan,<sup>1,2</sup> Wandi Zhang,<sup>1,2</sup> Sachet A. Shukla,<sup>1,4</sup> Alexander Vartanov,<sup>2</sup> Stacey M. Fernandes,<sup>2</sup> Gordon Saksena,<sup>4</sup> Kristian Cibulskis,<sup>4</sup> Bethany Tesar,<sup>2</sup> Stacey Gabriel,<sup>4</sup> Nir Hacohen,<sup>4,8</sup> Matthew Meyerson,<sup>4</sup> Eric S. Lander,<sup>4</sup> Donna Neuberg,<sup>3</sup> Jennifer R. Brown,<sup>2,7</sup> Gad Getz,<sup>4,9,10,\*</sup> and Catherine J. Wu<sup>1,2,7,10,\*</sup>

<sup>1</sup>Cancer Vaccine Center

<sup>2</sup>Department of Medical Oncology

<sup>3</sup>Department of Biostatistics and Computational Biology  
Dana-Farber Cancer Institute, Boston, MA 02115, USA

<sup>4</sup>Broad Institute, Cambridge, MA 02139, USA

<sup>5</sup>Department of Hematology, Yale Cancer Center, New Haven, CT 06510, USA

<sup>6</sup>Université Paris Diderot, Paris 75013, France

<sup>7</sup>Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA

<sup>8</sup>The Division of Allergy, Immunology, and Rheumatology, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

<sup>9</sup>Massachusetts General Hospital Cancer Center and Department of Pathology, Boston, MA 02114, USA

<sup>10</sup>These authors contributed equally to this work

\*Correspondence: gadgetz@broadinstitute.org (G.G.), cwu@partners.org (C.J.W.)

<http://dx.doi.org/10.1016/j.cell.2013.01.019>

## SUMMARY

Clonal evolution is a key feature of cancer progression and relapse. We studied intratumoral heterogeneity in 149 chronic lymphocytic leukemia (CLL) cases by integrating whole-exome sequence and copy number to measure the fraction of cancer cells harboring each somatic mutation. We identified driver mutations as predominantly clonal (e.g., *MYD88*, trisomy 12, and *del(13q)*) or subclonal (e.g., *SF3B1* and *TP53*), corresponding to earlier and later events in CLL evolution. We sampled leukemia cells from 18 patients at two time points. Ten of twelve CLL cases treated with chemotherapy (but only one of six without treatment) underwent clonal evolution, predominantly involving subclones with driver mutations (e.g., *SF3B1* and *TP53*) that expanded over time. Furthermore, presence of a subclonal driver mutation was an independent risk factor for rapid disease progression. Our study thus uncovers patterns of clonal evolution in CLL, providing insights into its stepwise transformation, and links the presence of subclones with adverse clinical outcomes.

## INTRODUCTION

Recent genomic studies have revealed that individual cancer samples are genetically heterogeneous and contain subclonal

populations (Carter et al., 2012; Ding et al., 2012; Gerlinger et al., 2012; Mullighan et al., 2008; Navin et al., 2011; Nik-Zainal et al., 2012; Shah et al., 2012). Indeed, tumors likely evolve through competition and interactions between genetically diverse clones (Snuderl et al., 2011). While the existence of intratumoral subclones has long been appreciated, little is known about the frequency, identity, and evolution of subclonal genetic alterations or their impact on clinical course.

To examine the evolution and impact of subclonal mutations, we focused on chronic lymphocytic leukemia (CLL), a slow-growing B cell malignancy with disease onset in older individuals. CLL shows a highly variable disease course, partly explained by the diverse combinations of somatic mutations uncovered by sequencing studies (Quesada et al., 2012; Wang et al., 2011). We hypothesized that the presence, diversity, and evolutionary dynamics of subclonal mutations in CLL also contribute to the variations observed in disease tempo and response to therapy (Schuh et al., 2012; Stilgenbauer et al., 2007). Importantly, the slow growth of CLL-B cells (relative to other malignancies) provides an extended window for observing the process of clonal evolution, as it may take months to years for a new clone to fully replace previous clones (Schuh et al., 2012; Wu, 2012).

Subclonal mutations in CLL have been detected by fluorescence *in situ* hybridization (FISH) (Shanafelt et al., 2008) and microarrays (Grubor et al., 2009), showing that they harbor driver lesions and evolve over time. Because these methods can only be used to detect a limited number of genetic alterations, more recent studies have used whole-genome sequencing to quantify thousands of somatic mutations per sample and track subclones

by clustering alterations of similar allelic frequency (Ding et al., 2012; Egan et al., 2012; Nik-Zainal et al., 2012; Schuh et al., 2012; Shah et al., 2012; Walter et al., 2012). However, because genome-wide sequencing is currently not feasibly applied to large sample collections, the patterns of clonal evolution and their effects on disease course have not been fully elucidated.

Whole-exome sequencing (WES) (Gnirke et al., 2009) of tumors is an affordable, rapid, and comprehensive technology for detecting somatic coding mutations. We sought to refine and apply a method for analysis of subclonal mutations using WES because: (1) the high sequencing depth obtained by WES (typically  $\sim 100\times$ – $150\times$ ) enables reliable detection of subclonal mutations required for defining subclones and tracking them over time (Cibulskis et al., 2013); (2) coding mutations likely encompass many of the important driver events that provide fitness advantage for specific clones, and finally; and (3) the relatively low cost of WES permits studies of large cohorts, which is key for understanding the relative fitness and temporal order of driver mutations and for assessing the impact of clonal heterogeneity on disease outcome.

To this end, we performed large-scale WES of 160 CLL tumor/normal pairs that represented the broad clinical spectrum of CLL. In particular, we examined the roles of CLL subclones and the mutations that they harbor by integrative analysis of coding mutations and somatic copy number alterations, which enabled estimation of the cancer cell fraction. This was performed in samples from 149 CLL patients, including 18 patients sampled at two time points, for which both exome sequencing data and copy number data were available. This analysis allowed us to study mutation frequencies, observe clonal evolution, and link subclonal mutations to clinical outcome.

## RESULTS

### Large-Scale WES Analysis of CLL Expands the Compendium of CLL Drivers and Pathways

We performed WES of 160 matched CLL and germline DNA samples (including 82 of the 91 samples previously reported [Wang et al., 2011]). This cohort included patients with both low- and high-risk features based on established prognostic risk factors (Table S1 available online). We applied MuTect (a highly sensitive and specific mutation-calling algorithm; Cibulskis et al., 2013) to the WES data to detect somatic single nucleotide variations (sSNVs) present in as few as 10% of cancer cells. Average sequencing depth of WES across samples was  $\sim 112\times$  (see Extended Experimental Procedures). In total, we detected 2,444 nonsynonymous and 837 synonymous mutations in protein-coding sequences, corresponding to a mean ( $\pm$ SD) somatic mutation rate of  $0.6 \pm 0.28$  per megabase (range, 0.03–2.3) and an average of 15.3 nonsynonymous mutations per patient (range, 2–53) (Table S2A).

Expansion of our sample cohort provided us with the sensitivity to detect 20 putative CLL cancer genes ( $q < 0.1$ ), which was accomplished through recurrence analysis to detect genes enriched with mutations beyond the background mutation rate (Figures 1A, top, and Figure S1) or genes with mutations that overlap with previously reported mutated sites (from COSMIC [Forbes et al., 2010]; Figure 1A, middle; Experimental Proce-

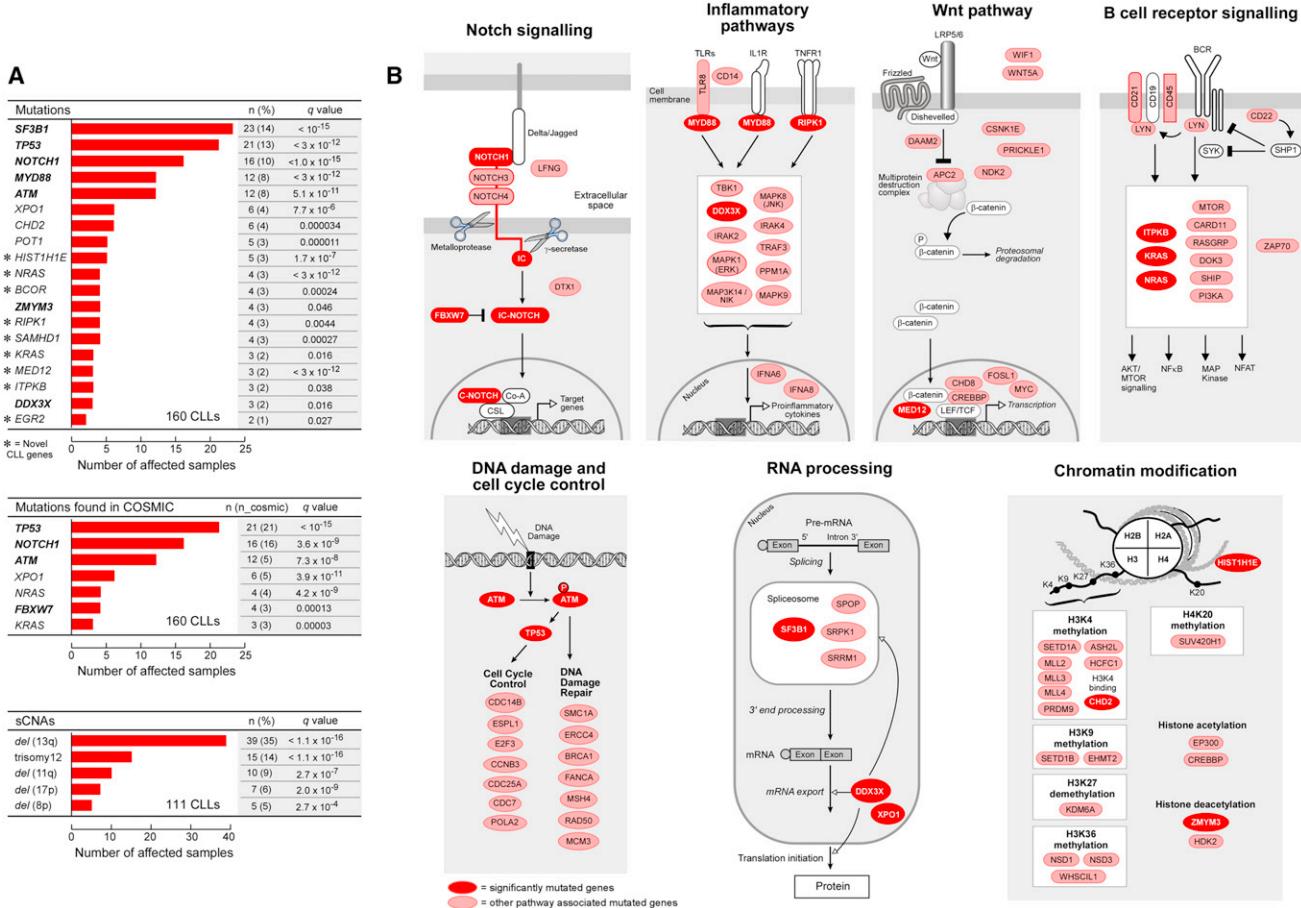
dures). These included eight of the nine genes identified in our initial report (*TP53*, *ATM*, *MYD88*, *SF3B1*, *NOTCH1*, *DDX3X*, *ZMYM3*, and *FBXW7*) (Wang et al., 2011). The missing gene, *MAPK1*, did not harbor additional mutations in the increased sample set, and therefore its overall mutation frequency now fell below our significance threshold. The 12 genes were mutated at lower frequencies and, hence, were not detected in the previously reported subset of samples. Three of the twelve additional candidate driver genes were identified in recent CLL sequencing efforts (*XPO1*, *CHD2*, and *POT1*) (Fabbri et al., 2011; Puente et al., 2011). The nine remaining genes represent candidate CLL drivers, with mutations occurring at highly conserved sites (Figure S2). These included six genes with known roles in cancer biology (*NRAS*, *KRAS* [Bos, 1989], *BCOR* [Grossmann et al., 2011], *EGR2* [Unoki and Nakamura, 2003], *MED12* [Mäkinen et al., 2011], and *RIPK1* [Hosgood et al., 2009]), two genes that affect immune pathways (*SAMHD1* [Rice et al., 2009] and *ITPKB* [Maréchal et al., 2011]), and a histone gene (*HIST1H1E* [Alami et al., 2003]).

Together, the 20 candidate CLL driver genes appeared to fall into seven core signaling pathways, in which the genes play well-established roles. These include all five pathways that we previously reported to play a role in CLL (DNA repair and cell-cycle control, Notch signaling, inflammatory pathways, Wnt signaling, RNA splicing, and processing). Two pathways were implicated by our analysis: B cell receptor signaling and chromatin modification (Figure 1B). We also noted that the CLL samples contained additional mutations in the genes that form these pathways, some of which are known drivers in other malignancies.

Because recurrent chromosomal abnormalities have defined roles in CLL biology (Döhner et al., 2000; Klein et al., 2010), we further searched for loci that were significantly amplified or deleted by analyzing somatic copy-number alterations (sCNAs). We applied GISTIC2.0 (Mermel et al., 2011) to 111 matched tumor and normal samples, which were analyzed by SNP6.0 arrays (Brown et al., 2012). Through this analysis, we identified deletions in chromosomes 8p, 13q, 11q, and 17p and trisomy of chromosome 12 as significantly recurrent events (Figure 1A, bottom). Thus, based on WES and copy number analysis, we altogether identified 20 mutated genes and five cytogenetic alterations as putative CLL driver events.

### Inference of Genetic Evolution with Whole-Exome Sequencing Data

In order to study clonal evolution in CLL, we performed integrative analysis of sCNAs and sSNVs using a recently reported algorithm ABSOLUTE (Carter et al., 2012), which jointly estimated the purity of the sample (fraction of cancer nuclei) and the average ploidy of the cancer cells. All samples were estimated to have near-diploid DNA content; these estimates were confirmed by fluorescence-activated cell sorting (FACS) analysis of seven CLL samples (Figure S3A). Our data were sufficient for resolution of these quantities in 149 of the 160 samples (Table S2B), allowing for discrimination of subclonal from clonal alterations, including sCNAs, sSNVs, and selected indels (see Extended Experimental Procedures). Our analysis approach is outlined in Figure 2A. For each sSNV, we estimated its *allelic fraction* by



**Figure 1. Significantly Mutated Genes and Associated Gene Pathways in 160 CLL Samples**

(A) Mutation significance analysis using the MutSig2.0 and GISTIC2.0 algorithms identifies recurrently mutated genes and recurrent sCNAs in CLL, respectively. Bold: significantly mutated genes identified in a previous CLL sequencing effort (Wang et al., 2011); \*, novel CLL genes identified in the present analysis (Figures S1 and S2); n, number of samples out of 160 CLLs harboring a mutation in a specific gene; n\_cosmic, number of samples harboring a mutation in a specific gene at a site previously observed in the COSMIC database.

(B) The significantly mutated genes fall into seven core signaling pathways, in which the genes play well-established roles. Red: genes with significant mutation frequencies; pink: additional pathway genes with mutations.

See also Figures S1, S2, and Table S2A.

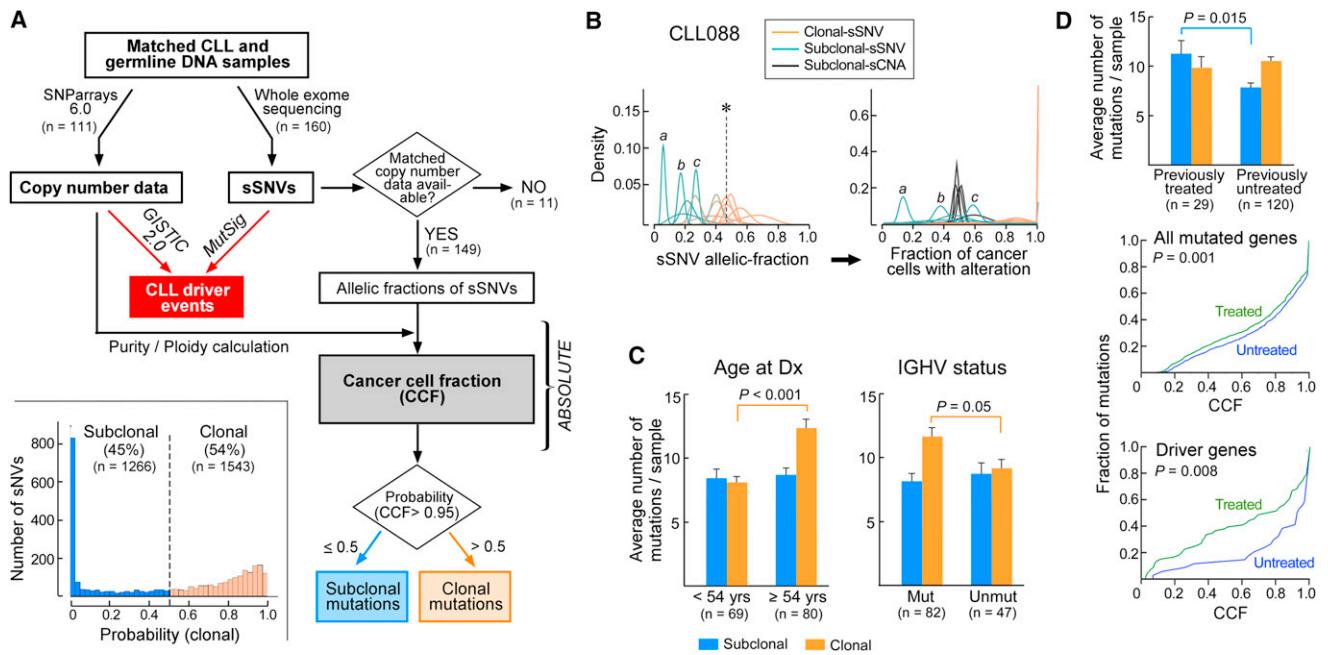
calculating the ratio of alternate to total number of reads covering the mutation site in the WES data. These estimates were consistent with independent deeper sequencing and RNA sequencing (Figures S3B and S3C; Tables S3 and S4). Next, we used ABSOLUTE to estimate the cancer cell fraction (CCF) harboring the mutation by correcting for sample purity and local copy-number at the sSNV sites (Experimental Procedures; Table S2C; Figure 2B). We classified a mutation as clonal if the CCF harboring it was >0.95 with probability >0.5 and subclonal otherwise (Figure 2A, inset). The results remained unchanged when more stringent cutoffs were used (Extended Experimental Procedures). For sSNVs designated as subclonal, median CCF was 0.49 with a range of 0.11–0.89.

