RESEARCH STRATEGY

SIGNIFICANCE

Executive Summary. The goal of this study is to develop an integrated computational platform to identify novel TEs in primary and PDX tumor samples, and to perform functional studies of TE somatic insertion candidates by *in silico* functional and experimental analysis. Our strategy is based on the scientific premise that further TE discovery is essential to address major gaps in knowledge with respect to the range of TEs that can initiate and propagate in human cancers and the mechanisms by which they operate.

Structural variations (SVs), such as deletions, duplications, insertions, inversions and translocations, are among the most significant determinants of human genetic diversity[1]. A large body of work suggests that SVs play important roles in human evolution and disease susceptibility[1,2]. As part of the 1000 Human Genome Project (1000GP) SV Group, we have participated in the development and optimization of a number of computational pipelines (e.g. Meerkat, Hydra-Multi, CNVNator, AGE and PEMer) to detect SVs in whole genome sequencing (WGS) datasets at high resolution, and have provided the research community with an unprecedented set of germline SVs from more than 2,500 healthy human genomes[3]. The 1000GP Consortium studies estimate that a typical genome contains 2,100–2,500 SVs affecting ~20 million bases, far more than single-nucleotide polymorphisms (SNPs)[3,4]. The 1000 GP SV subgroup also developed a computational platform called Mobile Element Locator Tool (MELT) for genotyping non-reference mobile element insertions in WGS datasets and published a catalogue of 16,631 germline mobile element insertions present across the 2,504 individuals[3].

One disease in which SVs often feature prominently is in cancer, a highly heterogeneous and complex disease driven by genetic changes in oncogenes, tumor suppressor genes, DNA repair genes, and other regulatory elements in the human genome[5,6]. Somatic insertions of the long interspersed element-1 (LINE-1, L1), the short interspersed element (SINE, Alu) and human endogenous retroviruses (HERVs) have been detected in a number of human cancers, suggesting that they have a profound impact on tumor heterogeneity and adaptation during cancer progression[7,8]. In a previous study, we developed the Transposable Element Analyzer (TEA) computational method to detect TE insertions in 43 cancer genomes (colorectal cancer, prostate cancer, ovarian cancer, multiple myeloma and glioblastomas) from the TCGA WGS dataset, and identified 194 high confidence somatic TE insertions (183 L1s, 10 Alus and 1 ERV)[7].

While such genomic profiling technologies as TEA have rapidly advanced our ability to probe and characterize TEs *in silico*, current computational platforms for novel TE discovery are limited by poor sensitivity and accuracy at breakpoints. Therefore, new solutions to detect the genomic and transcriptomic activity of TE while taking into account the idiosyncrasies of different TEs families are necessary. We will develop TeWG, a software pipeline that is focused on detecting the molecular signatures of retrotranspositions mediated by L1 reverse transcriptase. Preliminary results suggest that TeWG has higher sensitivity than existing tools, especially for short TEs such as ALUs. For comprehensive discovery, we propose to develop a novel computational platform that will create an ensemble of existing callers—TEA, MELT and the newly developed TeWG—to improve performance over any of the three callers by themselves. Moreover, we will develop TeXP, the first software pipeline that quantifies TE transcription levels by taking into account the effect of pervasive transcription. By estimating the alignment fingerprint of TE subfamilies and the signature of pervasive transcription, TeXP will be able to estimate, for the first time, the autonomous transcription level of TE subfamilies. These tools will produce a more comprehensive list of TEs in the genome than previously possible, enabling unprecedented detail that can be studied in our model system.

Studying the role of TEs in human cancers has long been a challenge owing to lack of good disease models[9]. Patient-derived xenograft (PDX) mouse models, which are developed through engraftment of the patient primary tumor into immune-deficient NSG mice, has allowed the field to develop more relevant patient-specific tumor models[10,11]. In this study, we propose to investigate TEs using PDX models for several reasons. First, The Jackson Laboratory and its partners have developed over 400 PDX models and demonstrated that these models retain many of the key characteristics of patients' primary tumors, including histology, genomic signature, cellular heterogeneity, and drug responsiveness, and therefore may serve as a powerful platform for studies of TEs and their role in cancer biology. Second, we have established a comprehensive tissue bio-banking system that preserves the patient primary tumor tissues and the normal control tissues, as well as the PDX tumor tissues derived from the same patients, offering large quantities of tumor sample on which follow up studies can be performed (see letter: Kim). This is in contrast to studies utilizing databases such as TCGA where access to the original patient samples for additional studies is extremely limited. Our bio-banked and matched tissue samples will facilitate powerful comparisons of the landscapes of TEs in the primary and PDX tumors as well as additional functional studies *in vitro* and *in vivo*. For example, we can use our bio-banked tumor samples to study the impact of TE insertions on tumor heterogeneity, adaptation, and drug responsiveness using the PDX models. Therefore, we believe that this study will have significant impact on our understanding the biological functions of TEs in cancers.

