
Justin Y. Newberg  
*Moffitt Cancer Center*

Michael A. Black  
*University of Otago*

Nancy A. Jenkins  
*University of Texas*

Neal G. Copeland  
*University of Texas*

Karen M. Mann  
*University of South Florida*, kmann4@usf.edu

Follow this and additional works at: https://digitalcommons.usf.edu/onc_facpub

See next page for additional authors

Part of the *Oncology Commons*

---

**Scholar Commons Citation**

https://digitalcommons.usf.edu/onc_facpub/8

This Article is brought to you for free and open access by the Oncologic Sciences at Digital Commons @ University of South Florida. It has been accepted for inclusion in Oncologic Sciences Faculty Publications by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact scholarcommons@usf.edu.
SB Driver Analysis: a Sleeping Beauty cancer driver analysis framework for identifying and prioritizing experimentally actionable oncogenes and tumor suppressors

Justin Y. Newberg1,*, Michael A. Black2, Nancy A. Jenkins3, Neal G. Copeland3,*, Karen M. Mann1,4,5,* and Michael B. Mann1,5,6,*

1Department of Molecular Oncology, Moffitt Cancer Center, Tampa, FL, USA, 2Department of Biochemistry, University of Otago, Dunedin, New Zealand, 3Genetics Department, University of Texas MD Anderson Cancer Center, Houston, TX, USA, 4Departments of Gastrointestinal Oncology and Malignant Hematology, Moffitt Cancer Center, Tampa, FL, USA, 5Department of Oncological Sciences, College of Medicine, University of South Florida, Tampa, FL, USA and 6Department of Cutaneous Oncology and Donald A. Adam Melanoma and Skin Cancer Research Center of Excellence, Moffitt Cancer Center, Tampa, FL, USA

Received December 14, 2017; Revised April 05, 2018; Editorial Decision May 08, 2018; Accepted May 10, 2018

ABSTRACT

Cancer driver prioritization for functional analysis of potential actionable therapeutic targets is a significant challenge. Meta-analyses of mutated genes across different human cancer types for driver prioritization has reaffirmed the role of major players in cancer, including KRAS, TP53 and EGFR, but has had limited success in prioritizing genes with non-recurrent mutations in specific cancer types. Sleeping Beauty (SB) insertional mutagenesis is a powerful experimental gene discovery framework to define driver genes in mouse models of human cancers. Meta-analyses of SB datasets across multiple tumor types is a potentially informative approach to prioritize drivers, and complements efforts in human cancers. Here, we report the development of SB Driver Analysis, an in-silico method for defining cancer driver genes that positively contribute to tumor initiation and progression from population-level SB insertion data sets. We demonstrate that SB Driver Analysis computationally prioritizes drivers and defines distinct driver classes from end-stage tumors that predict their putative functions during tumorigenesis. SB Driver Analysis greatly enhances our ability to analyze, interpret and prioritize drivers from SB cancer datasets and will continue to substantially increase our understanding of the genetic basis of cancer.

INTRODUCTION

Forward genetic screens using insertional mutagenesis have been instrumental in identifying large sets of mutations that drive cancer in mouse models of human disease (1–3). Insertional mutagenesis using either retroviruses or DNA transposons relies on the detection of these elements to identify loci that may contain genes or other genomic elements that contribute to tumor development. In 2005, Sleeping Beauty (SB) was first reported as a DNA transposon-based somatic insertional mutagenesis system capable of driving both hematopoietic and solid tumors in the mouse (4,5). SB mutagenesis proved to be advantageous over classic retroviral mutagenesis approaches due to its short-acting effects on targeted genes (through the use of minimal promoter elements) and its ability to create mutations in any cell type in the body. SB is a two-component system, consisting of an SB transposon and an SB transposase (SBase) enzyme that work together to facilitate mobilization of the transposon throughout the genome. SBase binds to the transposon and mediates its excision and reintegration at TA dinucleotides by a cut-and-paste mechanism. Following reintegration, the transposon can activate the expression of a downstream proto-oncogene via the internal promoter and splice donor site; alternatively, it can inactivate the expression of a tumor suppressor gene by inducing premature termination of tran-
scripts via internal splice acceptor and bi-directional polyA sites, essentially functioning as a gene trap.