Overall, we identified 1,543 clonal mutations (54% of all detected mutations, average of  $10.3 \pm 5.5$  mutations per sample; Table S1). These mutations were likely acquired either before or during the most recent complete selective sweep. This set there-

fore includes both neutral somatic mutations that preceded transformation and the driver and passenger event(s) present in each complete clonal sweep. A total of 1,266 subclonal sSNVs were detected in 146 of 149 samples called by ABSOLUTE (46%; average of  $8.5 \pm 5.8$  subclonal mutations per sample). These subclonal sSNVs exist in only a fraction of leukemic cells and, hence, occurred after the emergence of the “most recent common ancestor” and, by definition, also after disease initiation. The mutational spectra were similar in clonal and subclonal sSNVs (Figure S3D), consistent with a common set of mutational processes, giving rise to both groups.

#### Age and Mutated IGHV Status Are Associated with an Increased Number of Clonal Somatic Mutations

The identification of subclones enabled us to analyze several aspects of leukemia progression. We first addressed how clonal and subclonal mutations relate to the salient clinical



**Figure 2. Subclonal and Clonal sSNVs Are Detected in CLL in Varying Quantities Based on Age at Diagnosis, IGHV Mutation Status, and Treatment Status**

(A) The analysis workflow: “CLL driver events” (red box) were identified by mutation significance analysis using WES and SNP array data collected from matched germline and tumor DNA. For the 149 samples that had matched WES and copy number data, ABSOLUTE was applied to estimate the cancer cell fraction (CCF). Mutations were classified as subclonal (blue) or clonal (orange), based on the probability that their CCF is greater than 0.95 (clonal). Inset: histogram of the probability of being clonal for all sSNVs across 149 CLL samples.

(B) A representative example of the transformations generated by ABSOLUTE (for sample CLL088). First, probability density distributions of allelic fractions for each mutation are plotted (representative peaks for sSNVs a, b, and c shown). Second, these data are converted to CCF (right), incorporating purity and local copy number information. The probability of the event being clonal (i.e., affecting >0.95 of cells) is represented by a color spectrum: orange is high probability and blue is low probability. \*, allelic fraction of a clonal mutation at multiplicity of one (for example, a heterozygous mutation in a diploid region).

(C) Comparison of the number of subclonal and clonal sSNVs/sample based on patient age at diagnosis and IGHV mutation status (error bars represent standard error of mean).

(D) Comparison of the number of subclonal and clonal sSNVs/sample based on treatment status at time of sample collection (top, error bars represent standard error of mean). Cumulative distribution of the sSNVs by CCF is shown for samples from treated and untreated patients for all (middle) and only driver sSNVs (bottom).

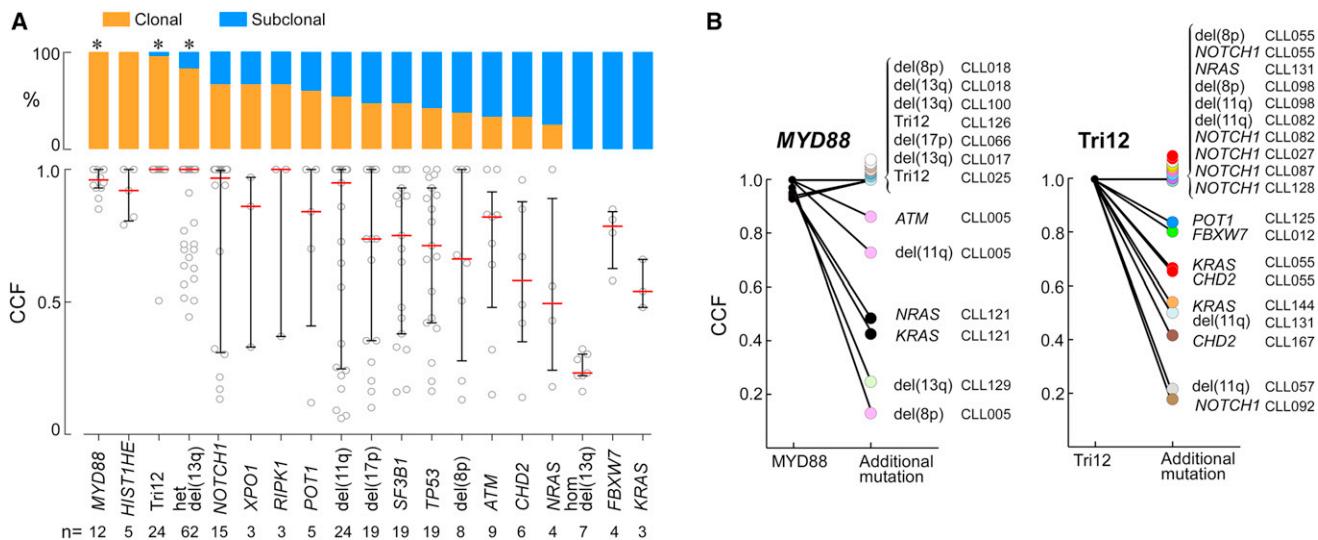
See also Figure S3 and Tables S1, S2B, and S2C.

characteristics of CLL. CLL is generally a disease of the elderly with established prognostic factors, such as the *IGHV* mutation (Döhner, 2005) and *ZAP70* expression. Patients with a high number of *IGHV* mutations (mutated *IGHV*) tend to have better prognosis than those with a low number (unmutated *IGHV*) (Döhner, 2005). This marker distinguishes between leukemias originating from B cells that have or have not yet, respectively, undergone the process of somatic hypermutation that occurs as part of normal B cell development. We examined the association of these factors, as well as patient age at diagnosis, with the prevalence of clonal and subclonal mutations. Age and mutated *IGHV* status (but not *ZAP70* expression) were found to associate with greater numbers of clonal (but not subclonal) mutations (age,  $p < 0.001$ ; mutated versus unmutated *IGHV*,  $p = 0.05$ ; Figure 2C; Table S1). Because CLL samples with mutated *IGHV* derive from B cells that have experienced a burst of mutagenesis as part of normal B cell somatic hypermutation, the increased number of clonal somatic mutations is likely related to aberrant mutagenesis that preceded clonal transforma-

mation (Deutsch et al., 2007; McCarthy et al., 2003). Furthermore, the higher number of clonal sSNVs in older individuals is consistent with the expectation that more neutral somatic mutations accumulate over the patient’s lifetime prior to the onset of cancer later in life (Stephens et al., 2012; Welch et al., 2012).

### Subclonal Mutations Are Increased with Treatment

The effect of treatment on subclonal heterogeneity in CLL is unknown. In samples from 29 patients treated with chemotherapy prior to sample collection, we observed a significantly higher number of subclonal (but not clonal) sSNVs per sample than in the 120 patients who were chemotherapy-naïve at the time of sample (Figure 2D, top and middle). Using an analysis of covariance model, we observed that receipt of treatment prior to sample among the 149 patients was statistically significant ( $p = 0.048$ ) but time from diagnosis to sample was not ( $p = 0.31$ ). Because patients that do not require treatment in the long-term may have a distinct subtype of CLL, we also restricted the comparison of the 29 pretreated CLLs to only the 42 that



**Figure 3. Identification of Earlier and Later CLL Driver Mutations**

(A) Distribution of estimated CCF with median and interquartile range shown in black (bottom) and percent of the mutations classified as clonal (orange) or subclonal (blue) for CLL drivers (top). \*, drivers with  $q$ -values  $<0.1$  for a higher proportion of clonal mutations compared with the entire CLL drivers set. Het, heterozygous deletion; Hom, homozygous deletion. The analysis includes all recurrently mutated genes (Figure 1A) with three or more events in the 149 samples, excluding sSNVs affecting the X chromosome currently not analyzable by ABSOLUTE and also excluding indels in genes other than NOTCH1.

(B) All CLL samples with *MYD88* (left) or trisomy 12 (right) and at least one additional defined CLL driver (i.e., 9 of 12 samples with mutated *MYD88* and 14 of 16 tumors with trisomy 12) are depicted. Each dot color denotes separate individual CLL samples.

See also Figure S4 and Table S2C.

were eventually treated after sample collection and again confirmed this finding ( $p = 0.02$ ). In these 42 patients, a higher number of subclonal mutations was not correlated with a shorter time to treatment (correlation coefficient = 0.03;  $p = 0.87$ ). Thus, therapy prior to sample was associated with a higher number of subclonal sSNVs detected increased with the number of prior therapies ( $p = 0.011$ , Table S1).

Cancer therapy has been theorized to be an evolutionary bottleneck, in which a massive reduction in malignant cell numbers results in reduced genetic variation in the cell population (Gerlinger and Swanton, 2010). It is likely that the overall diversity in CLL is diminished after therapeutic bottlenecks as well. Because most of the genetic heterogeneity within a cancer is present at very low frequencies (Gerstung et al., 2012)—below the level of detection afforded by the  $\sim 112 \times$  sequence coverage we generated—we were unable to directly assess reduction in overall genetic variation.

However, in the range of larger subclones that were observable by our methods ( $>10\%$  of malignant cells), we witnessed increased diversity after therapy (Figure 2D). Although the available data cannot definitively rule out extensive diversification following therapy, this increase likely results, at least in part, from outgrowth of pre-existing minor subclones (Schuh et al., 2012; Wu, 2012). This may result from the removal of dominant clones by cytotoxic treatment, eliminating competition for growth and allowing the expansion of one or more fit subclones to frequencies above our detection threshold. Further supporting our interpretation that fitter clones grow more effectively and become detectable after treatment, we observed an increased

frequency of subclonal driver events (which are presumably fitter) in treated relative to untreated patients (Figure 2D, bottom) (note that driver events include CLL driver mutations [Figure 1A] and sSNVs in highly conserved sites of genes in the Cancer Gene Census [Futreal et al., 2004]).

#### Infering the Order of Genetic Changes Underlying CLL

While general aspects of temporal evolution could not be completely resolved in single time point WES samples, the order of driver mutation acquisition could be partially inferred from the aggregate frequencies at which they are found to be clonal or subclonal. We considered the 149 samples as a series of “snapshots” taken along a temporal axis. Clonal status in all or most mutations affecting a specific gene or chromosomal lesion would suggest that this alteration was acquired at or prior to the most recent selective sweep before sampling and hence could be defined as a stereotypically early event. Conversely, predominantly subclonal status in a specific genetic alteration implies a likely later event that is tolerated and selected for only in the presence of an additional mutation.

This strategy was used to infer temporal ordering of the recurrent sSNVs and sCNAs (Figures 3A and S4). We focused on alterations found in at least three samples within the cohort of 149 CLL samples. We found that three driver mutations—*MYD88* ( $n = 12$ ), trisomy 12 ( $n = 24$ ), and hemizygous *del(13q)* ( $n = 70$ )—were clonal in 80%–100% of samples harboring these alterations, a significantly higher level than for other driver events ( $q < 0.1$ , Fisher exact test with Benjamini-Hochberg FDR [Benjamini and Hochberg, 1995]), implying that they arise earlier in typical CLL development. Mutations in *HIST1H1E*, although

clonal in five of five affected samples, did not reach statistical significance. Other recurrent CLL drivers—for example, *ATM*, *TP53*, and *SF3B1* (9, 19, and 19 mutations in 6, 17, and 19 samples, respectively)—were more often subclonal, indicating that they tend to arise later in leukemic development and contribute to disease progression. We note that the above approach assumed that different CLL samples evolve along a common temporal progression axis. We therefore examined specifically CLL samples that harbored one “early” driver mutation and any additional driver alteration(s). As expected, the early events had either similar or a higher CCF compared to “later” events (examples for trisomy 12 and *MYD88* given in Figure 3B).

#### Direct Observation of Clonal Evolution by Longitudinal Data Analysis of Chemotherapy-Treated CLL

To directly assess the evolution of somatic mutations in a subset of patients, we compared CCF for each alteration across two clinical time points in 18 of the 149 samples (median years between time points was 3.5; range 3.1–4.5). Six patients (“untreated”) did not receive treatment throughout the time of study. The remaining 12 patients (“treated”) received intervening chemotherapy (primarily fludarabine and/or rituxan-based) (Table S3). The two patient groups were not significantly different in terms of elapsed time between first and second sample (median 3.7 years for the six untreated patients compared to 3.5 years for the 12 treated patients,  $p = 0.62$ ; exact Wilcoxon rank-sum test), nor did it differ between time of diagnosis to first sample ( $p = 0.29$ ).

Analysis of the 18 sets of data revealed that 11% of mutations increased (34 sSNVs, 15 sCNAs), 2% decreased (six sSNVs, two sCNAs), and 87% did not change their CCF over time ( $q < 0.1$  for significant change in CCF, Table S5C). As suggested by our single time point analysis, we observed a shift of subclonal driver mutations (e.g., *del(11q)*, *SF3B1*, and *TP53*) toward clonality over time. Changes in the genetic composition of CLL cells with clonal evolution were associated with network level changes in gene expression related to emergence of specific subclonal populations (e.g., changes in signatures associated with *SF3B1* or *NRAS* mutation [Figures S5D and S5E; Table S6]). Finally, expanding sSNVs were enriched in genes included in the Cancer Gene Census (Futreal et al., 2004) ( $p = 0.021$ ) and in CLL drivers ( $p = 0.028$ ), consistent with the expected positive selection for the subclones harboring them.

Clustering analysis of CCF distributions of individual genetic events over the two time points (Extended Experimental Procedures) revealed clear clonal evolution in 11 of 18 CLL sample pairs. We observed clonal evolution in 10 of 12 sample pairs that had undergone intervening treatment between time points 1 and 2 (Figures 4B and S5A–S5C). This was contrasted with the six untreated CLLs, five of which demonstrated equilibrium between subpopulations that was maintained over several years (Figure 4A;  $p = 0.012$ , Fisher exact test). Of the 11 patients with subclonal evolution across the sampling interval, five followed a branched evolution pattern, as indicated by the disappearance of mutations with high CCF co-occurring with the expansion of other subclones (Figure 4B). This finding demonstrates that co-existing sibling subclones are at least as common in CLL as are linear nested subclones, as demonstrated in other hematological malignancies (Ding et al., 2012; Egan et al., 2012). We conclude that chemotherapy-treated CLLs often undergo clonal evolution, resulting in the expansion of previously minor subclones. Thus, these longitudinal data validate the insights obtained in the cross-sectional analysis, namely that (1) later driver events expand over time (Figure 3A), and (2) treatment results in the expansion of subclones enriched with drivers (and thus presumably have higher fitness) (Figure 2D).

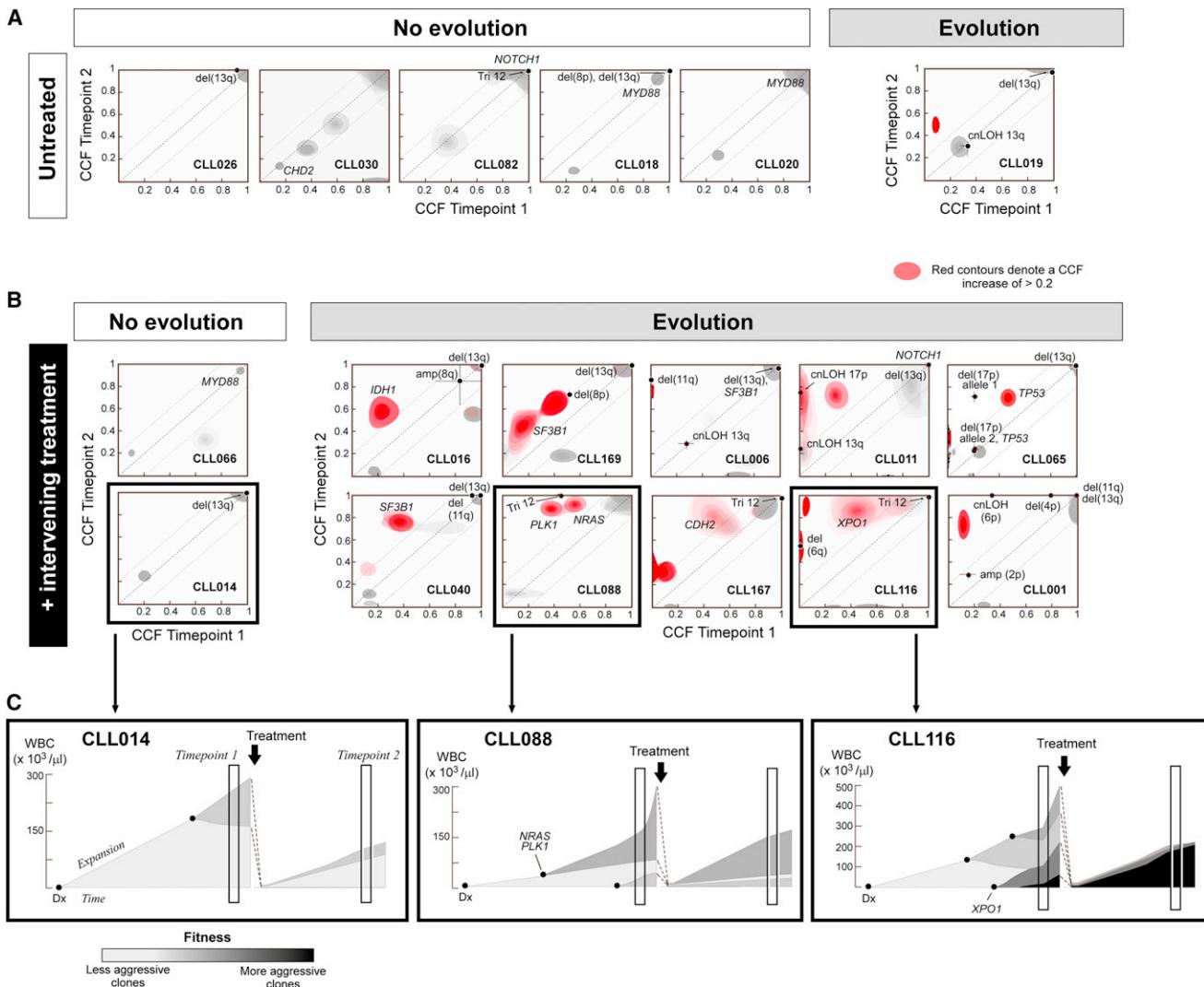
**Presence of Subclonal Drivers Adversely Impacts Clinical Outcome**

#### Presence of Subclonal Drivers Adversely Impacts Clinical Outcome

We observed treatment-associated clonal evolution to lead to the replacement of the incumbent clone by a fitter pre-existing subclone (Figure 4B). Therefore, we would expect a shorter time to relapse in individuals with evidence of clonal evolution following treatment. As a measure of relapse, we assessed failure-free survival from time of sample (“FFS\_Sample”) and failure-free survival from time of next therapy (“FFS\_Rx”) (Figure 5A), where failure is defined as retreatment (a recognized endpoint in slow growing lymphomas [Cheson et al., 2007]) or death. For the study of clonal evolution in CLL, retreatment as an endpoint is preferable to other measures, such as progression alone, as this is a well-defined event that reflects CLL disease aggressiveness. For example, disease progression alone in CLL may be asymptomatic without necessitating treatment; conversely, treatment is administered only in the setting of symptomatic disease or active disease relapse (Hallek et al., 2008).

