Published in final edited form as:

Nat Biotechnol. 2016 September ; 34(9): 962–972. doi:10.1038/nbt.3637.

Analyzing tumor heterogeneity and driver genes in single myeloid leukemia cells with SBCapSeq

Karen M Mann^{1,2,7}, Justin Y Newberg¹, Michael A Black³, Devin J Jones¹, Felipe Amaya-Manzanares¹, Liliana Guzman-Rojas¹, Takahiro Kodama¹, Jerrold M Ward^{2,6}, Alistair G Rust^{4,6}, Louise van der Weyden⁴, Christopher Chin Kuan Yew^{2,6}, Jill L Waters^{5,6}, Marco L Leung⁵, Keith Rogers², Susan M Rogers², Leslie A McNoe³, Luxmanan Selvanesan^{3,6}, Nicholas Navin⁵, Nancy A Jenkins^{1,2}, Neal G Copeland^{1,2}, and Michael B Mann^{1,2,7}

¹Cancer Research Program, Houston Methodist Research Institute, Houston, Texas, USA

²Institute of Molecular and Cell Biology, Singapore, Republic of Singapore ³Department of Biochemistry, University of Otago, Dunedin, New Zealand ⁴Experimental Cancer Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK ⁵Department of Genetics, University of Texas MD Anderson Cancer Center, Houston, Texas, USA

Abstract

A central challenge in oncology is how to kill tumors containing heterogeneous cell populations defined by different combinations of mutated genes. Identifying these mutated genes and

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

Correspondence should be addressed to N.G.C. (ncopeland@houstonmethodist.org) or M.B.M. (mbmann@houstonmethodist.org).

⁶Present addresses: Global VetPathology, Montgomery Village, Maryland, USA (J.M.W.), Tumour Profiling Unit, the Institute of Cancer Research, Chester Beatty Laboratories, London, UK (A.G.R.), National Heart Research Institute Singapore, Republic of Singapore (C.C.K.Y.), Illumina, Inc., San Diego, California, USA (J.L.W.) and Pacific Edge Limited, Dunedin, Otago, New Zealand (L.S.)

⁷These authors contributed equally to this work.

Accession codes. BioProject: data have been deposited under accession codes [PRJNA285079](#), [PRJNA285082](#), [PRJNA325944](#) and [PRJNA285671](#).

Note: Any Supplementary Information and Source Data files are available in the [online version of the paper](#).

AUTHOR CONTRIBUTIONS

K.M.M., M.B.M., N.G.C. and N.A.J. designed the study, directed the research, interpreted the data and wrote the manuscript. K.M.M. and M.B.M. performed experimental work, designed the SBCapSeq oligos, coordinated sequencing efforts and analyzed the data. M.A.B., J.M.W., A.G.R. and N.N. contributed to the experimental design. J.Y.N., M.A.B. and A.G.R. provided essential statistical and bioinformatics resources. J.Y.N. wrote the Python code for SBCapSeq and splink HiSeq workflows and gCIS analysis; performed statistical and bioinformatic analysis for SBCapSeq, splink HiSeq, and RNA-seq data; and managed compute resources and data archiving of SBCapSeq and splink HiSeq data. M.A.B. wrote the R code and performed statistical and bioinformatic analysis for microarray, RNA-seq and WGS data analysis. A.G.R. wrote the R and Perl code for splink 454 workflow and managed resources and data archiving of splink 454 data. K.R. and S.M.R. performed and directed necropsy and histopathological analysis. J.M.W. performed and directed veterinary pathology analysis, including tumor grading and diagnosis. L.v.d.W. optimized library preparation for splink 454 sequencing. C.C.K.Y. performed bioinformatic analysis for splink 454 data. L.A.M. and L.S. isolated RNA and performed microarray hybridizations. K.M.M. developed and optimized staining protocols for FACS analysis. J.L.W., M.L.L. and N.N. isolated single cells by FACS and performed WGA of single-cell genomes. D.J.J. performed and optimized library preparation for SBCapSeq method and performed library preparation for splink HiSeq sequencing. L.G.-R. and F.A.-M. performed and optimized capture hybridizations for the SBCapSeq method and performed Ion Torrent sequencing for SBCapSeq, RNA-seq and WGS experiments. T.K. optimized library preparation splink HiSeq sequencing. All coauthors contributed to editing the manuscript before submission. N.G.C. and N.A.J. provided laboratory resources and personnel for animal husbandry, specimen archiving, sequencing and computer management.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

understanding how they cooperate requires single-cell analysis, but current single-cell analytic methods, such as PCR-based strategies or whole-exome sequencing, are biased, lack sequencing depth or are cost prohibitive. Transposon-based mutagenesis allows the identification of early cancer drivers, but current sequencing methods have limitations that prevent single-cell analysis. We report a liquid-phase, capture-based sequencing and bioinformatics pipeline, *Sleeping Beauty* (SB) capture hybridization sequencing (SBCapSeq), that facilitates sequencing of transposon insertion sites from single tumor cells in a SB mouse model of myeloid leukemia (ML). SBCapSeq analysis of just 26 cells from one tumor revealed the tumor's major clonal subpopulations, enabled detection of clonal insertion events not detected by other sequencing methods and led to the identification of dominant subclones, each containing a unique pair of interacting gene drivers along with three to six cooperating cancer genes with SB-driven expression changes.

SB insertional mutagenesis is used for cancer gene discovery in animal models of both hematopoietic and solid tumors^{1,2}. SB transposons can identify early cancer progression drivers, which cooperate to initiate tumors^{3,4}, and potential metastatic drivers^{5,6}, providing information about human cancer that is not easily obtained with current sequencing technology and limited patient samples.

In mouse cancer models driven by insertional mutagenesis, transposon insertions in tumor cells are detected with PCR-based approaches to enrich for transposon-containing tumor DNA fragmented by enzymatic digestion or random shearing. These DNA fragments are then pooled, sequenced and mapped to the mouse genome. The method requires micrograms of input DNA, and PCR amplification of enzymatically fragmented DNA has a bias for recovery of insertions that occur within TA dinucleotides located <1 kb from the end of a DNA fragment. These limitations influence the detection of insertions and preclude the approach from application to small amounts of DNA isolated from laser-microdissected tumors or single cells.

We report the development of a transposon-sequencing method to identify SB transposon insertion sites from bulk tumor tissue and single tumor cells. Our approach relies on random shearing of genomic tumor DNA into discrete fragment sizes, which are then enriched for transposon sequences using liquid-capture hybridization with capture probes specific for the inverted repeats in the transposon termini. These capture probes are combined with blocking probes that are directed to the original transposon concatamer and prevent enrichment of unmobilized transposons from the donor site. Using Ion Torrent sequencing technology, SBCapSeq identifies insertions in a mouse model of AML in a semiquantitative manner whereby read depth can be used to identify clonally selected events. This is possible because captured fragments of discrete size are not subject to PCR bias or sequencing errors that occur in low-complexity amplicon-based libraries for NGS. We developed a bioinformatics pipeline to identify SB insertions in bulk tumor and single cells by mapping quality scores of sequence reads that contain both SB transposons and flanking genomic sequences that map uniquely to the mouse genome. Finally, we integrated data from whole-genome sequencing (WGS), SB capture hybridization and RNA-seq and found that SB insertions cause acute

ML (AML) through overexpression of genes in the MAP kinase (MAPK) and JAK–STAT signaling pathways and identified cooperating mutagenic events in single tumor cells.

RESULTS

SB model of ML

SB transposition in mice carrying an inducible SB transposase⁷ and the *Actb-Cre* transgene⁸ drives a rapidly fatal hematopoietic disease prominently affecting the spleen (Fig. 1a and Supplementary Fig. 1a). Addition of an inducible *Trp53*-null (*Trp53*^{+/-})⁹ or *Trp53* point mutation (*Trp53*^{R172H/+})¹⁰ accelerated disease progression ($P < 0.0001$, Fig. 1a). *Trp53*^{R172H/+} mice died earlier than *Trp53*^{+/-} and wild-type mice and accumulated Trp53 protein in their tumor cells (Supplementary Fig. 1b–d), results that are consistent with previous studies showing that missense mutations in *Trp53* abrogate the tumor suppressor functions of Trp53 and promote malignant transformation^{10–14}.

Mice from all three cohorts succumbed to an early, completely penetrant ML characterized by dense populations of undifferentiated cells in the red pulp (Fig. 1b) and subcapsular areas (Fig. 1c) of the spleen, with foci of tumor cells present in the liver (Supplementary Fig. 1e). Approximately half of all tumors stained positive for the myeloid cell marker myeloperoxidase (MPO) in poorly differentiated regions (Fig. 1d) and in the subcapsular area (Fig. 1e). Many of the MPO-positive tumors were also positive for PAX5 staining under the capsule¹⁵ (Supplementary Fig. 1f). PAX5 is found in human myeloid pro-genitors¹⁶. On the basis of these findings, we classified this disease as an undifferentiated ML, termed ‘ML without maturation’ in humans¹⁷.

Candidate cancer genes in mouse ML

We sequenced SB insertion sites from 168 mouse spleens and identified 27,889 nonredundant transposon insertion sites, which we used to statistically define 466 nonredundant candidate cancer genes (CCGs) by combining the data for all tumor cohorts analyzed together (Online Methods) and for cohorts based on *Trp53* genotype analyzed separately (Supplementary Table 1). Most CCGs were identified across all cohorts independently, suggesting that their roles in ML are *Trp53* independent.