Over the past ten years, SB forward genetic screens in both hematopoietic and solid tumor models have identified thousands of candidate cancer genes (6). SB datasets have evolved in both size and complexity with the use of multiple transposon donor strains and combinations of sensitizing mutations and tissue-specific promoter-driven Cre recombinase to refine SB-driven mouse models of human cancers. SB cancer gene discovery relies on high throughput sequencing of tumor genomes with accompanying statistical pipelines designed to address the unique complexities associated with SB insertional mutagenesis. Enrichment analysis of insertion tags in tumor cohorts statistically defines those genomic loci, termed common insertion sites (CISs) (7), containing insertions at a greater incidence than expected by chance, or relative to the background mutation rate observed across tumor genomes, and therefore likely contain one or more cancer drivers involved in promoting the initiation and/or progression of cancer. Locus-centric statistical approaches using Monte Carlo (MC) simulation (5,8–10), Gaussian Kernel Convolution (GKC) (7,11), or Poisson distribution statistics (TAPDANCE) (12,13) have been successful in defining CIS loci that are likely to harbor one or more candidate drivers. The GKC method is particularly effective at identifying CISs when there are densely clustered insertions in a locus; however, this method misses CISs when insertions are randomly distributed across loci. Locus-centric approaches filter insertions residing on donor chromosomes; in datasets with multiple donor chromosomes, computational limitations over-estimate the expectations for genes residing on these chromosomes, increasing the rate of false negatives. In addition, data output requires manual curation to identify candidate driver genes associated with a CIS. As the majority of CISs defined by locus-centric methods occur within or in close proximity to gene coding regions, Dupuy and colleagues developed a gene common insertion site (gCIS) analysis method (14) to statistically define drivers using transposon insertions mapped only to genic regions. However, computational requirements and the limited availability of this method precluded its widespread adoption. Importantly, all of these approaches rely on the end-user to classify SB insertions in CIS loci as activating or inactivating expression of candidate driver genes.

Given the wealth of published SB datasets (5,15–22), there is an opportunity to perform meta-analyses across SB cancer models and reanalyze tumors grouped by various biological characteristics. To enable meta-analyses of SB transposon data, we introduce a gene-centric driver analysis, SB Driver Analysis, along with its implementation in a simple-to-run command-line application. This statistical approach accommodates large datasets generated using multiple SB transposon donors by partitioning insertions present on donor chromosomes on a per-tumor basis, while adjusting expectations accordingly for each gene. This obviates the limitations of the locus-centric methods, which require the user to either run analyses only on tumors with the same donor chromosomes or to mask all donor chromosomes for each analysis, thereby censoring insertion data and limiting applicability of these methods in meta-analyses.

Here we describe the SB Driver Analysis enhancements that allow users to define drivers based on various mapping criteria (such as inclusion of insertions upstream of coding regions) and stringency (based on selected method of multiple hypothesis testing correction). We tune the stringency parameters to derive different types of driver analysis results: Discovery Drivers, which are genes statistically enriched with insertions in a population of tumors; Progression Drivers, a more (statistically) stringent defined subset of Discovery Drivers; and Trunk Drivers, a set of drivers associated with high read depth insertion sites, indicative of early initiating events from clonally expanded populations of cells. SB Driver Analysis is available for download at http://sbcddb.moffitt.org/software/. It relies on minimal dependencies (e.g. SciPy, NumPy) (23) and contains embedded annotations, with functionality that allows for user-defined annotations. Due to its flexibility, SB Driver Analysis is a powerful tool for prioritizing recurrent drivers across SB studies for comparative genomic analysis in human cancers.

MATERIALS AND METHODS

Datasets and annotations

Tumors from transposon screens sequenced using the PCR-based, 454 sequencing platform were mapped to TA dinucleotides in the mouse genome (mm9) using a previously established workflow (11). We then used more than 1 million SB insertions occurring within 17 primary tumors listed in the Sleeping Beauty Cancer Driver Database (SBDDDB) (6) to define cancer driver genes using SB Driver Analysis. Insertions from the myeloid leukemia (ML) dataset are included in Supplemental Table S1, and links to previously published BED files can be found in the software linked on the SBDDDB website (http://sbcddb.moffitt.org/software/). RefGene annotations in genePred format were downloaded from the UCSC Genome Browser (n = 24 341 genes). Genes associated with multiple chromosomes or strands, or those that were greater than 5 MB in size, were removed. Additionally, we performed a liftOver of transposon data to mm10, and downloaded the corresponding mm10 reference sequences and RefGene annotations (n = 24 371 genes). Note that annotations from sources such as RefSeq, GENCODE, or Ensembl may be used as long as they are converted to genePred format.

Defining non-redundant TA sites

SB insertions occur exclusively at TA dinucleotides (17). In order to ascertain the significance of insertions in genes, we tabulated every TA site in the mouse genome that maps non-redundantly within a 20-base sequence, as this corresponds to the length of sequences produced by the splinkerette pipeline used by many published transposon screens (Supplemental Figure S1). These TA sites were then tallied across each chromosome (Supplemental Tables S2 and S3) and in each gene (Supplemental Tables S4 and S5). For genes with multiple isoforms, we defined the gene boundaries to extend from the nucleotide position at the beginning...
of the most 5’ feature (UTR/exon) to the nucleotide position at the end of the most 3’ feature (UTR/exon). These boundaries define the gene-coding regions for our analyses, though it is possible that the start and end do not correspond to a single known transcript. For each gene, \( g \), the number of bases \( (B_g) \) and unique TA sites \( (T_g) \) were tallied. A small number of loci were defined by discontinuous coding regions, including \( Mecom \) which encodes a complex locus involving a dual-protein read-through transcript for previously separately annotated genes \( Evil \) and \( Mds1 \).