Within the 12 of 18 longitudinally analyzed samples that received intervening treatment, we observed that the ten samples with clonal evolution exhibited shortened FFS\_Rx (log-rank test;  $p = 0.015$ , Figure 5B). Importantly, the somatic driver mutations that expanded to take over the entire population upon relapse (“time point-2”) were often already detectable in the pretreatment (“time point-1”) sample (Figures 4B and S5B). Our results thus suggested that presence of detectable subclonal drivers in pretreatment samples can anticipate clonal evolution in association with treatment. Indeed, the 8 of 12 samples with presence of subclonal drivers in pretreatment samples exhibited shorter FFS\_Rx than the four samples with subclonal drivers absent ( $p = 0.041$ , Figure 5C). Together, the results of our longitudinally studied patient samples suggested that the presence of driver events within subclones may impact prognosis and clinical outcome.

We tested this hypothesis in the set of 149 patient samples, of which subclonal driver mutations were detected in 46% (Figure 6A; Extended Experimental Procedures; Table S4). Indeed, we found that CLL samples with subclonal driver mutations were associated with a shorter time from sample collection to treatment or death (FFS\_Sample,  $p < 0.001$ , Figure 6B; Tables S7A and S7C), which seemed to be independent of established markers of poor prognosis (i.e., unmutated *IGHV* or presence of *del(11q)* or *del(17p)*, Figure S6). Moreover, we tested specifically whether the presence of pretreatment subclonal drivers was associated with a shorter FFS\_Rx, as we observed in the longitudinal data. Therefore, we focused on the 67 patients who were treated after sample collection (median time to first therapy from time of sample was 11 months [range 1–45]). These patients



**Figure 4. Longitudinal Analysis of Subclonal Evolution in CLL and Its Relation to Therapy**

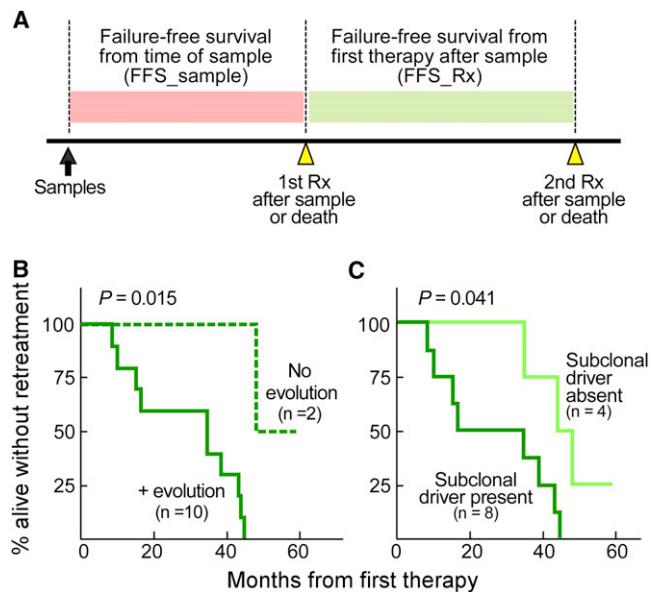
(A and B) Joint distributions of cancer cell fraction (CCF) values across two time points were estimated using clustering analysis (see Extended Experimental Procedures). Red: a mutation with an increase in CCF of greater than 0.2 (with probability >0.5). The dotted diagonal line represents  $y = x$  or where identical CCF values across the two time points fall; the dotted parallel lines denote the 0.2 CCF interval on either side. Likely driver mutations were labeled. Six CLLs with no intervening treatment (A) and 12 CLLs with intervening treatment (B) were classified according to clonal evolution status, based on the presence of mutations with an increase of CCF > 0.2.

(C) Hypothesized sequence of evolution, inferred from the patients' WBC counts, treatment dates, and changes in CCF for three representative examples. See also Figure S5 and Table S5C.

could be divided into two groups based on the presence ( $n = 39$ ) or absence ( $n = 29$ ) of a subclonal driver (62% and 64%, respectively, were treated with fludarabine-based immunochemotherapy,  $p = 0.4$ ). The 39 of these patients in which subclonal CLL drivers were detected required earlier retreatment or died (shorter FFS\_Rx; log-rank test,  $p = 0.006$ , Figure 6C; Table S7A), indicative of a more rapid disease course.

Regression models adjusting for multiple CLL prognostic factors (*I*GHV status, prior therapy, and high-risk cytogenetics) supported the presence of a subclonal driver as an independent risk factor for earlier retreatment (adjusted hazard ratio [HR] of 3.61 [ $\text{CI } 1.42\text{--}9.18$ ], Cox  $p = 0.007$ ; unadjusted HR, 3.20 [ $\text{CI } 1.14\text{--}8.26$ ], Cox  $p = 0.008$ ).

1.35–7.60], Figure 6D), comparable to the strongest known CLL risk factors. In similar modeling within a subset of 62 patients who had at least one driver (clonal or subclonal), the association of the presence of a subclonal driver with a shorter time to retreatment or death was also significant ( $p = 0.012$ , Table S7B), reflecting that this difference is not merely attributable to the presence of a driver. Additionally, an increased number of subclonal driver mutations per sample (but not clonal drivers) was also associated with a stronger HR for shorter FFS\_Rx (Table S7D). Finally, this association retained significance (Cox  $p = 0.033$ , Table S7E) after adjusting for the presence of mutations previously associated with poor prognosis (ATM,



**Figure 5. Genetic Evolution and Clonal Heterogeneity Result in Altered Clinical Outcome**

(A) Schema of the main clinical outcome measures that were analyzed: failure-free survival from time of sample (FFS<sub>Sample</sub>) and from initiation of first treatment after sampling (FFS<sub>Rx</sub>). Within the longitudinally followed CLLs that received intervening treatment (12 of 18), shorter FFS<sub>Rx</sub> was observed in CLL samples that (B) had evidence of genetic evolution ( $n = 10$ ) compared to samples with absent or minimal evolution ( $n = 2$ ; Fisher exact test) and that (C) harbored a detectable subclonal driver in the pretreatment sample ( $n = 8$ ) compared to samples with absent subclonal driver ( $n = 4$ ).

See also Table S3.

TP53, and SF3B1), suggesting that, in addition to the driver's identity, its subclonal status also affects clinical outcome.

## DISCUSSION

While intertumoral (Quesada et al., 2012; Wang et al., 2011) and intratumoral (Schuh et al., 2012; Stilgenbauer et al., 2007) genetic heterogeneity had been previously demonstrated in CLL, our use of WES-based algorithms enabled a more comprehensive study of clonal evolution and its clinical impact. We propose the existence of distinct periods in CLL progression. In the first period prior to transformation, passenger events accumulate in the cell that will eventually be the founder of the leukemia (in proportion to the age of the patient, Figure 2C), and are thus clonal mutations (Figure 7A). In the second period, the founding CLL mutation appears in a single cell and leads to transformation (Figure 7B); these are also clonal mutations, but unlike passenger mutations, these are recurrent across patients. We identified driver mutations that were consistently clonal (del(13q), MYD88, and trisomy 12 (Figure 3A), which appear to be relatively specific drivers of CLL or B cell malignancies (Beroukhim et al., 2010; Döhner et al., 2000; Ngo et al., 2010). In the third period of disease progression, subclonal mutations expand over time as a function of their fitness-integrating intrinsic factors (e.g., proliferation and apoptosis) and extrinsic

pressures (e.g., interclonal competition and therapy) (Figures 7C and 7D). The subclonal drivers include ubiquitous cancer genes, such as ATM, TP53, or RAS mutations (Figure 3A). These data suggest that mutations that selectively affect B cells may contribute more to the initiation of disease and precede selection of more generic cancer drivers that underlie disease progression—providing predictions that can be tested in human B cells or animal models of CLL.

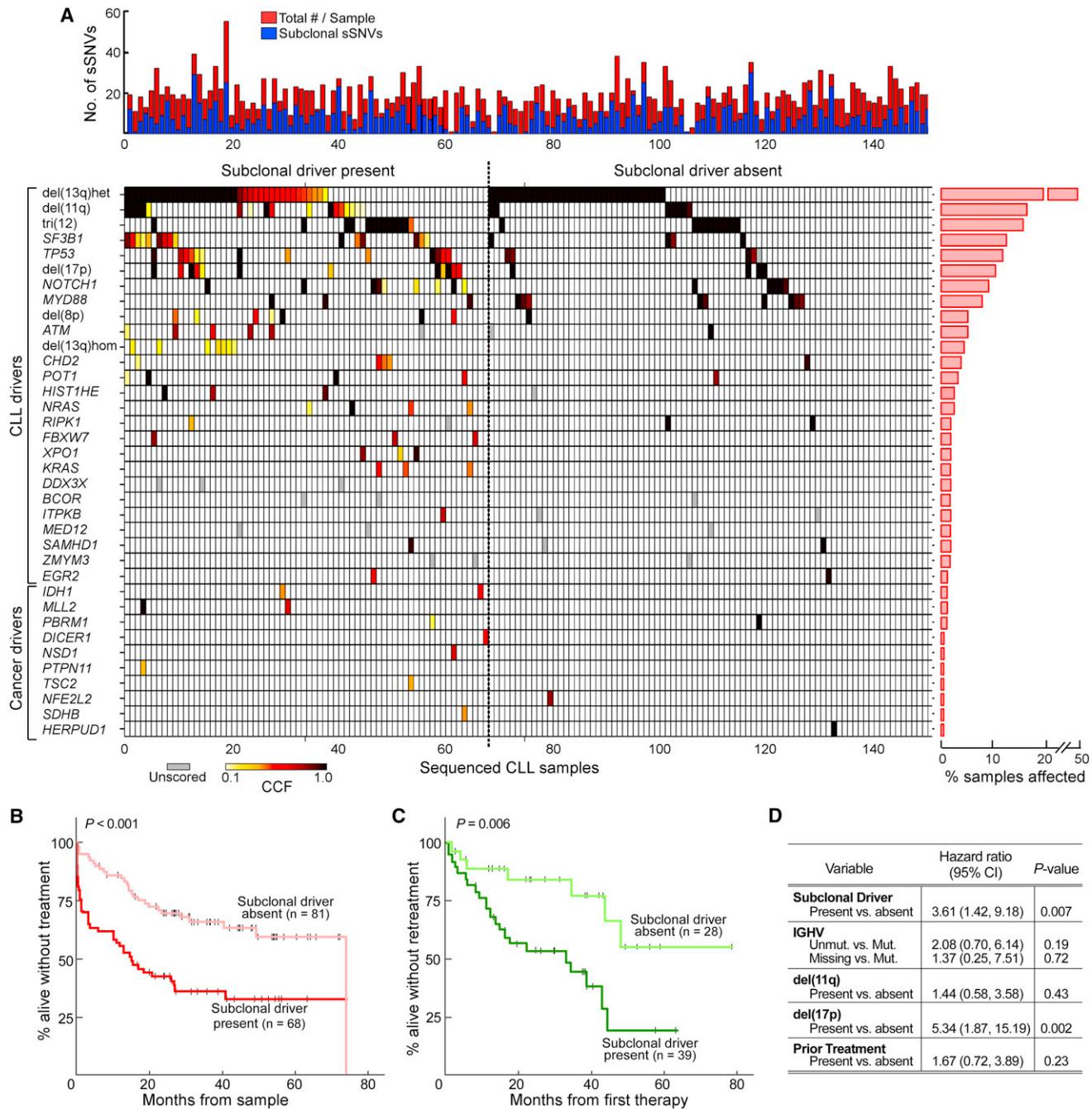
An important question addressed here is how treatment affects clonal evolution in CLL. In the 18 patients monitored at two time points, we observed two general patterns—clonal equilibrium, in which the relative sizes of each subclone were maintained, and clonal evolution, in which some subclones emerge as dominant (Figure 4). We propose that, in untreated samples, more time is needed for a new fit clone to take over the population in the presence of existing dominant clones (Figure 7D, top). In contrast, in treated samples, cytotoxic therapy typically removes the incumbent clones (Jablonski, 2001)—acting like a “mass extinction” event (Jablonski, 2001)—and shifts the evolutionary landscape (Nowak and Sigmund, 2004; Vincent and Gatenby, 2008) in favor of one or more aggressive subclones (Maley et al., 2006; Figure 7D, bottom). Thus, highly fit subclones likely benefit from treatment and exhibit rapid outgrowth (Greaves and Maley, 2012).

CLL is an incurable disease with a prolonged course of remissions and relapses. It has been long recognized that relapsed disease responds increasingly less well to therapy over time. We now show an association between increased clinical aggressiveness and genetic evolution, which has therapeutic implications. We found that the presence of pretreatment subclonal driver mutations anticipated the dominant genetic composition of the relapsing tumor. Such information may eventually guide the selection of therapies to prevent the expansion of highly fit subclones. In addition, the potential hastening of the evolutionary process with treatment provides a mechanistic justification for the empirical practice of “watch and wait” as the CLL treatment paradigm (CLL Trialists Collaborative Group, 1999). The detection of driver mutations in subclones (a testimony to an active evolutionary process) may thus provide a prognostic approach in CLL, which can now be rigorously tested in larger clinical trials.

In conclusion, we demonstrate the ability to study tumor heterogeneity and clonal evolution with standard WES. These innovations will allow characterization of the subclonal mutation spectrum in large, publicly available data sets (Masica and Karchin, 2011). The implementation described here may also be readily adopted for clinical applications. Even more importantly, our studies underscore the importance of evolutionary development as the engine driving cancer relapse. This knowledge challenges us to develop therapeutic paradigms that not only target specific drivers (i.e., “targeted therapy”) but also the evolutionary landscape (Nowak and Sigmund, 2004) of these drivers.

## EXPERIMENTAL PROCEDURES

One hundred forty-nine patients with CLL provided tumor and normal DNA for exome-sequencing and copy number assessment in this study. Sample



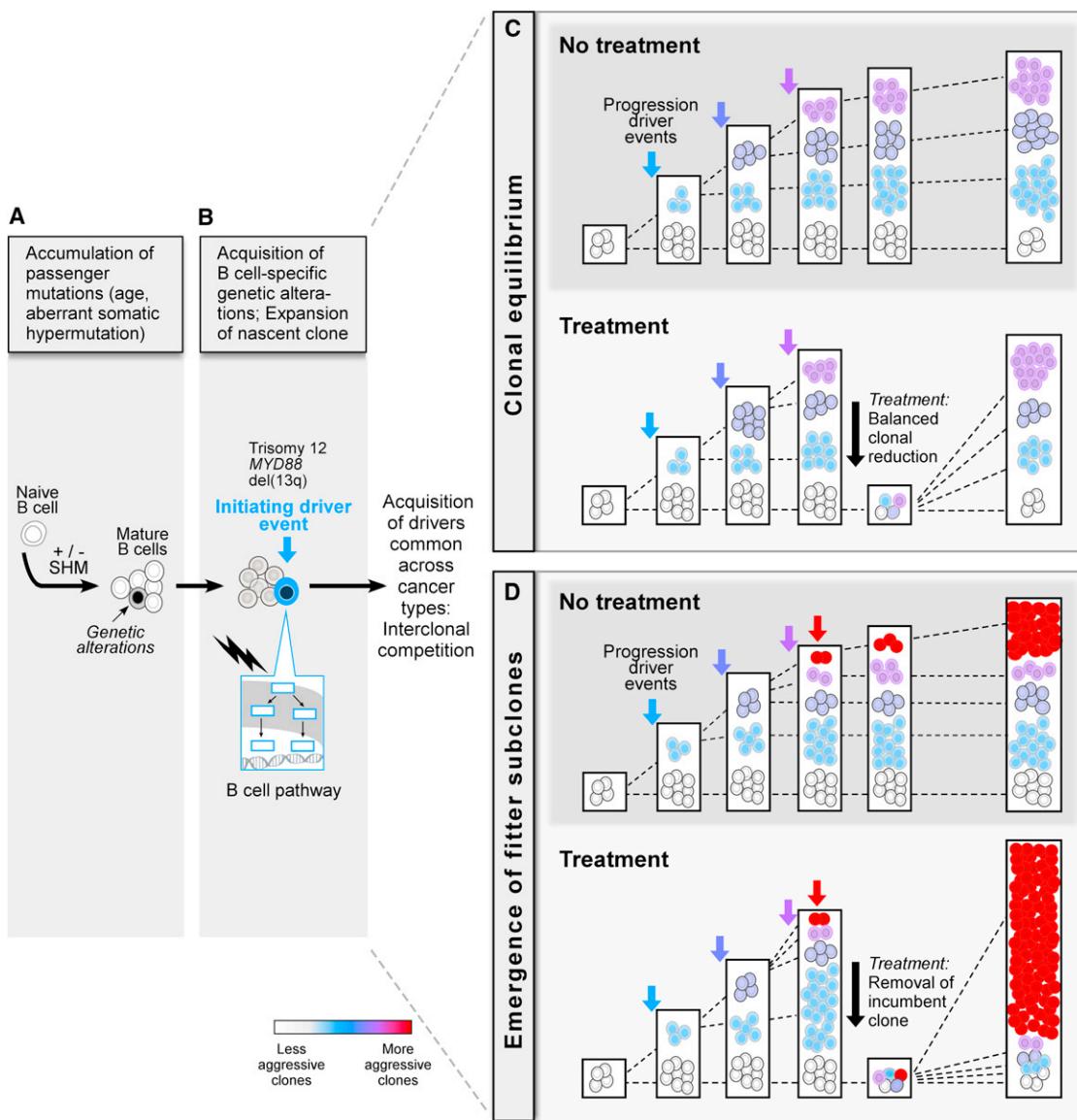
**Figure 6. Presence of Subclonal Drivers Mutations Adversely Impacts Clinical Outcome**

(A) Analysis of genetic evolution and clonal heterogeneity in 149 CLL samples. Top: total number (red) and the number of subclonal (blue) mutations per sample. Bottom: co-occurring CLL or cancer drivers (sSNVs in highly conserved sites in Cancer Gene Census genes) detected in the 149 samples. Color spectrum (light yellow to black) corresponds to estimated CCF; white boxes: not detected; gray: CCF not estimated (genes on the X chromosome and indels other than in NOTCH1).