Thirty-six CCGs are orthologs of human cancer genes found in the Cancer Gene Census¹⁸, a catalog of genes for which mutations have been causally implicated in cancer ($P = 7.15 \times 10^{-8}$, Bonferroni corrected). Another 110 CCGs, including *Flt3* and *Runx1*, have human orthologs with mutations identified by exome or WGS¹⁹ in at least one AML genome ($P = 3.5 \times 10^{-7}$, Bonferroni corrected). Notably, many CCGs overlapped with human AML CCGs that fall within the long-tail regions of nonsynonymous coding mutations identified by the Cancer Genome Atlas for AML¹⁹ and with CCGs identified in two different mouse models of leukemia that also used SB mutagenesis^{20,21} (Supplementary Note). We conclude from these findings that SB captures the genes mutated by the diversity of mutational processes in human AML.

Previous studies suggest that most CCGs identified by SB mutagenesis function during tumor progression rather than initiation³. To identify ML CCGs that function in tumor

initiation, we selected the SB insertion sites with the highest number of sequencing reads (Online Methods), reasoning that insertions that caused tumor initiation would be present in the largest number of tumor cells. We identified 35 putative ‘trunk drivers’, genes involved in initiation and early progression of ML *in vivo* (Fig. 1f). All statistically defined candidate cancer-associated genes identified in our analysis are drivers; our analysis does not define non-driver genes. AML patients who had mutations in one or more trunk drivers showed significantly $P < 0.03$ lower survival than patients lacking mutations in these genes (Supplementary Fig. 2a). Notably, 80% of trunk drivers were involved in MAPK or JAK-STAT signaling.

Pathway enrichment analysis using all 466 CCGs identified MAPK and JAK-STAT signaling as significantly $P < 0.001$ enriched pathways (Supplementary Table 2), similar to what is observed in human AML^{22,23}. Ninety-seven percent of our CCGs had a mutation in at least one gene involved in MAPK, JAK-STAT or Jun N-terminal kinase (JNK)-p38 MAPK pathways (Fig. 2a). Notably, mutations in human homologs of the Sleeping Beauty-myeloid leukemia (SB-ML) CCGs that mapped to JAK-STAT or MAPK signaling are also associated with decreased survival in human AML (Supplementary Fig. 2c,d). The most frequently mutated CCGs were *Erg*, *Ets1* and *Fli1*, which encode ETS-related transcription factors and were collectively mutated in 80% of tumors. Insertions were primarily located in the 5′ end of each gene in the same transcriptional orientation, where they were predicted to drive expression of a truncated chimeric transcript containing the oncogenic ETS domain (Fig. 2b, top). This hypothesis was subsequently confirmed by microarray analysis (Fig. 2b, bottom).

We also identified statistically significant ($P < 0.05$) co-occurrences of SB insertions in trunk drivers such as *Ghr* and *Jak2* and mutually exclusive insertions in trunk drivers such as *Erg* and *Ets1*. These statistically defined relationships (Fig. 2c) indicate that cooperating oncogenic mutations drive leukemogenesis in this mouse model of human AML.

Capture hybridization enhances transposon insertion detection

On average, each SB-induced ML contains 17.2 insertional mutations in CCGs, reflecting the high degree of intra- and intertumor heterogeneity of these tumors. This heterogeneity, together with the ability to detect insertional mutations in small fractions of tumor cells from bulk tumor using a PCR-based approach, makes it possible to identify mutations that occur late in tumor progression. However, the technology is not sensitive enough to sequence transposon insertions from single tumor cells—a necessary approach to delineate intratumor heterogeneity and cooperating insertion events.

The SBCapSeq liquid-capture hybridization method uses capture probes specific for sequences located at the ends of the transposon to enrich for DNA fragments containing both transposon and mouse DNA before sequencing on the Ion Torrent Proton platform (Fig. 3a,b and Supplementary Fig. 3a). A detailed protocol describing this method can be found at <http://dx.doi.org/10.1038/protex.2016.053>. To enable post-processing of sequence output from the Ion Torrent software suite, we designed a complementary SBCapSeq bioinformatics pipeline (Supplementary Fig. 3b). After stringent read filtering and mapping, we retained an average of 4×10^5 reads per library for further analysis. The capture bias for

right- or left-transposon inverted repeats was small (Supplementary Fig. 4), and most of the unmappable sequences were short reads that either mapped to the vector or did not meet the filtering criteria required for transposon or mouse genomic elements.

To evaluate the reproducibility and sensitivity of SBCapSeq, we constructed a DNA library from a single leukemic spleen and performed nine independent SBCapSeq reactions from the same sample. Sequence analysis demonstrated high reproducibility for detecting both reads and fragments, which are reads with identical start sites collapsed into a single unit (Fig. 3c). A single sequencing run detected 50% of SB insertions at a read depth of 225, and 80% of the insertions identified by 5 fragments were identified in two runs. The fact that the number of single reads or fragments never reached saturation (Fig. 3c) suggests that they represent passenger insertions in which the transposon is maintained but does not confer a biological advantage. We also assessed the reproducibility of SBCapSeq to identify the same SB insertions using two independent sequencing runs for one tumor and found an overall high correlation between shared insertions and sequencing depth across experiments (Supplementary Fig. 5a). When we specifically plotted insertions below the 95% precision cut-off, we found a negative correlation (Supplementary Fig. 5b,c). Using fragments instead of reads for the same two libraries showed an even higher correlation between experiments and a greater precision and recall, where the optimal F-measure and 95% precision are the same and fragments below the cut-off show no correlation (Supplementary Fig. 5d – f).

Comparison of SBCapSeq with other SB sequencing methods

Having shown that SBCapSeq identifies SB insertions with high reproducibility, we then compared insertion data generated by SBCapSeq for ten ML spleens previously analyzed by SB splinkerette PCR coupled with 454 sequencing (SB splink 454). SBCapSeq identified five times as many unique SB insertions as SB splink 454 (Fig. 4a), and at least 50% of these insertions mapped to genes (Supplementary Table 3). SBCapSeq also identified genic insertions detected with high read counts that were not detected by SB splink 454. The narrow window of fragment sizes (225 ± 25 bp) isolated for capture and sequencing on the Ion Torrent platform (compared to the 200- to 400-bp range published by Friedel *et al.*²⁴ or the 100- to 1,000-bp range published by Koudijs *et al.*²⁵) ensures a more uniform amplification of all fragments during pre- and post-capture amplification steps and downstream sequencing representation. Thus, SBCapSeq can overcome a major limitation of restriction-enzyme-based splinkerette for detecting insertions (PCR bias for amplifying smaller fragments)²⁶ and enables detection of previously unknown clonal insertion events. We also compared the ability of each method to detect CCGs in MAPK, JAK-STAT and JNK-p38 pathways (Fig. 4b). SBCapSeq identified seven insertions in CCGs that were missed by splink 454 in tumor C0117_SP, one of the tumors that we resampled (Fig. 4b), and detected 18 CCGs in the 10 resampled tumors with at least two independent insertions in a single tumor (possibly a major clone and a minor clone), as defined by read counts for individual insertions. In contrast, splink 454 identified only two CCGs with multiple insertions in a single tumor. SBCapSeq detected multiple insertions per tumor in *Erg* (Fig. 4c) and all insertions were in the correct position and orientation to activate *Erg* expression. SBCapSeq revealed a high degree of convergent tumor evolution that was not as apparent in the splink 454 data.

We next compared the SBCapSeq data from the ten ML samples to SB insertion data obtained using splink HiSeq²⁶, a platform widely used for identifying SB insertions given the massive amount of obtainable sequencing reads²⁵. We performed splinkerette PCR on randomly sheared genomic DNA using a modified version of the protocol published by Brett *et al*²⁶ (Online Methods). Hierarchical clustering of insertions from splink 454, splink HiSeq and SBCapSeq shared by at least two platforms showed that in all but one sample, the data from SBCapSeq were closely related to the splink HiSeq output (Supplementary Fig. 6a – j). Insertions from SBCapSeq and splink HiSeq plotted by read depth showed high correlation (Pearson's correlation = 0.8), indicating that SBCapSeq identifies the same insertion sites with the same representation as splink HiSeq but requires one-sixth of the DNA input for library preparation. Notably, 25% of the insertions detected with splink HiSeq were assigned to more than one sample, despite our use of staggered barcodes and double indexing to overcome errors caused by low library complexity^{27,28}. Notably, SBCapSeq yielded <0.01% cross-contamination of insertions in more than one sample, showing its higher sensitivity than Splink_HiSeq methods and potentially expanding the sequencing capability to single-cell sequencing.

