LETTER

https://doi.org/10.1038/s41586-018-0465-8

Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia

Shih-Han Lee[1,5], Irtisha Singh[2,3,5], Sarah Tisdale[1], Omar Abdel-Wahab[4], Christina S. Leslie[2] & Christine Mayr[1]*

DNA mutations are known cancer drivers. Here we investigated whether mRNA events that are upregulated in cancer can functionally mimic the outcome of genetic alterations. RNA sequencing or 3′-end sequencing techniques were applied to normal and malignant B cells from 59 patients with chronic lymphocytic leukaemia (CLL)[1–3]. We discovered widespread upregulation of truncated mRNAs and proteins in primary CLL cells that were not generated by genetic alterations but instead occurred by intronic polyadenylation. Truncated mRNAs caused by intronic polyadenylation were recurrent ($n = 330$) and predominantly affected genes with tumour-suppressive functions. The truncated proteins generated by intronic polyadenylation often lack the tumour-suppressive functions of the corresponding full-length proteins (such as DICER and FOXN3), and several even acted in an oncogenic manner (such as CARD11, MGA and CHST11). In CLL, the inactivation of tumour-suppressor genes by aberrant mRNA processing is substantially more prevalent than the functional loss of such genes through genetic events. We further identified new candidate tumour-suppressor genes that are inactivated by intronic polyadenylation in leukaemia and by truncating DNA mutations in solid tumours[4,5]. These genes are understudied in cancer, as their overall mutation rates are lower than those of well-known tumour-suppressor genes. Our findings show the need to go beyond genomic analyses in cancer diagnostics, as mRNA events that are silent at the DNA level are widespread contributors to cancer pathogenesis through the inactivation of tumour-suppressor genes.

In addition to DNA-based mutations, recent studies found that alterations in mRNA processing, including splicing, promote tumorigenesis[6]. In CLL, up to one-quarter of patients have mutations in ATM or SF3B1, but one-third have less than two mutated driver genes, and most patients (58%) only have a 13q deletion or have a normal karyotype[3,7–9]. Here, we investigated whether intronic polyadenylation (IPA) might serve as a new driver of tumorigenesis. Because 16% of genes in normal immune cells use IPA to generate truncated mRNAs that contribute to transcriptome diversity[2], we hypothesized that cancer-specific IPA would generate truncated proteins that lack essential domains, and thus, may phenocopy truncating (TR) mutations (Fig. 1a).

Using 3′-seq, a 3′-end sequencing method, on 44 samples including B cells from healthy donors and from patients with CLL, we identified 5,587 IPA isoforms, including 3,484 without previous annotation[1,2] (Extended Data Table 1 and Methods). We validated 4,630 IPA isoforms using RNA sequencing (RNA-seq) and additional 3′-seq data[2,10] (Extended Data Fig. 1a, b). To assess IPA usage in CLL, we first identified the normal B cell subset, the gene expression profile of which was most closely related to CLL cells. Lymphoid tissue-derived $CD5^+$ B cells were most similar (Extended Data Fig. 2), but clustered separately from CLL samples based on IPA site usage (Extended Data Fig. 1c). Using a generalized linear model (GLM), we identified 931 IPA events with significantly higher expression among 13 CLL samples, but low or absent expression in $CD5^+$ B cells[1,2] (Fig. 1b, Extended Data Fig. 1d). Because CLL IPAs are detectable by RNA-seq, we used an unrelated RNA-seq dataset to validate our CLL-IPA events[3] (Fig. 1c). We verified up to 71% of testable IPAs by this independent method and dataset (Extended Data Fig. 1d). For further analysis, we combined the datasets ($n = 59$ CLL samples) and focused only on CLL-IPAs that were present in more than 10% of the sample cohort resulting in 330 CLL-IPAs, derived from 306 genes (Fig. 1d, Supplementary Table 1). Although CLL-IPAs were detected in all CLL samples, one-third of the samples had a significantly higher number of CLL-IPAs (Fig. 1e, Extended Data Fig. 1e).

To investigate whether CLL-IPAs express truncated proteins, we performed western blots on 13 candidates. Whereas normal B cells only expressed the full-length proteins, the malignant B cells also expressed truncated proteins, the size of which was consistent with the predicted size of IPA-generated proteins (Fig. 2a, Extended Data Figs. 3 and 4).

To rule out that proteolytic cleavage truncates the proteins, we validated the presence of the IPA-generated truncated mRNAs (Extended Data Fig. 5a). Moreover, we were able to induce IPA isoform expression through the downregulation of splicing factors or through the inhibition of 5′ splice site recognition using an antisense oligonucleotide, indicating that deregulated mRNA processing can cause the expression of a truncated protein[11,12] (Extended Data Fig. 5b).

Many of the truncated proteins generated by CLL-IPAs are markedly similar to the predicted protein products produced by TR mutations, suggesting that CLL-IPAs may functionally mimic the outcome of genetic mutations (Fig. 2b, Extended Data Fig. 6a). To test this, we investigated the functional consequences of the expression of IPA and full-length protein isoforms of four candidates in malignant B cells. CARD11 is a positive regulator of the NF-κB pathway and is important for lymphocyte survival and proliferation[13]. We observed substantial CARD11 IPA protein production, compared to only slightly increased CARD11 IPA mRNA expression, indicating that the truncated protein is more stable and may activate the NF-κB signalling pathway more potently than the full-length protein[14] (Fig. 2a). To test this, we exclusively knocked down either full-length or CARD11 IPA in a malignant B cell line that expresses CARD11 IPA at comparable levels to those expressed by CLL cells (Extended Data Fig. 6b, c). We measured phosphorylated p65 (also known as RELA) to assess NF-κB activity and found significantly lower activity after knockdown of CARD11 IPA than of the full-length protein (Fig. 2c, Extended Data Fig. 6d). Thus, CARD11 IPA activates NF-κB more potently than full-length CARD11, suggesting that it may mimic activating mutations present in high-grade lymphomas[13]. CARD11 IPA may contribute to NF-κB activation in CLL, in which the signalling components are rarely mutated[15].

DICER IPA generates a truncated protein that partially lacks the RNase IIIB domain responsible for microRNA (miRNA) processing[16] (Fig. 2b). In contrast to full-length DICER, DICER IPA entirely lacks miRNA cleavage ability and mimicked TR mutations that remove both RNase III domains[16] (Fig. 2b, d, Extended Data Fig. 6e, f). Although DICER IPA does not act in a dominant-negative manner, its expression reduces functional DICER protein, thus potentially decreasing endogenous miRNA expression.

The tumour-suppressor gene (TSG) *MGA* is targeted by TR mutations in CLL and solid cancers[3,7,17] (Fig. 2b). MGA negatively regulates

[1]Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2]Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [3]Tri-I Program in Computational Biology and Medicine, Weill Cornell Graduate College, New York, NY, USA. [4]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [5]These authors contributed equally: Shih-Han Lee, Irtisha Singh. *e-mail: mayrc@mskcc.org

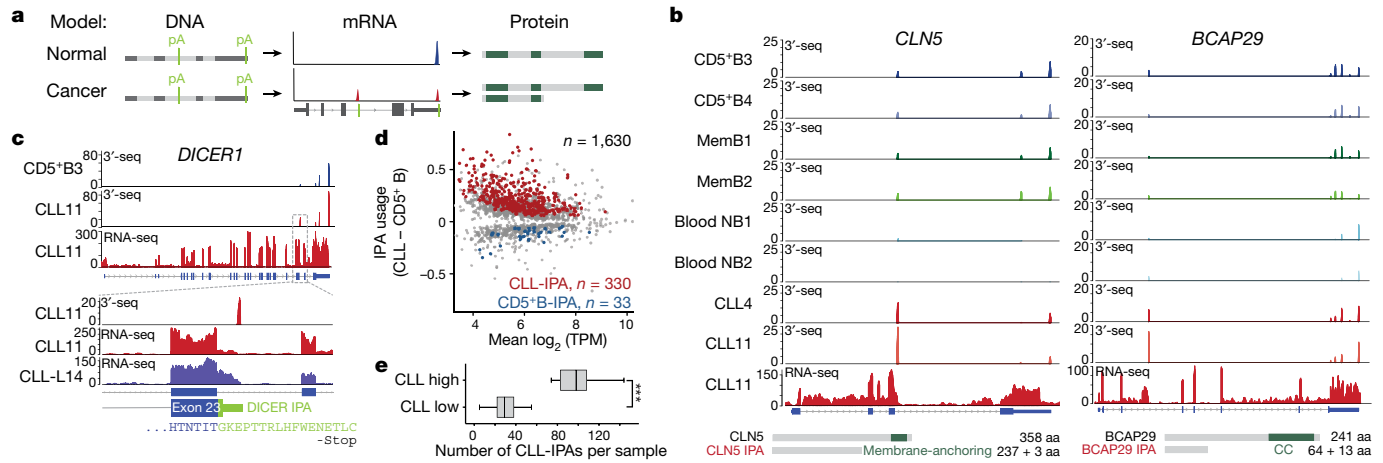


Fig. 1 | Hundreds of genes generate recurrent CLL-IPAs. a, Schematic showing full-length mRNA and protein expression in normal cells and the generation of a truncated mRNA and protein through cancer-specific IPA, despite no difference in DNA sequence. Polyadenylation sites (pA) are shown in light green. Loss of essential protein domains (dark green boxes) through cancer-gained IPA may inactivate TSGs, thus contributing to cancer pathogenesis. **b**, Representative CLL-IPAs (from $n = 330$) are shown. mRNA 3′ ends detected by 3′-seq are depicted as peaks, the heights of which correspond to transcript abundance shown in transcripts per million (TPM). The bottom panel shows RNA-seq reads and numbers correspond to read counts. Full-length and IPA-generated truncated proteins are depicted in grey, known domains are shown in green and the domains lost through IPA are named. For CLL-IPA, the number of retained and novel amino acids (aa) and amino acids of full-length proteins are given. CC, coil–coil; MemB, memory B cells, NB, naive B cells. **c**, Representative RNA-seq tracks from two independent CLL datasets are shown as in **b**; one is indicated by 'L' before the patient number (CLL-L14). B3 denotes donor 3. Zoomed-in view shows the exonized part of intron 23 of *DICER1* (green). **d**, Difference in relative abundance (usage) of IPA isoforms between CLL and normal $CD5^+$ B cells. A GLM was used to identify significant events. CLL-IPAs with significantly higher usage are shown in red (false discovery rate (FDR)-adjusted $P < 0.1$, usage difference ≥ 0.05, TPM in $CD5^+$ B < 8) and $CD5^+$ B-IPAs are shown in blue. Grey denotes IPAs present in CLL and $CD5^+$ B cells without significantly different usage. **e**, Number of CLL-IPAs per sample is shown as box plots, in which the horizontal line denotes the median; boxes denote the 25th and 75th percentiles; error bars denote the range. CLL high, $n = 21/59$, median of CLL-IPAs/sample = 98 versus CLL low, $n = 38/59$, median = 29. $^{***}P = 6 \times 10^{-10}$, two-sided Mann–Whitney *U*-test.

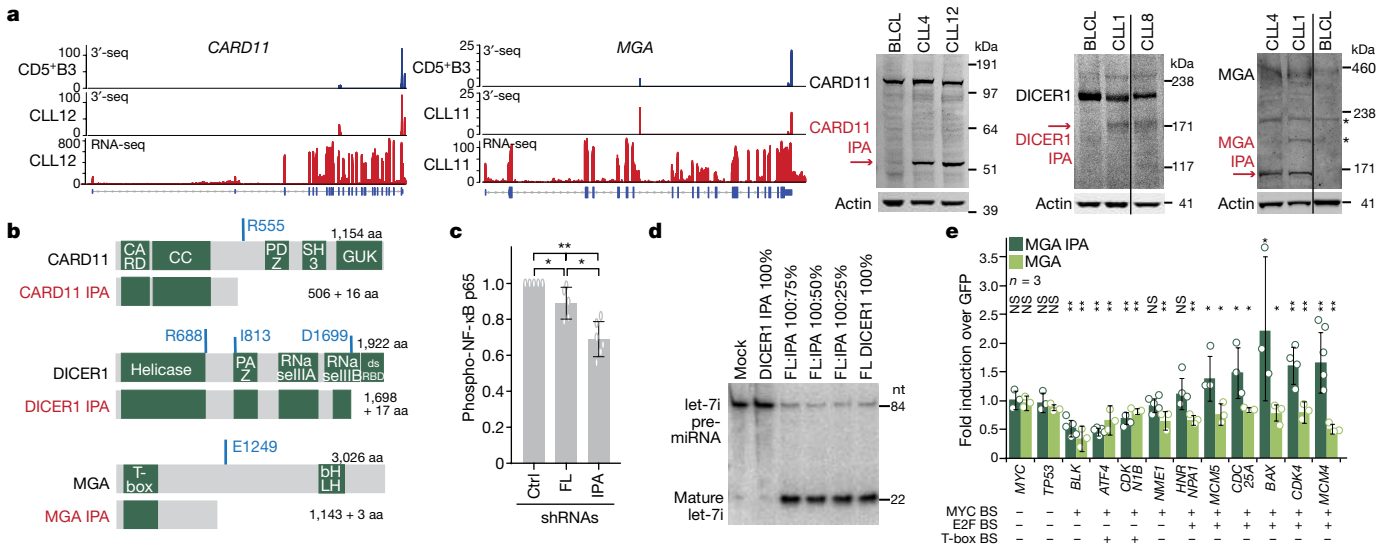


Fig. 2 | IPA-generated truncated proteins resemble the protein products of truncating DNA mutations and have cancer-promoting properties. **a**, RNA-seq and 3′-seq data of functionally validated CLL-IPAs ($n = 5$) as in Fig. 1b. The remaining tracks are shown in Extended Data Fig. 3. Endogenous full-length proteins are detected by western blot analysis in CLL and normal B cells (B lymphoblastoid cells; BLCL), whereas IPA-generated truncated proteins (red arrows) are only present in primary CLL cells. Actin was used as loading control. The experiment was replicated with similar results (*CARD11*, $n = 4$, *DICER*, $n = 3$, *MGA*, $n = 2$). For gel source data see Supplementary Fig. 1. Asterisks denote an unspecific band. **b**, Protein models are shown as in Fig. 1b. The amino acid positions of recurrent TR mutations are shown in blue. **c**, Endogenous phospho-NF-κB p65 levels are shown as normalized mean fluorescent intensity (MFI) values after short hairpin RNA (shRNA)-mediated knockdown of full-length (FL) CARD11 and CARD11 IPA. $n = 5$ (FL shRNA1 and control (ctrl) shRNA) or $n = 6$ (IPA; shRNA2 $n = 3$, shRNA3 $n = 3$) biologically independent experiments. Data are mean ± s.d. $^{**}P = 0.002$, two-sided Kruskal–Wallis test; *P* value of two-sided Mann–Whitney *U*-test was adjusted for multiple testing, *adjusted $P = 0.036$. **d**, miRNA cleavage assay, performed twice with similar results, showing processing of pre-let-7i into mature let-7i by V5-DICER. Mock indicates that no protein was added. V5-DICER IPA shows a complete loss of function, but no dominant-negative activity. nt, nucleotides. **e**, qRT–PCR of endogenous MYC target genes after expression of full-length or MGA IPA in Raji cells. Shown are *GAPDH*-normalized values as mean ± s.d. from three biological replicates, each performed in technical triplicates. $^{*}P < 0.05$, $^{**}P < 0.001$, two-sided *t*-test for independent samples. NS, not significant. Exact *P* values are shown in Supplementary Fig. 1. MGA represses all MYC target genes. BS, binding sites.