(B and C) Subclonal drivers are associated with adverse clinical outcome. (B) CLL samples containing a detectable subclonal driver ( $n = 68$ ) exhibited shorter FFS\_Sample compared to samples with absent subclonal drivers ( $n = 81$ ) (also see Figure S6). (C) Subclonal drivers were associated with shorter FFS\_Rx in 67 samples, which were treated after sampling.

(D) A Cox multivariable regression model designed to test for prognostic factors contributing to shorter FFS\_Rx showed that presence of a subclonal driver was an independent predictor of outcome.

See also Figure S6 and Tables S4 and S7.



**Figure 7. A Model for the Stepwise Transformation of CLL**

(A–D) Our data suggest distinct periods in the history of CLL. We observed an increase in clonal mutations in older patients and in the *IGHV* mutated subtype, likely corresponding to pretransformation mutagenesis (A). We identified earlier and later mutations in CLL, consistent with B cell-specific (B) and ubiquitous cancer events (C and D), respectively. Finally, clonal evolution and treatment show a complex relationship. Most untreated CLLs and a minority of treated CLLs maintain stable clonal equilibrium over years (C). However, in the presence of a subclone containing a strong driver, treatment may disrupt interclonal equilibrium and hasten clonal evolution (D).

material was obtained from patients enrolled in clinical research protocols at the Dana-Farber Harvard Cancer Center (DFHCC), approved by the DFHCC Human Subjects Protection Committee. Tumor and normal DNA from 11 additional patients were also analyzed by DNA sequencing alone (a total of 160 CLL samples). Eighty-two CLL samples were previously reported (Wang et al., 2011), and the raw BAM files for these samples were reprocessed and reanalyzed together with the new data to ensure the consistency of the results and to enable the detection of smaller subclones made possible with a newer version of the mutation caller. Written informed consent was obtained prior to sample collection according to the Declaration of Helsinki. DNA was extracted from blood- or marrow-derived lymphocytes (tumor) and autologous epithelial cells (saliva), fibroblasts, or granulocytes (normal).

Libraries for WES were constructed and sequenced on either an Illumina HiSeq 2000 or GA-IIx using 76 bp paired-end reads (Berger et al., 2011; Chapman et al., 2011). Output from Illumina software was processed by the Picard data processing pipeline to yield BAM files containing well-calibrated, aligned reads (Chapman et al., 2011; DePristo et al., 2011). sSNVs and indels were identified using MuTect (V119, [Cibulskis et al., 2013]) and indelocator (V61, [Wang et al., 2011]), respectively. Recurrent sSNV and indels in 160 CLLs were identified using MutSig2.0 (Lohr et al., 2012). For 111 of 149 matched CLL-normal DNA samples, copy number profiles were obtained using the Affymetrix Genome-wide Human SNP Array 6.0, with allele-specific analysis (HAPSEG [Carter et al., 2011]). Recurrent sCNAs were identified using the GISTIC2.0 algorithm (Mermel et al., 2011), after excluding germline copy

number variants. For CLL samples with no available SNP arrays (38 of 149 CLLs), sCNAs were estimated directly from the WES data based on the ratio of CLL sample read-depth to the average read-depth observed in normal samples for that region. We applied ABSOLUTE to estimate sample purity, ploidy, and absolute somatic copy numbers. These were used to infer the CCFs of point mutations from the WES data. Following the framework previously described (Carter et al., 2012), we computed the posterior probability distribution over CCF  $c$  as follows. Consider a somatic mutation observed in  $a$  of  $N$  sequencing reads on a locus of absolute somatic copy number  $q$  in a sample of purity  $\alpha$ . The expected allele-fraction  $f$  of a mutation present in one copy in a fraction  $c$  of cancer cells is calculated by  $f(c) = \alpha c / (2(1 - \alpha) + \alpha q)$ , with  $c \in [0.01, 1]$ . Then,  $P(c) \propto \text{Binom}(a|N, f(c))$ , assuming a uniform prior on  $c$ . The distribution over CCF was then obtained by calculating these values over a regular grid of 100  $c$  values and normalizing by dividing them by their sum, which is the constant of proportionality in the above equation. Mutations were thereafter classified as clonal based on the posterior probability that the CCF exceeded 0.95 and subclonal otherwise. Validation of allelic fraction was performed by using deep sequencing with indexed libraries recovered on a Fluidigm chip. Resulting normalized libraries were loaded on a MiSeq instrument (Illumina) and sequenced using paired-end 150 bp sequencing reads to an average coverage depth of 4,200 $\times$ .

Associations between mutation rates and clinical features were assessed by the Wilcoxon rank-sum test, Fisher exact test, or the Kruskal-Wallis test, as appropriate. Time to event data were estimated by the method of Kaplan and Meier, and differences between groups were assessed using the log-rank test. Unadjusted and adjusted Cox modeling was performed to assess the impact of the presence of a subclonal driver on clinical outcome measures alone and in the presence of clinical features known to impact outcome. A chi-square test with 1 degree of freedom and the  $-2 \log_{10}$  likelihood statistic were used to test the prognostic independence of subclonal status in Cox modeling.

A complete description of the materials and methods is provided in the [Extended Experimental Procedures](#).

## ACCESSION NUMBERS

The dbGaP accession number for the WES data reported in this paper is phs000435.v1.p1. The National Center for Biotechnology Information Gene Expression Omnibus accession number for the array data reported in this paper is GSE37168.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, six figures, and seven tables and can be found with this article online at <http://dx.doi.org/10.1016/j.cell.2013.01.019>.

## ACKNOWLEDGMENTS

D.A.L. dedicates this manuscript to the loving memory of his mother Nina, who passed away during the final stages of this work. We thank all members of the Broad Institute's Biological Samples, Genetic Analysis, and Genome Sequencing Platforms, who made this work possible (NHGRI-U54HG003067). D.A.L. is supported by an American Society of Hematology (ASH) Research Award for Fellows-in-Training and an ACS Postdoctoral Fellowship. J.R.B. is supported by NIH K23 CA115682, the Melton and Rosenbach Funds, and is an ASH Scholar and a LLS Clinical Research Scholar. C.J.W. acknowledges support from the Blavatnik Family Foundation, AACR (SU2C Innovative Research Grant), NHLBI (1RO1HL103532-01), NCI (1R01CA155010-01A1), and is a clinical investigator supported in part by the Damon-Runyon Cancer Research Foundation (CI-38-07).

Received: October 4, 2012

Revised: November 15, 2012

Accepted: January 6, 2013

Published: February 14, 2013

## REFERENCES

- Alami, R., Fan, Y., Pack, S., Sonbuchner, T.M., Besse, A., Lin, Q., Greally, J.M., Skoultchi, A.I., and Bouhassira, E.E. (2003). Mammalian linker-histone subtypes differentially affect gene expression in vivo. Proc. Natl. Acad. Sci. USA 100, 5920–5925.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B 57, 289–300.
- Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esguerra, R., Pflueger, D., Sougnez, C., et al. (2011). The genomic complexity of primary human prostate cancer. Nature 470, 214–220.
- Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature 463, 899–905.
- Bos, J.L. (1989). ras oncogenes in human cancer: a review. Cancer Res. 49, 4682–4689.
- Brown, J.R., Hanna, M., Tesar, B., Werner, L., Pochet, N., Asara, J.M., Wang, Y.E., Dal Cin, P., Fernandes, S.M., Thompson, C., et al. (2012). Integrative genomic analysis implicates gain of PIK3CA at 3q26 and MYC at 8q24 in chronic lymphocytic leukemia. Clin. Cancer Res. 18, 3791–3802.
- Carter, S.L., Meyerson, M., and Getz, G. (2011). Accurate estimation of homologue-specific DNA concentration ratios in cancer samples allows long-range haplotyping. Nature Precedings, hdl:10101/npre.2011.6494.1, <http://hdl.handle.net/10101/npre.2011.6494.1>.
- Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. 30, 413–421.
- Chapman, M.A., Lawrence, M.S., Keats, J.J., Cibulskis, K., Sougnez, C., Schinzel, A.C., Harview, C.L., Brunet, J.P., Ahmann, G.J., Adli, M., et al. (2011). Initial genome sequencing and analysis of multiple myeloma. Nature 471, 467–472.
- Cheson, B.D., Pfistner, B., Juweid, M.E., Gascoyne, R.D., Specht, L., Horning, S.J., Coiffier, B., Fisher, R.I., Hagenbeek, A., Zucca, E., et al.; International Harmonization Project on Lymphoma (2007). Revised response criteria for malignant lymphoma. J. Clin. Oncol. 25, 579–586.
- Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotech. Published online February 10, 2013. <http://dx.doi.org/10.1038/nbt.2514>.
- CLL Trialists Collaborative Group. (1999). Chemotherapeutic options in chronic lymphocytic leukemia: a meta-analysis of the randomized trials. CLL Trialists' Collaborative Group. J. Natl. Cancer Inst. 91, 861–868.
- DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498.
- Deutsch, A.J., Aigelsreiter, A., Staber, P.B., Beham, A., Linkesch, W., Guelly, C., Brezinschek, R.I., Fruhwirth, M., Emberger, W., Buettner, M., et al. (2007). MALT lymphoma and extranodal diffuse large B-cell lymphoma are targeted by aberrant somatic hypermutation. Blood 109, 3500–3504.
- Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature 481, 506–510.
- Döhner, H. (2005). The use of molecular markers in selecting therapy for CLL. Clin. Adv. Hematol. Oncol. 3, 103–104.

- Döhner, H., Stilgenbauer, S., Benner, A., Leupolt, E., Kröber, A., Bullinger, L., Döhner, K., Bentz, M., and Lichter, P. (2000). Genomic aberrations and survival in chronic lymphocytic leukemia. *N. Engl. J. Med.* **343**, 1910–1916.
- Egan, J.B., Shi, C.X., Tembe, W., Christoforides, A., Kurdoglu, A., Sinari, S., Middha, S., Asmann, Y., Schmidt, J., Braggio, E., et al. (2012). Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides. *Blood* **120**, 1060–1066.
- Fabbri, G., Rasi, S., Rossi, D., Trifonov, V., Khiabanian, H., Ma, J., Grunn, A., Fangazio, M., Capello, D., Monti, S., et al. (2011). Analysis of the chronic lymphocytic leukemia coding genome: role of NOTCH1 mutational activation. *J. Exp. Med.* **208**, 1389–1401.
- Forbes, S.A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., Kok, C.Y., Jia, M., Ewing, R., Menzies, A., et al. (2010). COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res.* **38**(Database issue), D652–D657.
- Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183.
- Gerlinger, M., and Swanton, C. (2010). How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. *Br. J. Cancer* **103**, 1139–1143.
- Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892.
- Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., and Beerenswinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* **3**, 811.
- Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189.
- Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. *Nature* **481**, 306–313.
- Grossmann, V., Tiacci, E., Holmes, A.B., Kohlmann, A., Martelli, M.P., Kern, W., Spanhol-Rossetto, A., Klein, H.U., Dugas, M., Schindela, S., et al. (2011). Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype. *Blood* **118**, 6153–6163.
- Grubor, V., Krasnitz, A., Troge, J.E., Meth, J.L., Lakshmi, B., Kendall, J.T., Yamrom, B., Alex, G., Pai, D., Navin, N., et al. (2009). Novel genomic alterations and clonal evolution in chronic lymphocytic leukemia revealed by representational oligonucleotide microarray analysis (ROMA). *Blood* **113**, 1294–1303.
- Hallek, M., Cheson, B.D., Catovsky, D., Caligaris-Cappio, F., Dighiero, G., Döhner, H., Hillmen, P., Keating, M.J.,Montserrat, E., Rai, K.R., and Kipps, T.J.; International Workshop on Chronic Lymphocytic Leukemia (2008). Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines. *Blood* **111**, 5446–5456.
- Hosgood, H.D., 3rd, Baris, D., Zhang, Y., Berndt, S.I., Menashe, I., Morton, L.M., Lee, K.M., Yeager, M., Zahm, S.H., Chanock, S., et al. (2009). Genetic variation in cell cycle and apoptosis related genes and multiple myeloma risk. *Leuk. Res.* **33**, 1609–1614.
- Jablonski, D. (2001). Lessons from the past: evolutionary impacts of mass extinctions. *Proc. Natl. Acad. Sci. USA* **98**, 5393–5398.
- Klein, U., Lia, M., Crespo, M., Siegel, R., Shen, Q., Mo, T., Ambesi-Impiombato, A., Califano, A., Migliazza, A., Bhagat, G., and Dalla-Favera, R. (2010). The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia. *Cancer Cell* **17**, 28–40.
- Lohr, J.G., Stojanov, P., Lawrence, M.S., Auclair, D., Chapuy, B., Sougnez, C., Cruz-Gordillo, P., Knobelspitz, B., Asmann, Y.W., Slager, S.L., et al. (2012). Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl. Acad. Sci. USA* **109**, 3879–3884.
- Mäkinen, N., Mehine, M., Tolvanen, J., Kaasinen, E., Li, Y., Lehtonen, H.J., Gentile, M., Yan, J., Enge, M., Taipale, M., et al. (2011). MED12, the mediator complex subunit 12 gene, is mutated at high frequency in uterine leiomyomas. *Science* **334**, 252–255.
- Maley, C.C., Galipeau, P.C., Finley, J.C., Wongsurawat, V.J., Li, X., Sanchez, C.A., Paulson, T.G., Blount, P.L., Risques, R.A., Rabinovitch, P.S., and Reid, B.J. (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat. Genet.* **38**, 468–473.
- Maréchal, Y., Quément, S., Polizzi, S., Pouillon, V., and Schurmans, S. (2011). Inositol 1,4,5-trisphosphate 3-kinase B controls survival and prevents anergy in B cells. *Immunobiology* **216**, 103–109.
- Masica, D.L., and Karchin, R. (2011). Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res.* **71**, 4550–4561.
- McCarthy, H., Wierda, W.G., Barron, L.L., Cromwell, C.C., Wang, J., Coombes, K.R., Rangel, R., Elenitoba-Johnson, K.S., Keating, M.J., and Abruzzo, L.V. (2003). High expression of activation-induced cytidine deaminase (AID) and splice variants is a distinctive feature of poor-prognosis chronic lymphocytic leukemia. *Blood* **101**, 4903–4908.
- Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41.
- Mullighan, C.G., Phillips, L.A., Su, X., Ma, J., Miller, C.B., Shurtleff, S.A., and Downing, J.R. (2008). Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science* **322**, 1377–1380.
- Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94.
- Ngo, V.N., Young, R.M., Schmitz, R., Jhavar, S., Xiao, W., Lim, K.H., Kohlhammer, H., Xu, W., Yang, Y., Zhao, H., et al. (2010). Oncogenically active MYD88 mutations in human lymphoma. *Nature* **470**, 115–119.
- Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al.; Breast Cancer Working Group of the International Cancer Genome Consortium (2012). The life history of 21 breast cancers. *Cell* **149**, 994–1007.
- Nowak, M.A., and Sigmund, K. (2004). Evolutionary dynamics of biological games. *Science* **303**, 793–799.
- Puente, X.S., Pinyol, M., Quesada, V., Conde, L., Ordóñez, G.R., Villamor, N., Escaramis, G., Jares, P., Beà, S., González-Díaz, M., et al. (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475**, 101–105.
- Quesada, V., Conde, L., Villamor, N., Ordóñez, G.R., Jares, P., Bassaganyas, L., Ramsay, A.J., Beà, S., Pinyol, M., Martínez-Trillo, A., et al. (2012). Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 47–52.
- Rice, G.I., Bond, J., Asipu, A., Brunette, R.L., Manfield, I.W., Carr, I.M., Fuller, J.C., Jackson, R.M., Lamb, T., Briggs, T.A., et al. (2009). Mutations involved in Aicardi-Goutières syndrome implicate SAMHD1 as regulator of the innate immune response. *Nat. Genet.* **41**, 829–832.
- Schuh, A., Becq, J., Humphray, S., Alexa, A., Burns, A., Clifford, R., Feller, S.M., Grocock, R., Henderson, S., Khrebtukova, I., et al. (2012). Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* **120**, 4191–4196.
- Shah, S.P., Roth, A., Goya, R., Olumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., et al. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399.