Finally, we compared SBCapSeq to WGS as a benchmark of its quantitative nature. We sequenced a single leukemic spleen (sample D5) and normal tail from a sibling mouse (sample D6) using both methods. WGS identified unique SB insertions from D5 in 439 genes, including 28 CCGs identified previously (Fig. 5a). The loci containing *Col4a3/Mff* and *Lmf1/Sox8* had the greatest number of reads for a single insertion (25 and 22, respectively), followed by four trunk drivers—*Notch1* (21 reads), *Ghr* (19 reads), *Erg* (15 reads) and *Ets1* (15 reads). *Notch1* had a second unique insertion detected at 6 reads, suggesting coevolution in separate bulk tumor cells; both insertions were in the correct orientation to activate *Notch1*. WGS data from control sample D6 revealed only four insertions located on nondonor chromosomes represented by >2 reads, and no coevolving events were detected. SBCapSeq detected all but three of the insertions identified by WGS in D5 with ≥ 2 reads, at an average 947-fold increased read depth, and the read depth across the two platforms for the ML CCGs mutated in this tumor was highly correlated (Fig. 5b). The output for SB insertion data is summarized in Figure 5c. Eighty-seven genes on nondonor chromosomes, including *Erg*, *Ghr*, *Notch1* and *Zmiz1*, showed evidence, based on the detection of two or more insertions in independent TA dinucleotides, of coevolving SB insertion events. For several genes (*Ankib1*, *Man2a1*, *Muc4*, *Plcx3*, *Ptprt* and *Ryr3*), a 'major' insertion was detected at ≥ 1,000 reads, in addition to one or more 'minor' insertions represented by ≥ 100 reads. This finding illustrates the high degree of intratumor heterogeneity in SB-induced tumors detectable by SBCapSeq. Considering shared insertions detected by SBCapSeq, WGS and/or splink HiSeq for this tumor, output from SBCapSeq and splink HiSeq were closely related (Supplementary Fig. 6k,l), with a Pearson correlation of 0.81, even with increased sequencing depth from SBCapSeq. Furthermore, SBCapSeq showed a higher correlation with WGS insertion data (Supplementary Fig. 6m) than did splink HiSeq (Supplementary Fig. 6n), suggesting that the fidelity of SBCapSeq for defining real SB insertions is also greater than that of splink HiSeq.

WGS confirms cytologically normal leukemic genome

WGS showed that the D5 leukemia has a ‘quiet’ genome containing only 1,080 significant single-nucleotide variations (SNVs) and 160 significant insertions or deletions (indels), most of which map to introns and non-coding regions (Supplementary Fig. 7a and Supplementary Table 4). Only five events occurred within exons, including a frameshift mutation in *Olfir429*; three nonsynonymous SNVs in *Krt5*, *Tdrd6* and *Tenm3*; and one synonymous SNV in *Lactb*. None are ML CCGs or have known roles in leukemia, suggesting they are passenger mutations. In addition, no significant copy-number alterations (CNAs) were observed in the D5 leukemic genome across a 5-kb window, no translocation events were detected (Supplementary Fig. 7b) and we could not find evidence for a significant mutational signature. WGS identified seven insertions containing a TACAG or TACTG sequence, which are the canonical SB mutagenic footprints²⁹ resulting from SB remobilization (Supplementary Table 4). Most of these footprints occurred on the donor chromosomes at a read depth on par with that of the SB insertions in trunk driver genes, indicating that we detected clonal footprints using WGS. Together, these data strongly suggest that SB mutagenesis is the predominant force driving tumor evolution.

Sequencing analysis of single-cell transposon insertions

We sorted live cells from the D5 leukemic spleen using the cell surface markers CD71 and Ter119, which we identified to be enriched on tumor cells (Supplementary Fig. 8a and Supplementary Note). Each single-cell genome was screened for the presence or absence of a subset of SB insertions that are highly represented in the D5 bulk tumor, including insertions in trunk drivers *Ets1* and *Notch1* and in *Ankib1*, *Phf2*, *Kdm4c* and *Zfp704* (Supplementary Fig. 8b). We identified 26 distinct SB insertion patterns in 84 single cells, with 12 cells lacking insertions in any of these genes. This heterogeneity of SB insertion patterns was unexpected, given that insertions in these genes were represented by high read counts in bulk tumor DNA. Twenty-six cells representing the different insertion patterns were then sequenced by SBCapSeq to obtain the full complement of SB insertions in each cell. A total of 3,819 unique insertions were identified in the 26 single-cell genomes, with a median of 147 unique insertions per cell; 16.3% (622/3,819) of the insertions mapped to the donor chromosome, whereas 83.7% (3,197/3,819) mapped to non-donor chromosomes. This is a much lower frequency of donor chromosome insertions than the 50% observed in bulk tumor DNA and probably reflects the remobilization of transposons from the donor chromosome that we can observe in these enriched tumor cells.

In total, 1,576 genes contained one or more SB insertion events among these 26 cells, and each cell had an average of 60 genic insertions. Only 27% of these genes were found to have insertions in the bulk tumor DNA by WGS or SBCapSeq. Forty-two of these genes were ML CCGs. Most insertions identified in single cells seem to represent unique insertion events and are probably passenger mutations. We also identified SB insertions in 58 ML CCGs in single cells not found in the bulk tumor by SBCapSeq or WGS, suggesting that these events were subclonal in this leukemia. Single-cell sequencing identified all but two of the insertional events detected by WGS with 2 reads and confirmed most events detected by SBCapSeq with 5 reads. This overlap suggests that we identified the major clonal subpopulations in this tumor by sequencing just 26 cells. The average number of insertional

mutations in CCGs per cell was three, and no single cell had more than six mutated CCGs (Supplementary Table 5), suggesting that as few as three cooperating mutations are sufficient to drive clonal expansion. This observation is consistent with human studies that postulate that a typical tumor contains 2–8 driver gene mutations³⁰.

We next performed hierarchical clustering, using the Hamming distance metric with Ward's linkage method, of SB insertions shared by three or more cells and assigned the 26 cells to two major clades (Fig. 6). One population, composed of half of the sequenced cells, was anchored by an insertion in *Erg*, and all but one of these cells also had an insertion in *Ghr*. The *Erg* insertion may have occurred first, or the cell lacking a *Ghr* insertion may represent a false negative. Insertions in several other loci, including *Kdm4c*, *Phf2* and *Ankib1*, were also coincident with insertions in *Erg* and *Ghr*. None of these are CCGs, suggesting that they could be passenger mutations that were present in the cell before clonal expansion. The second major clade comprised two subpopulations of cells. A cluster of three Ter119⁺ cells at the right of the heat map had insertions in *Ets1* and *Notch1*. Another Ter119⁺ cell (D5_SP_SC17) carried the *Ets1* insertion but lacked the *Notch1* insertion, whereas two Ter119⁺ cells (D5_SP_SC06 and D5_SP_SC09) had the *Notch1* insertion but lacked the *Ets1* insertion. D5_SP_SC17 may represent a false negative (a known error due to insufficient coverage during whole-genome amplification (WGA))³¹, given that the rest of the cells in this clade contained a *Notch1* insertion. Finally, one cell (D5_SP_SC95) had insertions in loci associated with both major clades, which suggests a failed single-cell sort. Six cells had few insertions in common with other cells and are likely to represent contaminating normal cells. When the analysis was expanded to include all insertions that mapped to ML CCGs and not just those shared in three or more cells, we found that most CCG insertions were unique to a single cell (Supplementary Fig. 8c). This indicates that SB mutagenesis continues at a high level during clonal expansion, probably generating additional subclones as the tumor progresses and cells experience ongoing selective pressure. Notably, single-cell sequencing confirmed the cooperative relationship between *Ets1* and *Notch1* identified by population-based analysis and uncovered a possible cooperation between *Erg* and *Ghr*. Mapped SB insertions in single cells and D5 bulk tumor cells identified by SBcapSeq and SB insertions identified in the ML population by splink 454 for the trunk drivers *Erg*, *Ghr*, *Ets1*, *Notch1* and *Akt1* showed activating driver events predicted to upregulate gene expression from the minimal promoter from the transposon (Supplementary Fig. 9).

Gene-expression changes arise from SB insertions

To analyze the impact of clonally selected SB insertions on gene expression, we sequenced a poly(A)-enriched (mRNA-seq) library of total RNA isolated from the D5 leukemic spleen to identify chimeric spliced mRNA-seq reads containing transposon splice-donor or splice-acceptor sequences fused to adjacent known exons (SBfusions). We identified 1,331 SBfusion mRNA-seq reads from 682 RefSeq genes (Supplementary Table 6), and the top 2% of genes containing SBfusion transcripts were supported by 7 SBfusion reads and enriched for ML trunk drivers, including *Erg*, *Notch1*, *Ghr* and *Ets1*.

To confirm and extend the transcriptome analysis, we sequenced ribo-depleted, whole-transcriptome RNA-seq (wtRNA-seq) libraries from the D5 leukemic spleen, as well as two littermate control non-leukemic spleens in which SB was not mobilized, on the Ion Proton sequencing platform. Using the same modified SBCapSeq workflow, we identified 838 SBfusion RNA-seq reads from 48 genes and found increases in gene expression in leukemic cells for a number of ML trunk drivers, including *Erg*, *Notch1*, *Ghr* and *Ets1* (Supplementary Table 6). CCGs with SB insertions in the sense orientation relative to the gene showed increased expression in the tumor over normal spleen; the same was true of insertions private to the D5 tumor predicted to activate gene expression (Supplementary Table 6). Insertions with high read counts led to greater changes in gene expression; however, insertions in CCGs and genes private to the D5 tumor oriented in the antisense direction rarely led to decreased gene expression in the tumor, indicating that they are passenger insertions and that this ML is largely an oncogene-driven disease.

Finally, we integrated the RNA-seq and insertion data for genes identified by single-cell sequencing to help distinguish between driver and passenger mutations. Figure 7 shows all of the genes identified in the two leukemic subpopulations, anchored by insertions in *Erg* and *Ghr*, or *Ets1* and *Notch 1*. Figure 7a shows expressed SB-fusion transcripts for trunk drivers *Erg* and *Ghr*, in addition to *Ankib1*, *Akap9*, *Fam65c*, *Cacnb4*, *Gak* and *Susd1*, which might therefore also contribute to tumorigenesis. Upregulated SBfusion transcripts were also observed for three genes located on the donor chromosome, including *Ncoa2*, *Vsig8* and *Ppp1r7*. Because these genes are on the donor chromosome, a role for these genes in leukemia induction is less certain. Figure 7b shows expressed fusion transcripts for trunk drivers *Ets1* and *Notch*, which are both transcriptional targets of *Runx1* (ref. 32), a major player in AML, and for *Akt1*. We also detected upregulated expression of *Kif1a* and *Poglut1*, which might contribute to leukemogenesis. These results provide additional strong evidence for the role of altered MAPK signaling in AML. Notably, six CCGs (*Akt1*, *Erg*, *Ets1*, *Ghr*, *Ncoa2* and *Notch 1*) represented in the bulk tumor at high read depth and confirmed as drivers by single-cell analysis were associated with a decrease in AML patient survival ($P=0.00021$, Supplementary Fig. 2b).