Team. The present team brings the requisite expertise in *in silico* structural variant discovery, computational pipeline development, genomic and transcriptomic profiling of human cancers, and *in vitro* and *in vivo* experimental validation studies. The two PIs (Charles Lee, Ph.D. and Mark Gerstein, Ph.D) are leaders in the field of SVs, have a history of productive scientific collaboration and bring complementary experience in SV detection, large-scale data analysis and functional characterizations. Each also brings significant experience in leading (1000GP SV group, Lee; modENCODE AWG, Gerstein; ENCODE networks group, Gerstein; PsychENCODE AWG, Gerstein; exRNA AWG, Gerstein) and participating in (1000GP, Lee/Gerstein; ENCODE, Gerstein; ICGC, Gerstein; KBase, Gerstein; GSP (Genome Sequencing Program), Gerstein) large-scale sequencing consortia. Under Dr. Lee's leadership, the 1000GP SV project identified SV events in ~2,500 healthy genomes and helped define the methodologies for identifying and characterizing SVs from "lower depth" (~4X) WGS datasets. The two co-investigators, Ankit Malhotra, Ph.D and Chengsheng Zhang, M.D., Ph.D. are also key participants of the 1000GP. Dr. Malhotra has extensive experience in development of computational platforms for genomic studies, whereas Dr. Zhang has extensive experience in experimental assay development and functional studies using *in vitro* and *in vivo* model systems.

INNOVATION

This proposal presents a number of innovative aspects including the novel computational platforms for identification and characterization of TEs in cancer genomes and utilization of a renewable 3D cancer source in the form of PDX models.

Development of novel computational platforms: The originality of this proposal lies in the integration of cutting-edge computational methodologies—pioneered by the group—into a comprehensive, cloud-ready platform for novel TE discovery, characterization and association with human cancers. This group will also pioneer the development of a solution to quantify the transcription level of TEs by taking into account the effect of pervasive transcription of TEs.

PDX models: To our knowledge, this project is the first to investigate TEs in cancers using PDX models. The PDX resource at JAX was conceived to enable more faithful and robust animal modeling of human cancers, in particular genomic analysis of patient tumors. Our approach to leverage the PDX model as a means of identifying TEs expressed both in the primary patient and PDX-derived tumor will allow us, early on, to advance the most robust TEs for further experimental analysis. In future studies, the candidate TEs we identify can be further dissected in the parent PDX model. Because the PDX model very closely mimics the patient tumor, it is an excellent resource to perform drug studies and investigate the impact of various drugs on the TE landscape. Our approach therefore offers the specificity, resolution and continuity to follow a putative oncogenic TE from *in silico* discovery to *in vivo* analysis of tumor heterogeneity, adaptation, and drug responsiveness.

Functional studies of the TE elements using the CRISPR/Cas-9 technology: We will establish a panel of cancer cell lines that bears the specific TE insertion, respectively using the CRISPR/Cas-9 technology. These *in vitro* reagents will compliment the PDX models for studying the activity and function of TEs in cancers.

APPROACH

AIM 1: Development of novel computational platforms to discover TEs in human genomes.

Rationale. Recent literature suggests that L1s are not the only autonomous TEs active in the human genome. HERVs, especially solo long–terminal repeats (LTRs), were recently described as polymorphic in human populations[12]. On the other hand, little is known about the mobilization of non-autonomous TEs. ALUs, SVAs and protein-coding mRNA (retroCNVs) mobilizations are thought to be rare events in the tumoral context; however, only a handful of publications have investigated the mobilization of these entities[7,13,14]. To date, most the pipelines for detecting the mobilization of TEs in humans focus on the mobilization of large L1Hs by using paired-end read alignments or transductions.

To understand the mechanistic regulation activity of TEs, we are creating an ensemble of software pipelines that will help us identify TEs from WGS datasets. The software ensemble would include TEA[7]and MELT[3]—software developed by our group and others to specifically detect mobile elements from WGS datasets. The ensemble would also entail the development of *TEWG*, a novel computational framework to detect genomic TE insertions based on the molecular signature of L1 retrotransposition. Moreover, to understand the transcriptional activity of TEs, we are developing a method capable of distinguishing pervasive (passive) from autonomous transcription of TEs. We will apply the novel software ensemble and transcriptomic framework to several PDXs (Specific Aim 2) and thousands of PCAWG (TCGA/ICGC) samples. To support these large datasets, both frameworks will be implemented on a cloud platform capable of providing the resources (both computational and storage-wise) for such a large study. **Ultimately, these studies will deliver the most comprehensive resource of somatic TEs in multiple human cancers and empower us to make novel biological inferences.**

Preliminary results (Aim 1)

TEs are one of the major mechanisms creating variation across human populations. As part of the 1000GP SV project, we have provided the research community with an unprecedented set of germline SVs from more than 2,500 human genomes that have been sequenced at low depth[4]. We detected a total of 15,834 insertions of TEs; of which 3,048 were LINE-1 insertions and 12,786 were Alu insertions[3]. Our group also has extensive experience in developing pipelines to detect structural variations in WGS datasets. Meerkat[15], Hydra-Multi[16], CNVNator[17], AGE[18] and Paired-End Mapper (PEMer)[19] are all computational pipelines developed by our group for mapping SVs at high resolution with confidence and then genotyping the discovered SVs in large populations.