- Shanafelt, T.D., Hanson, C., Dewald, G.W., Witzig, T.E., LaPlant, B., Abramzon, J., Jelinek, D.F., and Kay, N.E. (2008). Karyotype evolution on fluorescent *in situ* hybridization analysis is associated with short survival in patients with chronic lymphocytic leukemia and is related to CD49d expression. *J. Clin. Oncol.* 26, e5–e6.
- Snuderl, M., Fazlollahi, L., Le, L.P., Nitta, M., Zhelyazkova, B.H., Davidson, C.J., Akhavanfard, S., Cahill, D.P., Aldape, K.D., Betensky, R.A., et al. (2011). Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma. *Cancer Cell* 20, 810–817.
- Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R., et al.; Oslo Breast Cancer Consortium (OSBREAC) (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature* 486, 400–404.
- Stilgenbauer, S., Sander, S., Bullinger, L., Benner, A., Leupolt, E., Winkler, D., Kröber, A., Kienle, D., Lichter, P., and Döhner, H. (2007). Clonal evolution in chronic lymphocytic leukemia: acquisition of high-risk genomic aberrations associated with unmutated VH, resistance to therapy, and short survival. *Haematologica* 92, 1242–1245.
- Unoki, M., and Nakamura, Y. (2003). EGR2 induces apoptosis in various cancer cell lines by direct transactivation of BNIP3L and BAK. *Oncogene* 22, 2172–2185.
- Vincent, T.L., and Gatenby, R.A. (2008). An evolutionary model for initiation, promotion, and progression in carcinogenesis. *Int. J. Oncol.* 32, 729–737.
- Walter, M.J., Shen, D., Ding, L., Shao, J., Koboldt, D.C., Chen, K., Larson, D.E., McLellan, M.D., Dooling, D., Abbott, R., et al. (2012). Clonal architecture of secondary acute myeloid leukemia. *N. Engl. J. Med.* 366, 1090–1098.
- Wang, L., Lawrence, M.S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K., Werner, L., Sivachenko, A., DeLuca, D.S., Zhang, L., et al. (2011). SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N. Engl. J. Med.* 365, 2497–2506.
- Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., et al. (2012). The origin and evolution of mutations in acute myeloid leukemia. *Cell* 150, 264–278.
- Wu, C.J. (2012). CLL clonal heterogeneity: an ecology of competing subpopulations. *Blood* 120, 4117–4118.

## EXTENDED EXPERIMENTAL PROCEDURES

### Human Samples

Heparinized blood, skin biopsies and saliva were obtained from patients enrolled on clinical research protocols at the Dana-Farber Harvard Cancer Center (DFHCC) approved by the DFHCC Human Subjects Protection Committee. The diagnosis of CLL according to WHO criteria was confirmed in all cases by flow cytometry, or by lymph node or bone marrow biopsy. Peripheral blood mononuclear cells (PBMC) from normal donors and patients were isolated by Ficoll/Hypaque density gradient centrifugation. Mononuclear cells were used fresh or cryopreserved with FBS 10% DMSO and stored in vapor-phase liquid nitrogen until the time of analysis. Primary skin fibroblast lines were generated from skin punch biopsies as previously described (Wang et al., 2011). The patients included in the cohort represent the broad clinical spectrum of CLL (Table S1).

### Established CLL Prognostic Factor Analysis

Immunoglobulin heavy-chain variable (*IGHV*) homology (unmutated was defined as greater than or equal to 98% homology to the closest germline match) and ZAP-70 expression (high risk defined as >20% positive) were determined (Rassenti et al., 2008). Cytogenetics were evaluated by FISH for the most common CLL abnormalities (del(13q), trisomy 12, del(11q), del(17p), rearrangements of chromosome 14) (all probes from Vysis, Des Plaines, IL, performed at the Brigham and Women's Hospital Cytogenetics Laboratory, Boston MA). Samples were scored positive for a chromosomal aberration based on consensus cytogenetic scoring (Smoley et al., 2010).

### DNA Quality Control

We used standard Broad Institute protocols as recently described (Berger et al., 2011; Chapman et al., 2011). Tumor and normal DNA concentration were measured using PicoGreen® dsDNA Quantitation Reagent (Invitrogen, Carlsbad, CA). A minimum DNA concentration of 60 ng/ $\mu$ l was required for sequencing. In select cases where concentration was < 60 ng/ $\mu$ l, ethanol precipitation and re-suspension was performed. Gel electrophoresis confirmed that the large majority of DNA was high molecular weight. All Illumina sequencing libraries were created with the native DNA. The identities of all tumor and normal DNA samples (native and WGA product) were confirmed by mass spectrometric fingerprint genotyping of 24 common SNPs (Sequenom, San Diego, CA).

### Whole-Exome DNA Sequencing

Informed consent on DFCI IRB-approved protocols for whole-exome sequencing of patients' samples was obtained prior to the initiation of sequencing studies. DNA was extracted from blood or marrow-derived lymphocytes (tumor) and saliva, fibroblasts or granulocytes (normal), as previously described (Wang et al., 2011). Libraries for WES were constructed and sequenced on either an Illumina HiSeq 2000 or Illumina GA-IIx using 76 bp paired-end reads. Details of whole-exome library construction have been detailed elsewhere (Fisher et al., 2011). Standard quality control metrics, including error rates, percentage passing filter reads, and total Gb produced, were used to characterize process performance before downstream analysis. Average exome coverage depth was 112 $\times$ /107 $\times$  for tumor/germline. The Illumina pipeline generates data files (BAM files) that contain the reads together with quality parameters. Of the 160 CLL samples reported in the current manuscript, 82 were included in a previous study (Wang et al., 2011). 340 CLL and germline samples were sequenced overall. These include 160 CLL and matched germline DNA samples as well as time point 2 samples for 17 of 160 CLLs, and an additional sample pair and germline for a longitudinal sample pair not included in the 160 cohort (CLL020).

### Identification of Somatic Mutations

Output from Illumina software was processed by the "Picard" data processing pipeline to yield BAM files containing aligned reads (via MAQ, to the NCBI Human Reference Genome Build hg18) with well-calibrated quality scores (Chapman et al., 2011; DePristo et al., 2011). For 51 of the 160 CLL samples included in the analysis, sequencing was performed on capture libraries generated from whole-genome-amplified (WGA) samples. For those samples, 100 ng inputs of samples were whole-genome amplified with the QIAGEN REPLI-g Midi Kit (Valencia, CA).

From the sequencing data, somatic alterations were identified using a set of tools within the "Firehose" pipeline, developed at the Broad Institute (<http://www.broadinstitute.org/cancer/cga>). The details of our sequencing data processing have been described elsewhere (Berger et al., 2011; Chapman et al., 2011). Somatic single nucleotide variations (sSNVs) were detected using MuTect [V119, <http://www.broadinstitute.org/cancer/cga/mutect>, (Cibulskis et al., 2013)]; somatic small insertions and deletions (indels) were detected using Indelocator [v61, <http://www.broadinstitute.org/cancer/cga/indelocator>, (Wang et al., 2011)]. All mutations identified in longitudinal samples were confirmed by manual inspection of the sequencing data (Robinson et al., 2011). An estimated contamination threshold of 5% was used for all samples based on the highest contamination values seen in a formal contamination analysis done with ContEst based on matched SNP arrays (Cibulskis et al., 2011). Ig loci mutations were not included in this analysis. All somatic mutations detected in the 160 CLL samples are listed in Table S2A. WES data is deposited in dbGaP (phs000435.v1.p1).

### Significance Analysis for Recurrently Mutated Genes

The prioritization of somatic mutations in terms of conferring selective advantage was done with the statistical method MutSig2.0 (Lohr et al., 2012). In short, the algorithm takes an aggregated list of mutations and tries to detect genes that are affected

more than expected by chance, as those likely reflect positive selection (i.e driver events). There are two main components to MutSig2.0:

The first component attempts to model the background mutation rate for each gene, while taking into account various different factors. Namely, it takes into account the fact that the background mutation rate may vary depending on the base context and base change of the mutation, as well as the fact that the background rate of a gene can also vary across different patients. Given these factors and the background model, it uses convolutions of binomial distributions to calculate a *P* value, which represents the probability that we obtain the observed configuration of mutations, or a more significant one.

The second component of the algorithm focuses on the positional configuration of mutations and their sequence conservation (Lohr et al., 2012). For each gene, the algorithm permutes the mutations preserving their tri-nucleotide context, and for each permutation calculates two metrics: one that measures the degree of clustering into hotspots along the coding length of the gene, and one that measures the average conservation of mutations in the gene. These two null models are then combined into a joint distribution, which is used to calculate a *P* value that reflects the probability by chance that we can obtain by chance the observed mutational degree of clustering and conservation, or a more significant outcome.

The two *P* values that are produced by the two components are then combined using Fisher-Combine (Fisher, 1932) which yields a final *P* value which is used to sort the genes by degree of mutational significance. This is subsequently corrected for multi-hypothesis using the Benjamini Hochberg procedure.

### **Genome-wide Copy Number Analysis**

Genome-wide copy number profiles of 111 CLL samples and their patient-matched germline DNA were obtained using the Genome-wide Human SNP Array 6.0 (Affymetrix), according to the manufacturer's protocol (Genetic Analysis Platform, Broad Institute, Cambridge MA). SNP array data were deposited in dbGaP (phs000435.v1.p1). Allele-specific analysis also allowed for the identification of copy neutral LOH events as well as quantification of the homologous copy-ratios (HSCSs) [HAPSEG(Carter et al., 2011)]. Significant recurrent chromosomal abnormalities were identified using the GISTIC2.0 algorithm ((Mermel et al., 2011), v87). Regions with germline copy number variants were excluded from the analysis.

For CLL samples with no available SNP arrays (38/160), sCNAs were estimated directly from the WES data, based on the ratio of CLL sample read-depth to the average read-depth observed in normal samples for that region. 11/160 samples were excluded from this analysis due to inability to obtain copy number information from the WES data. See Figure 2A for outline of sample processing.

### **Validation Deep Sequencing**

Validation targeted resequencing of 252 selected somatic mutations sSNVs was performed using microfluidic PCR. Target specific primers with Fluidigm-compatible tails were designed to flank sites of interest and produce amplicons of 200 +/- 20bp. Molecular barcoded, Illumina-compatible oligonucleotides, containing sequences complementary to the primer tails were added to the Fluidigm Access Array chip (San Francisco, CA) in the same well as the genomic DNA samples (20 - 50 ng of input) such that all amplicons for a given genomic sample shared the same index, and PCR was performed according to the manufacturer's recommendations. Indexed libraries were recovered for each sample in a single collection well on the Fluidigm chip, quantified using picogreen and then normalized for uniformity across libraries. Resulting normalized libraries were loaded on a MiSeq instrument (Illumina) and sequenced using paired end 150bp sequencing reads. 95.2% of called sSNVs were detected in the validation experiment (Table S5A). For 91.8% of the mutations, the allelic fraction estimates were concordant (with the discordant events enriched in sites of lower WES coverage).

### **RNA Sequencing and dUTP Library Construction**

5 µg of total RNA was poly-A selected using oligo-dT beads to extract the desired mRNA. The purified mRNA is treated with DNase, and cleaned up using SPRI (Solid Phase Reversible Immobilization) beads according to the manufacturers' protocol. Selected Poly-A RNA was then fragmented into ~450 bp fragments in an acetate buffer at high heat. Fragmented RNA was cleaned with SPRI and primed with random hexamers before first strand cDNA synthesis. The first strand was reverse transcribed off the RNA template in the presence of Actinomycin D to prevent hairpinning and purified using SPRI beads. The RNA in the RNA-DNA complex was then digested using RNase H. The second strand was next synthesized with a dNTP mixture in which dTTPs had been replaced with dUTPs. After another SPRI bead purification, the resultant cDNA was processed using Illumina library construction according to manufacturers protocol (end repair, phosphorylation, adenylation, and adaptor ligation with indexed adaptors). SPRI-based size selection was performed to remove adaptor dimers present in the newly constructed cDNA library. Libraries were then treated with Uracil-Specific Excision Reagent (USER) to nick the second strand at every incorporated Uracil (dUTP). Subsequently, libraries were enriched with 8 cycles of PCR using the entire volume of sample as template. After enrichment, the library is quantified using pico green, and the fragment size is measured using the Agilent Bioanalyzer according to manufactures protocol. Samples were pooled and sequenced using either 76 or 101bp paired end reads.

### **RNASEq Data Analysis**

RNaseq BAMs were aligned to the hg18 genome using the TopHat suite. Each somatic base substitution detected by WES was compared to reads at the same location in RNaseq. Based on the number of alternate and reference reads, a power calculation

was obtained with beta-binomial distribution (power threshold used was greater than 80%). Mutation calls were deemed validated if 2 or greater alternate allele reads were observed in RNA-Seq at the site, as long as RNaseq was powered to detect an event at the specified location.

### FACS Validation of Ploidy Estimates with ABSOLUTE

Consistent with published studies of CLL (Brown et al., 2012; Edelmann et al., 2012), ABSOLUTE measured all CLL samples to be near diploid (Table S2B; median - 2, range 1.95–2.1). We confirmed the measurements using a standard assay for measuring DNA content. For this analysis, peripheral blood mononuclear cells from normal volunteers and CLL patients and cell lines are first stained with anti-CD5 FITC and anti-CD19 PE antibodies in a PBS buffer containing 1% BSA for 30 min on ice. After extensive washes, the cells were then stained with a PBS buffer contained 1% BSA, 0.03% saponin (Sigma) and 250ug/ml 7-AAD (Invitrogen) for 1 hr on ice, followed by analysis on a Beckman Coulter FC500 machine (Figure S3A).

### Estimation of Mutation Cancer Cell Fraction Using ABSOLUTE

We used the ABSOLUTE algorithm to calculate the purity, ploidy, and absolute DNA copy-numbers of each sample (Carter et al., 2012). Modifications were made to the algorithm, which are implemented in version 1.05 of the software, available for download at <http://www.broadinstitute.org/cancer/cga/absolute>. Specifically, we added the ability to determine sample purity from sSNVs alone, in samples where no sCNAs are present (the ploidy of such samples is 2N). In addition, estimates of sample purity and absolute copy-numbers are used to compute distributions over cancer cell fraction (CCF) values of each sSNV, as described (Experimental Procedures), and for sCNAs (described below).

The current implementation of ABSOLUTE does not automatically correct for sCNA subclonality when computing CCF distributions of sSNVs (this is an area of ongoing development). Fortunately, the few sCNAs that occurred in our CLL samples were predominantly clonal. Manual corrections were made for CLL driver sSNVs occurring at site of subclonal sCNAs (5 TP53 sSNVs and 1 ATM sSNV), based on the sample purity, allelic fraction and the copy ratio of the matching sCNA.

Each sSNV was classified as clonal or subclonal based on the probability that the CCF exceeded 0.95. A probability threshold of 0.5 was used throughout the manuscript. However, as the histogram in Figure 2A shows, the distribution of events around the threshold was observed to be fairly uniform and results were not significantly affected across a range of thresholds. For example, the results of our analyses were unchanged when we altered our definition of clonal mutations to be  $\text{Pr}(\text{CCF} > 0.95) > 0.75$ , and subclonal when  $\text{Pr}(\text{CCF} > 0.95) < 0.25$ , leaving uncertain mutations unclassified. Using these thresholds, CLLs with mutated *IGHV* and age were associated with a higher number of clonal mutations (*P* values of 0.05 and < 0.0001, respectively). CLLs treated prior to sample collection had a higher number of subclonal mutations (*p* = 0.01) and the subclonal set was enriched with putative drivers (*p* = 0.0019). Importantly, the results of the clinical analysis also remained unchanged. FFS\_Rx was shorter in samples in which a subclonal driver was detected (*p* = 0.007) and regression models examining known poor prognostic indicators in CLL yielded an adjusted *P* value of 0.009.

One of the recurrent CLL cancer genes, *NOTCH1*, had 15 mutations, 14 of which were the identical canonical 2 base-pair deletions. Unlike sSNVs, the observed allelic fractions of indel events were not modeled as binomial sampling of reference and alternate sequence reads according to their true concentration in the sample (Carter et al., 2012). This was due to biases affecting the alignment of the short sequencing reads, which generally favor reference over alternate alleles. To measure the magnitude of this effect, we examined the allelic fraction (AF) of 514 germline 2bp deletions called in 4 normal germline WES samples. We observed that the distribution (data not shown) of allelic-fractions for heterozygous events was peaked at 0.41, as opposed to the expected mode of 0.5, with nearly all AFs between 0.3 to 0.6. Therefore, the bias factor toward reference is peaked at 0.82 but may range from 0.6 to 1 (unlikely to be greater than 1). CCF distributions for the 14 somatic indels in *NOTCH1* were calculated using bias factors of 1.0 (no bias), 0.82 (bias point-estimate), and 0.6 (worst case observed). Reassuringly, the classification of *NOTCH1* indels as clonal or subclonal was highly robust and was essentially the same using the three values—only a single case (CLL155) was ambiguous and was classified as subclonal using 1.0 and 0.82, and clonal using 0.6. Taking a conservative approach, not classifying a mutation as sub-clonal unless there is clear evidence for it, we decided to call this event as clonal for downstream analysis.

Estimation of CCF values for subclonal sCNAs is implemented (ABSOLUTEv1.05) in a manner analogous to the procedure for sSNVs (Experimental Procedures), although the transformation is more complex, due to the need for assumptions of the subclonal structure and the error model of microarray based copy-number data. Segmental sCNAs are defined as subclonal based on the mixture model used in ABSOLUTE (Carter et al., 2012). Let the functions  $h(x)$  and  $h'(x)$  denote a variance stabilizing transformation and its derivative, respectively. For SNP microarray data, these are defined as:  $h(x) = \sinh^{-1}(bx)$ , where  $b = (e^{\sigma_{\eta}^2} - 1)^{\frac{1}{2}}/\sigma_{\epsilon}$ , and  $h'(x) = b/(1 + (bx)^2)^{\frac{1}{2}}$  (Huber et al., 2002). The values  $\sigma_{\epsilon}$  and  $\sigma_{\eta}$  denote additive and multiplicative noise scales, respectively, for the microarray hybridization being analyzed; these are estimated by HAPSEG (Carter et al., 2011). The calibrated probe-level microarray data become approximately normal under this transformation, which is used by HAPSEG to estimate the segmental allelic copy-ratios  $r_i$  and the posterior standard deviation of their mean (under the transformation),  $\sigma_i$  (Carter et al., 2011). An additional parameter  $\sigma_H$  is estimated by ABSOLUTE (Carter et al., 2012), which represents additional sample-level variance corresponding to regional biases not captured in the probe-level model.