DISCUSSION

Although transposon mutagenesis can be used to identify cancer-causing genes, it is challenging to identify trunk drivers that participate in tumor initiation. Trunk drivers are present in every tumor cell and may therefore be good drug targets. By sequencing transposon insertions from single tumor cells, SBCapSeq can identify both trunk drivers and potential cooperating cancer genes. We identified cooperating trunk drivers *Erg* and *Ghr* and *Notch1* and *Ets1* in two dominant tumor cell populations isolated from a SB model of human minimally differentiated, cytologically normal, acute ML. ERG is at the center of a regulatory program critical for hematopoietic stem cell maintenance³³. In murine models, high *Erg* expression induces both T cell acute lymphoblastic leukemia (T-ALL) and AML³⁴, and high *ERG* expression in both human AML³⁵ and early T cell precursor ALL is associated with poor clinical outcome. Growth hormone receptor (GHR) signaling has not been reported in AML, but growth hormone activates mitogenic signaling through JAK-STAT, MAPK-ERK and mTOR pathways³⁶ that are active in AML. NOTCH1 was found to

be hyperexpressed in seven patients diagnosed with AMLs largely at stages M0 and M1 without morphological maturation³⁷. Given our findings and the encouraging effects of γ -secretase inhibitors in preclinical models and now in clinical trials³⁸ for T-ALL patients with NOTCH1 hyperexpression, AML patients with increased NOTCH1 signaling may be an additional group worth studying in this context. Decreased ETS1 binding is required for normal erythroid maturation, and hyperactivated ETS1 expression may interfere with terminal erythroid differentiation³⁹. Consistent with this, the subpopulation of leukemic cells containing *Notch1* and *Ets1* insertions in our data set expressed Ter119, a marker of late differentiation of erythroid cells.

One unanticipated finding of our analysis was the large number of unique transposon insertions in every tumor cell. Recently, Sottoriva *et al.*⁴⁰ presented a ‘Big Bang’ model of human colorectal cancer (CRC) growth whereby tumors grow predominantly as a single expansion, producing many intermixed subclones not subject to stringent selection due to spatial constraints. Although our ML model is considerably more spatially unrestricted than CRC, we believe that it is still likely to follow the Big Bang model because of the high intratumor heterogeneity (ITH) with extensive subclone mixing. Until recently, it was thought that adult hematopoiesis is driven by the action of a small number of long-term multipotent hematopoietic stem cells; however, recent studies in mice have shown this is not the case^{41,42}. Lineage tracing in adult mice showed several-hundred-fold larger myeloid than lymphoid output, which could explain why the myeloid compartment is particularly susceptible to the effects of SB mutagenesis.

SBCapSeq is not limited to the study of ML. Using a conditional system⁷, SB transposition can be activated in virtually any hematopoietic cell using a lineage-specific Cre, with or without a sensitizing mutation, to model hematopoietic processes including disease. Furthermore, SBCapSeq can be extended to the study of solid tumors and cell-based screens. This capture-based approach can be adapted to any transposon-based system by modification of the capture and blocking oligos and therefore provides a tool to evaluate clonal dynamics and evolution in transposon model tumors.

URLs

Cancer Gene Census, <http://cancer.sanger.ac.uk/census/>; DNACopy package, <http://www.bioconductor.org/packages/release/bioc/html/DNACopy.html>; Ingenuity Pathway Analysis (IPA), <http://www.ingenuity.com/products/ipa/>; Kyoto Encyclopedia of Genes and Genomes (KEGG), <http://www.genome.jp/kegg/>; Mouse Genome Informatics (MGI) database, <ftp://ftp.informatics.jax.org/pub/reports/index.html>; The Cancer Genome Atlas (TCGA), <http://cancergenome.nih.gov/>; Database for Annotation, Visualization and Integrated Discovery (DAVID), <http://david.abcc.ncifcrf.gov/>.

METHODS

Methods and any associated references are available in the [online version of the paper](#).

ONLINE METHODS

Mice

The following alleles were used to construct the SB-driven mouse model of myeloid leukemia: *Actb-Cre* (FVB/N-Tg(ACTB-cre)2Mrt/J, ref. 8); *Trp53*^{fllox/+} (FVB.129P2-*Trp53*^{tm1Bm}/Nci, ref. 9); *Trp53*^{LSL-R172H/+} (129S4-*Trp53*^{tm2Tyj}/Nci, ref. 10); T2/Onc2(TG.6070) (TgTn(sb-T2/Onc2) 6070Njen; ref. 43); T2/Onc2(TG.6113) (TgTn(sb-T2/Onc2)6113Njen; ref. 43); and *Rosa26*-LSL SBBase or SBBase^{LSL}; (Gt(ROSA)26Sor^{tm2(sb11)}Njen; ref. 44). The resulting cohorts of mice were on mixed genetic backgrounds consisting of C57BL/6J, 129, C3H and FVB. Genotyping by PCR assays with primers specific to the alleles was performed. No sample size estimate was used. No randomization was performed; mice were assigned to groups on the basis of genotype. Investigators were not blinded to genotype or group allocation. All mice were bred and maintained in accordance with approved procedures directed by the respective Institutional Animal Care and Use Committees in the A*STAR Biological Resource Centre and Houston Methodist Research Institute. Gross necropsies were performed, and all masses were documented and prepared for subsequent analysis. See Supplementary Note for details of mouse crosses. Male and female mice were used for experiments. Both male and female mice were assigned to cohorts based on genotypes by 5 weeks of age and were maintained until leukemic disease necessitated euthanasia.

Histological analysis

Histological analysis of spleen was performed on 5-µm sections of formalin-fixed, paraffin-embedded (FFPE) specimens stained with hematoxylin and eosin. Undifferentiated myeloid leukemia diagnosis used additional immunohistochemical staining on FFPE tissues after antigen retrieval, endogenous peroxidase inhibition and overnight incubation with antibody markers specific for various hematopoietic lineages⁴⁵, including myeloperoxidase (rabbit-anti-MPO; R&D Systems; pH 6; 1:2,000 dilution), PAX5 (mouse-anti-PAX5; Novocastra; pH 6; 1:100 dilution), CD45R (biotinylated rat-anti-CD45R; BD Pharmingen; pH 6; 1:500 dilution), CD3 (rabbit-anti-CD3; DAKO pH 9; 1:1,500 dilution), TdT (rabbit-anti-TdT; Novus Biologicals; pH 9; 1:100 dilution) and TRP53 (rabbit-anti-P53; Novocastra; pH 9; 1:500). As expected, robust nuclear staining of SB transposase (SBBase) was confirmed by immunohistochemistry on FFPE tissues after antigen retrieval (pH 9) and endogenous peroxidase inhibition followed by overnight incubation with mouse antibody to SBBase (anti-SBBase; R&D Systems; pH 9; 1:200 dilution). After incubation with primary antibody, chromogen detection (with HRP polymer, anti-rabbit or anti-mouse, with Envision System from Dako) and hematoxylin counterstaining were performed per the manufacturers' instructions. Nuclear staining of TRP53 was quantified from scanned sections using Slide Path software (SlidePath Ltd.). FACS isolation of single cells was performed using antibodies against mouse CD71 (BD Biosciences) or Ter119 (BD Biosciences). Genomic DNA (gDNA) was isolated from flash frozen necropsy specimens using Qiagen Gentra Puregene DNA isolation kit protocol for tissue.

Single-cell isolation and WGA

For isolation of single cells, fresh leukemic spleen cells were physically dissociated, purified by passing through a 40- μ M nylon mesh strainer and incubated with individual or multiplexed antibody cocktail to prepare for fluorescence-activated cell sorting (FACS). Forty-eight single cells were collected by FACS Aria II (BD Biosciences) analysis from cells stained with two antibodies: (i) Ter119 (BD Biosciences), an erythroid-specific marker commonly used to identify erythroid differentiation stages and (ii) CD71 (BD Biosciences), transferrin receptor, a non-erythroid-specific marker commonly used to identify erythroid differentiation stages. Single cells were collected into 96-well plates with 2.5 μ L lysis buffer (200 mM KOH and 50 mM DTT). WGA was carried out using multiple displacement amplification with Φ 29 bacteriophage polymerases supplied by Qiagen as REPLI-g UltraFast Mini Kit (Qiagen, 150035) as described^{31,46} with minor modifications. After WGA, a panel of 22 mouse-specific primer pairs designed to target all chromosomes was used to estimate the amplification efficiency of the collected single cells. In total, 84 of 96 sorted, WGA single cells (40/48 from Ter119 and 44/48 CD71) were carried forward for further analysis.