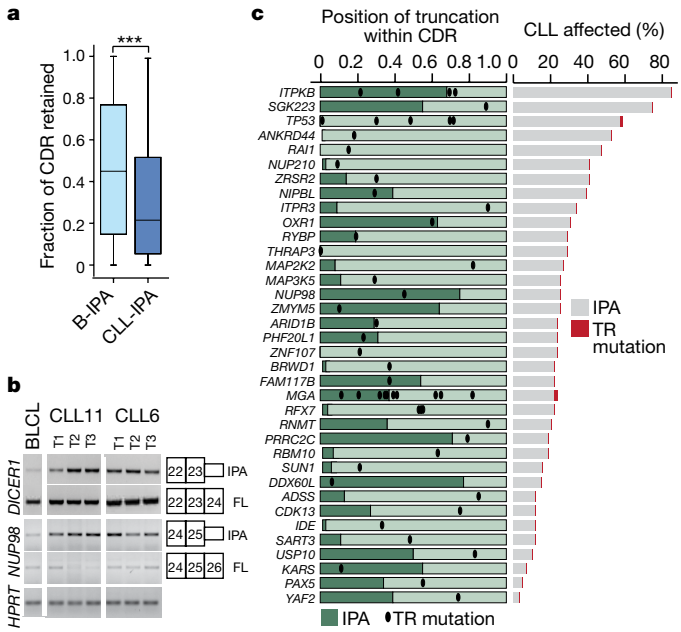


Fig. 3 | TSGs are enriched among CLL-IPAs. CLL-IPAs and TR mutations in CLL target the same genes but in different patients. **a**, The fraction of the retained coding region (CDR) is shown for genes that generate CLL-IPAs ($n = 306$, median fraction of retained CDR = 0.21; 112 amino acids) and B-IPAs ($n = 2{,}690$, median fraction of retained CDR = 0.45; 221 amino acids). ***$P = 1 \times 10^{-16}$, two-sided Mann–Whitney *U*-test. Box plots are as in Fig. 1e. **b**, RT–PCR analysis showing the expression of full-length and IPA isoforms for two TSGs (*DICER1* and *NUP98*) in samples from two patients with CLL that were collected over several years. CLL11: T1, 17 months after diagnosis, T2, 24 months, T3, 44 months; CLL6: T1, 16 months, T2, 49 months, T3, 91 months (42 months after treatment). Shown are the exons that contain primers for amplifications of the products. BLCL serve as control cells. The expression of *HPRT* was used as a loading control. **c**, Genes that are targeted by TR mutations in CLL and CLL-IPAs are shown ($n = 36$). Dark green bars indicate the fraction of retained CDRs for each IPA-generated protein. Black dots indicate the positions of TR mutations in CLL. CLL-IPAs occur mostly in the vicinity of or upstream of TR mutations. $P = 0.004$, two-sided Wilcoxon rank-sum test. Right, the fraction of CLL samples affected is shown for each gene and represents the fraction of CLL samples (out of 59) with significantly upregulated expression of the IPA isoform (CLL-IPA, grey; TR mutations, red).

the MYC transcriptional program and represses genes with MYC- and E2F-binding sites in a Polycomb-dependent manner[18,19]. Expression of MGA from constructs validated MGA IPA detected in CLL cells and confirmed the repressive effect of MGA on MYC target gene expression in malignant B cells (Fig. 2e, Extended Data Fig. 6g). Notably, on genes with binding sites for both MYC and E2F, MGA IPA acts as dominant-negative regulator of full-length MGA as it significantly induced the expression of 5 out of 6 genes in cells that endogenously express full-length MGA (Fig. 2e). However, as MGA IPA retains the N-terminal T-box, it still acts as a repressor on T-box target genes (Fig. 2e).

Lastly, the IPA isoform of the transcriptional repressor FOXN3[20] derepressed its oncogenic targets *MYC* and *PIM2* (Extended Data Figs. 3, 6h–j). In summary, the CLL-IPA-generated proteins can contribute to cancer pathogenesis in various ways. Their generation can reduce the expression of functional TSGs (DICER and FOXN3 IPA) or they behave as dominant-negatives, thus acting in an oncogenic manner (MGA IPA).

Because all functionally validated CLL-IPAs produced dysfunctional proteins, we investigated whether this is a general feature. We compared the retained fraction of amino acids of IPA isoforms present in normal B cells (B-IPA, $n = 2{,}690$) with CLL-IPAs. Although the protein size of full-length proteins targeted by IPA was similar, CLL-IPAs lose significantly more amino acids than B-IPAs (Fig. 3a, Extended Data Fig. 7a). This suggests that IPA in normal cells contributes to proteome diversity[2], whereas CLL-IPAs tend to produce dysfunctional proteins.

Because genes targeted by TR mutations are often TSGs[5] (Extended Data Fig. 7b), we investigated whether TSGs are overrepresented among CLL-IPAs. Compared to control groups with matched protein sizes, there was a significant enrichment of TSGs among CLL-IPAs ($P = 3 \times 10^{-5}$; Extended Data Fig. 7c–f). Importantly, IPA-generated truncated proteins usually lack either more or a comparable number of amino acids compared to truncated proteins generated by TR mutations, suggesting the IPA isoforms are probably inactive (Extended Data Fig. 7c). However, for CLL-IPAs to inactivate TSGs, they must also be stably expressed. For 11 out of 12 tested CLL-IPAs, we observed stable expression at the mRNA or protein level over a four-year time span (Fig. 3b, Extended Data Fig. 5c, d), indicating that they have the potential to inactivate TSGs.

In addition to TSGs in general, we found that genes inactivated by TR mutations in CLL are enriched among CLL-IPAs[3,7,8] (Fig. 3c, Extended Data Fig. 7g). Notably, the fraction of samples affected by CLL-IPA was substantially larger than the number of CLL samples affected by TR mutations (3.0–85% versus 0.13–2.0%; Fig. 3c, right). This indicates that TR mutations and CLL-IPAs target the same genes in different patient groups, thus substantially expanding the proportion of patients with protein truncations in potential drivers.

To rule out the possibility that CLL-IPAs are caused by somatic mutations, we examined the presence of DNA mutations in the CLL-IPA genes. Two genes were targeted by TR mutations and IPA in the same patient. Notably, the two inactivation mechanisms are predicted to generate different truncated protein products, suggesting that they occurred independently[3] (Extended Data Fig. 7h, i). The mutation data also enabled us to associate CLL-IPAs with specific somatic mutations. CLL samples with a high number of IPA were enriched in SF3B1 mutations, but they were independent of IGVH mutational status (Extended Data Fig. 7j–l).

Because of the enrichment of known TSGs among CLL-IPAs, we examined whether CLL-IPAs may enable us to identify novel TSGs. We selected CLL-IPAs present in at least 20% of CLL samples ($n = 199$, generated from 190 genes; Fig. 4a, Supplementary Tables 1 and 2). We next investigated whether these genes are inactivated by TR mutations in solid cancers using mutations from more than 86,000 tumours, compiled by the MSK cbio portal[4]. We observed that 72% of these genes are frequently affected by TR mutations in solid tumours and call them novel TSG candidates (136 out of 190; Fig. 4b). This is a significant

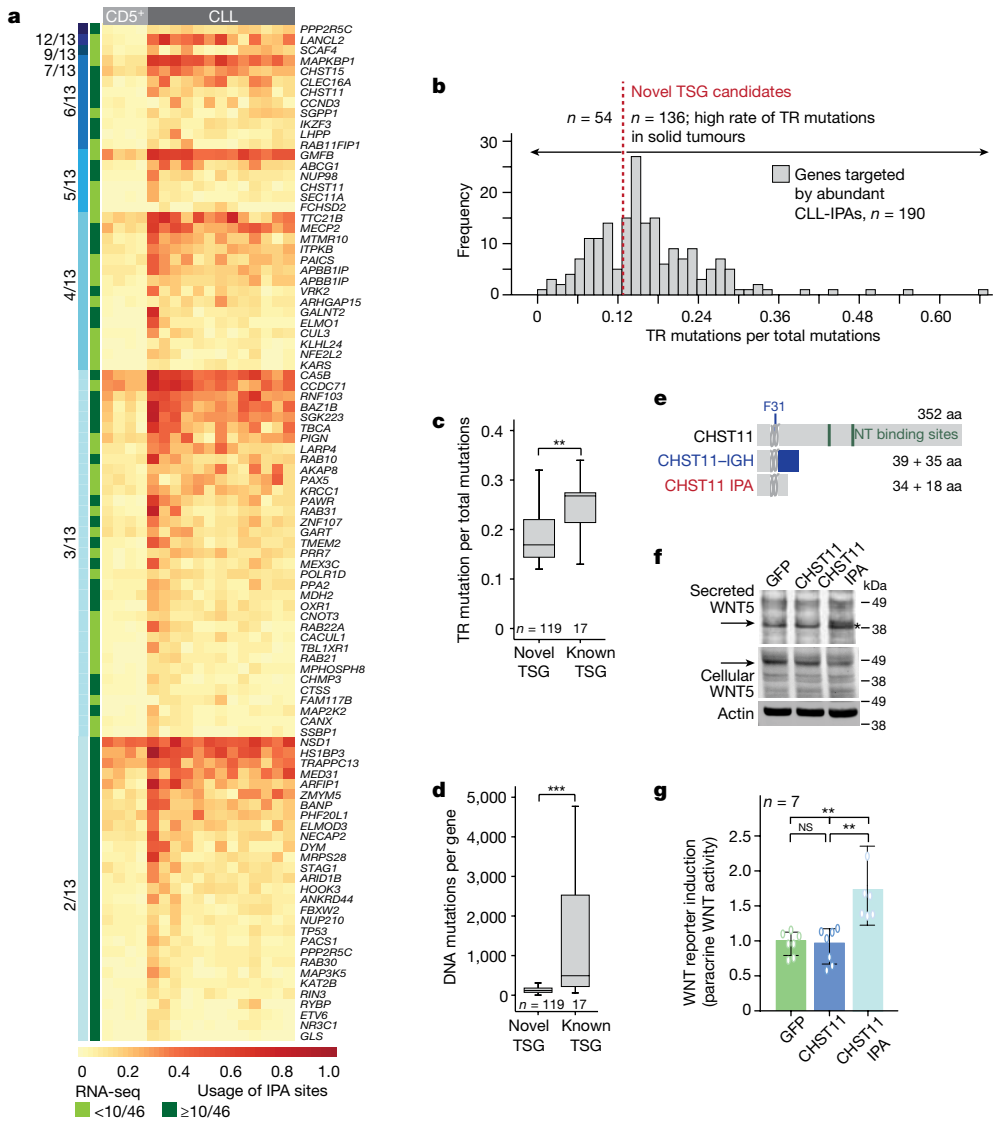


Fig. 4 | Novel TSG candidates are inactivated in CLL at the mRNA level and in solid tumours at the DNA level. a, Colour-coded IPA usage for a subset of CLL-IPAs (97 out of 199 of samples with significant expression of IPA in ≥20% of CLL samples). Gene names and number of affected CLL samples per CLL-IPA are indicated (blue bars, 3′-seq, green bars, RNA-seq). **b**, Truncating mutation rates (number of TR mutations per total mutations) in solid tumours, obtained from the MSK cbio portal for genes that generate abundant CLL-IPAs, partially shown in **a**. The bimodal distribution was separated at the local minimum (TR mutation per total mutations = 0.12, red line) into two gene groups: those rarely targeted by TR mutations and those with high TR mutation rates in solid cancers, defined as novel TSG candidates. **c**, TR mutation rates of known and novel TSG candidates. $^{**}P = 0.0002$, two-sided Mann–Whitney *U*-test. Box plots are as in Fig. 1e. **d**, As shown in **c**, but for overall mutation rates. $^{***}P = 1 \times 10^{-10}$, two-sided Mann–Whitney *U*-test. **e**, CHST11 protein models, shown as in Fig. 2b. Loops depict membrane domains. A chromosomal translocation in CLL results in fusion of the immunoglobulin heavy chain locus (IGH) with a truncated CHST11[23]. NT, nucleotide. **f**, Western blot of WNT5B, performed once, shown as in Fig. 2a, from cell lysates of or conditioned medium obtained from B cells stably expressing green fluorescent protein (GFP), GFP-tagged CHST11 or GFP-tagged CHST11 IPA. Conditioned medium from cells expressing CHST11 IPA contains unglycosylated WNT5B[25]. Asterisk, unspecific band. **g**, Conditioned medium from samples described in **f** was added to HEK293T cells expressing a WNT reporter, and normalized luciferase activity is shown. Data are mean ± s.d. from $n = 7$ biologically independent experiments. $^{**}P = 0.002$, two-sided Kruskal–Wallis test; *P* value of two-sided Mann–Whitney *U*-test was adjusted for multiple testing, **adjusted $P = 0.002$.

enrichment over background and the list contains 17 known TSGs and 119 novel TSG candidates[5] (Extended Data Fig. 8a, b). Again, CLL-IPAs lack more or a comparable number of amino acids as the proteins produced by TR mutations, suggesting that CLL-IPAs inactivate the functions of these genes (Extended Data Fig. 8a).

Although the TR mutation rates of the novel TSG candidates were comparable with known TSGs found at the lower end of the spectrum, their protein size and overall mutation rates were substantially lower (Fig. 4c, d, Extended Data Fig. 8c). This may explain why these potentially cancer-relevant genes have been overlooked thus far[21]. As they are targeted at the mRNA level in leukaemia and at the DNA level in solid cancers, they should be considered as a novel class of TSG candidates.

To support this, we functionally validated a highly recurrent CLL-IPA isoform that affected a poorly known cancer gene. *CHST11* encodes a Golgi-associated carbohydrate sulfotransferase that modifies chondroitin on the surface of WNT-expressing cells. The modification results in the binding of secreted WNT and prevents its paracrine action[22]. CHST11 IPA lacks catalytic activity, but retains the cytoplasmic tail[23] (Fig. 4e, Extended Data Fig. 8d). As exclusive expression of the cytoplasmic tail of Golgi enzymes inhibited localization of full-length enzymes[24], we hypothesized that CHST11 IPA may act in a dominant-negative manner. We expressed CHST11 and CHST11 IPA, collected the conditioned media, and detected secreted WNT in medium only after expressing CHST11 IPA[25] (Fig. 4f, Extended Data Fig. 8e, f). The

conditioned medium activated a WNT reporter in HEK293T cells (Fig. 4g), demonstrating that CHST11 IPA enabled paracrine WNT action on neighbouring cells through dominant-negative action. Thus, in addition to mutations in the WNT pathway[26], CLL-IPAs may also contribute to WNT activation in CLL.

A member of this new class of TSGs was recently found in breast cancers, in which tumour-specific expression of MAGI3 IPA generates a truncated protein with dominant-negative activity[27] (Extended Data Fig. 9a). Combined with our findings on T-lineage acute lymphoblastic leukaemia (T-ALL), in which we detected more than 100 IPA isoforms (Extended Data Fig. 9b), these data indicate that cancer-upregulated IPA isoforms are not restricted to CLL.

In summary, we found that TSGs can be inactivated, either fully or partially, by IPA. Even partial loss of TSG function was shown to contribute crucially to tumorigenesis[28]. As CLL-IPAs are not generated by DNA mutations in their corresponding transcription units, DNA and mRNA alterations occur in different patient groups. In CLL, the fraction of patients with TSGs that are inactivated by CLL-IPAs is considerably larger than those with TSGs disrupted by TR mutations (Fig. 3c); thus, CLL-IPAs substantially expand the number of patients with affected drivers. Moreover, these data identify a class of TSGs that is predominantly inactivated at the mRNA rather than the DNA level[27]. Thus, our study demonstrates that cancer-gained changes in mRNA processing can functionally mimic the effects of somatic mutations and shows the need to go beyond genomic analyses in cancer diagnostics.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0465-8.