For a subclonal segment  $i$ , let  $q_c$  denote the absolute copy number in the unaffected cells, and  $q_s$  denote the absolute copy number in the altered cells. Both of these values are unknown but we used a simplifying assumption that the difference between  $q_c$  and  $q_s$  is

one copy with  $q_c$  being closer to the modal copy-number. Therefore, for subclonal deletions (copy ratios below the ratio of modal copy number),  $q_s$  was set to the nearest copy number below the measured value, and  $q_c = q_s+1$ . For subclonal gains (ratios above the modal number),  $q_s$  was set to the nearest copy number above the measured value, and  $q_c = q_s-1$ . Because the CLL genomes analyzed here were universally near diploid, this was nearly equivalent to assuming that subclonal deletions had  $q_s = 0$  in the affected cells and gains  $q_s = 2$ , with  $q_c = 1$  in both cases (in allelic units). However, we note that these assumptions would not be strictly correct in genomes after doubling, or in cases of high-level amplification. In these cases, calculation of posterior CCF distributions will require integration over  $q_s$  and  $q_c$ , averaging over the set of plausible subclonal genomic configurations.

Let  $r_c$  and  $r_s$  be the theoretical copy ratio values corresponding to  $q_c$  and  $q_s$  (accounting for sample purity, ploidy, and the modeled attenuation rate of the microarray (Carter et al., 2011; Carter et al., 2012)). Let  $d = r_s - r_c$ , then, for CCF  $c$ , let  $r_x(c) = dc + r_c$ . Then  $P(c) \propto \mathcal{N}(h(r_x(c))|h(r_i), (\sigma_i + \sigma_H)^2)|h'(r_x(c))|$ . The absolute value of the derivative is required due to the change of coordinates from  $x$  to  $h(x)$ . The distribution over CCF is obtained by calculating these values over a regular grid of 100  $c$  values and normalizing. We note that, when copy numbers are estimated directly from sequencing data, the calculation is simpler, as there is no attenuation effect and  $h(x) = x$ . These calculations were used to generate the 95% confidence intervals on the CCF of subclonal driver sCNAs shown in Figure 4 and Figures S5A and S5B.

### Cancer Gene Census List and Conservation Annotations

Conservation of a specific mutated site was adapted from UCSC conservation score track. A scale of 0-100 was linearly converted from the -6 to 6 scale used in the phastCons track (Siepel et al., 2005). To confirm that driver mutations are more likely to occur in conserved sites, we quantified the conservation in the COSMIC database (Forbes et al., 2010) hotspots and compared it to non-COSMIC hotspots coding location. We matched conservation information for 5085 sites that had greater than 3 exact hits reported in mutations deposited in the COSMIC database, and compared it to conservation found for a set of non-overlapping 5085 randomly sampled coding sites. The conservation was higher in the COSMIC sites than in the non-COSMIC coding sites set (mean conservation 82.39 and 62.15, respectively,  $p < 1e-50$ ). We noted that the distribution of events was not uniform, and nearly one half of COSMIC hotspots had a conservation measure greater than 95 (49.65%, compared to 15.5% in the non-COSMIC set,  $p < 1e-50$ ). For our calculations, we used a cut off of > 95 to designate conserved sites likely to contain higher proportion of cancer drivers. We complemented the analysis for putative driver event enrichment by matching the altered genes to the Cancer Gene Census (Futreal et al., 2004).

### Clustering Analysis of sSNVs in 18 CLL Sample Pairs

In order to better resolve the true cancer cell fraction (CCF) of sSNVs detected in longitudinal samples, we employed a previously described Bayesian clustering procedure (Escobar and West, 1995). This approach exploits the assumption that the observed subclonal sSNV CCF values were sampled from a smaller number of subclonal cell populations (subclones). All remaining uncertainty (including the exact number of clusters) was integrated out using a mixture of Dirichlet processes, which was fit using a Gibbs sampling approach, building on a previously described framework (Escobar and West, 1995).

The inputs to this procedure are the posterior CCF distributions for each sSNV being considered. We note that the CCF distributions for sCNAs could be added into the model, however we did not attempt this in the present study. CCF distributions are represented as 100-bin histograms over the unit interval; the two-dimensional CCF distributions used for the 2D clustering of longitudinal samples were obtained as the outer product of the matched histogram pairs for each mutation, resulting in 10,000-bin histograms (Figure S5). We note that the use of histograms to represent posterior distributions on CCF, although computationally less efficient than parametric forms, has the advantage that CCFs of different mutation classes may be easily combined in the model, even though their posteriors may have very different forms. We also note that the algorithm implementation is identical for the single sample and paired (longitudinal) sample cases, although only the latter was used in the present study.

At each iteration of the Gibbs sampler, each mutation is assigned to a unique cluster and the posterior CCF distribution of each cluster is computed using Bayes' rule, as opposed to drawing a sample from the posterior (a uniform prior on CCF from 0.01 to 1 is used). When considering the probability of a mutation to join an existing cluster, the likelihood calculation of the mutation arising from the cluster is integrated over the uncertainty in the cluster CCF. This allows for rapid convergence of the Gibbs sampler to its stationary distribution, which was typically obtained in fewer than 100 iterations for the analysis presented in this study. We ran the Gibbs sampler for 1,000 iterations, of which the first 500 were discarded before summarization.

Because of the small number of clonal mutations in some WES samples, we make an additional modification to the standard Dirichlet process model by adding a fixed clonal cluster that persists even if no mutation is assigned to it. This reflects our prior knowledge that clonal mutations must exist, even if they are the minority of detected mutations. For the samples analyzed here, this modification had very little effect.

A key aspect of implementing the Dirichlet process model on WES data sets is re-parameterization of prior distributions on the number of subclones  $k$  as priors on the concentration parameter  $\alpha$  of the Dirichlet process model. Importantly, this must take into account the number of mutations  $N$  input to the model, as the effect of  $\alpha$  on  $k$  is strongly dependent on  $N$  (Escobar and West, 1995). We accomplish this by constructing a map from a regular grid over  $\alpha$  to expected values of  $k$ , given  $N$ , using the fact that:  $P(k|\alpha, N) = c_N(k)N!/\alpha^k(\Gamma(\alpha)/\Gamma(\alpha+N))$  (Escobar and West, 1995), where the  $c_N(k)$  factors correspond to the unsigned Stirling numbers of the first kind. With this map in hand, we perform an optimization procedure to find parameters  $a$  and  $b$  of a prior Gamma

distribution over  $\alpha$  resulting in the minimal Kullback-Leibler divergence with the specified prior over  $k$  (the divergence was computed numerically on the histograms). Once the prior over  $\alpha$  has been represented as a Gamma distribution, learning about  $\alpha$  (and therefore  $k$ ) from the data can be directly incorporated into the Gibbs sampling procedure, resulting in a continuous mixture of Dirichlet processes (Escobar and West, 1995). This allows consistent parameterization of prior knowledge (or lack thereof) on the number of subclonal populations in the face of vastly different numbers of input mutations, which is necessary for making consistent inferences across differing data sets (e.g., WES versus WGS). We note that taking uncertainty about  $\alpha$  into account is necessary for inferences on the number of subclonal populations to be strictly valid, as implementations with fixed values of  $\alpha$  result in an implicit prior over  $k$  that depends upon  $N$  (this is especially important for smaller values of  $N$ ). For the application presented in this study (Figure 4), we specified a weak prior on  $k$  using a negative binomial distribution with  $r = 10$ ,  $\mu = 2$  (these values favored 1-10 clusters). We note that these are the only two user-specified parameters of the clustering analysis.

Upon termination of the Gibbs sampler, we summarized the posterior probability over the CCF of each sSNV by averaging the posterior cluster distribution for all clusters to which the sSNV was assigned during sampling. This allowed shrinkage of the CCF probability distributions (as shown in Figure 4; pre-clustering results are shown in Figure S5A and S5B), without having to choose an exact number of subclonal clusters. Note that the 18 longitudinal sample pairs contain 1 CLL sample pair not initially included in the 160 CLLs (CLL020).

### Gene Expression Profiling

Total RNA was isolated from viably frozen PBMCs or B cells from CLL patients that were followed longitudinally (Midi kit; QIAGEN, Valencia CA), and hybridized to the U133Plus 2.0 array (Affymetrix, Santa Cruz, CA) at the DFCI Microarray Core Facility. All expression profiles were processed using RMA, implemented by the PreprocessData set module in GenePattern (<http://www.broadinstitute.org/cancer/software/genepattern/>) (Irizarry et al., 2003; Reich et al., 2006). 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 (Johnson et al., 2007) (Reich et al., 2006). Visualizations in GENE-E (<http://www.broadinstitute.org/cancer/software/GENE-E>) were based on logarithmic transformation (log2) of the data and centering each gene (zero mean). These data can be accessed at <http://www.ncbi.nlm.nih.gov/geo/info/linking.html> with accession number GSE37168.

### RNA Pyrosequencing for Mutation Confirmation

Quantitative targeted sequencing to detect somatic mutation within cDNA was performed, as previously described (Armistead et al., 2008). In brief, biotinylated amplicons generated from PCR of the regions of transcript surrounding the mutation of interest were generated. Immobilized biotinylated single-stranded DNA fragments were isolated per manufacturer's protocol, and sequencing undertaken using an automated pyrosequencing instrument (PSQ96; QIAGEN, Valencia CA), followed by quantitative analysis using Pyrosequencing software (QIAGEN).

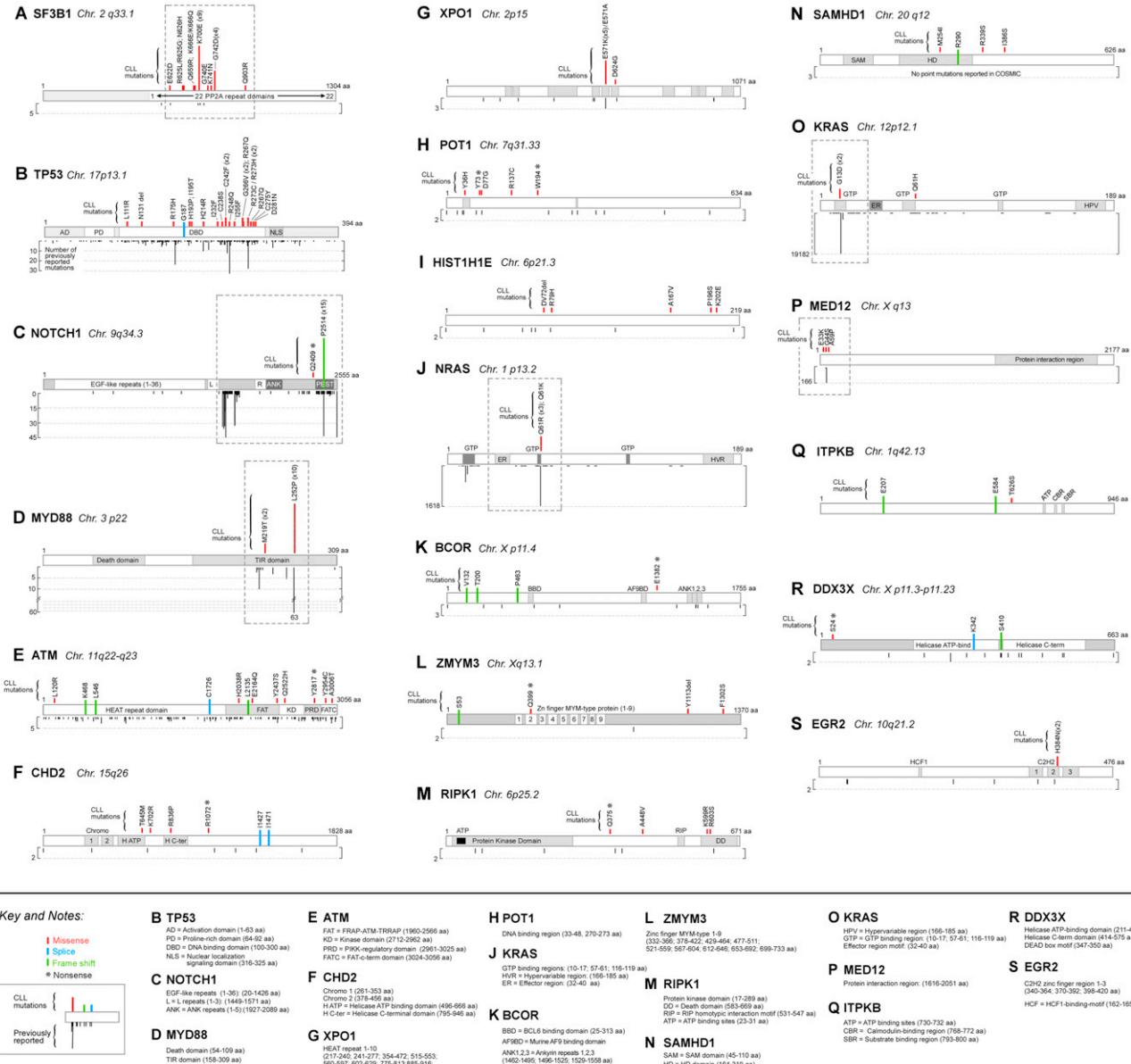
### Statistical Methods

Statistical analysis was performed with MATLAB (MathWorks, Natick, MA), R version 2.11.1 and SAS version 9.2 (SAS Institute, Cary, NC). Categorical variables were compared using the Fisher Exact test, and continuous variables were compared using the Student's t test, Wilcoxon rank sum test, or Kruskal Wallis test as appropriate; the association between two continuous variables was assessed by the Pearson correlation coefficient. The time from the date of sample to first therapy or death (failure-free survival from sample time or FFS\_Sample) was calculated as the time from sample to the time of the first treatment after the sample or death and was censored at the date of last contact. FFS\_Rx (failure-free survival from first treatment after sampling) was defined as the time to the 2<sup>nd</sup> treatment or death from the 1<sup>st</sup> treatment following sampling, was calculated only for those patients who had a 1<sup>st</sup> treatment after the sample and was censored at the date of last contact for those who had only one treatment after the sample. Time to event data were estimated by the method of Kaplan and Meier, and differences between groups were assessed using the log-rank test. Unadjusted and adjusted Cox modeling was performed to assess the impact of the presence of a subclonal driver and a driver irrespective of the CCF on FFS\_Sample and FFS\_Rx. A chi-square test with 1 degree of freedom and the -2 Log-likelihood statistic was used to test the prognostic independence of subclonal status in Cox modeling using a full model and one without subclonal status included. We also formally tested for non-proportionality of the hazards in Figure 6B First, we plotted the log(-log(survival)) versus log(time) for the two categories, and demonstrated that curves do not cross, which supports the fact that they are proportional. Second, we also tested for non-proportionality by including a time varying covariate for each variable in the model. None of these were significant indicating that the hazards are proportional. Models were adjusted for known prognostic factors for CLL treatment including the presence of a 17p deletion, the presence of a 11q deletion, *IGHV* mutational status, and prior treatment at the time of sample. Cytogenetic abnormalities were primarily assessed by FISH and if unknown, genomic data were included. For unknown *IGHV* mutational status an indicator was included in adjusted modeling and was not found to be significant. All P-values are two-sided and considered significant at the 0.05 level unless otherwise noted.

### SUPPLEMENTAL REFERENCES

Armistead, P.M., Mohseni, M., Gerwin, R., Walsh, E.C., Iravani, M., Chahardoli, B., Rostami, S., Zhang, W., Neuberg, D., Rioux, J., et al. (2008). Erythroid-lineage-specific engraftment in patients with severe hemoglobinopathy following allogeneic hematopoietic stem cell transplantation. *Exp. Hematol.* 36, 1205–1215.

- Cibulskis, K., McKenna, A., Fennell, T., Banks, E., DePristo, M., and Getz, G. (2011). ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* 27, 2601–2602.
- Edelmann, J., Holzmann, K., Miller, F., Winkler, D., Bühl, A., Zenz, T., Bullinger, L., Kühn, M.W., Gerhardinger, A., Bloehdorn, J., et al. (2012). High-resolution genomic profiling of chronic lymphocytic leukemia reveals new recurrent genomic alterations. *Blood* 120, 4783–4794.
- Escobar, M., and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.* 90, 577–588.
- Eskandarpour, M., Huang, F., Reeves, K.A., Clark, E., and Hansson, J. (2009). Oncogenic NRAS has multiple effects on the malignant phenotype of human melanoma cells cultured in vitro. *Int. J. Cancer* 124, 16–26.
- Fisher, R.A. (1932). Statistical methods for research workers, Fourth Edition (Edinburgh: Oliver and Boyd).
- Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T.M., Young, G., Fennell, T.J., Allen, A., Ambrogio, L., et al. (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* 12, R1.
- Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18(Suppl 1), S96–S104.
- Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264.
- Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118–127.
- Rassenti, L.Z., Jain, S., Keating, M.J., Wierda, W.G., Grever, M.R., Byrd, J.C., Kay, N.E., Brown, J.R., Gribben, J.G., Neuberg, D.S., et al. (2008). Relative value of ZAP-70, CD38, and immunoglobulin mutation status in predicting aggressive disease in chronic lymphocytic leukemia. *Blood* 112, 1923–1930.
- Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J.P. (2006). GenePattern 2.0. *Nat. Genet.* 38, 500–501.
- Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26.
- Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050.
- Smoley, S.A., Van Dyke, D.L., Kay, N.E., Heerema, N.A., Dell' Aquila, M.L., Dal Cin, P., Koduru, P., Aviram, A., Rassenti, L., Byrd, J.C., et al. (2010). Standardization of fluorescence *in situ* hybridization studies on chronic lymphocytic leukemia (CLL) blood and marrow cells by the CLL Research Consortium. *Cancer Genet. Cytogenet.* 203, 141–148.
- Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., et al. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* 478, 64–69.