Mapping transposon insertion sites using SBCapSeq

For selective SB insertion site sequencing by liquid-hybridization capture, gDNA (0.5 μ g per sample) of either bulk tumor specimens or single cell WGA genomes was used for library construction using the AB Library Builder System, including random fragmentation and ligation of barcoded Ion Xpress sequencing adapters. Adapter-ligated templates were purified by Agencourt AMPure beads and fragments with insert size of 200 ± 30 bp were excised, purified by Agencourt AMPure beads, amplified by eight cycles of adapter-ligation-mediated PCR (aLM-PCR), and purified by Agencourt AMPure beads with elution in 50 μ L TE (1x Tris-EDTA, pH8). Capture hybridization of single or multiplex barcoded libraries (60 ng per sample) was performed using the custom xGen Lockdown Probes (IDT). The 120-mer capture oligo probes used were: IRDRL (5' Target_r1_1): 5'-Biotin—TAC AGT TGA AGT CGG AAG TTT ACA TAC ACT TAA GTT GGA GTC ATT AAA ACT CGT TTT TCA ACT ACT CCA CAA ATT TCT TGT TAA CAA ACA ATA GTT TTG GCA AGT CAG TTA GGA CAT CTA-3'; and, IRDRR (3' Target_r1_1): 5'-Biotin-TGG TGA TCC TAA CTG ACC TAA GAC AGG GAA TTT TTA CTA GGA TTA AAT GTC AGG AAT TGT GAA AAA GTG AGT TTA AAT GTA TTT GGC TAA GGT GTA TGT AAA CTT CCG ACT TCA ACT GTA-3'. The 120-mer blocking oligo probes used were: Blocker_upr1_1: 5'-Biotin-CCA TTC GCC ATT CAG GCT GCG CAA CTG TTG GGA AGG GCG ATC GGT GCG GGC CTC TTC GCT ATT ACG CCA GCT GGC GAA AGG GGG ATG TGC TGC AAG GCG ATT AAG TTG GGT AAC GCC AGG-C3 Spacer-3'; Blocker_upr1_2: 5'-Biotin-GCT GCA AGG CGA TTA AGT TGG GTA ACG CCA GGG TTT TCC CAG TCA CGA CGT TGT AAA ACG ACG GCC AGT GAG CGC GCG TAA TAC GAC TCA CTA TAG GGC GAA TTG GAG CTC GGA TCC CTA-C3 Spacer-3'; Blocker_dnr1_1: 5'-Biotin-TAG GGA TCC TCT AGC TAG AGT CGA CCT CGA GGG GGG GCC CGG TAC CCA GCT TTT GTT CCC TTT AGT GAG GGT TAA TTT CGA GCT TGG CGT AAT CAT GGT CAT AGC TGT TTC CTG TGT GAA-C3 Spacer-3'; Blocker_dnr1_2: 5'-Biotin-ATT GTT ATC CGC TCA CAA TTC CAC ACA ACA TAC GAG CCG GAA GCA TAA AGT GTA AAG CCT GGG GTG CCT AAT GAG TGA GCT

AAC TCA CAT TAA TTG CGT TGC GCT CAC TGC CCG CTT TCC-C3 Spacer-3'; and Blocker_dnr1_3: 5'-Biotin-AGT CGG GAA ACC TGT CGT GCC AGC TGC ATT AAT GAA TCG GCC AAC GCG CGG GGA GAG GCG GTT TGC GTA TTG GGC GCT CTT CCG CTT CCT CGC TCA CTG ACT CGC TGC GCT CGG TCG TTC-C3 Spacer-3'. All capture and blocking oligos purchased from IDT as Ultramer DNA Oligos with Standard Desalting. Captured library fragments were further amplified by 12 cycles of aLM-PCR and run using an Agilent 2100 Bioanalyzer to estimate enrichment. Captured libraries were quantified by Qubit Fluorometer and quantitative real-time PCR (qRT-PCR) were used to dilute libraries for template preparation and Ion Sphere Particle (ISP) loading using Ion OneTouch 2 System and sequencing on the Ion Proton platform with PI_{v3} semiconductor wafer chips per manufacturers recommended instructions. High-throughput sequencing of up to eight multiplex captured libraries was carried out per chip to achieve at least 5 million reads per barcode. Reads containing the transposon IRDR element were identified by the sequence TAC AGT TGA AGT CGG (or its reverse compliment). If vector sequence (derived from the donor concatamer) was detected adjacent to the IRDR, these reads were filtered away. Remaining reads were then split at the TA end of the IRDR sequence, producing mouse and transposon sequences. The mouse sequences were mapped to a custom genomic reference containing the mm9 genome and the T2/Onc2 transposon sequence (mm9+pT2/Onc). Reads with MAPQ < 30 (at least a 1/1,000 chance of mapping incorrectly), with secondary alignments, or that mapped to T2/Onc2 were discarded. Transposon sequences from remaining reads were aligned to the mm9+pT2/Onc2 reference. If they mapped with MAPQ ≥ 30 and no secondary alignments, then they were considered to have a distinct IRDR (left or right). Reads without a distinct IRDR that did not map to the same locations as reads with a recognizable IRDR were discarded. If there was an error in two bases at the TA-end of the mouse read, the read was discarded. Alignments were performed using Bowtie 2 with the '-very-sensitive' flag enabled. Insertion sites were then defined by the genomic position of the TA end of the mouse sequences. If an insertion mapped to a repeat TA sequence, then the site was assigned to the 5' end of the repeat (with respect to the chromosome) and its reads were merged with other sites in the same repeat. Insertion sites with mouse DNA on the 5' end of the site and IRDR transposon were labeled as having a forward orientation, as were reads with 3' mouse DNA and IRDR transposon. Likewise, sites with 3' mouse DNA and IRDR transposon or 5' mouse DNA with IRDR transposon were assigned a reverse orientation. Unique reads were identified at each site on the basis of the position (in the mouse or transposon) of the read start. Sites with no discernable unique reads or orientations were discarded. Insertion sites in *En2*, *Foxf2*, or *Serinc3* were discarded, as these contain sequences found in the transposon.

Mapping transposon insertion sites using splink 454

SB insertion reads were generated by 454 GS Titanium sequencing (Roche) of pooled splinkerette PCR reactions with nested, barcoded primers was performed^{3,47,48}. Pre- and post-processing of 454 reads to assign sample DNA barcodes, filter out local hopping events from donor chromosomes, and map and orient the SB insertion sites across the entire nuclear genome of the mouse was performed. All SB insertions from donor chromosomes were filtered out before identification of common insertion sites using the Gaussian kernel convolution (GKC)^{48,49} and gCIS²⁶ methods.

Mapping transposon insertion sites using splink HiSeq

For selective SB insertion site sequencing by HiSeq (Illumina), we modified a published splinkerette PCR protocol²⁶. gDNA (3 µg per sample) was acoustically sheared to 200 bp using Covaris S220 Focused-ultrasonicator and end repaired using Fast DNA End Repair Kit (Thermo Fisher Scientific). Adapters (linker+: 5'-GTA ATA CGA CTC ACT ATA GGG CTC CGC TTA AGG GAC-3' and linker-: 5'-Phos-GTC CCT TAA GCG GAG-C3spacer-3') were annealed and ligated to repaired DNA. Purified repaired DNA was PCR amplified using linker (5'-GTA ATA CGA CTC ACT ATA GGG C-3'), IRR (5'-GGA TTA AAT GTC AGG AAT TGT GAA AA-3'), and IRL (5'-AAA TTT GTG GAG TAG TTG AAA AAC GA-3') primers. Nested secondary PCR was performed using indexed SB-2ndR primer (5'-CAA GCA GAA GAC GGC ATA CGA GAT (8 bp index) GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC TTA GGG CTC CGC TTA AGG GAC-3'), SB-RF primer (AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T (1–4 bp stagger) GGC TAA GGT GTA TGT AAA CTT CCG ACT TCA ACT G-3'), and SB-LF primer (AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T (1–4 bp stagger) AAC TTA AGT GTA TGT AAA CTT CCG ACT TCA ACT G-3'). All oligos were purchased from Integrated DNA Technologies with Standard Desalting. DNA fragments between 175 and 450 bp amplicon libraries were excised from the agarose gel and quantified by Qubit Fluorometer, Agilent 2100 Bioanalyzer and qRT-PCR. Excised libraries were then pooled, loaded onto two flow cell lanes and sequenced on the HiSeq2500 platform. Mapping of SB insertions to the mouse genome used a custom splink HiSeq bioinformatics workflow modeled to perform similarly to the splink 454 analysis pipeline⁴⁸ (J.Y.N., K.M.M., N.A.J., N.G.C. and M.B.M., unpublished data).

WGS

For low coverage WGS from leukemic spleen, gDNA (1 µg) was used for library construction on the AB Library Builder System, including random fragmentation and ligation of barcoded Ion Xpress sequencing adapters. Adapter-ligated templates were purified by Agencourt AMPure beads and fragments with insert size of 200 bp were selected using Pippin prep (Sage Science), purified by Agencourt AMPure beads, amplified by eight cycles of aLM-PCR, and purified by Agencourt AMPure beads with elution in 50 µl of TE (1x Tris-EDTA, pH 8). Ion-WGS library fragments were run using an Agilent 2100 Bioanalyzer to estimate enrichment and quantified by Qubit Fluorometer and qRT-PCR to dilute libraries for template preparation and Ion Sphere Particle (ISP) loading using Ion OneTouch 2 System and sequencing on the Ion Proton platform with PI_{v2} semiconductor wafer chips per manufacturers recommended instructions. Raw sequencing read files were generated by Ion Torrent Suite. Data from four independent PI_{v2} chips were combined to achieve 31.1 GB representing 14.2x coverage of the mouse genome. Subsequent bioinformatic processing using BAM and/or FASTQ files was performed at Houston Methodist Research Institute using a custom built SBCaptureSeq analysis pipeline (see below). For high-coverage WGS from leukemic spleen and normal tail (from littermate of the same gender), gDNA (5 µg per sample) was used for paired-end library preparation (random fragmentation, ligation of sequencing adapters, and size selection), cluster generation (DNA binding to flow cell, primer extension, amplification, linearization,

blocking and primer hybridization), and sequenced to 100x average genome coverage using 10 lanes per specimen on the Illumina HiSeq 2500 platform. Raw image files were processed using Illumina Consensus Assessment of Sequence and Variation (CASAVA 1.8.4) for base calling with default parameters and the sequences of each individual were generated as 100-bp paired-end reads. Library preparation and sequencing were performed by SeqWright. Bioinformatic processing using supplied FASTQ files was performed at Houston Methodist Research Institute using both a custom built mouse whole-genome analysis pipeline and a custom built SBCaptureSeq analysis pipeline.