Received: 18 July 2017; Accepted: 17 July 2018;
Published online: 27 August 2018

1. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* **27**, 2380–2396 (2013).
2. Singh, I. et al. Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nat. Commun.* **9**, 1716 (2018).
3. Landau, D. A. et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).
4. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
5. Davoli, T. et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).
6. Dvinge, H., Kim, E., Abdel-Wahab, O. & Bradley, R. K. RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer* **16**, 413–430 (2016).
7. Puente, X. S. et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475**, 101–105 (2011).
8. Quesada, V. et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 47–52 (2011).
9. Stilgenbauer, S., Bullinger, L., Lichter, P., Döhner, H. & the German CLL Study Group (GCLLSG). Genetics of chronic lymphocytic leukemia: genomic aberrations and V(H) gene mutation status in pathogenesis and clinical course. *Leukemia* **16**, 993–1007 (2002).
10. Gruber, A. J. et al. A comprehensive analysis of 3′ end sequencing datasets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res.* **26**, 1145–1159 (2016).
11. Vorlová, S. et al. Induction of antagonistic soluble decoy receptor tyrosine kinases by intronic polyA activation. *Mol. Cell* **43**, 927–939 (2011).
12. Zarnack, K. et al. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152**, 453–466 (2013).
13. Lenz, G. et al. Oncogenic CARD11 mutations in human diffuse large B cell lymphoma. *Science* **319**, 1676–1679 (2008).
14. Bertin, J. et al. CARD11 and CARD14 are novel caspase recruitment domain (CARD)/membrane-associated guanylate kinase (MAGUK) family members that interact with BCL10 and activate NF-κB. *J. Biol. Chem.* **276**, 11877–11882 (2001).
15. Mansouri, L., Papakonstantinou, N., Ntoufa, S., Stamatopoulos, K. & Rosenquist, R. NF-κB activation in chronic lymphocytic leukemia: a point of convergence of external triggers and intrinsic lesions. *Semin. Cancer Biol.* **39**, 40–48 (2016).
16. Rakheja, D. et al. Somatic mutations in DROSHA and DICER1 impair microRNA biogenesis through distinct mechanisms in Wilms tumours. *Nat. Commun.* **5**, 4802 (2014).
17. De Paoli, L. et al. MGA, a suppressor of MYC, is recurrently inactivated in high risk chronic lymphocytic leukemia. *Leuk. Lymphoma* **54**, 1087–1090 (2013).
18. Hurlin, P. J., Steingrimsson, E., Copeland, N. G., Jenkins, N. A. & Eisenman, R. N. Mga, a dual-specificity transcription factor that interacts with Max and contains a T-domain DNA-binding motif. *EMBO J.* **18**, 7019–7028 (1999).
19. Ogawa, H., Ishiguro, K., Gaubatz, S., Livingston, D. M. & Nakatani, Y. A complex with chromatin modifiers that occupies E2F- and Myc-responsive genes in G0 cells. *Science* **296**, 1132–1136 (2002).
20. Huot, G. et al. CHES1/FOXN3 regulates cell proliferation by repressing PIM2 and protein biosynthesis. *Mol. Biol. Cell* **25**, 554–565 (2014).
21. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
22. Nadanaka, S., Kinouchi, H., Taniguchi-Morita, K., Tamura, J. & Kitagawa, H. Down-regulation of chondroitin 4-*O*-sulfotransferase-1 by Wnt signaling triggers diffusion of Wnt-3a. *J. Biol. Chem.* **286**, 4199–4208 (2011).
23. Schmidt, H. H. et al. Deregulation of the *carbohydrate* (*chondroitin 4*) *sulfotransferase 11* (*CHST11*) gene in a B-cell chronic lymphocytic leukemia with a t(12;14)(q23;q32). *Oncogene* **23**, 6991–6996 (2004).
24. Milland, J., Russell, S. M., Dodson, H. C., McKenzie, I. F. & Sandrin, M. S. The cytoplasmic tail of α1,3-galactosyltransferase inhibits Golgi localization of the full-length enzyme. *J. Biol. Chem.* **277**, 10374–10378 (2002).
25. Kessenbrock, K. et al. A role for matrix metalloproteinases in regulating mammary stem cell function via the Wnt signaling pathway. *Cell Stem Cell* **13**, 300–313 (2013).
26. Wang, L. et al. Somatic mutation as a mechanism of Wnt/β-catenin pathway activation in CLL. *Blood* **124**, 1089–1098 (2014).
27. Ni, T. K. & Kuperwasser, C. Premature polyadenylation of MAGI3 produces a dominantly-acting oncogene in human breast cancer. *eLife* **5**, e14730 (2016).
28. Berger, A. H., Knudson, A. G. & Pandolfi, P. P. A continuum model for tumour suppression. *Nature* **476**, 163–169 (2011).

Acknowledgements This work was funded by the NCI grant U01-CA164190 (to C.M. and C.S.L.), a Starr Cancer Foundation grant (to C.M. and C.S.L.), the Innovator Award of the Damon Runyon-Rachleff Cancer Foundation and the Island Outreach Foundation (DRR-24-13; to C.M.), the NIH Director's Pioneer Award (DP1-GM123454, to C.M.), the Pershing Square Sohn Cancer Research Alliance (to C.M.) and the MSK Core Grant (P30 CA008748). We are grateful to V. K. Modi for access to lymphatic tissue, to D. A. Landau for providing CLL RNA-seq data and sample identities to validate our findings, and to C. Wu and D. Neuberg for clinical outcome analyses. We thank J. Mendell and V. Narry Kim for providing the V5-DICER construct and the DICER knockout cells, J. Chaudhuri for critical reading of the manuscript, and the members of the Mayr laboratory for discussions.

Reviewer information *Nature* thanks M. Muschen and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.-H.L. organized and collected the samples and generated the libraries. S.-H.L. and S.T. performed and analysed all validation and functional experiments and contributed to study design. I.S. performed all of the computational analyses with respect to identification of IPA isoforms and their integration with published CLL datasets with input from C.S.L. and C.M. O.A.-W. provided the CLL samples and some of the CLL RNA-seq data. C.M. conceived the study and integrated CLL-IPAs with mutation analysis of solid cancers. C.M., S.-H.L. and S.T. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0465-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0465-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.M.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Samples for 3′-seq and RNA-seq analyses. Samples were obtained from untreated patients with CLL seen at Memorial Sloan Kettering Cancer Center, New York (Extended Data Table 1a). All patients provided written informed consent before participating in the study. The sample collection was approved by the Institutional Review Board of Memorial Sloan Kettering Cancer Center. Peripheral blood mononuclear cells from CLL samples with a minimum white blood cell count of 75,000 per microlitre were isolated by Ficoll (GE Healthcare) gradient centrifugation at 400 r.c.f. for 30 min, followed by two washes in PBS at room temperature. Cells were treated with red blood cell lysis buffer (155 mM NH_4Cl, 12 mM $NaHCO_3$, 0.1 mM EDTA) for 5 min at room temperature and were washed twice with PBS. Pure CLL B cells were obtained from peripheral blood mononuclear cells using B-CLL isolation kit (Miltenyi Biotec). This selected untouched CLL cells using a cocktail of magnetic beads coated with CD2, CD3, CD4, CD14, CD15, CD16, CD56, CD61, CD235a, FcεRI and CD34. The purity of CLL B cells ($CD5^+$ and $CD19^+$) was analysed by FACS and the cells were immediately dissolved in TRI Reagent (Ambion) for RNA extraction, followed by 3′-seq or RNA-seq library preparation.

For longitudinal analyses, samples from two patients were investigated at different time points during the course of the disease. CLL11, time point 1 (T1) 17 months after diagnosis, T2, 24 months after diagnosis, T3, 44 months after diagnosis. The patient was not treated with chemotherapy during the sample collection period. CLL6: T1, 16 months after diagnosis, T2, 49 months, T3, 91 months (42 months after chemotherapeutic treatment).

In addition to the newly generated CLL 3′-seq data, we also used 3′-seq data from normal tissues, cell lines and immune cell subsets that we generated previously[1,2] (Extended Data Table 1b).

We performed RNA-seq on 11 CLL samples (Extended Data Table 1a) and obtained access to a previously published RNA-seq dataset from 44 patients with CLL[3] that was provided by D. A. Landau. RNA-seq data from normal immune cells were obtained from samples we generated previously[2] (Extended Data Table 1c). For validation of 3′-seq data, we also used publicly available RNA-seq (tonsil-derived NB, GSE45982 (GSM1129340–GSM1129347)[29], blood-derived NB, ERR431624, ERR431586[30], $CD3^+$ T cells, GSM1576415[31]) and 3′-seq data[10].

For RNA-seq-based identification of IPA isoforms expressed in T-ALL, we used publicly available RNA-seq data from 10 primary T-ALL samples and 2 whole human thymus extracts (GSE57982)[32].

FACS sorting of immune cell populations. Cells were washed with ice-cold PBS once, incubated with appropriate fluorochrome-conjugated antibodies for 30 min at 4 °C and washed twice with ice-cold PBS containing 0.5% FCS. The following antibodies were used: anti-CD3-PE (mouse, BD Biosciences, 555333), anti-CD5-FITC (mouse, BD Biosciences, 555352), anti-CD14-PECy7 (mouse, ebioscience, 25-0149-42), anti-CD19-APC (mouse, BD Biosciences, 555415), anti-CD27-PE (mouse BD Biosciences, 555441), anti-CD38-APC (mouse, BD Biosciences, 555462), anti-CD38-FITC (mouse, BD Biosciences, 555459). Surface protein expression was detected by a BD FACSCalibur cell analyser (BD Biosciences) and data were analysed using the FlowJo software.

3′-seq and RNA-seq analyses. 3′-seq libraries were generated as previously described and sequenced with Illumina HiSeq using single-end 50-nucleotide reads[1,2]. RNA-seq libraries were prepared at the Weill Cornell and the MSKCC Genomics core facilities.

Analysis of 3′-seq data was performed as described previously[1] with a few modifications that have been extensively described[2]. In brief, a gene is considered to be expressed if either the IPA isoform (≥5 TPM) or the full-length isoform (≥5.5 TPM) were expressed in 75% of the samples of a particular cell type. We focused our analysis on robustly expressed transcript isoforms and filtered 3′-seq peaks according to their usage. Robustly expressed 3′UTR isoforms that are part of the atlas are expressed with at least 3 TPM in at least one sample and each peak combines at least 10% of all reads that map to the 3′UTR. Robustly expressed IPA isoforms that are part of the atlas are expressed with 5 TPM or more and had ≥0.1 IPA site usage in at least one sample. IPA site usage is the relative expression of each IPA isoform with respect to the total expression of 3′UTR isoforms (all reads that fall into robust 3′UTR peaks are summed up). We only analysed IPA isoforms of protein coding genes.

Validation of IPA isoforms using external data sources. To obtain evidence of IPA isoforms from independent methods, we first used RNA-seq data obtained from the same RNA or from the same cell type to identify IPA isoforms. We used the coordinates of the IPA events obtained from 3′-seq and tested the RNA-seq read counts in windows of 100 nucleotides located upstream and downstream of the IPA peak using a GLM[2] (Extended Data Fig. 1a). The windows were separated by 51 nucleotides centred on the first nucleotide of the polyadenylation signal. Not all IPA isoforms could be tested. For example, if the defined windows overlapped with an annotated exon, the IPA event was excluded from further analysis. An IPA isoform was considered present if we detected a significant difference in read counts within the upstream and downstream windows (adjusted $P < 0.1$) using DESeq. This analysis was also used to validate CLL-gained IPA events in an independent CLL dataset.

We further regarded an IPA isoform as validated if reads that overlap with IPA peaks had at least four untemplated adenosines in the RNA-seq data and a polyadenylation signal (or one of its variants)[33] was detected within 50 nucleotides upstream of the read. In addition, we considered IPA isoforms as validated if we detected read evidence in independent 3′-seq datasets[10]. As no previous 3′-seq data exist for many of our cell types, we also included highly expressed (≥10 TPM and ≥0.1 IPA site usage) IPA isoforms with an upstream polyadenylation signal (AAUAAA and its variants)[33] in our downstream analysis.

Identification of the normal counterpart of CLL and of CLL-IPAs. Hierarchical clustering was performed on the normal human B cell subsets derived from lymphoid tissues or peripheral blood and CLL samples using RNA-seq derived mRNA expression levels (quantile normalized $\log_2$ reads per kilobase of transcript per million mapped reads (RPKM)). Genes expressed with greater than 5.5 RPKM in 75% of normal B cells or any of the CLL samples went into the analysis. The 20% most variable genes by median absolute deviation across the dataset were used for the clustering. The heat map was generated using aheatmap (http://cran.r-project.org/package=NMF) with row scaling. This analysis showed that lymphoid-tissue derived $CD5^+$ B cells are most closely related in their gene expression profile to CLL cells (Extended Data Fig. 2).

We performed hierarchical unsupervised clustering of CLL and control samples based on IPA site usage to test whether IPA site usage separates normal and malignant B cells (Extended Data Fig. 1c). The top 20% most variable genes by median absolute deviation across all the $CD5^+$ B and CLL samples were used. This analysis showed two main clusters: Four CLL samples (CLL4, CLL7, CLL11 and CLL12) clustered separately from the rest of the samples. However, within the rest of the samples, the control group ($CD5^+$ B) clustered separately. The four CLL samples that differed the most from the rest of the samples had a high number of significantly upregulated IPA isoforms (CLL high: median number of CLL-IPAs per sample, $n = 100$; range, $n = 42$–274), whereas the remaining samples had a low number of CLL-IPAs (CLL low: median, $n = 9$; range, $n = 5$–28; Extended Data Fig. 1e).

To identify CLL-upregulated IPA isoforms, we applied a GLM[1,2,34] and tested usage of each IPA isoform between the normal B cell group and each CLL sample. We only considered IPA isoforms that were significantly upregulated in CLL (FDR-adjusted $P < 0.1$, usage difference between CLL and $CD5^+$ B ≥ 0.05) and were either not or lowly expressed in $CD5^+$ B cells (TPM < 8, corresponding to 75% quantile for $CD5^+$ B TPM). This resulted in 931 significantly upregulated IPA events observed in 13 CLL samples. $n = 454$ IPA events were detected in only a single sample and were regarded as non-recurrent, whereas 477 IPA events occurred in more than one sample (≥2 out of 13), and were considered recurrent events by 3′-seq (Extended Data Fig. 1d). The recurrent events resulted in 168 recurrent CLL-IPA isoforms.

As CLL-IPAs are detectable by RNA-seq, we used an independent RNA-seq dataset containing 46 CLL samples for validation[3]. We verified up to 71% of testable IPAs by this independent method and dataset. Because of the high validation rate, we combined the two datasets ($n = 59$ CLL samples) and focused on CLL-IPAs present in more than 10% of the whole CLL sample cohort. This resulted in 330 CLL-upregulated IPA isoforms, derived from 306 genes (Supplementary Table 1). The list of 330 CLL upregulated IPA isoforms contains the 168 CLL-IPAs identified in at least 2 out of 13 3′-seq samples, but contains also CLL-IPA isoforms detected in one 3′-seq and in at least five additional RNA-seq samples (≥6 out of 59 total samples).

We detected 33 IPA events that showed significantly higher IPA site usage in $CD5^+$ B cells compared with CLL. IPA site usage was required to be higher than in 2 CLL samples (TPM <10, corresponding to 75% quantile for CLL TPM; FDR-adjusted $P < 0.1$, usage difference between CLL and $CD5^+$ B ≥ 0.05; Supplementary Table 1).

The fraction of CLL patients affected by IPA or TR mutations shown in Fig. 3c, Extended Data Figs. 7c and 8a were calculated as follows: If the CLL-IPA isoform was testable by RNA-seq, all 59 CLL samples were considered. If the CLL-IPA isoform was not being tested by RNA-seq (because, for example, the upstream exon is located too close to the IPA isoform), then only the 13 CLL samples analysed by 3′-seq were taken into account for calculating the fraction of samples with significant expression of the IPA isoform.

Cell lines. B lymphoblastoid cells (BLCL) are Epstein–Barr virus-immortalized human blood B cells[1]. MEC1 cells are malignant B cells from B-prolymphocytic leukaemia and were provided by O.A.-W. Raji and TMD8 cells are malignant B cells from lymphomas and were a gift from H.-G. Wendel. HEK293 and HEK293T cells (embryonic kidney), HeLa cells (cervical cancer) and A549 cells (lung adenocarcinoma) were purchased from ATCC. Wild-type and DICER-knockout HCT116

cells were provided by V. Narry Kim[35]. BLCL, MEC1 and Raji cells were cultured in RPMI with 20% FBS and 1% penicillin–streptomycin. HEK293, HEK293T, HeLa and A549 cells were cultured in DMEM with 10% FBS and 1% penicillin–streptomycin, whereas HCT116 cells were cultured in McCoy's medium with 10% FBS and 1% penicillin–streptomycin.

Western blotting. Cells were lysed on ice for 30 min with RIPA buffer (50 mM Tris pH 7.4, 150 mM NaCl, 1% NP-40, 1% Na-deoxycholate, 1 mM EDTA, 0.05% SDS), containing freshly added proteinase inhibitor cocktail (Thermo Scientific). For MGA, NUP98, SGK223 and DICER immunobloting, cell lysates were run using 3–8% Tris-Acetate NuPAGE gels with Tris-Acetate running buffer (Life Technologies). For CARD11, AKAP10, BAZ1B, SENP1, CUL3 and RIPK1, 4–12% Bis-Tris NuPAGE gels (Life Technologies) were run with MOPS running buffer and all other proteins were run with MES running buffer (Natural Diagnostics). The separated proteins were transferred to nitrocellulose membranes (Bio-Rad, 1620252), blocked with Odyssey Blocking Buffer (Li-Cor, 927-40000) for 1 h at room temperature, followed by incubation with primary antibodies at 4 °C overnight. After two washes using PBS and 0.1% Tween 20 (PBST), the blots were incubated with IRDye-conjugated secondary antibodies for 50 min at room temperature. After one wash with PBST and two washes with PBS, proteins were detected with Odyssey CLx imaging system (Li-Cor).