All bases and TA sites contained between these well annotated isoforms were included in the tally. The unique TA dinucleotide sites across the mouse mm9 and mm10 reference genomes are shown in Supplemental Tables S2 and S3, respectively. After application of these criteria, 24 172 annotated genes in the mm9 (Supplemental Table S4) or 24 218 annotated genes in the mm10 (Supplemental Table S5) genomes remained (genome build for \( SB \) Driver Analysis is chosen to match the reference genome used for mapping sequencing reads).

**Overlaying \( SB \) insertion data and gene annotations**

\( SB \) insertion sites, stored in six-column Browser Extensible Data (BED) format and defined by established preprocessing approaches for Splink 454 (11) were mapped to genes, and the observed number of tumors with an insertion in each gene was tallied. BED detail format files can be modified to contain either a seventh column or a header track relating tumors to transposons and score thresholds that distinguish between low- and high-depth reads, or a tumor annotation file can be included when performing the analysis. All genes beginning with ‘Gm’ (predicted genes) or known \( SB \) or mapping hotspots (\( Dpp10, En2, Foxf2, Serinc3, Sfi1 \)) were excluded (mm9, \( n = 23 \) 039 genes remain). Insertions in genes positioned on the transposon donor chromosome were ignored on a per tumor basis. For each gene, the total number of tumors in which the gene was not on the donor chromosome \( (N_g) \) and the number of unique TA sites across all non-donor chromosomes \( (U_g) \) were tallied. The workflow for annotating data is summarized in Figure 1A.

**Identification of statistically significant drivers: genes with more insertions than expected by chance**

In order to ascertain in a given gene, \( g \), whether a population of tumors contains more insertions than expected by chance, we performed a chi-squared test for each gene. We defined an expectation (i.e. the expected number of tumors, \( E_g \), with insertions in gene \( g \)), as

\[
E_g = \sum_t 1 - \left( 1 - \frac{T_g}{U_g} \right)^I_t
\]

where \( I_t \) is the number of observed non-donor transposon insertion sites in a tumor, \( t \), \( T_g \) is the number of unique TA sites in a gene, and \( U_g \) is the number TA sites in the genome as described previously. This expectation was used, along with the number of tumors with observed non-donor insertions in the gene \( (O_g) \), and the total number of tumors \( (N_g) \) in which the gene does not reside on the donor chromosome, to calculate the chi-squared statistic

\[
\chi^2_g = \sum_{i \in \{f, r, k\}} \frac{(O_i - E_i)^2}{E_i} = \frac{(O_g - E_g)^2}{E_g} + \frac{(O_r - E_r)^2}{E_r}
\]

where

\[
O_g = N_g - O_g \quad E_g = N_g - E_g
\]

from which a \( P \) value was determined (assuming one degree of freedom). A multiple-testing correction procedure, either family-wise error rate (FWER, e.g. Holm–Bonferroni) (24) or false discovery rate (FDR) (25), was applied across all genes. Since the \( P \) value relates to deviations from either side of the expectation, genes with \( O_g < E_g \) were flagged as non-significant. Furthermore, genes were flagged as non-significant when \( O_g < 3 \) \( O_g < 2 \) if \( N_g < 15 \) tumors or \( O_g / N_g < 0.05 \). Drivers were ordered and ranked based on \( \chi^2 \) value (since these were bounded in the floating point precision limit, whereas sometimes \( P \) values were below the floating point precision limit, making it impossible to order those genes that had equivalent \( P \) values of 0). The workflow for identifying drivers is summarized in Figure 1B.

**Classification of oncogenes and tumor suppressors**

For statistically significant genes, we tallied the number of forward, \( F_g \), and reverse, \( R_g \), insertions, and calculated the ratio of forward to reverse insertions, \( r_g \). When multiple insertions were detected in a gene within an individual tumor, only the highest-read depth site was tallied. A binomial test was used to determine the probability, \( p_{g,bs} \), of detecting \( F_g \) given \( F_g + R_g \) insertions. If there were at least three forward insertions, we next evaluated the spatial distribution of insertions by comparing the distribution of forward insertions across the gene to a uniform distribution using a Kolmogorov–Smirnov test, which determines the probability, \( p_{g,bs} \), the insertions were drawn from a uniform distribution. If \( p_{g,bs} < 0.1 \), we labeled the distribution as non-uniform. Finally, we used these metrics in a sequential decision process to assign a label, \( L_g \), denoting whether a gene exhibits an activating, inactivating, or indeterminate insertion pattern:

\[
L_g = \begin{cases} 
\text{Indeterminate default} & (r_g > 0.8) \text{ & } (p_{g,bs} < 0.1) \text{ & } (p_{g,bs} < 0.1) \\
\text{Activating} & (L_g \neq \text{activating}) \text{ & } (r_g < 0.6) \text{ & } (R_g > 2) 
\end{cases}
\]

The workflow for driver gene classification is summarized in Figure 1C.