Transcriptome sequencing

Total RNA was isolated from flash-frozen necropsy specimens using mirVana miRNA Isolation Kit (Ambion by Life Technologies, AM1560) from a shaved portion of leukemic spleen or the entire normal spleen (from two siblings that did not inherit the *Actb-Cre* transgene). For poly(A) RNA isolation (mRNA-seq), 5 µg total RNA isolated from D5 leukemic spleen was mixed with ERCC RNA Spike-In (Ambion by Life Technologies, 4456739) and used for poly(A) enrichment using Dynabeads mRNA DIRECT Micro Purification Kit (Ambion by Life Technologies, 61021). 200 ng poly(A)-enriched mRNA was used for whole-transcriptome library construction according to the Ion Total RNA-seq Kit v2 (Life Technologies, 4475936), protocol for non-barcoded libraries, Ion Sphere Particle (ISP) loading using Ion OneTouch 2 System and sequencing on the Ion Proton platform with PIv2 semiconductor wafer chips per the manufacturer's instructions. For whole-transcriptome RNA-seq (wtRNA-seq), total RNA (5 µg per sample) plus ERCC RNA Spike-In (Ambion by Life Technologies, 4456739) was used for selective ribosomal RNA (rRNA) depletion using the RiboMinus Eukaryote System v2 (Ambion by Life Technologies, A15026). rRNA-depleted total RNA (500 ng per sample) was used for whole-transcriptome library construction according to the Ion Total RNA-seq Kit for the AB Library Builder System (Life Technologies, 4482416) protocol for barcoded libraries, Ion Sphere Particle (ISP) loading using Ion OneTouch 2 System and sequencing on the Ion Proton platform with PIv3 semiconductor wafer chips per manufacturers recommended instructions. Each RNA-seq library was sequenced individually on a PIv2 chip to achieve >4 GB (~70 million reads) per specimen. The Life Technologies Torrent Suite software was used to perform checking of the raw sequence data before the generation of sequencing read files in FASTQ format. Minimally processed mRNA-seq and wtRNA-seq reads were processed using Bowtie⁵⁰ and Tophat⁵¹ algorithms to align reads to a custom version of the mouse mm9 genome, which included the pT2/Onc transposon sequence as an additional chromosome (denoted mm9+pT2/Onc).

Microarray gene-expression analysis

Gene-expression profiling of SB-ML mouse spleens was performed using Affymetrix microarrays. RNA was extracted using a NORGEN Biotek Animal Tissue RNA Purification kit (25700) per the manufacturer's instructions. RNA was then labeled using an Affymetrix 3' IVT Express kit (901229) using 100 ng total RNA for each sample per the manufacturer's instructions. After labeling, samples were hybridized to Affymetrix GeneChip Mouse Genome 430 2.0 Arrays, and scanned at the University of Otago Genomics and Bioinformatics Facility. Raw data were processed in R (version 2.15)⁵² using

the ‘rma’ function of the ‘affy’ package⁵³ (quantile normalization, no background correction). The ‘affyQCReport’ package (<http://rpackages.ianhowson.com/bioc/affyQCReport/>) or R⁵² was used to perform quality assessment of the microarray data.

Identification of common insertion sites

Identification of statistically significant common insertion sites (CISs), cooperating loci, and SB insertion abundance profiles was performed using a BED-formatted file containing the genomic locations of the filtered list of SB insertion sites from the SB-ML genomes.

Trunk driver analysis

The BED file containing all SB insertions from the 168 ML spleens sequenced by splink 454 was filtered to consider only insertions represented by five or more reads in three or more tumors. When genes from a ML genome were found to contain 1 SB insertion in the same RefSeq gene, only the SB insertion with the highest sequence read count was used for gCIS calculations. gCIS statistical analysis was then performed to identify the genes significantly enriched for insertions at high read counts with a corrected $P < 0.05$. 35 genes met these criteria and were termed trunk driver genes on the basis of the inferred high clonality across the tumor population.

Pathway analysis

Three independent databases, IPA (Ingenuity Pathway Analysis), DAVID (Database for Annotation, Visualization and Integrated Discovery) and GATHER (Gene Annotation Tool to Help Explain Relationships) were used to interrogate all ML CCGs for the enrichment of specific signaling pathways and processes.

Variant calling in WGS

Genomic variants (SNVs, indels and CNAs) were identified in the WGS data using an analysis pipeline centered on the VarScan2 (version 2.3.6)⁵⁴. This involved alignment to a custom genomic reference containing the mm9 genome and the T2/Onc2 transposon sequence (mm9+pT2/ Onc) using BWA followed by creation of mpileup files using samtools (version 0.1.19–44428cd)⁵⁵. These files were then used as the input to VarScan2 for calling SNVs and indels followed by ANNOVAR (accessed 23 August 2013)⁵⁶ software to assign potential functional relevance. The indel data were used to identify putative ‘footprints’ of previous SB transposon presence, by filtering for sequence motifs known to be indicative of transposon excision. CNV regions were identified using VarScan2 and further analyzed in R⁵² using the DNACopy package (version 1.36.0)⁵⁷ to find areas of gain and loss.

Detecting SB-fusion events in RNA-seq data

Sequencing reads from poly(A) and ribo-depleted libraries were aligned to the Lun-splice donor, En2 splice acceptor, and non-En2 splice acceptor regions of the T2/Onc2 sequence in Bowtie2 (ref. 50) using the ‘–very-sensitive-local’ flag. The CIGAR string associated with mapped reads was used to clip away the transposon portion, and resulting sequences <20 bp were discarded. Remaining sequences were then aligned to the mm9+pT2/Onc2 reference using Bowtie 2 with the ‘–very-sensitive’ flag enabled. Reads that failed to map, mapped

with a MAPQ less than 30, or mapped to T2/Onc2 were discarded. The coordinates of the mouse-genome fusion, the associated region in the transposon, and the orientation of the read relative to the genome were recorded in BED format.

Defining novel SB-fusion transcripts

Using the SB insertion and fusion information, along with reference transcript annotations (downloaded from UCSC in genePred format and converted to GTF format), we identified three types of novel transcripts that contained SB insertion and fusion information. To do this, we first identified SB insertions as being activating or inactivating: activating when the insertion orientation (defined with respect to the chromosome) is sense with the gene, and inactivating when the insertion orientation is antisense with the gene. The first type of novel transcript was defined as having a splice acceptor fusion in the exon upstream of an activating insertion (SB_5p). The second was defined as containing a splice donor in the exon downstream of an activating insertion (SB_3p). The third type was defined as having an En2 splice acceptor fusion in the exon upstream of an inactivating insertion (SB_5palt). Insertions within 15 kb upstream of a gene-coding region were associated with the transcripts from the downstream gene.

Measuring transcript abundance in normal and SB tumor samples

Reads from ribo-depleted RNA libraries were mapped to mm9+pT2/Onc2 using Tophat2 (ref. 58) with default settings. Mapped reads were used with the reference transcript annotations to determine differential transcript abundance between normal and tumor samples in Cuffdiff2 (ref. 59) and transcript abundances in the normal sample were reported in FPKM values. Cuffdiff2 was run again with the SB-fusion transcripts incorporated, and transcript abundances for the tumor samples were reported in FPKM values. Abundance values, SB insertions, and SB fusions were collated by transcript, along with other transcript annotations.

Code availability and software

Source code and scripts used are available in Supplementary Code. Unless otherwise noted, the aforementioned analysis pipelines, report generation, and figure visualization performed using bash, R, and Python scripts. The circular genomic plots in Supplementary Figure 5a,b were generated using the R package circlize⁶⁰. Hierarchical clustering was performed in Python 2.7.10 with the scipy 0.13.0b1 toolbox using the Hamming distance metric with Ward's linkage method. The R packages ggplot2 (ref. 61) and Gviz⁶² were used to generate the graphics in Figure 7.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank D. Adams, T. Whipp, R. Rance and the Wellcome Trust Sanger Institute sequencing and informatics teams for 454 sequencing; the Institute for Molecular and Cell Biology Histopathology Core; P. Cheok, N. Lim, D. Chen and C. Wee for assistance with tumor monitoring and animal husbandry at IMCB (Singapore), H.

Lee and E. Freiter for assistance with animal husbandry at HMRI (Houston), R. Zahr (Integrated DNA Technologies, Inc.) for assistance with SBCapture probe design and D. Adams and C. Print for valuable discussions and critical reading of the manuscript. Histology work was performed by the Advanced Molecular Pathology Laboratory, IMCB, A*STAR. This work was supported by the Cancer Prevention Research Institute of Texas (N.G.C. and N.A.J.), the Biomedical Research Council, Agency for Science, Technology, and Research, Singapore (N.G.C. and N.A.J.), Cancer Research UK (A.G.R.), the Medical Research Council, UK (A.G.R.) and the Wellcome Trust (A.G.R.).