The following primary antibodies were used: anti-actin (mouse, Sigma, A4700; rabbit, Sigma, A2066), anti-AKAP10 (mouse, clone 51, Santa Cruz Biotechnology, sc-136512), anti-CARD11 (rabbit, Cell Signaling, 4440S), anti-DICER (rabbit, a gift from W. Filipowicz), anti-DNM1L (mouse, Abcam, ab56788), anti-MGA (rabbit, H-286, Santa Cruz Biotechnology, sc-382569), anti-SFRS15 (SCAF4; mouse, Abnova, H00057466-B01), anti-WSTF (BAZ1B; mouse, clone G-5, Santa Cruz Biotechnology, sc-514287), anti-NUP98 (rabbit, Novus Biologicals, NB100-93325), anti-SGK223 (mouse, Santa Cruz Biotechnology, sc-398164), anti-SENP1 (rabbit, Bethyl Labs, A302-927A-T), anti-CUL3 (rabbit, Bethyl Labs, A301-108A-T), anti-PAWR (Abcam ab92590), anti-RIPK1 (Cell Signaling 4926), anti-GAPDH (goat, V-18, Santa Cruz Biotechnology) and anti-WNT5a/b (rabbit, clone C27E8, Cell Signaling 2530). The secondary antibodies used included anti-mouse IRDye 700 (donkey, Rockland Immunochemicals, 610-730-002), anti-rabbit IRDye 680 (donkey, Li-Cor Biosciences, 926-68073), anti-rabbit IRDye 800 (donkey, Li-Cor Biosciences, 926-32213) and anti-mouse IRDye 800 (donkey, Li-Cor Biosciences, 926-32212).

RT–PCR of IPA isoforms. Total RNA was isolated using Tri reagent solution (Invitrogen AM9738) and digested with DNase I (Invitrogen AM1906). RNA was reverse transcribed using the qScript cDNA SuperMix (Quanta Biosciences 101414-106). RT–PCR reactions were carried out using purified Taq polymerase using a 50 °C annealing temperature and 30 s extension at 72 °C. The linear range of amplification was determined by independent PCRs for each primer set. Primers were designed to be intron-spanning and are listed in Supplementary Table 3.

Induction of IPA isoforms. Endogenous U2AF1, U2AF2 and hnRNPC were knocked down using pLKO-puro lentiviral vector-based shRNAs (Sigma). Virus was produced using the helper plasmids pCMV-VSVG and pCMV-dR8.2 and cells were transduced in six-well plates, selected with puromycin (2 μg ml^{-1}) for 5 days and then collected for RT–PCR or western blot analysis.

To induce IPA isoform expression of DICER, an antisense morpholino oligonucleotide (GeneTools) targeting the 5′ splice site of DICER exon 23 was added directly to sub-confluent HeLa cells at the indicated concentrations in the presence of 6 μM EndoPorter-PEG delivery peptide (GeneTools) and harvested at the indicated time points. The control morpholino was used at 12 μM concentration.

Knockdown of CARD11 full-length and IPA isoforms. Isoform-specific shRNA primers were cloned into the TRC2-pLKO-GFP plasmid using KpnI and EcoRI. Lentivirus was produced as described above and centrifuged at 25,000 r.p.m. for 1 h 45 min at 4 °C (Sorvall WX Ultracentrifuge). Pellets were resuspended and dissolved in cold PBS overnight at 4 °C. The virus titre was estimated by transducing wild-type HEK293T cells. The 12-well culture plate was coated overnight with 5 μg ml^{-1} fibronectin. TMD8 cells were spin-infected and cultivated for three days, followed by western blot analysis of FACS-sorted GFP-positive cells.

Constructs. The V5-DICER construct was obtained from J. Mendell. To generate the DICER-IPA expression plasmid, the DICER-IPA cDNA was amplified from BLCL and cloned into the pCK-V5 plasmid using the BamHI and ApaI restriction sites.

The human *MGA* cDNA (Dharmacon, clone BC136659) was used to PCR-amplify the coding region of full-length *MGA* (8,571 nucleotides plus 6 nucleotides of endogenous Kozak sequence) as well as MGA IPA (3,430 nucleotides (end of exon 9) plus GTGAGTATTAA (intronic sequence that will be translated, followed by a stop codon; see Extended Data Fig. 6a)). MGA IPA was cloned into the pcDNA3.1 expression vector (Life Technologies) using NheI and XhoI sites. GFP fused-MGA IPA was generated by inserting MGA IPA downstream of eGFP using the restriction sites BsrGI and XhoI in the pcDNA3.1-GFP vector. MGA was cloned into pcDNA3.1-GFP using Gibson Assembly Cloning (New England Biolabs) from three pieces.

The full-length *FOXN3* mRNA was amplified from BLCL cDNA. To obtain GFP–FOXN3, it was cloned into pcDNA3.1-GFP[36] using BsrGI and XhoI restriction sites. FOXN3 IPA was PCR-amplified from two fragments. Fragment 1 was amplified from BLCL cDNA and corresponds to amino acids 1–180, whereas fragment 2 was amplified from genomic DNA from PBMC and corresponds to the 32 amino acids generated from intronic sequence, followed by a stop codon. FOXN3 IPA was fused with GFP at the C terminus as described above.

Full-length CHST11 was amplified from *BLCL* cDNA, whereas CHST11 IPA was amplified from genomic DNA. Both were fused to GFP at the C terminus as described above. The integrity of all constructs was confirmed by sequencing.

Functional validation of CLL-IPAs. *CARD11 IPA.* To assess NF-κB activation, lentiviral-transduced TMD8 cells (described above) were used. Cells were fixed with 4% formaldehyde at room temperature for 15 min. After two washes with excess PBS, fixed cells were resuspended with ice-cold PBS and permeablized with 90% methanol for 20 min on ice. Cells were then washed with cold PBS twice and resuspended with the incubation buffer (PBS + 0.5% BSA). Cells were aliquoted and incubated with anti-phospho-NF-κB p65 (1:1,500 dilution, Cell Signaling 3033) for 1.5 h at room temperature. Cells were washed with incubation buffer twice and incubated with fluorochrome-conjugated secondary antibody solution (1:10,000 Alexa Fluor 647 A27040, Invitrogen) for 15 min at room temperature. After two washes with incubation buffer, cells were analysed using a FACS Calibur.

DICER IPA. Full-length V5-DICER and V5-DICER IPA were immunoprecipitated from HEK293T cells as described before[16]. In brief, 48 h after transfection, cells were washed with cold PBS and lysed with IP buffer (20 mM Tris-HCl pH 8.0, 150 mM NaCl, 1 mM EDTA, 0.5% NP-40 and 1× EDTA-free protease inhibitor (Thermo Fisher)) for 30 min on ice with occasional vortexing. The cell lysate was then centrifuged at 20,000*g* for 10 min at 4 °C and the supernatant was collected. The cell lysate was incubated with 3 μg of anti-V5 tag antibody (Invitrogen R960-25) for 30 min on ice, then 900 μg of protein G Dynabeads were added and the reaction was rotated for an additional 2 h at 4 °C. After five washes with IP buffer and twice in DICER assay buffer (20 mM Tris-HCl pH 8.0, 100 mM KCl, 0.2 mM EDTA), 90% of beads were resuspended in DICER assay buffer for miRNA cleavage assay and the remaining beads were stored in 2× Laemmli sample buffer (Sigma) for western blot analysis.

The miRNA cleavage assay was performed as described previously[16]. In brief, synthesized pre-miRNA let-7i oligo (Dharmacon) was incubated with immunoprecipitated beads prepared as described above in the enzymatic mixture (10 μl of immunoprecipitated beads in DICER assay buffer, 2 μl of 20 mM $MgCl_2$, 0.2 μl of 0.4 μM pre-miRNA, 0.1 μl of 100 mM DTT, 0.5 μl of RNaseOUT (Invitrogen) and 7.2 μl of RNase-free water) at 37 °C for 30 min with interval mix. The reaction was stopped by chilling samples on ice and analysed by northern blot. To investigate whether DICER IPA acts as a dominant-negative version of full-length DICER, different ratios of V5-DICER and V5-DICER IPA were mixed and tested with respect to miRNA cleavage.

Reaction mixtures (10 μl) were added to 10 μl RNA loading buffer (95% formamide, 0.025% SDS, 0.025% bromophenol blue, 0.025% xylene cyanol FF, 0.5 mM EDTA) and denatured at 95 °C for 5 min followed by quenching on ice. Samples were run on a 15% TBE/Urea gel followed by transfer to a Hybond-N$^+$ nylon membrane (GE Healthcare RPN303B) using a semi-dry transfer apparatus (Hoefer TE70X). After transfer, membranes were briefly dried and then UV cross-linked twice with 1,200 μJ cm^{-2} each cycle. Cross-linked membranes were pre-hybridized for 1 h at 37 °C in ULTRAhyb-Oligo hybridization buffer (Ambion AM8663) in a rotary oven. DNA probes against the intended target RNA were synthesized as oligos and labelled with γ^{32}P-ATP in the presence of T4 polynucleotide kinase (NEB M0201S) for 30 min at 37 °C. Labelled probes were purified through G-25 microspin columns containing Sephadex resin (GE Healthcare 27-5325-01). Membranes were hybridized with labelled probe overnight at 37 °C in a rotary oven. The next day, membranes were washed twice in 2× SSC/0.1% SDS for 5 min each at 37 °C followed by one wash in 0.1× SSC/0.1% SDS for 5 min at 37 °C. Membranes were exposed to phosphorimager screens and scanned.

To assess whether expression of DICER IPA influences miRNA expression in vivo, endogenous let-7 miRNA expression levels were measured by northern blot analysis of total RNA (22 μg) from wild-type and DICER knockout HCT116 cells. DICER knockout HCT116 cells were transfected with different amounts of V5-DICER and V5-DICER IPA. Cells were harvested 3 days after transfection with Lipofectamine 2000 to assess DICER protein expression and corresponding endogenous let-7 levels.

FOXN3 IPA. The fork-head domain of FOXN3 is necessary for transcriptional repression of FOXN3 target genes. Thus, truncation of the fork-head domain predicts derepression of the target genes. Known target genes are *PIM2* and *MYC*[20,37]. MEC1 cells were nucleofected with pcDNA 3.1 vector containing GFP, GFP–FOXN3 or GFP–FOXN3 IPA using SF Cell Line 4D-Nucleofector X Kit (Lonza, Program FF-120). After 48 h, GFP$^+$ cells were FACS sorted, lysed immediately (Cells-to-cDNA II Kit, Ambion) and RNA was extracted. cDNA was synthesized

by qScript cDNA SuperMix (Quanta Biosciences) and quantitative PCR was performed using FastStart universal SYBR green master mix (Roche) on a 7900HT Fast Real-Time PCR System (Applied Biosystems). The experiment was performed from five biologically different replicates.

MGA IPA. Raji cells were nucleofected with pcDNA3.1 vector containing GFP, GFP–MGA or GFP–MGA IPA using Cell Line Nucleofector Kit V (Lonza, Program M-013). After 48 h, GFP$^+$ Raji cells were FACS-sorted and lysed immediately in lysis buffer (Cells-to-cDNA II Kit, Ambion) and RNA was extracted. cDNA synthesis and qRT–PCR was as described for FOXN3. qRT–PCR was done in technical triplicates from three biologically different experiments. MYC target genes were previously published[38,39]. E2F-binding sites in MYC target genes were identified using the Encode Transcription Factor ChIP–seq track, or they were previously described[19,39–41]. T-boxes were described for *ATF4* and *CDKN1B*[42,43].

CHST11 IPA. 3′-seq data were used to identify overexpressed WNT proteins in CLL cells compared to normal B cells. The expression of WNT was validated in MEC1 cells by qRT–PCR. WNT5B was the WNT with the highest expression in MEC1 cells.

For WNT detection in media, MEC1 cells stably expressing GFP, GFP–CHST11 or GFP–CHST11 IPA were counted and washed once with RPMI without FCS. Twenty million cells were cultured in 10 ml RPMI plus 1% pen–strep in one 10-cm culture dish. After 18 h, conditioned medium was collected by centrifugation at 280*g* for 5 min and passed through a 0.45-μm filter. The supernatant was concentrated by an Amicon Ultra-4 centrifugal filter (Millipore, UFC800324) at 3,000*g* at 10 °C for 2 h. The concentrated medium (~50 μl) was collected and subjected to western blot analysis using anti-WNT5a/b antibody (Cell Signaling 2530). The corresponding cell pellets were also collected for western blot analysis.

To assess paracrine WNT activity in MEC1 cells expressing CHST11 IPA, MEC1 cells were nucleofected with pcDNA3.1 vector containing GFP, GFP–CHST11 or GFP–CHST11 IPA. After 24 h, GFP$^+$ cells were FACS sorted and cultivated for three days. The conditioned medium was collected and added to HEK293T cells which were transiently transfected with a WNT reporter plasmid (Addgene 12456, M50, Super 8x TOPFlash) or WNT reporter control plasmid with mutated TCF/LEF binding sites (Addgene 12457, M51, Super 8x TOPFlash mutant)[44]. The conditioned medium was added 24 h after transfection. Luciferase activity was measured 24 h after the addition of conditioned medium using a Glomax 96 Microplate Luminometer as described previously[45].

Intersection of somatic mutations in CLL with IPA. CLL RNA-seq samples ($n = 44$) with available somatic DNA mutation and prognostic data were available to us to map IPA isoform expression[3]. The somatic mutations were obtained using exome sequencing that included extended exon boundaries[46]. We intersected the occurrence of somatic mutations with IPA isoforms in these samples. We focused on truncating mutations (nonsense mutations, frame-shift mutations and splice-site mutations) in expressed genes as they were likely to have a similar outcome as IPA.

The IGVH status of CLL samples was assessed at MSKCC for the CLL samples studied by 3′-seq. The IGVH status of 44 RNA-seq samples was published[3].

Positions of TR mutations. The positions of TR mutations in CLL were obtained from the published CLL somatic mutation datasets[3,7,8]. The positions of TR mutations in solid cancers of TSGs and of genes targeted by CLL-IPAs were obtained from the MSK cbio portal (date of reference, 23 February 2018, containing >86,000 cancer samples with 97% derived from solid tumours)[4]. The position with the highest number of TR mutations was used (hot spot) and is indicated by the symbol. The symbol is lacking if the genes had TR mutations without a hot spot.

Number of amino acids of full-length or IPA-generated truncated proteins. To calculate the number of amino acids of full-length proteins, we used the longest Ref-seq annotated mRNA isoform, obtained the number of coding nucleotides and divided this number by three to obtain the total number of amino acids. To calculate the number of amino acids of the IPA-generated truncated proteins we counted the number of nucleotides from the start codon to the end of the exon located upstream of the IPA isoform and divided this number by three to obtain the number of retained amino acids. This number also provided information about the reading frame of the protein at the exon/intron junction located upstream of the IPA isoform. We then used the correct reading frame and translated the intronic nucleotides until an in-frame stop codon was detected. The amino acids translated from intronic sequence were added to the retained amino acids to obtain the size of the IPA-generated truncated proteins.

The fraction of retained CDR is the number of amino acids retained (up to the end of the exon located upstream of the IPA isoform) divided by the number of amino acids calculated from the longest mRNA isoform encoding the full-length protein.

Identification of known and novel TSGs. For known TSGs, we used the 301 TSGs reported by Davoli et al.[5] that were expressed in CLL samples. Davoli used a computational method (TUSON Explorer) to predict 301 TSGs from genomic sequencing data obtained from more than 8,200 cancers (>90% are derived from solid tumours).

For novel TSGs, we used the data from the MSK cbio portal (see above). It was previously reported that the variable with the highest predictive power for TSGs was the proportion of TR mutations to all mutations[5]. We calculated this ratio for the 190 genes that generated CLL-IPAs in more than 20% of samples and identified a bimodal distribution with a separation point at 12% TR mutations to all mutations. The genes that generated CLL-IPAs in more than 20% of samples and had a TR mutation rate ≥ 12% in the data from MSK cbio portal were called novel TSG candidates (Supplementary Table 2).