**Detection of oncogenic insertions upstream of gene boundaries**

We repeated the above steps, this time altering the number of TA sites associated with genes by approximating the number of TA sites in a promotor region upstream of the 5’ end of a putative proto-oncogene gene. For this reason, a 15
Figure 1. Overview of SB Driver Analysis pipeline. The pipeline consists of four general steps. (A) First, number of TA dinucleotides is tallied for each gene in a selected mouse reference genome, followed by the number of tumors in which a gene has an insertion. (B) Using the annotated genes, an expected number of tumors is compared to the observed number of tumors with an insertion in a gene using a chi-squared test to identify whether or not more insertions were observed than would be expected by chance. (C) Genes enriched for insertions are then further characterized based on the insertion pattern in the gene boundaries, or in the gene boundary plus a predefined 5' end promoter (see Table 1). (D) Drivers identified when a promoter is included in gene definitions are merged with results when a promoter is not considered, and a list of drivers and their associated properties are produced in table form. Note that black boxes represent user inputs and SB Driver Analysis output, while white rectangles represent processes that are performed on data. Expectations and observations are related to the number of tumors containing insertions in a gene, and these are dependent upon dataset size (number of insertions, number of tumors) and number of unique TA sites in a gene. Either a family-wise error rate (FWER) or false discovery rate (FDR) multiple-test correction is applied to the gene enrichment determination step.

kb promoter was selected because this placed the transposon promoter in close proximity to the 5' end of the gene coding region. In the case of the gene Rtil a 50 kb promoter was used since past transposon validation studies have flagged this extended promoter as a biologically meaningful region of interest (26). The number of TA sites per gene with promoter, \( T_{g,p} \), was approximated as

\[
T_{g,p} = T_g \frac{P + B_g}{B_g}
\]

where \( P \) is the number of bases used to approximate the promoter, and \( T_g \) and \( B_g \) are TA sites per gene and bases per gene, respectively, as defined previously. Gene expectations (\( E_g \)) were re-derived using \( T_{g,p} \), and observations (\( O_g \)) were re-tallied to account for the sites in promoters. The statistical tests to identify drivers and classify insertion patterns were re-applied, and drivers flagged with activating patterns were added to the list of drivers identified in the absence of a promoter approximation (note that insertions in the promoter are not expected to influence tumor suppressive behavior defined by inactivating or indeterminate patterns). The workflow for the merging of run modes (merging of 0 kb and 15 kb driver lists) and a consolidated driver list report is summarized in Figure 1D.

Trunk driver analysis
Analysis was performed on a subset of insertions with read depths above an empirically determined cutoff. The recurrence criterion was relaxed such that \( O_g / N_g < 0.015 \), or \( O_g < 3 (O_g < 2 \text{ if } N_g < 15) \) were used to define non-significant genes. The cutoffs were chosen on a per-dataset bases, and were designed to select for as few drivers that were present in as many tumors as possible. In practice, this meant on the order of dozens of drivers were identified in upwards of 70% of the tumors in the datasets.

RESULTS
Defining drivers using statistical significance stringency
Driver identification from insertional mutagenesis screens relies on the determination of statistical enrichment of insertions across the genome. We applied SB Driver Analysis across 17 independent datasets from the SBCDDB (6) to define drivers across tumor models using a defined framework. For each tumor dataset, we applied the three statistical analysis methods described in Table 1. First, we performed driver analysis on all insertions using the FDR or FWER multiple testing corrections to identify drivers defined by different stringency metrics. Discovery Drivers are determined using the FDR correction, and Progression Drivers are defined using the more stringent FWER criterion, and these represent a subset of the Discovery Drivers. Because both of these approaches produce large gene lists, we applied SB Driver Analysis with the FWER correction to high read depth insertion sites to identify Trunk Drivers, as high read depths are indicative of early initiating events. These outputs demonstrate a wide range in the number of drivers across datasets (Figure 2). The digestive system tumors (INT-Kras, HCA, PDAC, INT-Trp53 and GAS) contained the greatest number of drivers, while the hematopoietic tumors (e.g. ML, LYM) have the least number of drivers. ML and LYM also contained a significant number of activating drivers (Figure 2, blue region in pie
Trunk drivers 
Progression drivers (FWER) 
Progression drivers (FDR) 
Activating pattern 
Inactivating pattern 
Indeterminate pattern