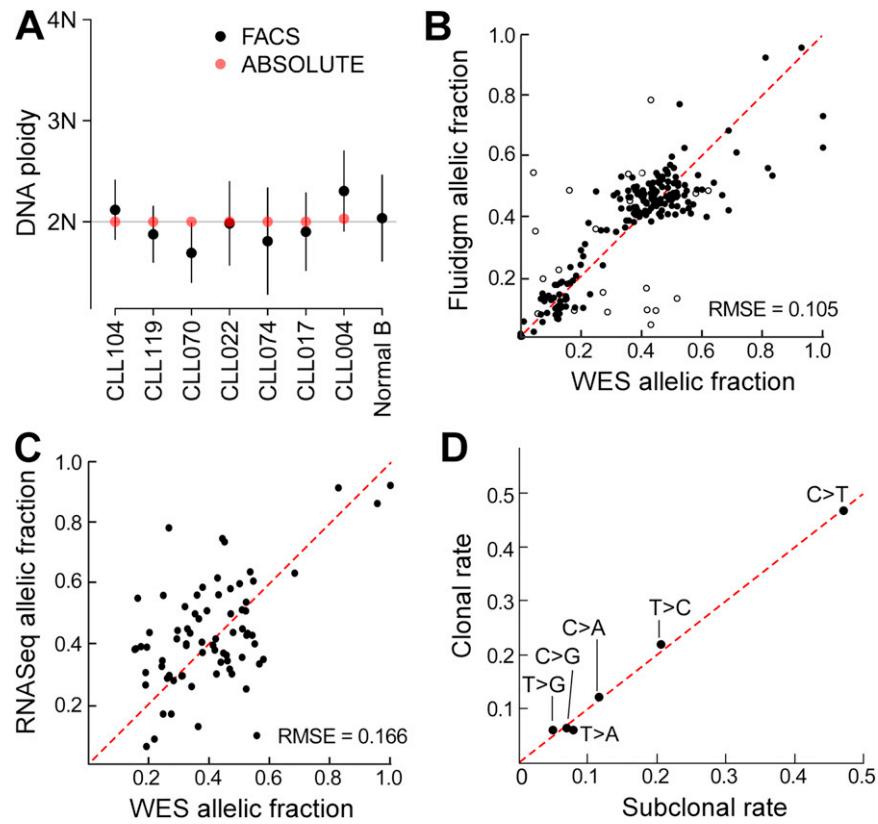
**Figure S1. Significantly Mutated Genes in 160 CLL Samples, Related to Figure 1**

(A–S) Type (missense, splice-site, nonsense) and location of mutations in the significantly mutated genes discovered among the 160 CLL samples (top) compared to previously reported mutations in literature or in the COSMIC database (v76) (bottom). Dashed boxes in (A), (C), (D), (J), (O), and (P) indicate mutations localizing to a discrete gene territory. Please refer to previous publication for mutation information for *FBXW7* (Wang et al., 2011).

		CHD2		RIPK1		Q375*	
		T645M	K702R				
Human	625	DDSLLYKTLIIDFSNHLRLLTGTPQLNSLKESELSSHHFIMPEKFPEWEDFEEDHGKGRENGYQSLHVLEPFLLRRVKKD	704	Human	319	QSLQDCCVANPSRSRSNATEQPSQSLHSSSQGLMGPVEESEWPSASLHEPQE-----NEPSLSGRQDEANWHIGER	390
Mouse	625	DDSLLYKTLIIDFSNHLRLLTGTPQLNSLKESELSSHHFIMPEKFPEWEDFEEDHGKGRENGYQSLHVLEPFLLRRVKKD	704	Mouse	320	FSLQDCCVSLPERSNS---EQPSQSLHSSSQGLMGPVEESEWPSASLHEPQE-----NDPSQAEQLEQSAYHATGIF	389
Zebrafish	620	DDSLLYKTLIIDFSNHLRLLTGTPQLNSLKESELSSHHFIMPEKFPEWEDFEEDHGKGRENGYQSLHVLEPFLLRRVKKD	699	Zebrafish	318	KLLNTNAL-1-PEDPSLCSRDSPPLRSD-----ISPVRAESEDLSLSPRSCEDSVLEADATPSCPMLLKELEQYYNNHKPGSR	393
		R836P					
Human	785	SLIRSSGKLILLDLMLLTLLRERCNWVLIFPQMVRLMDILALAYLT1KRVPPFLRDLGKIGEIRPKQHNAFQSDEDCPFL	664				
Mouse	785	SLIRSSGKLILLDLMLLTLLRERCNWVLIFPQMVRLMDILALAYLT1KRVPPFLRDLGKIGEIRPKQHNAFQSDEDCPFL	664				
Zebrafish	780	SLVNGGGKIVLVDLMLLTLLRERCNWVLIFPQMVRLMDILALAYLT1KRVPPFLRDLGKIGEIRPKQHNAFQSDEDCPFL	659				
		R1072*					
Human	1025	NFTAMTDEEKELERPKHWDNDEIPIKEPKRKEEEROKLEEYIYMLPRKRSRSTKKAQTNDSDTESESROAQRSSAEESE	1104				
Mouse	1025	NFTAMTDEEKELERPKHWDNDEIPIKEPKRKEEEROKLEEYIYMLPRKRSRSTKKAQTNDSDTESESROAQRSSAEESE	1104				
Zebrafish	1020	NFTMDEESTPDEKEGRWDEDEIPIPEPKRKEEEROKLEEYIYMLPRKRSRSTKKAQTNDSDTESESROAQRSSAEESE	1098				
		XPO1		E571K/E571A		A448V	
Human	536	NKAIATASINIMIVQQPPRFRAWWKFLKTVNKLPEMHTEDQDMACDTFFIKIAQKCRRHVFQVQVGRMPFIDEIL	615	Human	391	MDRQTKQQ---PRQMVAYNREERRRVSHDPPFAQRPYIENFONTEKGKTA---YSSAAASHONAVHOPSGLTSQPVLYQ	464
Mouse	536	NKAIATASINIMIVQQPPRFRAWWKFLKTVNKLPEMHTEDQDMACDTFFIKIAQKCRRHVFQVQVGRMPFIDEIL	615	Mouse	390	AEKQTRVY---PRQMVAYNREERRRVSHDPPFAQRPYIENFONTEKGKTA---DPSITTRGIAVQQLSPWATQ---TVW	460
Zebrafish	536	NKAIATASINIMIVQQPPRFRAWWKFLKTVNKLPEMHTEDQDMACDTFFIKIAQKCRRHVFQVQVGRMPFIDEIL	615	Zebrafish	394	IVDQSDDGSMGSPAQPS-----STDASSOSVSKVNTWPKAPESTKEELPRAAASPDSLHPE-EASQFSPVQPS	462
C. elegans	546	NKAIATASINIMIVQQPPRFRAWWKFLKTVNKLPEMHTEDQDMACDTFFIKIAQKCRRHVFQVQVGRMPFIDEIL	625				
S. pombe	537	NKAIATASINIMIVQQPPRFRAWWKFLKTVNKLPEMHTEDQDMACDTFFIKIAQKCRRHVFQVQVGRMPFIDEIL	616				
		D624G					
Human	616	NNINTTICDQIPOQVHTTFYEAVGYMIGAQTDQFVQEHLIERLIEYLMLLPNPWNWSDI1QIQQATKNVDILKDPETVQLQGSILKTN	695	Human	545	NYMEIGGTSSLDNSTNTNFKEERPAAKYQAIFFNTTSLTDKHDLP1IRENLGKHNWNCARKLQFTSQS1DEIDHEDYERDGL	624
Mouse	616	NNINTTICDQIPOQVHTTFYEAVGYMIGAQTDQFVQEHLIERLIEYLMLLPNPWNWSDI1QIQQATKNVDILKDPETVQLQGSILKTN	695	Mouse	534	NYMWDVSLN----QPNPNTCKEESTSHQJAIFFNTTSLTDKHDLP1IRENLGKHNWNCARKLQFTSQS1DEIDHEDYERDGL	609
Zebrafish	616	NNINTTICDQIPOQVHTTFYEAVGYMIGAQTDQFVQEHLIERLIEYLMLLPNPWNWSDI1QIQQATKNVDILKDPETVQLQGSILKTN	695	Zebrafish	530	NTMNLRIISDSSP5LSTGTGTSRYKELLMYESHTEESLELYEVNQANQVQARLGLSEIDVEEHDYERDGL	609
C. elegans	626	ENLTCIICLSDHAQVWVTEAAGHIIASQDNLHEDL1KMDIPHRNTWIIAQAASSTSVLEEVPMQVSMLVILKTN	705				
S. pombe	617	RELAKTTEDLTPQTETYEACGTYMSIQNPKRHLQERLIDFIMALPNQAMENIQAQNAQVQLEGPVTQKILANLTKTN	696				
		POT1		Y36H		K599R R603S	
Human	1	MGLVPAVNVVTTPLNQLEGGIVIVNVVGVVPPXPFLSKGPDYICSVVTTIVEQTNVNLTKLCLLGSNGVNEALIYIYDVR	80	Human	545	NYMEIGGTSSLDNSTNTNFKEERPAAKYQAIFFNTTSLTDKHDLP1IRENLGKHNWNCARKLQFTSQS1DEIDHEDYERDGL	624
Mouse	1	MGLVPAVNVVTTPLNQLEGGIVIVNVVGVVPPXPFLSKGPDYICSVVTTIVEQTNVNLTKLCLLGSNGVNEALIYIYDVR	80	Mouse	534	NYMWDVSLN----QPNPNTCKEESTSHQJAIFFNTTSLTDKHDLP1IRENLGKHNWNCARKLQFTSQS1DEIDHEDYERDGL	609
		R137C		Y373*	D77G	Z	Z
Human	116	PRTSSKTFNFTEDDEPMEVMEARLWRVASTHMSPSWLTLLKCDVPQYQFYDTCQLLGKRAEVGDASPLFLKWKDGTRTPFSWR	195				
Mouse	116	PRTSSKTFNFTEDDEPMEVMEARLWRVASTHMSPSWLTLLKCDVPQYQFYDTCQLLGKRAEVGDASPLFLKWKDGTRTPFSWR	195				
		W194*					
Human	116	PRTSSKTFNFTEDDEPMEVMEARLWRVASTHMSPSWLTLLKCDVPQYQFYDTCQLLGKRAEVGDASPLFLKWKDGTRTPFSWR	195				
		D624G		R290fs		R339S	
Human	616	NNINTTICDQIPOQVHTTFYEAVGYMIGAQTDQFVQEHLIERLIEYLMLLPNPWNWSDI1QIQQATKNVDILKDPETVQLQGSILKTN	695	Human	289	GRPFPSHEMFDFGRPIF1PARPEVKWTHQEQQSGMFMFHELININGK1PVMEQYGLIPEEDICF1IKEQ1VGPLESPVDSLWPFYK	288
Mouse	616	NNINTTICDQIPOQVHTTFYEAVGYMIGAQTDQFVQEHLIERLIEYLMLLPNPWNWSDI1QIQQATKNVDILKDPETVQLQGSILKTN	695	Mouse	321	GRPFPSHEMFDFGRPIF1PARPEVKWTHQEQQSGMFMFHELININGK1PVMEQYGLIPEEDICF1IKEQ1VGPLESPVDSLWPFYK	400
Zebrafish	616	NNINTTICDQIPOQVHTTFYEAVGYMIGAQTDQFVQEHLIERLIEYLMLLPNPWNWSDI1QIQQATKNVDILKDPETVQLQGSILKTN	695				
C. elegans	626	ENLTCIICLSDHAQVWVTEAAGHIIASQDNLHEDL1KMDIPHRNTWIIAQAASSTSVLEEVPMQVSMLVILKTN	705				
S. pombe	617	RELAKTTEDLTPQTETYEACGTYMSIQNPKRHLQERLIDFIMALPNQAMENIQAQNAQVQLEGPVTQKILANLTKTN	696				
		HIST1H1E		KRAS		G13D	
Human	1	MSETAAFAAAPAPAAEKTPTVKKKAKRASGAKAKRASGPPYSELT1KVAKSKERGSVGSIAALKALAAGYDVEKNNRSI	80	Human	1	MTETKLVVVAGGGVGSALTIQIQLIQNHVFDEYDPTEDSYRKQVVIDGETLDDILDATQGEQYESAMRDQYMRTEGFLC	80
Mouse	1	MSETAAFAAAPAPAAEKTPTVKKKAKRASGAKAKRASGPPYSELT1KVAKSKERGSVGSIAALKALAAGYDVEKNNRSI	80	Mouse	1	MTETKLVVVAGGGVGSALTIQIQLIQNHVFDEYDPTEDSYRKQVVIDGETLDDILDATQGEQYESAMRDQYMRTEGFLC	80
		A167V		Zebrafish	1	MTETKLVVVAGGGVGSALTIQIQLIQNHVFDEYDPTEDSYRKQVVIDGETLDDILDATQGEQYESAMRDQYMRTEGFLC	80
Human	161	PAAAAGAKKAKSPKKAKAAPPKKAKPSFAKAKVFKFAAFTKPKAJPKKAKPKKAKAARRKKK	219				
Mouse	161	PAAAAGAKKAKSPKKAKAAPPKKAKPSFAKAKVFKFAAFTKPKAJPKKAKAARRKKK	219				
		NTRAS		KRAS		Q61H	
Human	1	MTEYKLVVGAGGGVGSALTIQIQLIQNHVFDEYDPTEDSYRKQVVIDGETLDDILDATQGEQYESAMRDQYMRTEGFLC	80	Human	1	MTETKLVVVAGGGVGSALTIQIQLIQNHVFDEYDPTEDSYRKQVVIDGETLDDILDATQGEQYESAMRDQYMRTEGFLC	80
Mouse	1	MTEYKLVVGAGGGVGSALTIQIQLIQNHVFDEYDPTEDSYRKQVVIDGETLDDILDATQGEQYESAMRDQYMRTEGFLC	80	Mouse	1	MTETKLVVVAGGGVGSALTIQIQLIQNHVFDEYDPTEDSYRKQVVIDGETLDDILDATQGEQYESAMRDQYMRTEGFLC	80
Zebrafish	1	MTEYKLVVGAGGGVGSALTIQIQLIQNHVFDEYDPTEDSYRKQVVIDGETLDDILDATQGEQYESAMRDQYMRTEGFLC	80	Zebrafish	1	MTETKLVVVAGGGVGSALTIQIQLIQNHVFDEYDPTEDSYRKQVVIDGETLDDILDATQGEQYESAMRDQYMRTEGFLC	80
		Q61H		KRAS		G13D	
Human	1	Q61H		Human	1	MTETKLVVVAGGGVGSALTIQIQLIQNHVFDEYDPTEDSYRKQVVIDGETLDDILDATQGEQYESAMRDQYMRTEGFLC	80
Mouse	1	Q61H		Mouse	1	MTETKLVVVAGGGVGSALTIQIQLIQNHVFDEYDPTEDSYRKQVVIDGETLDDILDATQGEQYESAMRDQYMRTEGFLC	80
Zebrafish	1	Q61H		Zebrafish	1	MTETKLVVVAGGGVGSALTIQIQLIQNHVFDEYDPTEDSYRKQVVIDGETLDDILDATQGEQYESAMRDQYMRTEGFLC	80
		MED12		ITPKB		E207fs	
Human	1	MAAFGILSYEHRLPLRRLPRLRGPPDYQFDQKQEDDELTALENVKQGFNNQNPASGDEHGSARNVSNPAKISSNFSIIIAEK	80	Human	155	QAQSSAIQAFRSPRLGRARCPGCPCPFRSSQQPGVRV1UQCARSERKRTKSGWQFCPETSGDTSGRKGPSLCS8QVKGK	234
Mouse	1	MAAFGILSYEHRLPLRRLPRLRGPPDYQFDQKQEDDELTALENVKQGFNNQNPASGDEHGSARNVSNPAKISSNFSIIIAEK	80	Mouse	160	QAQSSAIQAFRSPRLGRARSPSPCPFRSSQQPGVRV1UQCARSERKRTKSGWQFCPETSGDTSGRKGPSLCS8QVKGK	239
Zebrafish	1	MAAFGILSYEHRLPLRRLPRLRGPPDYQFDQKQEDDELTALENVKQGFNNQNPASGDEHGSARNVSNPAKISSNFSIIIAEK	80				
		P196S		E33K		E548fs	
Human	161	PAAGAGAKKAKSPKKAKAAPPKKAKPSFAKAKVFKFAAFTKPKAJPKKAKAARRKKK	219	Human	554	KPFLRKACSPSNIPAPA---VIIDMTGQEDGALETTQGSPRGNLPLRKLSSSASSTGFSSYEDDEEISDSDPERTLDN	630
Mouse	161	PAAGAGAKKAKSPKKAKAAPPKKAKPSFAKAKVFKFAAFTKPKAJPKKAKAARRKKK	219	Mouse	547	KASLKEACSPSNIPAPAIVIITDGMQEDGELI---TGSFRGPNLPLRKLSSSASSTGFSSYEDDEEISDSDPERTLDN	626
		P200fs		G44S		T626S	
Human	159	AV---ATAEAKGLDRLPRA---SDKQSPLNING-ASYLRLP-WNPYFMECATPFLYQFPLDNLSPNYSLNMYKA---LLPQSQYSLA	230	Human	1	QYIYDPIVEATGNNCPPHIESFDSVEMGEIINMIELTRYTRTPVQKHAIIPIIWERDILMACAQTGSKTAFLPLPILS	240
Mouse	159	AV---ATAEAKGLDRLPRA---SDKQSPLNING-ASYLRLP-WNPYFMECATPFLYQFPLDNLSPNYSLNMYKA---LLPQSQYSLA	230	Mouse	161	QYIYDPIVEATGNNCPPHIESFDSVEMGEIINMIELTRYTRTPVQKHAIIPIIWERDILMACAQTGSKTAFLPLPILS	240
Zebrafish	89	LIVPGCACTGTLVPPNGCFLPGLPSLNGSPMSFVH---TOMYFLDSDSKYALNMYSQKQPPYFL	165				
		P463fs		A59P		S24*	
Human	413	TAVDQEDRGSSPPLLEKQTVTQDTPKDLSSKVVVDASKDHMKM---APTYLWHSRASGLVLGSNSSEIPKETLSPPG	491	Human	1	MSHVAVENALGLDQFAGDLNNSDNQSGGSTASKGRYIIPPHLNRREATFGYDTSKGSWSSSKDKDAYSFSFGSRSDSRG	80
Mouse	413	TAVDQEDRGSSPPLLEKQTVTQDTPKDLSSKVVVDASKDHMKM---APTYLWHSRASGLVLGSNSSEIPKETLSPPG	491	Mouse	1	MSHVAVENALGLDQFAGDLNNSDNQSGGSTASKGRYIIPPHLNRREATFGYDTSKGSWSSSKDKDAYSFSFGSRSDSRG	80
Zebrafish	404	GSSSPVTT-SDSKPQGPIKEDPADKPLDLSAKIIMEFEGHNTGKELAALKLGSPARYGLPNPHELKEYLSPSS	482				
		E1382		P463fs		K348, splice	
Human	1353	NSKEPKGSKRLCKTKEHLIPEQSRHQLLQFQGKQFPESSDYLSPY-AKQEPF---FDRQLQLL	1427	Human	241	QYIYDPIGEAALKAMENGRGKRRQKQYTF1S1VLP1FTRALQVYEEARFNSYRVRPCVYGCAD1QGQ105DLSLPERGCHL	320
Mouse	1322	NSKEPKGSKRLCKTKEHLIPEQSRHQLLQFQGKQFPESSDYLSPY-AKQEPF---FDRQLQLL	1397	Mouse	241	QYIYDPIGEAALKAMENGRGKRRQKQYTF1S1VLP1FTRALQVYEEARFNSYRVRPCVYGCAD1QGQ105DLSLPERGCHL	320
Zebrafish	1385	P-ERPKGSKRLCKTKEHLIPEQSRHQLLQFQGKQFPESSDYLSPY-AKQEPF---FDRQLQLL	1463				
		ZMYM3		S401fs		K348, splice	
Human	1	MOPSDPFSPPFPPLTLPKPLACGLDVLMDPFSPPFSPPSACLDLTTAGLEKDPG-VLQGATEL	79	Human	321	VATPGRVLVDMMERKGKIGLDFCYRVLVLEADDRMLONGFEPQ1RIRVEQDTIPPKGVURHTMOSATFPKEIQLARLDFL	400
Mouse	1	MOPSDPFSPPFPPLTLPKPLACGLDVLMDPFSPPFSPPSACLDLTTAGLEKDPG-VLQGATEL	79	Mouse	321	VATPGRVLVDMMERKGKIGLDFCYRVLVLEADDRMLONGFEPQ1RIRVEQDTIPPKGVURHTMOSATFPKEIQLARLDFL	400
Zebrafish	1	MOPSDPFSPPFPPLTLPKPLACGLDVLMDPFSPPFSPPSACLDLTTAGLEKDPG-VLQGATEL	79				
		Q389*		F1302S		C44S	
Human	395	RPIIPQGD PADATRCSC1QCRTGEVIL---HEVENSNSVWHLRCSDCFSKFRANKGLKTCQDCGAYI1TQDTPSP---ELLJF	472	Human	401	IFTAVLGRVGSNTENTYQKVVWVFSEDKRSFLDLDLNATGKDSLT1LVPVETKKGADSLEDFLYHEGVACTSINGRSQRD	480
Mouse	397	RPIIPQGD PADATRCSC1QCRTGEVIL---HEVENSNSVWHLRCSDCFSKFRANKGLKTCQDCGAYI1TQDTPSP---ELLJF	474	Mouse	401	IFTAVLGRVGSNTENTYQKVVWVFSEDKRSFLDLDLNATGKDSLT1LVPVETKKGADSLEDFLYHEGVACTSINGRSQRD	480
Zebrafish	265	-----KEVQELQAG-TTF1BKVNLMSDUSVH1LCSBUDCNEQVTSNKLNGNSMHEGQ1C1---M7DAQCSQ51	328				
		F1302S		EGR2		H304N	
Human	1288	NRMPNPLRCRCPVYKEFYLSLKCPESLRTRNDVFLQYPERC1AESPLWYSPVIMDRSMLSMLNRLIAVREY---EELGR	1364	Human	318	RP1LRRPKYPNRSPSKTPVHERPYPCPAEGCDRRFSRSDELTRIRH1HTGHKPFCR1CMONFSRSDELH1HTRHTGEK	397
Mouse	1288	NRMPNPLRCRCPVYKEFYLSLKCPESLRTRNDVFLQYPERC1AESPLWYSPVIMDRSMLSMLNRLIAVREY---EELGR	1364	Mouse	315	RP1LRRPKYPNRSPSKTPVHERPYPCPAEGCDRRFSRSDELTRIRH1HTGHKPFCR1CMONFSRSDELH1HTRHTGEK	394
Zebrafish	1433	NRMPNPLRCRCPVYKEFYLSLKCPESLRTRNDVFLQYPERC1AESPLWYSPVIMDRSMLSMLNRLIAVREY---EELGR	1512	Zebrafish	277	RP1LRRPKYPNRSPSKTPVHERPYPCPAEGCDRRFSRSDELTRIRH1HTGHKPFCR1CMONFSRSDELH1HTRHTGEK	356