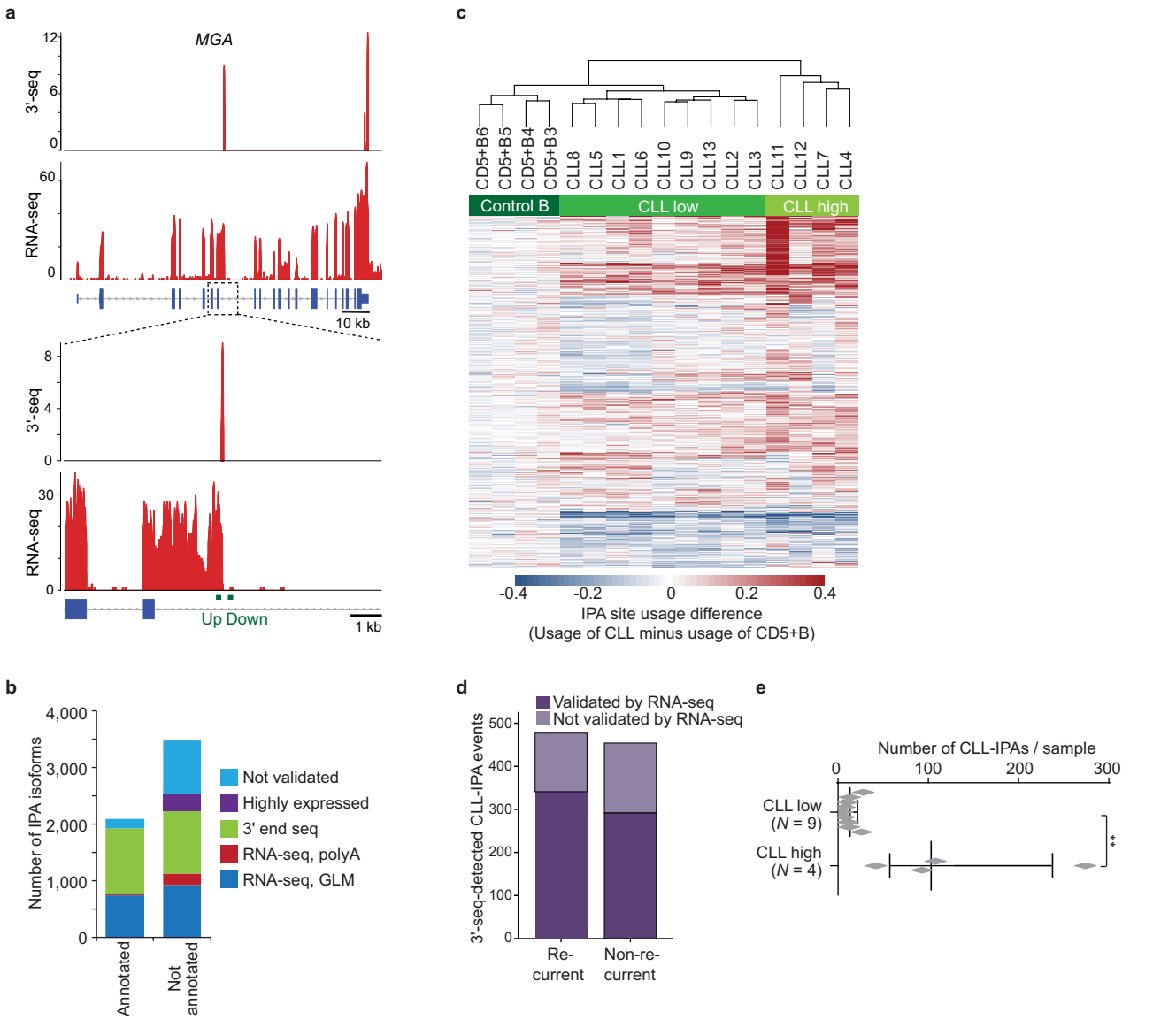
To assess whether known TSGs are enriched among CLL-IPAs a χ^2 test was performed. To exclude that this association occurred by chance, five control lists containing genes with similar coding region length and expression were generated and tested for enrichment of TSGs.

Others statistical methods. To perform enrichment statistics, we used a χ^2 test and calculated the *P* value using a two-sided Fisher's exact test. To assess the functional differences between full-length proteins and IPA-generated truncated proteins (MGA and FOXN3), we used a two-sided *t*-test for independent samples. When comparing three groups (CARD11 and CHST11), a two-sided Kruskal–Wallis test was used. For subsequent pair-wise comparisons, a two-sided Mann–Whitney *U*-test was applied and the *P* values were adjusted with Bonferroni multiple testing correction. For all other tests that assessed the differences of features between two groups, we used a two-sided Mann–Whitney *U*-test. To investigate the spatial relationship between the IPA-generated truncated proteins and hot spot TR mutations, we performed a two-sided Wilcoxon rank-sum test.

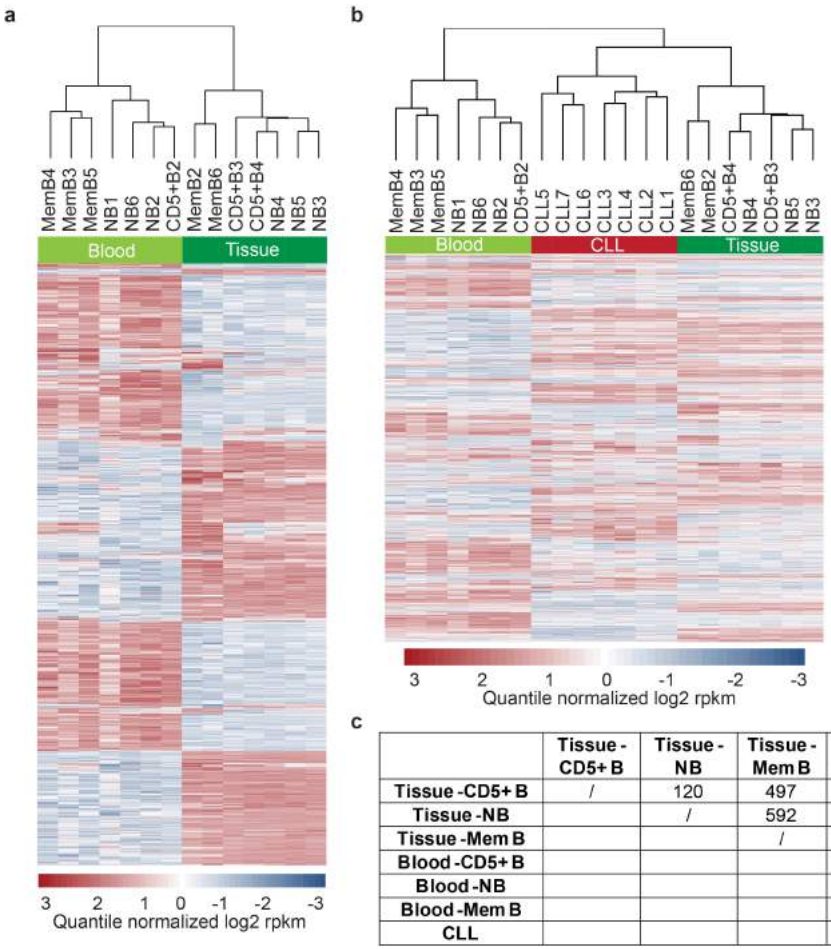
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. All 3′-seq and RNA-seq data generated and analysed for this study have been deposited in the Gene Expression Omnibus (GEO) database under accession numbers GSE111310 and GSE111793. The code to analyse the data are available at https://bitbucket.org/leslielab/apa_2018/ and the processed data are available in Supplementary Table 1 (for Figs. 1b–d, 2a, 4a, Extended Data Figs. 3 and 4) and Supplementary Table 2 (for Extended Data Fig. 8a), and in the Source Data files (for Figs. 1e, 2c, e, 3a, c, 4b–d, g, Extended Data Figs. 2c, 6j, 7c and 8a). Data on DNA mutations from patients with CLL were provided by D. A. Landau and need to be requested from him. The mutation data on solid cancers were obtained through the MSK cbio portal. The data can be accessed at http://www.cbioportal.org.

29. Béguelin, W. et al. EZH2 is required for germinal center formation and somatic EZH2 mutations promote lymphoid transformation. *Cancer Cell* **23**, 677–692 (2013).
30. Ranzani, V. et al. The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4. *Nat. Immunol.* **16**, 318–325 (2015).
31. Hoek, K. L. et al. A cell-based systems biology assessment of human blood to monitor immune responses after influenza vaccination. *PLoS ONE* **10**, e0118528 (2015).
32. Trimarchi, T. et al. Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia. *Cell* **158**, 593–606 (2014).
33. Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33**, 201–212 (2005).
34. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).
35. Kim, Y. K., Kim, B. & Kim, V. N. Re-evaluation of the roles of DROSHA, Export in 5, and DICER in microRNA biogenesis. *Proc. Natl Acad. Sci. USA* **113**, E1881–E1889 (2016).
36. Berkovits, B. D. & Mayr, C. Alternative 3′ UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522**, 363–367 (2015).
37. Karanth, S., Zinkhan, E. K., Hill, J. T., Yost, H. J. & Schlegel, A. FOXN3 regulates hepatic glucose utilization. *Cell Rep.* **15**, 2745–2755 (2016).
38. Li, Z. et al. A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl Acad. Sci. USA* **100**, 8164–8169 (2003).
39. Zeller, K. I. et al. Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proc. Natl Acad. Sci. USA* **103**, 17834–17839 (2006).
40. Ren, B. et al. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev.* **16**, 245–256 (2002).
41. Taubert, S. et al. E2F-dependent histone acetylation and recruitment of the Tip60 acetyltransferase complex to chromatin in late G1. *Mol. Cell. Biol.* **24**, 4546–4556 (2004).
42. Jenner, R. G. et al. The transcription factors T-bet and GATA-3 control alternative pathways of T-cell differentiation through a shared set of target genes. *Proc. Natl Acad. Sci. USA* **106**, 17876–17881 (2009).
43. Jezkova, J. et al. Brachyury regulates proliferation of cancer cells via a p27Kip1-dependent pathway. *Oncotarget* **5**, 3813–3822 (2014).
44. Veeman, M. T., Slusarski, D. C., Kaykas, A., Louie, S. H. & Moon, R. T. Zebrafish prickle, a modulator of noncanonical Wnt/Fz signaling, regulates gastrulation movements. *Curr. Biol.* **13**, 680–685 (2003).
45. Mayr, C. & Bartel, D. P. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**, 673–684 (2009).
46. Landau, D. A. et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).



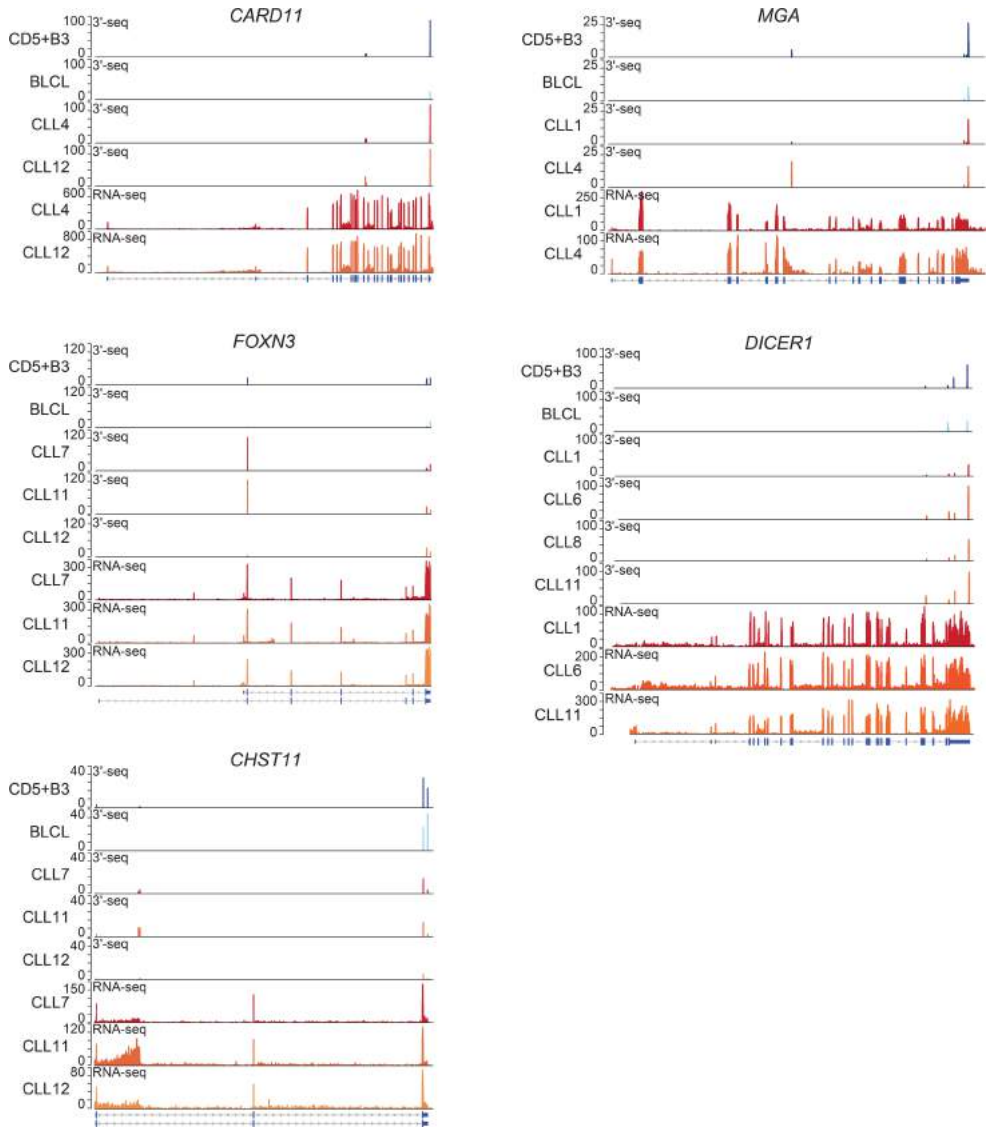
Extended Data Fig. 1 | Validation of IPA isoforms by independent methods and identification of CLL-IPAs used for further analysis. **a**, RNA-seq data were used to validate the presence of IPA isoforms using a GLM. Within two 100-nucleotide windows (green bars) separated by 51 nucleotides and located up- and downstream of the IPA peak, the RNA-seq reads were counted. The IPA peak was considered validated if adjusted $P < 0.1$ (see Methods). Out of $n = 5{,}587$ tested IPA isoforms, $n = 1{,}662$ were validated by this method. Shown is *MGA* as a representative example. **b**, As only a fraction of IPA isoforms were validated by the method from **a**, additional methods were used to obtain independent evidence for the presence of the IPA isoforms. Independent evidence was obtained using untemplated adenosines from RNA-seq data or through the presence of the IPA isoform in other 3′-seq protocols[10]. As the majority of immune cell types used in this study have not been investigated using other 3′-seq protocols and IPA isoform expression is cell type-specific[2], highly expressed IPA isoforms (>10 TPM) were not excluded from further analysis even if no read evidence was found by other protocols. **c**, Hierarchical clustering based on IPA site usage separates the 3′-seq dataset into four groups. It separates $CD5^+$ B from CLL samples and clusters CLL samples into three different groups. Shown is the usage difference of the 20% most variable IPA isoforms across the dataset ($n = 342$). Four out of thirteen CLL samples cluster away from the rest of the samples and are characterized by a high number of IPA isoforms (CLL high). **d**, The GLM (FDR-adjusted $P < 0.1$, IPA usage difference ≥ 0.05, IPA isoform expressed in $CD5^+$ B < 8 TPM) identified 477 recurrent (significantly upregulated in at least 2 out of 13 CLL samples by 3′-seq) and 454 non-recurrent (significantly upregulated in 1 out of 13 CLL samples by 3′-seq). IPAs were validated in an independent RNA-seq dataset containing 46 new CLL samples. Among the recurrent IPAs, 71% of testable IPAs were verified using another GLM (see **a**). Among the non-recurrent IPAs, 64% of testable IPAs were verified. **e**, Plotting the number of CLL-IPAs per sample separates the CLL samples investigated by 3′-seq into two groups: 4 out of 13 samples generate a high number of CLL-IPAs (CLL high, median of CLL-IPAs/sample, $n = 100$, range, 42–274), whereas the rest of the samples generate lower numbers (CLL low, median, $n = 9$, range, 5–28). Centre bar denotes the median; error bars denote the interquartile range. $^{**}P = 0.003$, two-sided Mann–Whitney *U*-test.