Figure 2. SB Driver Analysis identifies gene sets characteristic of different tumor types. When applied to high read depth insertion sites, driver analysis using a FWER multiple testing correction identified dozens of Trunk Driver genes (black bars). SB Driver Analysis applied to all insertions identified a larger set of Progression Drivers (using FWER correction, dark gray bars) and Discovery Drivers (using FDR correction, light gray bars). The magnitude of drivers varies greatly with the dataset. For each set of drivers, the breakdown of pattern types is shown in the pie charts (blue = activating, red = inactivating, orange = indeterminate). BRCA, breast cancer; GAS, gastric cancer; HCA, hepatocellular adenoma; INT, Intestinal cancers; KA, keratoacanthoma; LYM, lymphoma; MB, medulloblastoma; MEL, melanoma; ML, myeloid leukemia; OS, osteosarcoma; PCA, prostate cancer; PDAC, pancreatic ductal adenocarcinoma; RMS, rhabdomyosarcoma; SCC, cutaneous squamous cell carcinoma. Note that there are four different intestine datasets, distinguished by the sensitizing mutation used to model the tumors. * denotes datasets whose BED files are either publicly available or included with release of this paper.

chart), highlighting that these tumor types are driven by co-operating proto-oncogenes (17). Strikingly, the solid tumors (e.g., BRCA, MEL, PDAC, SCC, GAS, HCA, INT, KA, MB, OS, PCA, RMS) have a preponderance of inactivating and indeterminate drivers, (Figure 2, red and yellow regions of pie charts, respectively), demonstrating that these tumors are driven by co-operating tumor suppressors. These data support the observation from human cancers that hematopoietic tumors require fewer cooperating events for full transformation compared with solid tumors (27).

Notably, the lengths of genes identified by SB Driver Analysis tend to be larger than the average gene length of all genes in the mouse genome, a phenomenon we also observed in cancer genes defined from human cancers (Supplemental Figure S2). The application of the different analysis methods is referenced in Supplemental Tables S6–S9.

Determination of read depth cutoffs for Trunk Driver analysis

While Trunk or early Progression Drivers have been described for some SB screens (19,28), the empirical criteria and statistical methods used to define these Trunk Drivers have not been consistently applied. We investigated how read depth cutoffs and recurrence criteria impact the overall number of defined high-priority Trunk Drivers in a given dataset (Figure 3). We found that for many of the datasets, the number of Trunk Drivers is invariant to frequency; when there are fewer than 200 tumors, recurrence is influenced by the \( N \geq 3 \) criterion, whereas when there are greater than 200 tumors the \( O_g/N_g \leq 0.015 \) criterion becomes the deciding factor. However, in all of the datasets, the number of drivers detected was highly influenced by the read depth cutoff. Therefore, we fixed a frequency to 1.5% for all of the datasets and chose dataset-specific read depth cutoffs to generate a defined list of Trunk Drivers found in at least 50% of the tumors.

Comparison of SB Driver Analysis to the Gaussian kernel convolution method

We next compared SB Driver Analysis with the Gaussian kernel convolution (GKC) method, which is the most commonly reported method for CIS detection in solid tumors. Using the pancreatic ductal adenocarcinoma (PDAC) dataset, we found statistically significant overlap of Trunk Drivers defined by SB Driver Analysis (Figure 4A) and genes identified by GKC using only high read depth insertion sites. We then extended this analysis to each of the 17 cancer studies, comparing genes identified by GKC with genes detected by Trunk Driver analysis (Figure 4B), Progression Driver analysis (Figure 4C) and Discovery Driver analysis (Figure 4D). Progression and Discovery Driver analyses utilize all insertions regardless of read depth. Overlap between methods tended to be higher when there were more genes identified by the methods; however, in the INT-Apc dataset, GKC consistently identified more drivers, as many of the genes were detected below the recurrence threshold used by SB Driver Analysis. For each dataset, there was a weak-to-moderate correlation between SB Driver Analysis and GKC \( P \)-values, with a mean Pearson correlation of 0.40 (min = 0.08, max = 0.66). Importantly, after excluding the INT-Apc set, 91% of GKC genes with significant \( P \)-values (\( P < 10^{-12} \)) were also identified by SB Driver Analysis. 9% of GKC genes were excluded by SB Driver Analysis due to differences in reference annotations,
Drivers

Progression Drivers are a subset of Discovery Drivers; Trunk Drivers are often also Progression and Discovery Drivers.

<table>
<thead>
<tr>
<th>Analysis method</th>
<th>Analysis method description</th>
<th>Promoter cutoffs</th>
<th>Minimum read depth filtering</th>
<th>Minimum tumor threshold</th>
<th>Multiple hypothesis testing P value correction</th>
<th>Driver classifications</th>
<th>Driver applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery Drivers</td>
<td>Discovery significant progression drivers</td>
<td>0 kb</td>
<td>No</td>
<td>5% or ≥3, whichever is larger</td>
<td>FDR</td>
<td>Keep ‘activating’, ‘inactivating’, and ‘indeterminate’</td>
<td>Pathway enrichment analysis; co-occurrence analysis; comparative oncogenomic analysis</td>
</tr>
<tr>
<td>Progression Drivers</td>
<td>Genome significant progression drivers</td>
<td>15 kb or custom 0 kb</td>
<td>No</td>
<td>5% or ≥3, whichever is larger</td>
<td>FWER</td>
<td>Keep ‘activating’, ‘inactivating’, and ‘indeterminate’</td>
<td>Pathway enrichment analysis; co-occurrence analysis; comparative oncogenomic analysis</td>
</tr>
<tr>
<td>Trunk Drivers</td>
<td>Genome significant trunk drivers</td>
<td>15 kb or custom 0 kb</td>
<td>Yes</td>
<td>1.5% or ≥3, whichever is larger</td>
<td>FWER</td>
<td>Keep ‘activating’, ‘inactivating’, and ‘indeterminate’</td>
<td>Prioritized drivers for co-occurrence analysis and validation studies</td>
</tr>
</tbody>
</table>