References

1. Mann MB, Jenkins NA, Copeland NG, Mann KM. Sleeping Beauty mutagenesis: exploiting forward genetic screens for cancer gene discovery. *Curr. Opin. Genet. Dev.* 2014; 24:16–22. [PubMed: 24657532]
2. Mann KM, Jenkins NA, Copeland NG, Mann MB. Transposon insertional mutagenesis models of cancer. *Cold Spring Harb. Protoc.* 2014; 2014:235–247. [PubMed: 24591685]
3. Mann MB, et al. Transposon mutagenesis identifies genetic drivers of *Brat*^{V600E} melanoma. *Nat. Genet.* 2015; 47:486–495. [PubMed: 25848750]
4. Takeda H, et al. Transposon mutagenesis identifies genes and evolutionary forces driving gastrointestinal tract tumor progression. *Nat. Genet.* 2015; 47:142–150. [PubMed: 25559195]
5. Genovesi LA, et al. *Sleeping Beauty* mutagenesis in a mouse medulloblastoma model defines networks that discriminate between human molecular subgroups. *Proc. Natl. Acad. Sci. USA.* 2013; 110:E4325–E4334. [PubMed: 24167280]
6. Pérez-Mancera PA, et al. The deubiquitinase USP9X suppresses pancreatic ductal adenocarcinoma. *Nature.* 2012; 486:266–270. [PubMed: 22699621]
7. Dupuy AJ, et al. A modified *Sleeping Beauty* transposon system that can be used to model a wide variety of human cancers in mice. *Cancer Res.* 2009; 69:8150–8156. [PubMed: 19808965]
8. Lewandoski M, Meyers EN, Martin GR. Analysis of *Fgf8* gene function in vertebrate development. *Cold Spring Harb. Symp. Quant. Biol.* 1997; 62:159–168. [PubMed: 9598348]
9. Jonkers J, et al. Synergistic tumor suppressor activity of *BRCA2* and p53 in a conditional mouse model for breast cancer. *Nat. Genet.* 2001; 29:418–425. [PubMed: 11694875]
10. Olive KP, et al. Mutant p53 gain of function in two mouse models of Li-Fraumeni syndrome. *Cell.* 2004; 119:847–860. [PubMed: 15607980]
11. Doyle B, et al. p53 mutation and loss have different effects on tumorigenesis in a novel mouse model of pleomorphic rhabdomyosarcoma. *J. Pathol.* 2010; 222:129–137. [PubMed: 20662002]
12. Hanel W, et al. Two hot spot mutant p53 mouse models display differential gain of function in tumorigenesis. *Cell Death Differ.* 2013; 20:898–909. [PubMed: 23538418]
13. Lang GA, et al. Gain of function of a p53 hot spot mutation in a mouse model of Li-Fraumeni syndrome. *Cell.* 2004; 119:861–872. [PubMed: 15607981]
14. Morton JP, et al. Mutant p53 drives metastasis and overcomes growth arrest/senescence in pancreatic cancer. *Proc. Natl. Acad. Sci. USA.* 2010; 107:246–251. [PubMed: 20018721]
15. Cesta MF. Normal structure, function, and histology of the spleen. *Toxicol. Pathol.* 2006; 34:455–465. [PubMed: 17067939]
16. Anderson K, et al. Ectopic expression of PAX5 promotes maintenance of biphenotypic myeloid progenitors coexpressing myeloid and B-cell lineage-associated genes. *Blood.* 2007; 109:3697–3705. [PubMed: 17218387]
17. Swerdlow SH, et al. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. 4. Vol. 2. The International Agency for Research on Cancer of the World Health Organization; 2008.
18. Futreal PA, et al. A census of human cancer genes. *Nat. Rev. Cancer.* 2004; 4:177–183. [PubMed: 14993899]
19. Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* 2013; 368:2059–2074. [PubMed: 23634996]
20. Tang JZ, et al. Transposon mutagenesis reveals cooperation of ETS family transcription factors with signaling pathways in erythro-megakaryocytic leukemia. *Proc. Natl. Acad. Sci. USA.* 2013; 110:6091–6096. [PubMed: 23533276]

21. Vassiliou GS, et al. Mutant nucleophosmin and cooperating pathways drive leukemia initiation and progression in mice. *Nat. Genet.* 2011; 43:470–475. [PubMed: 21441929]
22. Milella M, et al. Therapeutic targeting of the MEK/MAPK signal transduction module in acute myeloid leukemia. *J. Clin. Invest.* 2001; 108:851–859. [PubMed: 11560954]
23. Ricciardi MR, et al. Quantitative single cell determination of ERK phosphorylation and regulation in relapsed and refractory primary acute myeloid leukemia. *Leukemia.* 2005; 19:1543–1549. [PubMed: 16001087]
24. Friedel RH, et al. Clonal expansion analysis of transposon insertions by high-throughput sequencing identifies candidate cancer genes in a PiggyBac mutagenesis screen. *PLoS One.* 2013; 8:e72338. [PubMed: 23940809]
25. Koudijs MJ, et al. High-throughput semiquantitative analysis of insertional mutations in heterogeneous tumors. *Genome Res.* 2011; 21:2181–2189. [PubMed: 21852388]
26. Brett BT, et al. Novel molecular and computational methods improve the accuracy of insertion site analysis in *Sleeping Beauty*-induced tumors. *PLoS One.* 2011; 6:e24668. [PubMed: 21931803]
27. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 2012; 40:e3. [PubMed: 22021376]
28. Lou DI, et al. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl. Acad. Sci. USA.* 2013; 110:19872–19877. [PubMed: 24243955]
29. Liu G, et al. Target-site preferences of *Sleeping Beauty* transposons. *J. Mol. Biol.* 2005; 346:161–173. [PubMed: 15663935]
30. Vogelstein B, et al. Cancer genome landscapes. *Science.* 2013; 339:1546–1558. [PubMed: 23539594]
31. Wang Y, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature.* 2014; 512:155–160. [PubMed: 25079324]
32. Tanaka Y, et al. The transcriptional programme controlled by Runx1 during early embryonic blood development. *Dev. Biol.* 2012; 366:404–419. [PubMed: 22554697]
33. Taoudi S, et al. ERG dependence distinguishes developmental control of hematopoietic stem cell maintenance from hematopoietic specification. *Genes Dev.* 2011; 25:251–262. [PubMed: 21245161]
34. Thoms JA, et al. ERG promotes T-acute lymphoblastic leukemia and is transcriptionally regulated in leukemic cells by a stem cell enhancer. *Blood.* 2011; 117:7079–7089. [PubMed: 21536859]
35. Marcucci G, et al. Cancer and Leukemia Group B Study. High expression levels of the ETS-related gene, ERG, predict adverse outcome and improve molecular risk-based classification of cytogenetically normal acute myeloid leukemia: a Cancer and Leukemia Group B Study. *J. Clin. Oncol.* 2007; 25:3337–3343. [PubMed: 17577018]
36. Hayashi AA, Proud CG. The rapid activation of protein synthesis by growth hormone requires signaling through mTOR. *Am. J. Physiol. Endocrinol. Metab.* 2007; 292:E1647–E1655. [PubMed: 17284572]
37. Sliwa T, et al. Hyperexpression of NOTCH-1 is found in immature acute myeloid leukemia. *Int. J. Clin. Exp. Pathol.* 2014; 7:882–889. [PubMed: 24695418]
38. Wu Y, et al. Therapeutic antibody targeting of individual Notch receptors. *Nature.* 2010; 464:1052–1057. [PubMed: 20393564]
39. Marzali G, et al. Role of Ets-1 in erythroid differentiation. *Blood Cells Mol. Dis.* 2002; 29:553–561. [PubMed: 12547252]
40. Sottoriva A, et al. A Big Bang model of human colorectal tumor growth. *Nat. Genet.* 2015; 47:209–216. [PubMed: 25665006]
41. Busch K, et al. Fundamental properties of unperturbed haematopoiesis from stem cells *in vivo*. *Nature.* 2015; 518:542–546. [PubMed: 25686605]
42. Sun J, et al. Clonal dynamics of native haematopoiesis. *Nature.* 2014; 514:322–327. [PubMed: 25296256]
43. Dupuy AJ, Akagi K, Largaespada DA, Copeland NG, Jenkins NA. Mammalian mutagenesis using a highly mobile somatic *Sleeping Beauty* transposon system. *Nature.* 2005; 436:221–226. [PubMed: 16015321]

44. Starr TK, et al. A transposon-based genetic screen in mice identifies genes altered in colorectal cancer. *Science*. 2009; 323:1747–1750. [PubMed: 19251594]
45. Rehg JE, Bush D, Ward JM. The utility of immunohistochemistry for the identification of hematopoietic and lymphoid cells in normal tissues and interpretation of proliferative and inflammatory lesions of mice and rats. *Toxicol. Pathol.* 2012; 40:345–374. [PubMed: 22434870]
46. Leung ML, Wang Y, Waters J, Navin NE. SNES: single nucleus exome sequencing. *Genome Biol.* 2015; 16:55. [PubMed: 25853327]
47. Mann KM, et al. Australian Pancreatic Cancer Genome Initiative. *Sleeping Beauty* mutagenesis reveals cooperating mutations and pathways in pancreatic adenocarcinoma. *Proc. Natl. Acad. Sci. USA*. 2012; 109:5934–5941. [PubMed: 22421440]
48. March HN, et al. Insertional mutagenesis identifies multiple networks of cooperating genes driving intestinal tumorigenesis. *Nat. Genet.* 2011; 43:1202–1209. [PubMed: 22057237]
49. de Ridder J, Uren A, Kool J, Reinders M, Wessels L. Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens. *PLoS Comput. Biol.* 2006; 2:e166. [PubMed: 17154714]
50. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009; 10:R25. [PubMed: 19261174]
51. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics*. 2009; 25:1105–1111. [PubMed: 19289445]
52. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; 2013.
53. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 2004; 20:307–315. [PubMed: 14960456]
54. Koboldt DC, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012; 22:568–576. [PubMed: 22300766]
55. Li H, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; 25:2078–2079. [PubMed: 19505943]
56. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010; 38:e164. [PubMed: 20601685]
57. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*. 2007; 23:657–663. [PubMed: 17234643]
58. Kim D, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 2013; 14:R36. [PubMed: 23618408]
59. Trapnell C, et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* 2013; 31:46–53. [PubMed: 23222703]
60. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. *Bioinformatics*. 2014; 30:2811–2812. [PubMed: 24930139]
61. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer; 2009.
62. Henare K, et al. Dissection of stromal and cancer cell-derived signals in melanoma xenografts before and after treatment with DMXAA. *Br. J. Cancer*. 2012; 106:1134–1147. [PubMed: 22415295]

Figure 1.