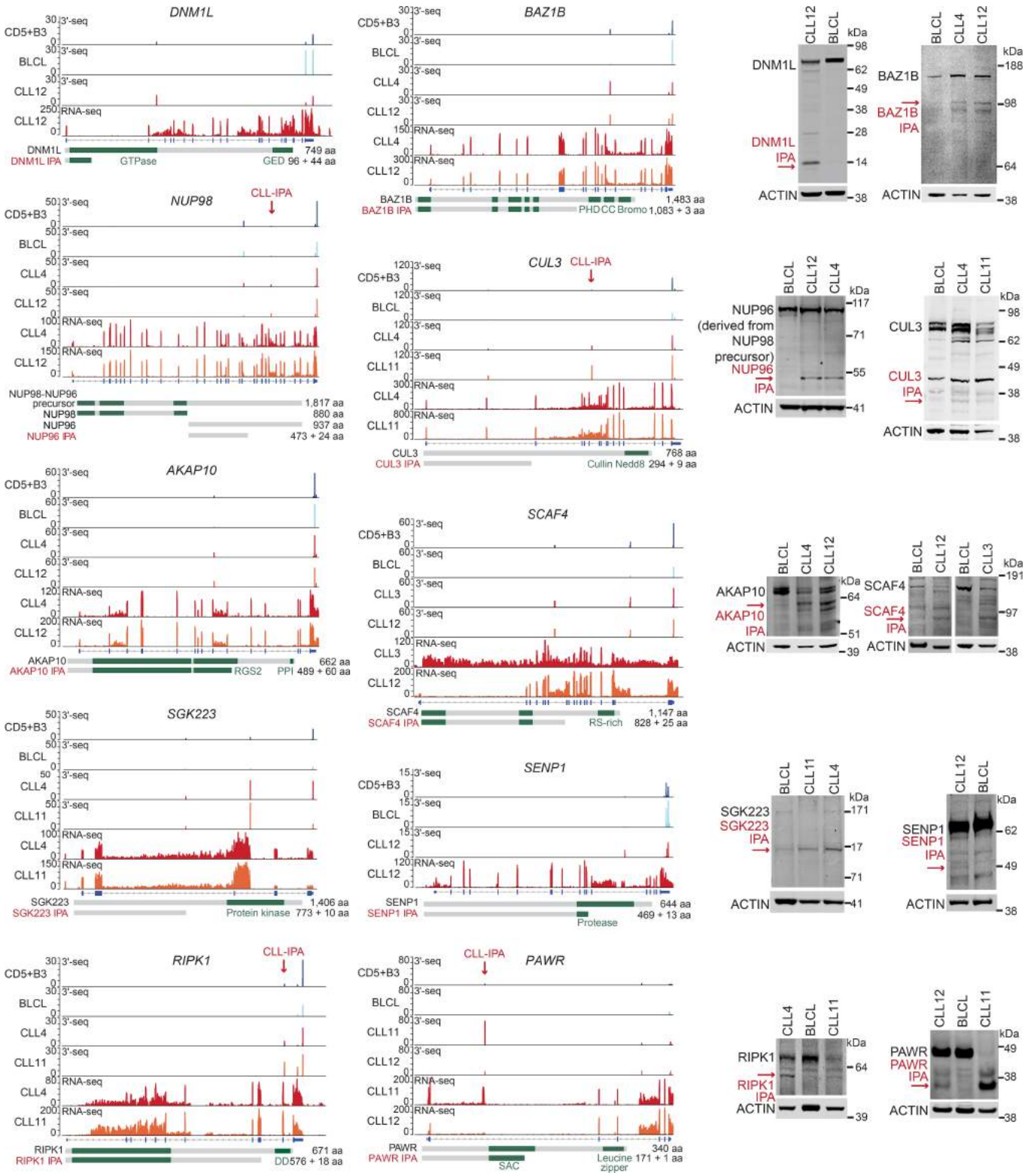
c

	Tissue - CD5+ B	Tissue - NB	Tissue - Mem B	Blood - CD5+ B	Blood - NB	Blood - Mem B	CLL
Tissue -CD5+ B	/	120	497	3,565	1,202	1,381	586
Tissue -NB		/	592	4,805	2,149	2,305	867
Tissue -Mem B			/	2,982	1,620	1,266	679
Blood -CD5+ B				/	3	77	1,605
Blood -NB					/	149	2,211
Blood -Mem B						/	2,298
CLL							/

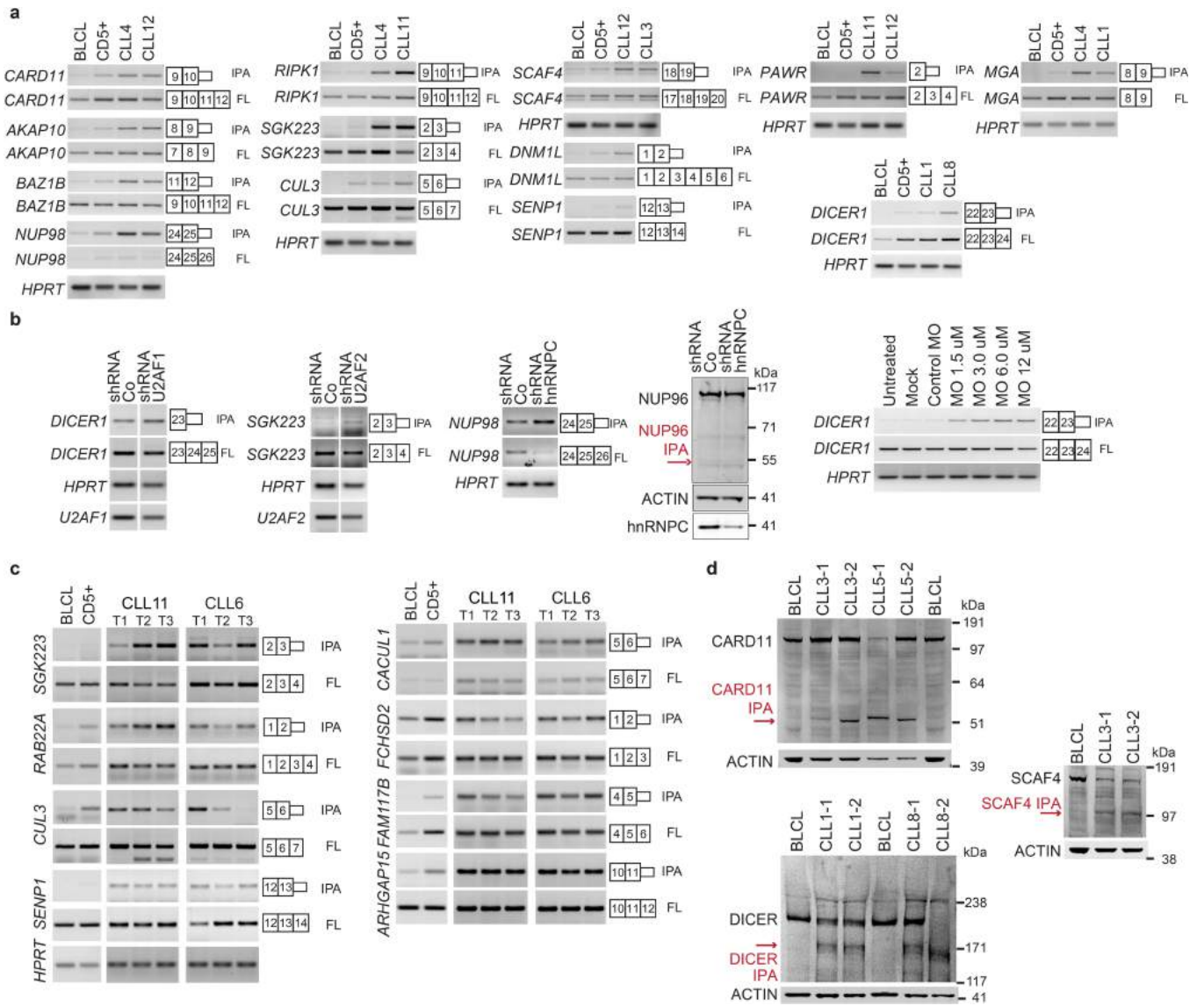
Extended Data Fig. 2 | The normal B cell counterpart of CLL cells are CD5$^+$ B cells derived from lymphoid tissue. a, Hierarchical clustering of normal human B cells (naive B (NB), memory B (MemB) and CD5$^+$ B) derived from lymphoid tissues or peripheral blood based on mRNA expression obtained from RNA-seq. The heat map shows the 20% most variable genes across the dataset ($n = 1{,}887$). The gene expression profiles of B cell subsets derived from peripheral blood or lymphoid tissue differ substantially, although the same markers were used for purification. **b**, As in **a**, but RNA-seq data from CLL samples were added to the analysis. The heat map shows the 20% most variable genes across the dataset ($n = 2{,}078$). CLL samples cluster with tissue-derived and not with blood-derived normal immune cells. **c**, Number of all differentially expressed genes from the analysis shown in **b**.



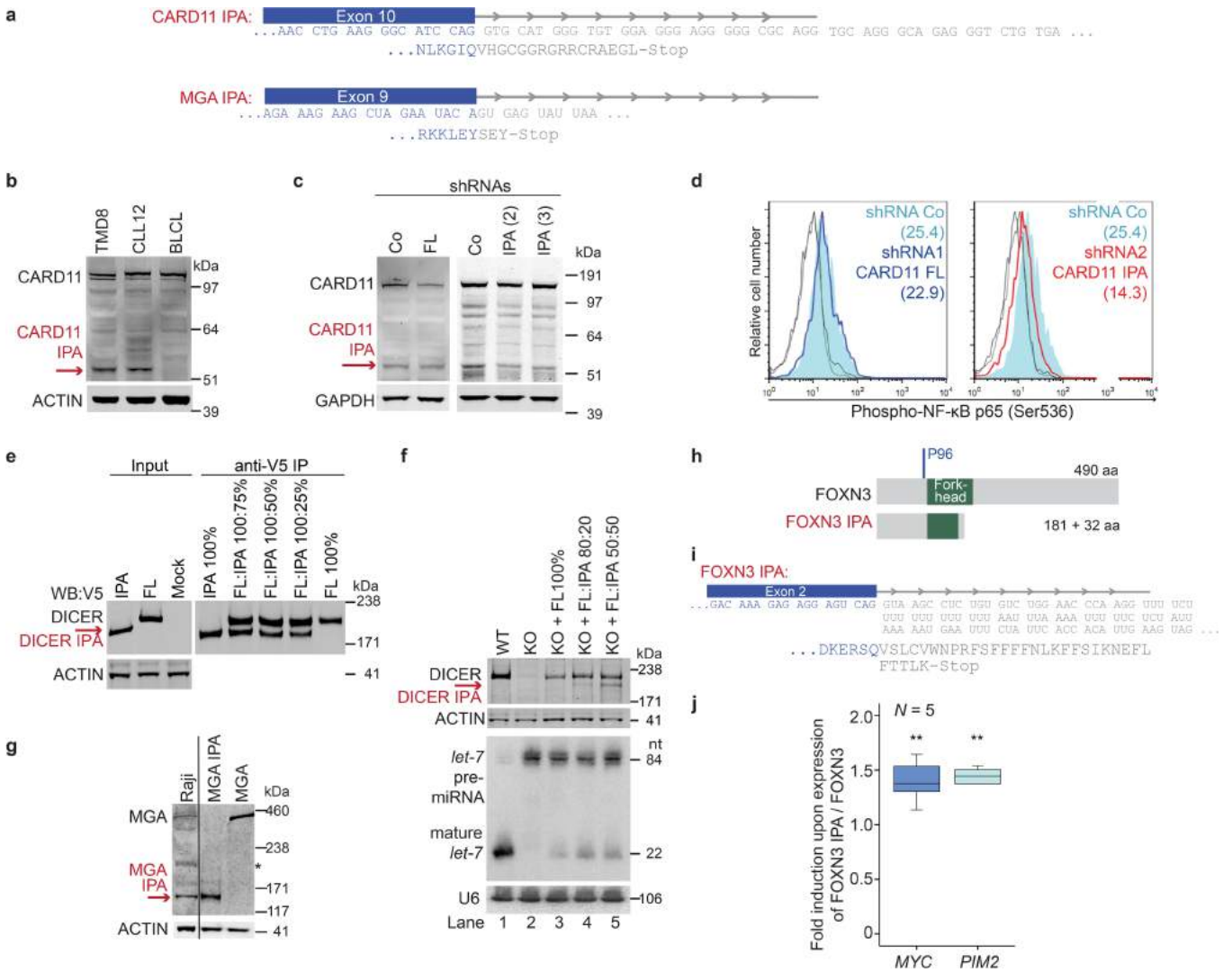
Extended Data Fig. 3 | The 3′-seq and RNA-seq tracks of functionally validated CLL-IPAs. Five CLL-IPAs were functionally validated. Their 3′-seq and RNA-seq tracks are shown here and in Fig. 2a. Data are shown as in Fig. 1b. The corresponding RT–PCRs are shown in Extended Data Fig. 5a.



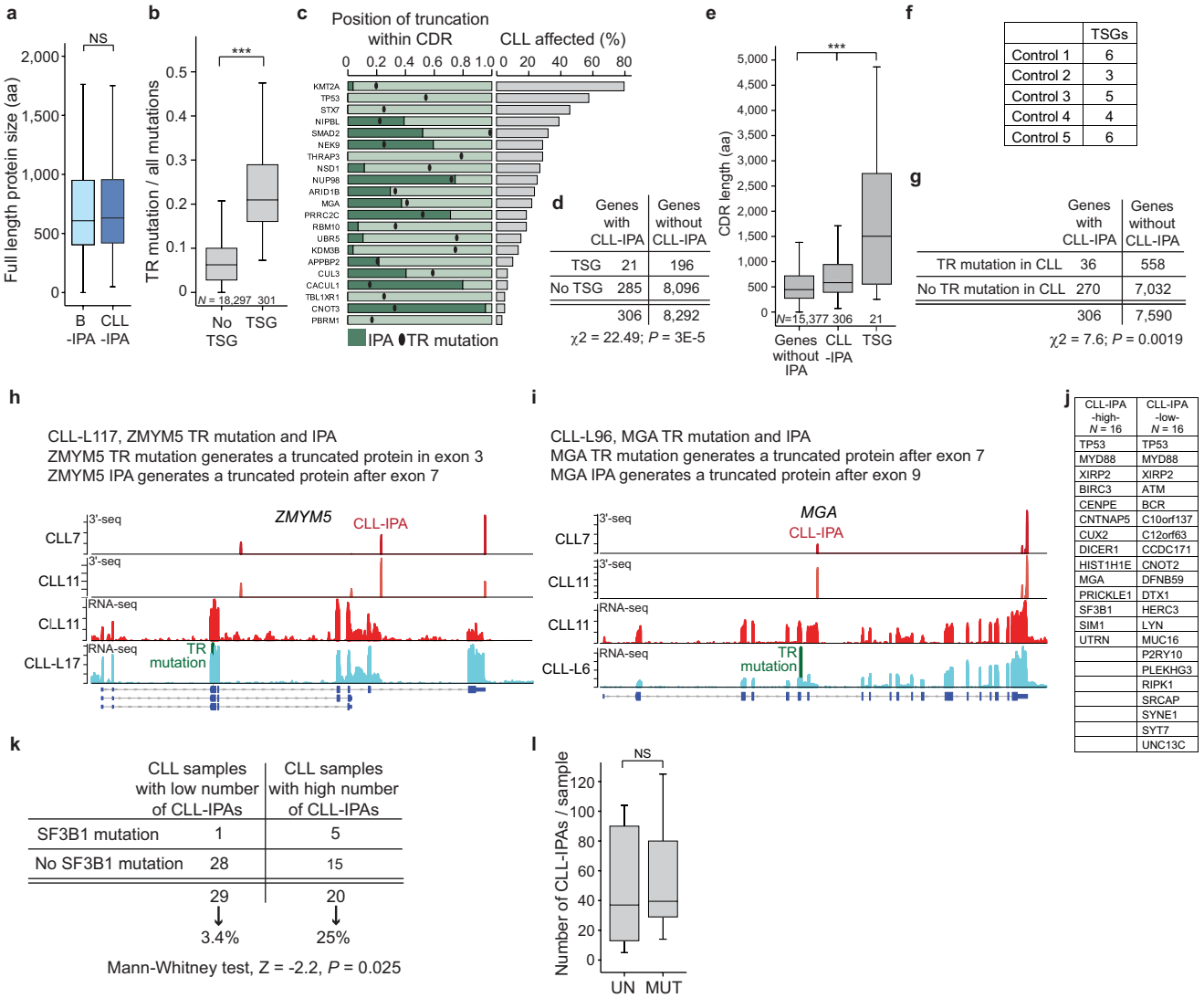
Extended Data Fig. 4 | CLL-IPAs generate truncated mRNAs and proteins. Gene models and western blots of 10 candidates depicted as in Figs. 1b and 2a show that CLL B cells generate full-length and IPA-generated truncated proteins. BLCL were used as control B cells and were included in the 3′-seq tracks. Actin was used as loading control on the same blots. For gel source data see Supplementary Fig. 1.