Classifying and visualizing SB insertion patterns within drivers

SB Driver Analysis separately classifies drivers as activating, inactivating, or indeterminate based on the pattern of the highest read depth SB insertions per tumor within each gene-coding region. Analysis of the insertion patterns for drivers across 17 tumor cohorts confirmed that drivers predicted to be inactivated in one cohort were generally inactivated across all cohorts in which they were defined as drivers; likewise, drivers predicted to be activated had this designation across cohorts. There were, however, notable exceptions to this observation. Figure 5 shows a composite of insertion patterns for Progression Drivers from all 17 tumor cohorts. Activating patterns were determined for Hras, Erg and a subset of tumors with Zmiz1 insertions (Figure 5A), while inactivation patterns were predominant for drivers Pten and Tcf12 and a subset of tumors with Zmiz1 insertions (Figure 5B). The insertion patterns in Pten and Zmiz1 were indeterminate in a subset of tumors, meaning that the pattern did not deviate significantly from the expected uniform distribution (Figure 5C). Drivers with indeterminate insertion patterns require investigator curation to determine the most likely biological classification as a tu-
Figure 3. Tuning SB Driver Analysis parameters in different datasets. Adjustment of read depth cutoff and recurrence frequency affects the number of detected Trunk Drivers and the number of tumors containing at least one resulting Trunk Driver. (A) The number of Trunk Drivers as a function of driver recurrence frequency (x-axis, in percentage) and insertion read depth cutoff (y-axis). For visualization purposes, any number of genes ≥ 60 was set to red. The crosshairs denote the frequency and cutoff used in the SBCDDB. (B) Percentage of samples with altered trunk driver genes as a function of frequency (x-axis) and read depth cutoff (y-axis). Dataset abbreviations are the same as in the Figure 2 legend.

SB Driver Analysis enables meta-analysis of Trunk Drivers across tumor models

SB Driver Analysis applied across all tumors (n = 1852) in the 17 datasets identified 31 Trunk Driver genes that were altered in a majority (60.4%) of the tumors (n = 1119) (Figure 6). The 31 Trunk Drivers were cataloged alongside the rest of the genes tested by Trunk Driver analysis (Supplemental Table S8). Most of the Trunk Drivers contained inactivating insertion patterns, indicating they are TSGs, but four exhibited activating patterns (proto-oncogenes). Note that three of these four (Erg, Ets1, Flt3) were predominantly altered in blood (LYM, ML) tumors, while Zmiz1 was frequently altered in SCC. 16 of the 31 Trunk Drivers appear in the Cancer Gene Census (version 83) (29), which contains a growing catalogue of 551 genes causally implicated in the initiation and/or progression of cancer (30), including mouse orthologs for Apc, Arid1b, Crebbp, Erg, Flt3, Foxp1, Kmt2c, Lpp, Ncoa2, Nfib, Pten, Ptprk, Smad4, Sndl, Tcf12 and this overlap is greater than expected by chance (P < 0.001, Fisher’s exact test). Nine of the remaining 15 Trunk Drivers that are not yet indexed by the CGC, including Cep350 (19), Ets1 (17), Nfia (16), Pard3 (31), Re11/Riam (26), Usp9x (18,20), Wac (32), Zmiz1 (33), have been independently implicated in contributing to cancer through rigorous in vitro and in vivo experiments. The remaining few Trunk Drivers, including Ankrd11, Chuk, Ctnna1, Dennd1a, Nipbl, Pum1, and the over 800 additional Progression and Discovery Drivers produced by this meta-analysis (Supplemental Table S9), represent high-confidence candidate cancer drivers that should be prioritized for experimental validation. Notably absent from this list are genes like Hras, Cdkn2a or Notch1, which are significant in several different tumor types analyzed individually but were not detected as Trunk Drivers in this meta-analysis because they did not meet the recurrence threshold of 1.5% of tumors.