SB mutagenesis drives ML development in mice. (a) Three cohorts of mice with wild-type *Trp53* (*Trp53*^{+/+}), *Trp53*^{+/-} or *Trp53*^{R172H/+} were generated and aged for tumors. All mice developed hematopoietic disease but *Trp53*^{+/-} and *Trp53*^{R172H/+} mice had significantly decreased survival compared to the WT cohort ($P < 0.0001$, log-rank). (b,c) Histological analysis of tumor cell differentiation in the red pulp of the spleen (b) and in the subcapsular area (c), shown at 1,000x magnification. Histological analysis was performed on all mice in each cohort. Tumor cell differentiation representative of all cohorts is shown in (b) for the red pulp in a *Trp53*^{+/+} animal and (c) for the subcapsular area in a *Trp53*^{+/-} animal. (d,e) MPO staining in poorly differentiated regions (d) and PAX5 staining of tumor cells in the subcapsular area (e), shown at 400x magnification. Scale bars, 100 μ m (b–e). MPO and PAX5 staining was performed on all animals in each cohort and half showed positive MPO staining (d) in poorly differentiated regions and half showed positive PAX5 staining (e) in the subcapsular area. Representative images from *Trp53*^{+/+} mice are shown. (f) Statistical analysis of SB insertions present in 3 tumors at a read depth of 5 (based on splink 454 sequencing) identified 35 CCGs that are highly represented in SB-ML. Genes are plotted on the basis of mean read depth (y axis) relative to the median number of reads (x axis); SB insertions are predicted to activate 80% of these genes (blue) and inactivate 20% (red). CIS, common insertion site.

Figure 2.

Trunk drivers are enriched in key signaling pathways and exhibit significant interactions. **(a)** Statistically defined ML trunk drivers are enriched in MAPK, JAK-STAT and JNK-p38 signaling. The incidence of SB insertions in CCGs across the ML tumor population is depicted as an oncoprint; CCGs are shown on the y axis; individual tumors are listed on the x-axis. Bars indicate the presence of an SB insertion on nondonor (black) or donor (gray) chromosomes. **(b)** Top, schematic of SB insertions in *Erg*, which are predicted to drive expression of a truncated transcript encoding the ETS domain. Bottom, microarray analysis of *Erg* expression in tumors with (yes) and without (no) SB insertions in *Erg*. Box

boundaries indicate interquartile range; whiskers indicate maximum and minimum values; center lines indicate medians. (c) Statistically defined relationships among trunk drivers and other ML CCGs. Blue circles represent trunk drivers predicted to be activated by SB insertions; red circles represent genes predicted to be inactivated; gray circles represent genes predicted to co-occur with the trunk drivers. Trunk driver CCGs are shown in black text; ML CCGs not defined as trunk drivers are shown in blue. Solid lines depict co-occurrence relationships; dotted lines indicate mutually exclusive relationships. Line weights indicate the significance of the interaction (thick, $P < 0.001$, medium, $P < 0.01$, and thin, $P < 0.05$; Fisher's exact test).

Figure 3.

The SBCapSeq method for sequencing transposon insertion sites from tumors. **(a)** Capture hybridization probes (orange bars) were designed to hybridize to the first 120 nucleotides from the 5' end of the IRDL transposon inverted repeat and the last 120 nucleotides from the 3' end of the IRDR of the SB transposon. Capture probes contain biotin moieties that allow for capture by streptavidin beads. Blocking oligos (blue bars) were designed from vector sequence to minimize capture of unmobilized transposon remaining at the donor site where the transposon concatamer first integrated. Blocking oligos contain a bulky adduct (yellow circle) that prevents isolation by bead capture. IRDL, inverted repeat direct repeat left; IRDR, inverted repeat direct repeat right; SA, splice acceptor; MSCV, murine stem cell virus minimal promoter; SD, splice donor; En2-SA, engrailed 2 splice acceptor. **(b)** Liquid capture hybridization performed on DNA libraries prepared from sheared genomic tumor DNA enriched for DNA fragments containing transposon sequences. Fragments are

isolated and sequenced on the Ion Torrent platform. A custom SBCapSeq bioinformatics pipeline then maps SB insertions to the mouse genome. (c) Sequence analysis of nine independent SB capture hybridization reactions from a single ML tumor library to assess reproducibility of the method for detecting reads (left) or fragments (right) with SB insertions in genes.

Figure 4. Comparison of SBCapSeq to SB splink 454. (a) Numbers of unique SB insertions detected with each method for ten ML tumors chosen randomly for direct comparison. (b) SB insertions in early progression drivers identified by splink 454 and/or SBCapSeq. (c) SBCapSeq and splink 454 analysis of SB insertions in *Erg* in multiple ML. The mapped locations of all insertions in the analyzed tumor genomes are shown on the left, and the presence of an insertion in a given tumor is denoted by a black bar if detected by splink 454 and a red or pink bar if identified by SBCapSeq. The number of sequencing reads supporting the insertion appear inside the bars. SBC, SBCapSeq.

Figure 5. WGS confirms SB insertions drive ML. **(a)** Distribution and representation of SB insertions throughout the D5 genome. Red bars indicate SB insertions identified by WGS with >1 read; pink bars represent SB insertions represented by a single read. Clonal insertions identified by SBCapSeq with ≥1,000 reads are represented by green bars; insertions with <1,000 reads are shown as orange bars. The cut-off was determined on the basis of the read depth at which the number of insertions in the tumor exceeded the number of insertions detected in tail DNA by fourfold. Data from ribo-depleted (whole-transcriptome sequencing, dark blue), and poly(A)-selected (mRNA, light blue) RNA-seq show coincidence of fusion transcripts with the SB insertions identified from DNA analysis and detection of fusion transcripts from subclonal SB insertion events. Genes with clonal SB insertion events are listed in the center of the plot. Black (significant after Bonferroni correction; $P < 0.05$) and gray bars highlight SB footprints identified by WGS. **(b)** Correlation of ML CCGs identified with high SB read counts by both SBCapSeq and WGS. Activated CCGs are depicted by blue dots; inactivated CCGs are depicted by red dots. Early progression drivers are shown in blue (activating mutations) or black (inactivating mutations) text. Asterisk denotes a secondary *Notch1* SB

insertion. SB insertions that are private to this tumor appear as gray dots if they map to the donor chromosome and black dots if they map to nondonor chromosomes. CIS, common insertion site. (c) SB insertion data from SBCapSeq and WGS for the D5 leukemic genome (D5_SP) and control tail (D6_TL). The number of CCGs was determined by comparing these genes to the 466 defined ML CCGs.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Nat Biotechnol. Author manuscript; available in PMC 2018 September 05.

early passenger events or selected SB events that were not appreciated in earlier population-based analyses. Cells present in the other cluster (top right) were sorted with Ter119 and are anchored by insertions in *Ets1* and *Notch1*. The cell CD71-D5_SP_SC95 grouped with both populations and may represent a failed single-cell sort. Bars at the right of the graph represent the same SB insertions identified in the bulk tumor by WGS and SBCapSeq.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 7.

Clonally selected SB insertion events affect gene expression in SB-ML. **(a,b)** RNA-seq analysis of SB-fusion transcripts confirmed the direct impact of clonally selected SB insertions identified in single cells on gene expression. Genes with SB insertions present in the single-cell cluster anchored by insertions in *Erg* and *Ghr* **(a)** or *Ets1* and *Notch1* **(b)** are shown. All genes contained an insertion found in 3 cells at a read depth of 2. Asterisks denote genes on the donor chromosome. Transcripts identified from bulk analysis of D5 leukemic spleen cells by RNA-seq are represented as the log₁₀ ratio of SB-fusion transcripts compared to wild-type transcripts, defined by fragments per kilobase of transcript per million mapped reads from ribo-depleted (FsWTS) or poly(A)-selected (FsPolyA) RNA-seq. Green bars indicate enrichment of SB-fusion transcripts with SB inserted in the sense direction; yellow bars indicate SB fusion transcripts with SB inserted in the antisense direction. Gray bars represent genes for which the ratio of SB-fusion to normal transcripts was opposite to that predicted by the directionality of the SB insertion. Red and purple boxes denote evidence for SB insertions by WGS and SBCapSeq, respectively. Dark and light blue boxes denote evidence for RNA transcripts containing the splice donor or one of the splice acceptors present in the SB transposon for FsPolyA and FsWTS, respectively. Numbers indicate single cells sharing a single insertion for each gene.