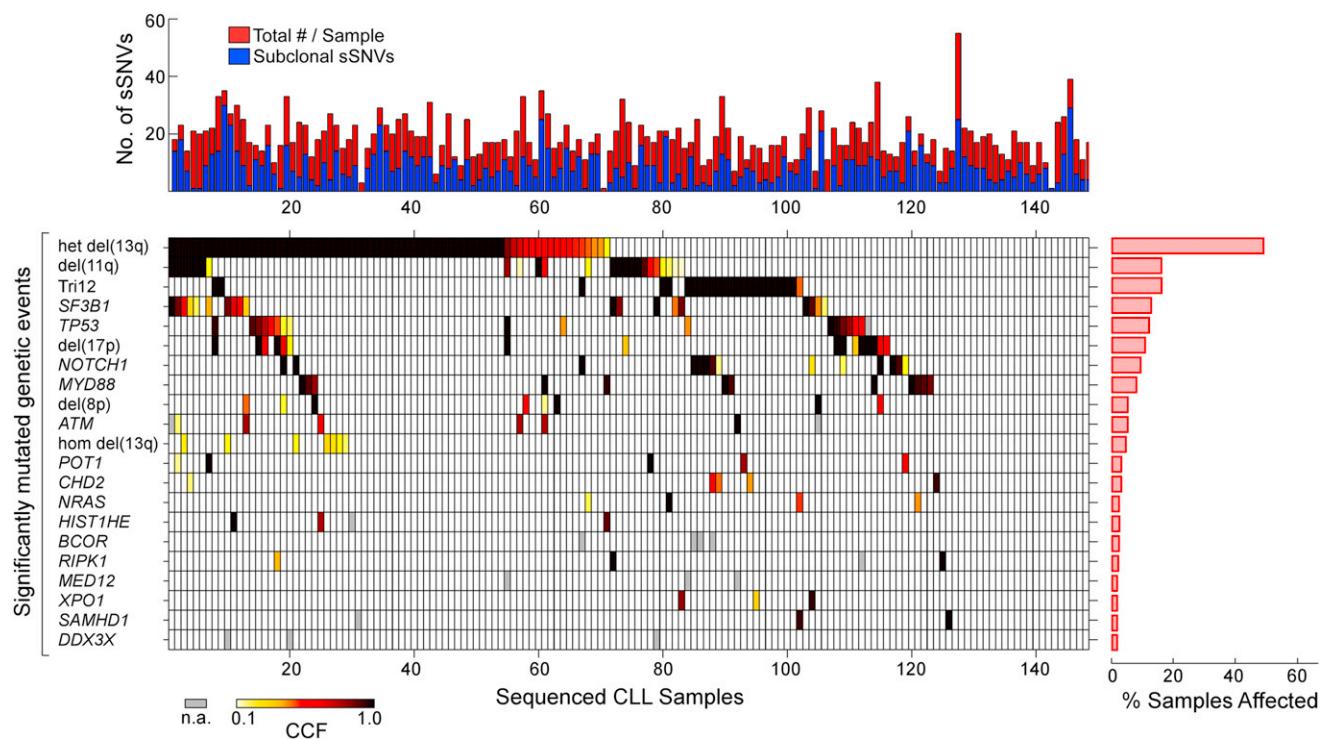
**Figure S2. Mutation Sites in 14 Significantly Mutated Genes Are Localized to Conserved Regions of Genes, Related to Figure 1**

Where available, alignments of gene sequences around each mutation are shown for human, mouse, zebrafish, C. elegans, and S. pombe genes (USCS Genomic Bioinformatics: <http://genome.ucsc.edu>).



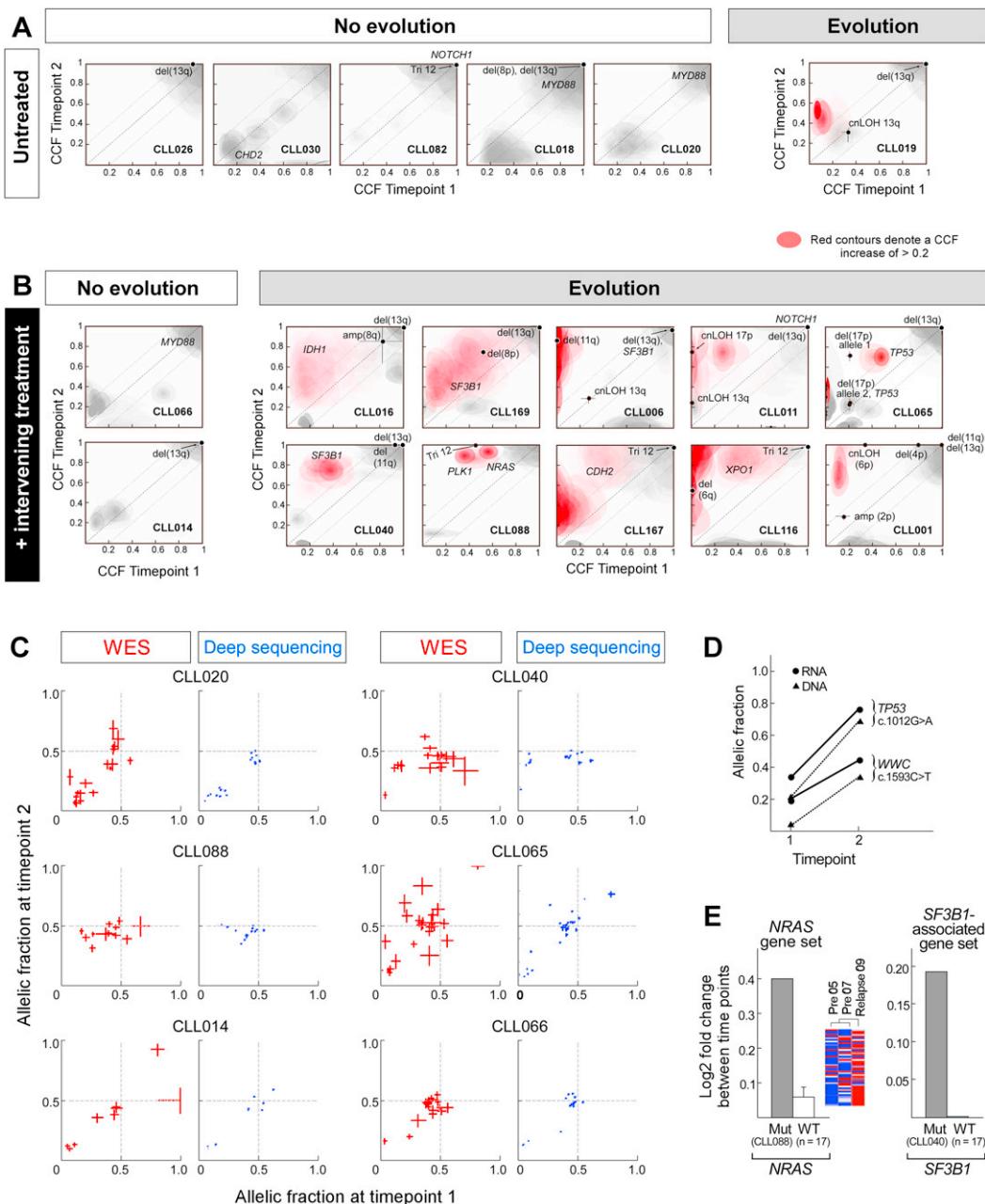
**Figure S3. Whole-Exome Sequencing Allelic Fraction Estimates Are Consistent with Deep Sequencing and RNA Sequencing Measurements, Related to Figure 2**

- (A) Comparison of ploidy estimates by ABSOLUTE with flow analyses for DNA content of 7 CLL samples and one normal B cell control (not analyzed by ABSOLUTE). Vertical lines indicate 95% confidence intervals of ploidy measurements by FACS.
- (B) Comparison of measurements of allelic fraction of 252 gene mutations detected by WES compared to detection using Fluidigm-based amplification following by deep sequencing (average 4,200 $\times$  coverage) using a MiSeq instrument. Significantly different estimates were assigned open circles.
- (C) Comparison of allelic fraction measured for 74 validated sites from 16 CLL samples by WES or RNA sequencing.
- (D) Comparison of mutational spectrum between subclonal and clonal sSNVs (detected in 149 CLLs). Rates were calculated as the fraction of the total number of sSNVs in the set with a particular mutation variant.



**Figure S4. Co-Occurrence of Mutations, Related to Figure 3**

The commonly occurring mutations, sorted in the order of decreasing frequency of affected. The top panel - the total number of mutations (red) and the number of subclonal mutations (blue) per sample. Bottom panel - co-occurring CLL driver events (y axis) are marked per individual CLL sample (x axis). Color spectrum (light yellow to black) corresponds to CCF; white boxes - no driver mutation identified; gray - mutations whose CCF was not estimated (i.e., mutations involving the X chromosome and indels other than in NOTCH1, currently not evaluated with ABSOLUTE).



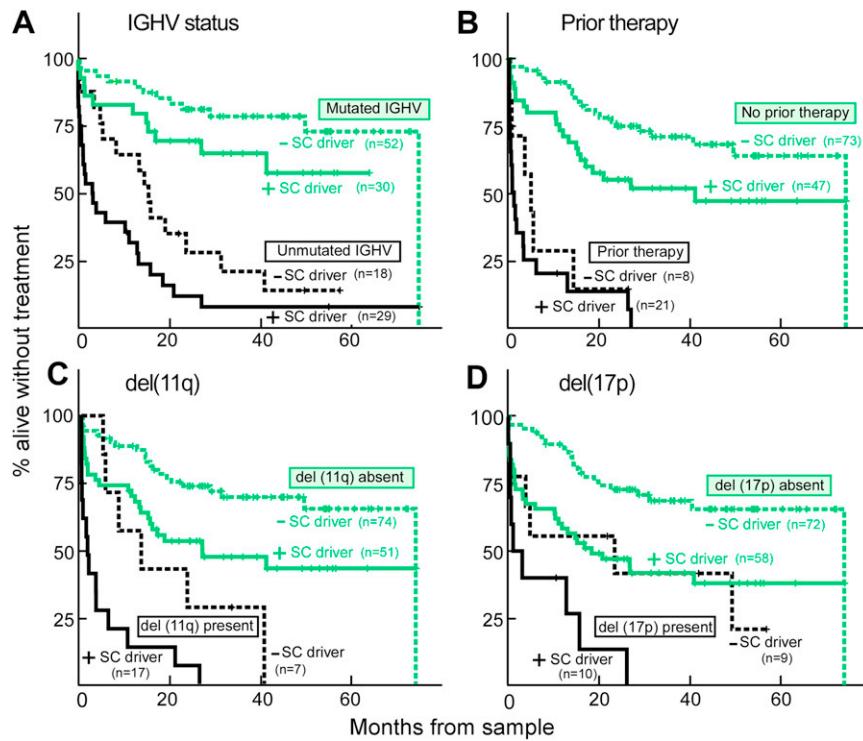
**Figure S5. Characterization of CLL Clonal Evolution through Analysis of Subclonal Mutations at Two Time Points in 18 Patients, Related to Figure 4**

(A and B) Unclustered results for 18 longitudinally studied CLLs, comparing CCF at two time points. Red color denotes a mutation with an increase in CCF greater than 0.2 (with probability >0.5). Six CLLs with no interval treatment (A) and 12 CLLs with intervening treatment (B) were classified as nonevolvers or evolvers, based on the presence of mutations with a statistically significant increase in CCF.

(C) Deep sequencing validation of 6 of the 18 CLLs. For each set of samples, allelic frequency (AF) by WES (red) (with 95% CI by binofit shown by cross bars) is shown on the left and AF by deep sequencing (blue) (with 95% CI by binofit shown by cross bars) is shown on the right. Deep sequencing was performed to an average coverage of 4,200 $\times$ .

(D) RNA pyrosequencing demonstrates a change in mRNA transcript levels that are consistent with changes in DNA allelic frequencies.

(E) Genetic changes correlate with transcript level of predefined gene sets expected to be altered as a result of the genetic lesion. These include change in expression level in the nonsense-mediated mRNA decay (NMD) pathway gene set, expected to be increased in association with splicing abnormalities, such as SF3B1 mutations (Table S6B). In addition, changes in expression level of the NRAS Q61 gene set (Table S6A) accompany the shift in allelic frequency for the NRAS mutations. Error bars represent one SD.



**Figure S6. The Presence of a Subclonal Driver Is Associated with Shorter FFS\_Sample When Added to Known Clinical High Risk Indicators, Related to Figure 6**

(A-D) FFS\_Sample plots of the patient groups based on presence or absence of a subclonal driver ("+/- SC driver") and their (A) IGHV mutation status; (B) exposure to prior therapy; (C) presence or absence of del(11q); and (D) presence or absence of del(17p).