DISCUSSION

Defining and prioritizing cancer drivers from genomics-based data is critical to enhancing our understanding of mor suppressor gene (TSG) or proto-oncogene (ONC). To facilitate this, we have developed visualization scripts that permit the rapid drawing of global and/or individual SB insertion patterns for each driver gene to promote new research hypotheses using SB insertion datasets (Supplemental Figure S4). Importantly, a driver gene can have multiple independent driver classifications depending on the tissue type or biological context. The mapped insertions for Zmiz1 highlight its role as a proto-oncogene in cutaneous SCC (Figure 5D), a tumor suppressor in pancreatic cancer (Figure 5E), and of undetermined influence in myeloid leukemia (Figure 5F).
the molecular mechanisms underlying cancer initiation and progression. We have developed SB Driver Analysis to enhance cancer driver identification and prioritization using Sleeping Beauty (SB) insertional mutagenesis. We applied this analysis to 17 independent SB models of human cancer and showed for the first time that we can define drivers across tumor types using a single methodology, allowing for direct comparison of the outputs. SB Driver Analysis automates driver identification using a gene-centric rather than locus-centric statistical approach, minimizing time-consuming manual annotation required of existing SB analysis platforms. It is the first transposon analysis tool to automatically classify transposon insertion patterns as activat-
Driver analysis may help to improve the pattern classification scheme. Leads to under-representation of Trunk Drivers for which Trunk Driver analysis with unbalanced tumor cohort sizes lead to data or an unclear pattern. These patterns are found in some tumor models for Tcf12 (yellow), Pten (blue), and Zmiz1 (red). Some insertion patterns cannot be determined due to a lack of insertion data or an unclear pattern. The patterns are found in some tumor models for Pten and Zmiz1. (D–F) Insertion maps showing the locations of various mapped SB insertions (triangles) in Zmiz1 transcripts across three primary tumor models. Right facing arrows (above transcripts) show forward insertions (sense strand events), while left facing arrows (below transcripts) correspond to reverse insertions (antisense strand events). (D) In cutaneous squamous cell carcinoma, Zmiz1 appears as a proto-oncogene with an activating pattern. Most insertions are on the sense strand and occur upstream of exon 9, which may indicate oncogenic behavior. (E) In pancreatic ductal adenocarcinoma, Zmiz1 appears as a tumor suppressor. The presence of insertions across the Zmiz1 locus, equally in both the forward and reverse orientation, indicates that this locus is selectively inactivated. (F) In myeloid leukemia, the distribution of SB insertion events represents an indeterminate pattern, as the insertions appear uniformly scattered across the gene. However, more insertions are present on the sense strand and all occur upstream of exon 9, which may indicate oncogenic behavior; hinting that incorporation of exon annotations in driver analysis may help to improve the pattern classification scheme.

**Figure 5.** Representative insertion patterns in driver genes. (A) Activating patterns exhibit defined groupings of forward insertions, with the transposon providing a promoter and splice donor. Hras (green), Ets1 (purple), and Zmiz1 (red) are examples of genes that exhibit forward insertion patterns. (B) Insertions that are scattered uniformly across the gene in both the forward and reverse orientation are indicative of inactivating patterns. These patterns are found in some tumor models for Tcf12 (yellow), Pten (blue), and Zmiz1 (red). (C) Some insertion patterns cannot be determined due to a lack of insertion data or an unclear pattern. These patterns are found in some tumor models for Pten and Zmiz1. (D–F) Insertion maps showing the locations of various mapped SB insertions (triangles) in Zmiz1 transcripts across three primary tumor models. Right facing arrows (above transcripts) show forward insertions (sense strand events), while left facing arrows (below transcripts) correspond to reverse insertions (antisense strand events). (D) In cutaneous squamous cell carcinoma, Zmiz1 appears as a proto-oncogene with an activating pattern. Most insertions are on the sense strand and occur upstream of exon 9, which may indicate oncogenic behavior. (E) In pancreatic ductal adenocarcinoma, Zmiz1 appears as a tumor suppressor. The presence of insertions across the Zmiz1 locus, equally in both the forward and reverse orientation, indicates that this locus is selectively inactivated. (F) In myeloid leukemia, the distribution of SB insertion events represents an indeterminate pattern, as the insertions appear uniformly scattered across the gene. However, more insertions are present on the sense strand and all occur upstream of exon 9, which may indicate oncogenic behavior; hinting that incorporation of exon annotations in driver analysis may help to improve the pattern classification scheme.

**SB** forward genetic screens have identified hundreds of genes in tumor datasets that must be prioritized for follow-on validation studies using various parameters or biological functions. We showed that **SB** Driver Analysis can prioritize the driving events that occur early in tumorigenesis. Tumor development is considered to be an evolutionary process whereby early selected insertion events occur along the main branch or trunk of the tumor evolutionary tree (17,19,22). Thus, we applied **SB** Driver Analysis to high read depth insertions present in individual tumor datasets to define and prioritize Trunk Drivers, as clonal insertions are likely to be represented in sequencing data by high read depths. We then extended Trunk Driver analysis to perform a meta-analysis across 17 **SB** tumor datasets. Similar to what has been observed in human meta-analyses for recurrent mutations, **SB** trunk driver meta-analysis reaffirmed trunk drivers present in individual tumor datasets, while missing a few key trunk drivers from individual datasets (Hras, Cdkn2a and Notch1) that fell below the recurrence threshold. Interestingly, many of the Trunk Drivers from this analysis appear in intestinal tumors, this may be in part due to the fact that there is a disproportionate number of intestinal tumors in this analysis relative to other types of tumors. Hras, Cdkn2a or Notch1, are not recurrently mutated in intestinal tumors; therefore, these data suggest that Trunk Driver analysis with unbalanced tumor cohort sizes leads to under-representation of Trunk Drivers for which there is overwhelming evidence in individual tumor cohorts. This highlights the need to consider weighting tumor contributions to the overall results, normalizing results to dataset sizes, or comparing results from this type of meta-analysis to results from analyses performed on various subgroupings of tumors.