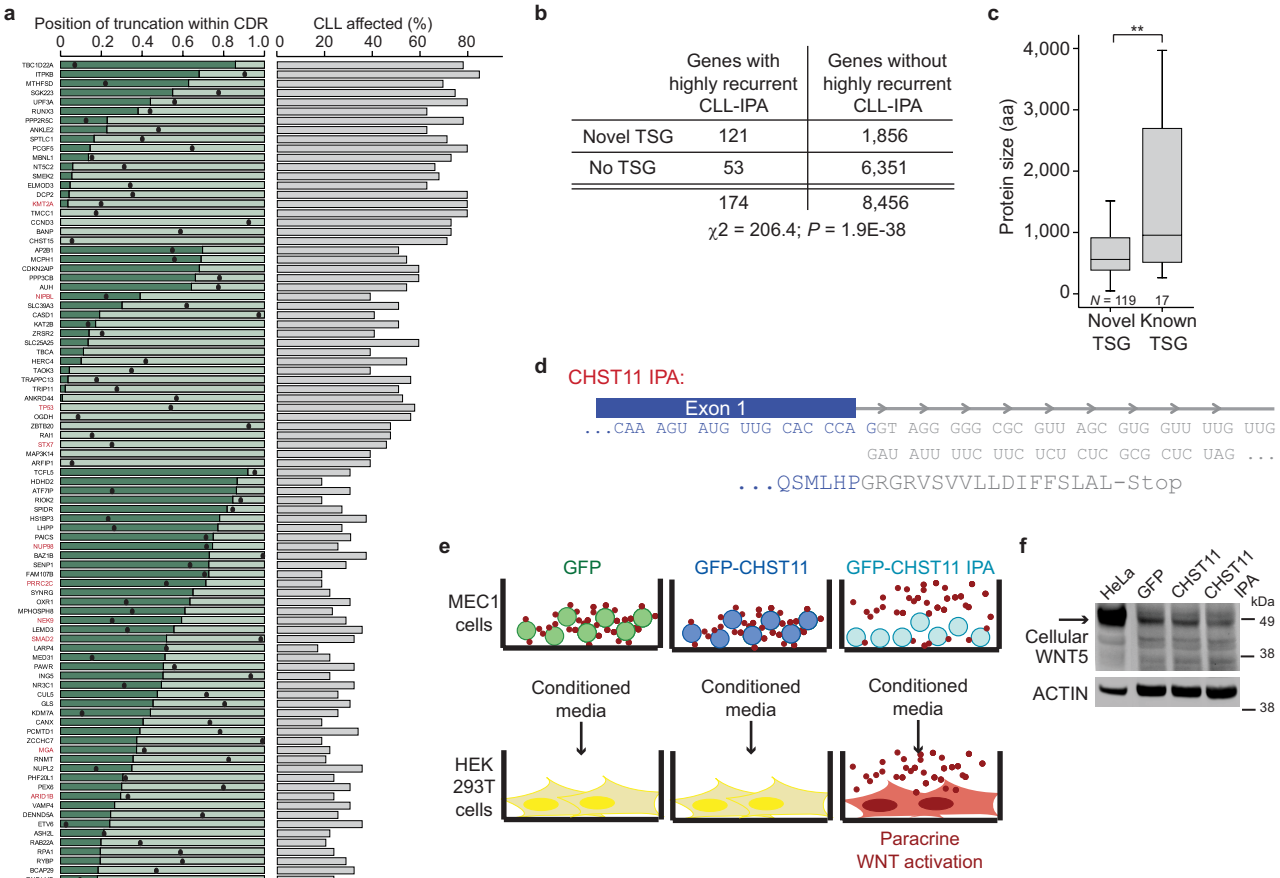
Extended Data Fig. 5 | Validation of the IPA-generated truncated mRNAs and validation of their stable expression over time. a, Detection of full-length and IPA-generated truncated mRNAs by RT–PCR in normal B cells ($CD5^+$ B, BLCL) and CLL cells used in the western blot validations shown in Fig. 2a and Extended Data Fig. 4. All experiments were performed twice with similar results. Primers to amplify the mRNA isoforms are located in the first and last exons shown in the gene models and are listed in Supplementary Table 3. *HPRT* was used as loading control. **b**, Induction of truncated mRNAs and proteins through shRNA-mediated knockdown of splicing factors. All experiments were performed twice with similar results. U2AF1 was knocked down in HeLa cells, U2AF2 was knocked down in HEK293 cells and hnRNPC was knocked down in A549 cells. Shown as in **a**, except for NUP96, which is shown as in Extended Data Fig. 4. NUP96 is derived from NUP98 precursor. Induction of DICER1 IPA by transfection of increasing amounts of anti-sense morpholinos (MO) directed against the 5′ splice site of intron 23 of *DICER1* in HeLa cells. Shown are RT–PCRs. **c**, RT–PCRs, performed once, on expression of full-length and IPA isoforms for eight CLL-IPAs in samples from two patients with CLL and control B cells ($CD5^+$ B, BLCL). The samples were collected over a time interval of over 6 years. CLL11: T1, 17 months after diagnosis, T2, 24 months, T3, 44 months; CLL6: T1, 16 months, T2, 49 months, T3, 91 months (42 months after treatment). Samples from all time points (except CLL6, T3) were obtained from untreated patients. The primers for amplifications of the products were located in the first and last exons shown in the gene models and are listed in Supplementary Table 3. Expression of *HPRT* serves as loading control. The same gel picture of HPRT is shown in Fig. 3b for CLL samples and in **a**, far right panel, for BLCL and $CD5^+$ control samples. All tested CLL-IPA isoforms were detectable at several time points during the course of the disease. Compared with $CD5^+$ B cells, expression of FCHSD2 IPA was not significantly upregulated in CLL. **d**, Western blots of full-length and IPA-generated truncated proteins from CARD11, DICER and SCAF4. All experiments were performed twice with similar results. Actin was used as loading control. Shown are samples from normal B cells (BLCL) and two patients with CLL, both at two different time points 0.5–10 months apart. For gel source data, see Supplementary Fig. 1.



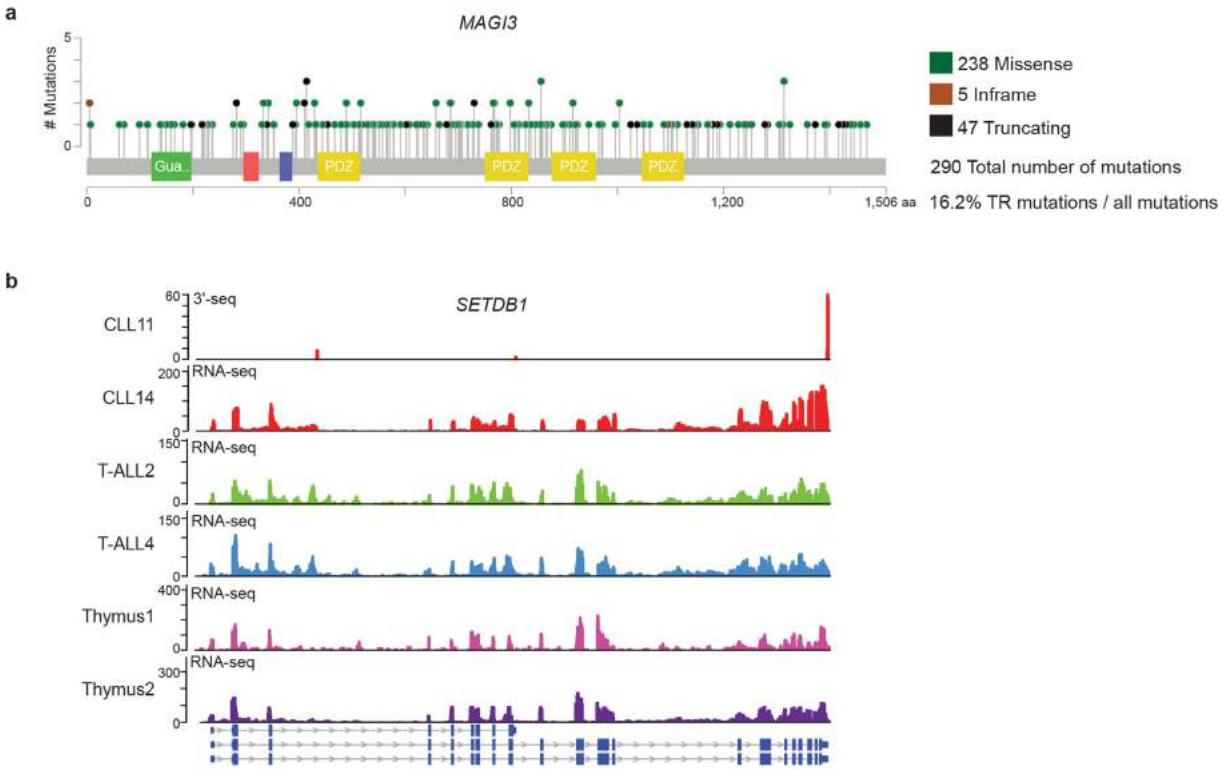
Extended Data Fig. 6 | IPA-generated truncated proteins resemble the protein products of truncating DNA mutations and have cancer-promoting properties. a, CARD11 IPA results in translation of intronic nucleotides (grey) until an in-frame stop codon is encountered. This results in the generation of 16 new amino acids (grey) downstream of exon 10. In the case of MGA IPA, three new amino acids downstream of exon 9 are generated. **b**, Western blot showing that TMD8 cells express similar amounts of CARD11 IPA as CLL samples. The western blot is shown as in Fig. 2a and was performed twice. Actin was used as loading control. **c**, Western blot (as in **b**) showing full-length CARD11 as well as CARD11 IPA in TMD8 cells expressing a control shRNA (Co), an shRNA that exclusively knocks down the full-length protein and two different shRNAs that exclusively knock down the CARD11 IPA isoform. The experiment was performed twice with similar results. GAPDH was used as loading control. **d**, Endogenous phospho-NF-κB p65 levels were measured by FACS in TMD8 cells expressing the indicated shRNAs from **c**. Mean fluorescent intensity values are shown in parentheses in FACS plots of a representative experiment out of three. **e**, Immunoprecipitation of V5-DICER or V5-DICER IPA from HEK293T cells using an anti-V5 antibody. The experiment was performed twice with similar results. 2.5% of input was loaded. **f**, The extent of miRNA processing depends on the expression levels of full-length DICER, but not IPA. Shown are wild-type (WT) and DICER knockout (KO) HCT116 cells. Re-expression of different amounts of full-length DICER1 protein in the knockout cells (measured by western blot of DICER1 in the top panel) results in different levels of endogenous let-7 expression (measured by northern blot in the bottom panel; compare lanes 3 and 4). Expression of DICER IPA has no influence on miRNA processing (compare lanes 4 and 5). Actin and U6 were used as loading controls. The experiment was performed twice with similar results. **g**, Western blot of MGA. MGA and MGA IPA were cloned and expressed in HEK293T cells to confirm the predicted protein size. The experiment was performed twice with similar results. Shown is also the endogenous MGA expression in Raji cells. Actin was used as loading control on the same blot. Asterisk denotes an unspecific band. **h**, Protein models of full-length and FOXN3 IPA are shown as in Fig. 2b. The IPA-generated protein truncates the fork-head domain and is predicted to lose the repressive activity. **i**, As in **a**, but for FOXN3. FOXN3 IPA generates 32 new amino acids downstream of exon 2. **j**, FOXN3 IPA significantly derepresses expression of the oncogenic targets *MYC* and *PIM2*. Fold change in mRNA level of endogenous genes in MEC1 B cells after transfection of GFP–FOXN3 IPA compared with transfection of full-length GFP-FOXN3. *HPRT*-normalized values are shown as box plots (as in Fig. 1e) from $n = 5$ biologically independent experiments, each performed in technical triplicates. $^{**}P = 0.002$, two-sided t-test for independent samples. For gel source data, see Supplementary Fig. 1.



Extended Data Fig. 7 | Inactivation of TSGs by CLL-IPAs independently of DNA mutations. a, The distribution of full-length protein size of genes that generate CLL-IPAs ($n = 306$) and B-IPAs ($n = 2,690$) is shown in amino acids. Box plots are as in Fig. 1e. $P = 0.87$, two-sided Mann–Whitney *U*-test. **b**, TR rate (ratio of TR mutations compared to total mutations) is shown for known TSGs obtained previously[5]. Box plots are as in Fig. 1e. $P = 1 \times 10^{-155}$, two-sided Mann–Whitney *U*-test. **c**, Known TSGs, obtained previously[5] that are targeted by CLL-IPAs ($n = 21$) are shown. Dark green bars indicate the fraction of retained CDRs for each IPA-generated protein. Black dots indicate the hot spot positions of TR mutations obtained from MSK cbio portal. CLL-IPAs mostly occur upstream or within 10% (of overall amino acid length) of the mutations. $P = 0.04$, two-sided Wilcoxon rank-sum test. **d**, Contingency table for enrichment of TSGs among genes that generate CLL-IPAs. *P* value was obtained from two-sided Fisher's exact test. TSGs were obtained previously[5]. **e**, TSGs and genes that generate CLL-IPA isoforms have longer CDRs than genes that do not generate IPA isoforms. Box plots are as in Fig. 1e. $P = 1 \times 10^{-80}$, two-sided Kruskal–Wallis test. **f**, Five control gene lists ($n = 306$, each) with a similar size distribution as CLL-IPAs and expressed in CLL were tested for enrichment of TSGs. Shown is the number of TSGs found. A χ^2 test did not show a significant enrichment of TSGs among the control genes. **g**, Contingency table for enrichment of TR mutation genes in CLL among genes that generate CLL-IPAs. *P* value was obtained from two-sided Fisher's exact test. **h**, ZMYM5 is truncated by a TR mutation and an IPA isoform in the same patient, but the aberrations are predicted to result in different truncated proteins. A 10-bp deletion in exon 3 results in a frameshift leading to the generation of a truncated ZMYM5 protein, whereas ZMYM5 IPA (not yet annotated) produces a truncated protein containing 352 more amino acids in the same patient. The genes shown in **h** and **i** are the only genes with simultaneous presence of a TR mutation and CLL-IPA out of $n = 268$ tested. The position of the TR mutation is indicated in green. CLL7 and CLL11 3′-seq and RNA-seq tracks are shown for comparison reasons. **i**, MGA is truncated by a TR mutation and an IPA isoform in the same patient. The TR mutation affects the 5′ splice site of intron 7, thus generating two additional amino acids downstream of exon 7, whereas the IPA isoform encodes a truncated MGA protein containing three more amino acids downstream of exon 9. Mutation and 3′-seq analysis were performed once. CLL7 and CLL11 are shown for comparison reasons. **j**, Shown are additional recurrent ($n > 1$) DNA mutations found by exome sequencing of CLL patient samples stratified by a high or low number of CLL-IPAs per patient. Only the top and bottom 16 samples with high or low CLL-IPAs are shown to normalize the number of samples analysed. This analysis is only descriptive and no test was performed. **k**, Significant enrichment of SF3B1 mutations in the group of CLL samples with abundant CLL-IPA isoforms. Two-sided Mann–Whitney *U*-test was performed. **l**, Abundance of CLL-IPAs is not associated with IGVH mutational status. Shown is the number of CLL-IPAs per sample for patients with mutated (MUT, $n = 30$) or unmutated (UN, $n = 21$) IGVH genes. Box plots are as in Fig. 1e. $P = 0.4$, two-sided Mann–Whitney *U*-test.



Extended Data Fig. 8 | Novel TSG candidates and validation of CHST11 IPA as cancer-promoting isoform. a, As in Fig. 3c, but shown are known (red gene names) and novel TSG candidates (black gene names) among the abundant CLL-IPAs. CLL-IPAs seem to inactivate these genes as they mostly occur upstream or within 10% (of overall amino acid length) of the mutations. $P = 1 \times 10^{-8}$, two-sided Wilcoxon rank-sum test performed on all 136 TSGs; $P = 1 \times 10^{-8}$, two-sided Wilcoxon rank-sum test performed on the novel TSGs, $n = 119$. Position of the TR mutation was determined using the data obtained from the MSK cbio portal and indicates the hot spot mutation. Right, the fraction of CLL samples affected represents the fraction of CLL samples (out of 59) with significant expression of the IPA isoform. Genes were included if they were affected in at least 20% of samples investigated either by 3′-seq or RNA-seq. **b**, Contingency table for enrichment of novel TSGs among highly recurrent CLL-IPAs. *P* value was obtained from two-sided Fisher's exact test. **c**, TSGs have larger protein sizes. Box plots are as in Fig. 1e. $^{**}P = 0.005$, two-sided Mann–Whitney *U*-test. The increased overall mutation rate of known TSGs correlates with larger protein size. $P = 1 \times 10^{-6}$, Spearman's correlation coefficient, $r = 0.74$. **d**, CHST11 IPA generates 18 new amino acids (grey) downstream of exon 1. **e**, Experimental set-up to measure paracrine WNT activity produced by MEC1 B cells either expressing GFP, GFP–CHST11 or GFP–CHST11 IPA and using a WNT reporter expressed in HEK293T cells. Primary CLL cells and the CLL cell line MEC1 express several WNTs, including WNT5B. In the presence of CHST11 WNT (red dots) binds to sulfated proteins on the surface of WNT producing cells, whereas WNT is secreted into the medium in the presence of CHST11 IPA. WNT-conditioned medium activates a WNT reporter in HEK293T cells. This set-up refers to Fig. 4f, g. **f**, Western blot, performed once, for WNT5 shown as in Fig. 4f, but including HeLa cells as positive control for WNT5 expression. Actin was used as loading control on the same blot.