We report the application of **SB** Driver Analysis to data generated from 454 sequencing of amplification based (splinkerette PCR) libraries, and we are working to adapt **SB** Driver Analysis to hybridization-based (SBCapSeq) libraries sequenced on Illumina or Ion Torrent deep sequencing platforms (1,11,17–20,22). Notably, we have deployed **SB** Driver Analysis to conduct genome-wide cancer driver discovery from over one million **SB** insertion events from 2354 tumors in 956 mice from primary cancer models sequenced on the 454 platform (available at http://sbcddb.moffitt.org/) (6), demonstrating the scalability of **SB** Driver Analysis to meta-analysis approaches. Based on statistical stringency we can refine the numbers of drivers without altering the types of drivers (i.e. activating versus inactivating) across tumor types. Future works are focused on using **SB** Driver Analysis to prioritize progression drivers and their biological contexts.

While we have focused our application of **SB** Driver Analysis to genome-wide approaches, the analysis framework can be restricted to discrete regions of the genome, including single or few whole chromosomes or sub-regions of chromosomes, in order to define drivers that may reside within a locus/loci of interest (e.g. syntenic regions
Figure 6. Trunk Driver incidence by tumor type. (A) Bar chart of 31 Trunk Drivers from meta-analysis of 17 tumor datasets using SB Driver Analysis. Each stacked bar corresponds to a Trunk Driver, and bar heights denote the number of tumors in which the trunk driver gene has an insertion. Colors within the bars denote datasets to which tumors belong, see key in panel B. (B) Oncoprint of 31 Trunk Drivers from meta-analysis of 17 tumor datasets highlighting Trunk Driver co-occurrences across the tumor cohorts. Rows and columns represent genes and tumors, respectively. Rectangles to the right of gene symbols denote activating (blue) and inactivating (red) SB insertion patterns, representing oncogenes and TSGs respectively. Values to the right of each driver profile denote the percentage of tumors containing the Trunk Driver. Values in parentheses to the right of the tumor cohort key denote the percentage of tumors in the dataset with at least one high read depth driver insertion. Across datasets, 60.4% (n = 1119/1852) tumors contain high read depth insertions in one or more of the 31 Trunk Driver genes.

of the mouse genome that harbor a disease-associated locus in humans). More broadly, the SB Driver Analysis framework we report may be applied to detect and determine statistical significance of any genomic feature with a well-defined DNA motif, including non-SB transposon insertional mutagenesis data from eukaryotic (e.g. piggyBac) (34–40) and prokaryotic (41) cells or within human cancer genomes exhibiting simple nucleotide mutational signatures (42,43) (e.g. ultraviolet light induced cyclobutane pyrimidine dimers) (44).

SB Driver Analysis described here and the companion SBCDDB (http://sbcddb.moffitt.org/) provide a unique set of bioinformatics and genomics tools that will be invaluable for understanding the tumor driver landscapes of SB-driven models of human cancers. SB Driver Analysis greatly strengthens our ability to detect actionable cancer drivers, prioritize cancer drivers for validation studies, and contributes positively to our understanding of the genetic basis of human cancers.
DATA AVAILABILITY
A Python implementation of the SB Driver Analysis source code and documentation described in this paper is available at http://sbcddb.moffitt.org/software/.

SBDDDB, https://sbcddb.moffitt.org/;
UCSC Genome Browser, https://genome.ucsc.edu/;
RefGene annotations, http://hgdownload.soe.ucsc.edu/downloads.html#mouse;
BED (Browser Extensible Data) format, https://genome.ucsc.edu/FAQ/FAQformat.html#format9;
Ensembl genome browser, https://www.ensembl.org/index.html/;

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
The authors would like to thank Dr David Steffen and the members of the Copeland/Jenkins lab for their helpful suggestions and comments during the development of SB Driver Analysis. N.G.C. and N.A.J. are CPRIT Scholars in Cancer Research. We thank the Research Information Technology Department at Moffitt Cancer Center for managing the high performance computational resources in support of this work.

FUNDING
Moffitt Cancer Center (to M.B.M.); Moffitt Cancer Center (to K.M.M.); Cancer Prevention Research Institute of Texas [R1112 to N.G.C., R1113 to N.A.J.]. Funding for open access charge: H. Lee Moffitt Cancer Center and Research Institute.

Conflict of interest statement. None declared.

REFERENCES

Nucleic Acids Research, 2018, Vol. 46, No. 16 e94