Extended Data Fig. 9 | Cancer-upregulated IPA isoforms are also detected in breast cancer and T-ALL. a, *MAGI3* is a TSG that is preferentially targeted by IPA in breast cancer[27]. Shown is the mutation profile obtained from MSK cbio portal. **b**, Expression of IPA isoforms in T-ALL detected by RNA-seq. Shown are 3′-seq and RNA-seq tracks of a representative mRNA (out of $n = 101$) from CLL samples, T-ALL samples and normal thymus. The T-ALL RNA-seq data were obtained previously[32]. We detected $n = 381$ IPA isoforms in at least one T-ALL sample, $n = 133$ in at least one thymus sample, $n = 104$ in at least one T-ALL and one thymus sample, and $n = 101$ in at least two T-ALL samples, but not in any of the thymus samples.

Extended Data Table 1 | Samples investigated by 3′-seq and RNA-seq

a

	CLL low vs CLL high	Number of CLL-IPAs	Age at diagnosis	Rai stage at sample collection	WBC count at sample collection	IgVH status	Cyto-genetics	Treated before sample collection	Treated after sample collection	Diagnosis to sample collection (time; mo)	Treatment-free survival (yr)	RNA-seq	3′-seq
CLL1	L	13	62	III	153	UN	Del 11q	N	Y	10	1	Y	Y
CLL2	L	7	62	III	300	NA	Del 17p	N	Y	112	9	Y	Y
CLL3	L	26	54	IV	139	NA	Tri12, t(14;19)	N	Y	84	8	Y	Y
CLL4	H	93	72	0	173	UN	Normal	N	Y	37	4	Y	Y
CLL5	L	11	55	III/IV	193	UN	Tri8, del 13q	N	Y	46	4.5	Y	Y
CLL6	L	28	39	I	137	MUT	Del 13q	N	Y	49	3.3	Y	Y
CLL7	H	108	54	IV	111	MUT	Del 13q	N	Y	108	8	Y	Y
CLL8	L	12	72	III	365	UN	Tri12	N	Y	109	9	N	Y
CLL9	L	5	63	III	200	UN	Del 13q, t(6;19)	N	Y	30	2	N	Y
CLL10	L	11	51	III	77	UN	Del 11q	N	Y	70	6	N	Y
CLL11	H	274	39	0	100	UN	Del 11q, 13q, 14q	N	Y	44	5.5	Y	Y
CLL12	H	42	49	II	178	NA	NA	N	N	240	23.3	Y	Y
CLL13	L	7	66	I	125	UN	Del 11q, del 13q	N	Y	5	0.5	N	Y
CLL14	H	160	45	NA	NA	NA	NA	N	NA	112	NA	Y	N
CLL15	L	49	NA	NA	NA	NA	NA	N	NA	NA	NA	Y	N

N, No; Y, Yes; NA, not analyzed

b

Sample	Derived from	Sample name	Markers for sorting	No. of samples
CD5+B	Tonsil	CD5+B3-CD5+B6	CD5+, CD19+	4
NB	Tonsil	NB3-NB4	CD19+, CD27-	2
NB	Blood	NB1-NB2	CD19+, CD27-	2
MemB	Tonsil	M1-M2	CD19+, CD27+	2
GC	Tonsil	GC1-GC2	CD19+, CD38+	2
PC	BM	PC1-PC3	CD138+	3
T	Blood	T2-T3	CD3+	2

BM, bone marrow

c

Sample	Derived from	Sample name	Markers for sorting	No. of samples
CD5+B	Tonsil	CD5+B3-CD5+B4	CD5+, CD19+	2
CD5+B	Blood	CD5+B2	CD5+, CD19+	1
NB	Tonsil	NB3-NB5	CD19+, CD27-	3
NB	Blood	NB1-NB2, NB6	CD19+, CD27-	3
MemB	Tonsil	M2, M6	CD19+, CD27+	2
MemB	Blood	M3-M5	CD19+, CD27+	3
GC	Tonsil	GC1-GC4	CD19+, CD38+	4
PC	BM	PC4-PC21	CD138+	18

a, CLL sample characteristics. **b**, Normal human immune cells investigated by 3′-seq. **c**, Normal human immune cells investigated by RNA-seq.

nature research

Corresponding author(s): Christine Mayr

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	For 3'-seq and RNA-seq analyses, read counts from new samples and publicly available samples (see Data availability statement) were obtained. Information on DNA mutations of cancers other than CLL were obtained from the cbioportal. Information on DNA mutations of CLL samples were obtained from published data and anonymized identities were obtained from Dan A. Landau.
Data analysis	3'-seq and RNA-seq: Peak-calling and quantification were performed using in-house R packages biosignals and TagSeq. These packages depend on the following Bioconductor and CRAN packages: GenomicRanges, Rsamtools, GenomicAlignments, data.table, foreach. The source code of these packages is available in a public repository on BitBucket (https://bitbucket.org/leslielab/apa_2018/). All the GLM based differential analysis was performed using DEXSeq and DESeq. The remaining analyses were done using R version 3.1.2 (2014-10-31). Western blot: Odyssey CLx imaging system (Li-Cor). Northern blot: Fuji phosphorimager. FACS: Flow Jo. Statistics: R and SPSS. DNA sequencing: Vector NTI (Invitrogen) and Chromas (Technelysium Pty Ltd).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All 3'-seq and RNA-seq data generated and analyzed for this study have been deposited in the Gene Expression Omnibus database under accession numbers GSE111310 and GSE111793.
The code to analyze the data is available under https://bitbucket.org/leslielab/apa_2018/ and the processed data are available in Supplementary Table S1 (for Fig. 1b-d, 2a, 4a, Extended Data Fig. 4, and 5) and Supplementary Table S2 (for Extended Data Fig. 9a), and in the Source data files (for Fig. 1e, 2c, 2e, 3a, 3c, 4b-d, 4g, Extended Data Fig. 3c, 7j, 8c, and 9a). Data on DNA mutations from CLL patients were provided by Dan A. Landau (Weill-Cornell Medical College) and need to be requested from him. The mutation data on solid cancers was obtained through the MSK cbio portal. The data can be accessed through www.cbioportal.org.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. CLL samples investigated by 3'-seq were chosen based on availability. The samples needed to be untreated and contain a minimum of 75,000 WBC/ul. We only used fresh samples. We sequenced samples from N = 13 different CLL patients and obtained significantly different IPA isoforms. This indicates that the sample number was sufficient. As controls, we used N = 4 CD5+ normal B cells from lymphatic tissues obtained from different healthy donors. The 3'-seq was confirmed by obtaining concordant results from RNA-seq data from the same samples. To increase the sample size, we included 46 previously published RNA-seq samples for which exome sequencing data and IGVH mutational data were available.
Data exclusions	We did not exclude samples from the analysis.
Replication	3'-seq was performed in a single experiment, but we used normal B cells from N = 4 and CLL samples from N = 13 different donors. The data was reproducible among normal and malignant B cells obtained from different individuals. Furthermore, 3'-seq data were validated by several different methods. Genome-wide IPA isoforms were validated by RNA-seq and other genome-wide methods and 80% (4,456/5,587) were validated. In addition to genome-wide validation, selected truncated mRNAs generated from CLL-IPAs were validated by RT-PCRs (N = 18) and by western blots (N = 13). One CLL-IPA isoform was present, but did not show a difference in expression to normal B cells investigated by RT-PCR and another CLL-IPA isoform could not be validated by western blot analysis. All other tested CLL-IPA isoforms were successfully validated. Several shRNAs per knock-down were used. All experiments were performed as several biological replicates.
Randomization	The experimental groups are defined as cancer vs normal samples.
Blinding	There was no blinding during data collection as the normal and cancer cells are derived from different sources. 3'-seq and RNA-seq libraries from cancer and normal samples were generated at the same time. As the analysis aimed to identify differences between normal and cancer B cells, the data analysis was not performed in a blinded fashion.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
[ ]	[x] Unique biological materials
[ ]	[x] Antibodies
[ ]	[x] Eukaryotic cell lines
[x]	[ ] Palaeontology
[x]	[ ] Animals and other organisms
[x]	[ ] Human research participants

Methods

n/a	Involved in the study
[x]	[ ] ChIP-seq
[ ]	[x] Flow cytometry
[x]	[ ] MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials

Fresh primary CLL B cells were obtained from peripheral blood of untreated CLL patients. The material is no longer available. The BLCL cell line was made by us and is available upon request.

Antibodies

Antibodies used

For sorting of primary B cell subsets and FACS analysis of B cells and CLL cells, the following antibodies were used:
anti-CD3-PE (mouse, BD Biosciences, 555333, Lot#2317603, 1:100 dilution), anti-CD5-FITC (mouse, BD Biosciences, 555352, Lot#3046601, 1:100 dilution), anti-CD14-PECy7 (mouse, ebioscience, 25-0149-42, Lot#E10278-1635, 1:300 dilution), anti-CD19-APC (mouse, BD Biosciences, 555415, Lot#2347818, 1:100 dilution), anti-CD27-PE (mouse BD Biosciences, 555441, Lot#3051680, 1:50 dilution), anti-CD38-APC (mouse, BD Biosciences, 555462, Lot#3057748, 1:100 dilution), anti-CD38-FITC (mouse, BD Biosciences, 555459, Lot#2289722, 1:50 dilution).

For FACS analysis:
anti-phospho-NF-κB p65 (rabbit, Cell Signaling 3033, Lot#16, used 1:1,500 dilution, validated by citations and western blots in HeLa cells); secondary antibody: Alexa Fluor 647 (goat, Invitrogen, A27040, Lot#1834794, 1:10,000 dilution). Data in this paper: Figure 2c, Extended Data Figure 7d.

For Western blot analysis:
anti-ACTIN (mouse, clone AC-40, Sigma, A4700 , used 1:2,000 dilution; rabbit, Sigma, A2066, Lot#106M4770V, 1:7,000 dilution, validated by protein size prediction and citations, data in this paper: Figure 2a, 4f, Extended Data Figure 5, 6b, 6d, 7b, 7e-g, 9f), anti-GAPDH (goat, V-18, Santa Cruz Biotechnology, Lot#A1316, 1:500 dilution, validated by protein size prediction and citations, data in this paper: Extended Data Figure 7c), anti-AKAP10 (mouse, clone 51, Santa Cruz Biotechnology, sc-136512, Lot#F0410, 1:500 dilution, validated by protein size prediction and citations, data in this paper: Extended Data Figure 5), anti-CARD11 (rabbit, clone 93H1, Cell Signaling, 4440S, Lot#1, 1:1,000 dilution, validated by RT-PCR, protein size prediction, exclusive knockdown of endogenous isoforms in human TMD8 cells and citations, data in this paper: Figure 2a, Extended Data Figure 6d, 7b, c), anti-DICER1 (rabbit, gift from Dr. Witold Filipowicz (FMI Basel), 1:7,000 dilution, validated by the Filipowicz lab, citations, RT-PCR, protein size prediction and endogenously induced isoform expression in HeLa cells, data in this paper: Figure 2a, Extended Data Figure 6d, 7f), anti-DNM1L (mouse, Abcam, ab56788, Lot# GR237898-1, 1:1,000 dilution, validated by protein size prediction and citations, data in this paper: Extended Data Figure 5), anti-MGA (rabbit, H-286, Santa Cruz Biotechnology, sc-382569, Lot#A1516, 1:200 dilution, validated by protein size prediction and ectopic expression of isoforms in HEK293T cells, data in this paper: Figure 2a, Extended Data Figure 7g), anti-SFRS15 (SCAF4; mouse, Abnova, H00057466-B01, Lot#08163WULZ, 1:1,500 dilution, validated by protein size prediction, data in this paper: Extended Data Figure 5, 6d), anti-WSTF (BAZ1B; mouse, clone G-5, Santa Cruz Biotechnology, sc-514287, Lot#B0415, 1:500 dilution, validated by protein size prediction and citations, data in this paper: Extended Data Figure 5), anti-NUP98 (rabbit, Novus Biologicals, NB100-93325, Lot#A1, 1:2,000 dilution, validated by RT-PCR, protein size prediction, endogenously induced isoform expression in A549 cells (confirmed by RT-PCR and Western blot) and citations, data in this paper: Extended Data Figure 5, 6b), anti-SGK223 (mouse, clone A-6, Santa Cruz, sc-398164, Lot#G1715, 1:100 dilution, validated by RT-PCR, protein size prediction and endogenously induced isoform expression in HEK293 cells (confirmed by RTPCR), data in this paper: Extended Data Figure 5), anti-SENP1 (rabbit, Bethyl Labs, A302-927A-T, Lot#A302-927A-T-1, 1:1,000 dilution, validated by RT-PCR and protein size prediction, data in this paper: Extended Data Figure 5), anti-CUL3 (rabbit, Bethyl Labs, A301-108A-T, Lot#A301-108A-T-1, 1:1,000 dilution, validated by RT-PCR and protein size prediction, data in this paper: Extended Data Figure 5), anti-PAWR (rabbit, Abcam ab92590, clone EPR3991, 1:5,000 dilution, validated by RT-PCR and protein size prediction, data in this paper: Extended Data Figure 5), anti-RIPK1 (rabbit, Cell Signaling 4926, Lot#2, 1:1,000 dilution, validated by RT-PCR, protein size prediction and citations, data in this paper: Extended Data Figure 5), anti-V5 tag antibody (mouse, Invitrogen R960-25, Lot#1949337, 1:2,500 dilution, validated by citations, data in this paper: Extended Data Figure 7e), anti-GFP (chicken, Abcam ab13970, Lot#GR236651-17, 1:2,000 dilution, validated by citations, data in this paper: Extended Data Figure 7g), anti-WNT5a/b (rabbit, clone C27E8, Cell Signaling 2530, Lot#4, 1:1,000 dilution, validated by citations, data in this paper: Figure 4f, Extended Data Figure 9f).

The secondary antibodies used included anti-mouse IRDye 700 (donkey, Rockland Immunochemicals, 610-730-002), anti-rabbit IRDye 680 (donkey, Li-Cor Biosciences, 926-68073), anti-rabbit IRDye 800 (donkey, Li-Cor Biosciences, 926-32213), and anti-mouse IRDye 800 (donkey, Li-Cor Biosciences, 926-32212).

For protein immunoprecipitation:
anti-V5 tag antibody (mouse, Invitrogen R960-25, Lot#1949337, validated by citations, data in this paper: Extended Data Figure 7e).

Validation	Validations of the antibodies are indicated above.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	B lymphoblastoid cells (BLCL) are Epstein Barr virus-immortalized human blood B cells. They were immortalized by us and were described before (Lianoglou et al., Genes Dev 2013). MEC1 cells are malignant B cells from B-Prolymphocytic leukemia and were provided by Dr. Abdel-Wahab (MSKCC). They are also available at ATCC. Raji and TMD8 cells are malignant B cells from lymphomas and were a gift from Dr. Hans-Guido Wendel (MSKCC) and are also available from ATCC. HEK293 and HEK293T cells (embryonic kidney), HeLa cells (cervical cancer) and A549 cells (lung adenocarcinoma) were purchased from ATCC. The parental and DICER KO HCT116 cells were published and were provided by V. Narry Kim (Seoul National University).
Authentication	None of the cell lines used have been authenticated.
Mycoplasma contamination	All cell lines were tested for mycoplasma contamination and found to be negative.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Human TMD8 cells were fixed with 4% formaldehyde at room temperature for 15 mins. After two washes with excess PBS, fixed cells were resuspended with ice-cold PBS and permeablized with 90% methanol for 20 mins on ice. Cells were then washed with cold PBS twice and resuspended with the incubation buffer (PBS + 0.5% BSA). Cells were aliquoted and incubated with anti-phospho-NF-κB p65 (1:1500 dilution, Cell signaling #3033) for 1.5 hrs at room temperature. Cells were washed with the incubation buffer twice and incubated with the fluorochrome-conjugated secondary antibody solution (1:10,000 Alexa Fluor 647 A27040, Invitrogen) for 15 mins at room temperature. After two washes with the incubation buffer, cells were analyzed using a FACS Calibur.
Instrument	Becton Dickinson FACSCalibur Flow Cytometer
Software	FlowJo
Cell population abundance	GFP-positive cells were used for the analysis. The cell numbers of analyzed cells were at least 2,500 up to 7,000 cells, with the percentage range of 16.3-44% of total live cells.
Gating strategy	FSC-H (>200) and SSC-H (50-400) were set to gate live cells. By using the unstained cell control, GFP (FL1-H)-positive cells were gated at FL1-H > 60. All the GFP-positive cells were analyzed for FL4-H intensity.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.