Existing software tools to detect TE insertions TEA is a computational method to detect somatic TE insertions in the human genome at nucleotide resolution using paired-end WGS datasets. In a published study[7], we analyzed TE insertions using a TCGA WGS datasets from 43 cancer genomes, including five cancer types—colorectal, prostate, ovarian, multiple myeloma and glioblastoma. We discovered 194 high-confidence somatic TE insertions (183 L1s, 10 Alus and 1 ERV). Colorectal tumors showed the highest frequency of somatic L1 insertions but they were not found in blood or brain cancers. A majority (38/39, or 97%) of the detected sites were validates using PCR and Sanger sequencing techniques.

MELT (melt.igs.umaryland.edu) was developed by Eugene Gardner and Scott Devine at the University of Maryland, Baltimore. MELT was used in Phase 3 of the 1000GP SV subgroup to discover, annotate and genotype non-reference mobile element insertions in WGS datasets. Using MELT, the 1000GP SV subgroup published a catalogue of 16,631 germline MEIs present across the 2,504 individuals[3].

Retrotransposition is known to create most Alus and processed pseudogenes in the human genome. We have extensive experience in processed pseudogene identification and annotation[20]. As a contributor to the GENCODE project, we developed the GENCODE pseudogene resource[21], which contains both computationally predicted and manually annotated pseudogenes. We will leverage our experience in SV detection and pseudogene annotation to build a platform to detect genomic insertions mediated by L1 reverse transcriptase. The software pipeline will be named TEWG. As a pilot analysis, we analyzed 63 samples from PCAWG and focused on the mobilization of ALUs. We found 1062 putative somatic insertions in 63 tumor samples, yielding an average of 17 Alu insertions per tumor, suggesting a higher sensitivity when compared to other methodologies.

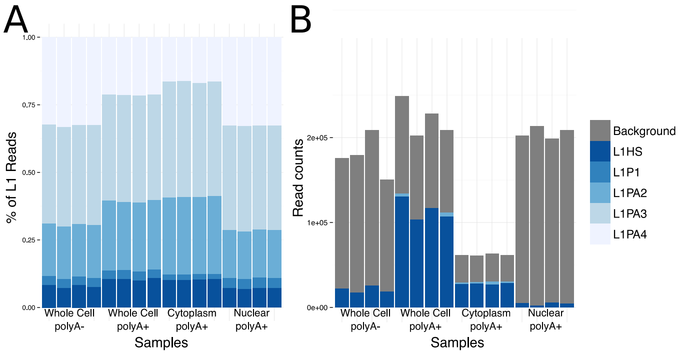

Figure 1. Alignment profile of RNA-seq reads mapping to L1 subfamilies in four MCF-7 cell compartments. Each cell compartment (Whole-Cell PolyA-, Whole Cell PolyA+, Cytoplasm polyA+ and Nuclear polyA+) has four RNA-seq replicates. A) Percentage of L1 reads mapping to L1Hs, L1P1, L1PA2, L1PA3 and L1PA4 subfamilies across the 16 RNA-seq experiments. B) Absolute number of reads counts originating from each L1 subfamily active transcription (blue tones) and the number of reads originating from pervasive transcription (grey).

RNA Analysis to study TE activation and transcription We also have considerable expertise in analyzing RNA-Seq data, including experience in developing and setting up pipelines for the processing of RNA-seq data; specially for RNA-seq data for ENCODE[22] and for the PsychENCODE[23]. To better understand the transcriptional activity of TEs, we analyzed more than 50 RNA-seq experiments from ENCODE cell lines datasets from 11 human cancer cell lines and investigated the autonomous transcription of TEs. In agreement with previous work, we found that MCF-7, a breast cancer–derived cell line, has a remarkably high number of reads overlapping L1 subfamilies[24]. Interestingly, we observed that the proportion of reads mapping to different L1 subfamilies varies across different cell compartments, suggesting that different cell compartments have distinct compositions of L1 pervasive and autonomous transcription **(Figure 1A).** We further investigated the transcription level of L1 subfamilies in distinct cell compartments of well-established cancer cell lines. In RNA-seq experiments from MCF-7, approximately 200,000 reads mapped to L1 subfamilies in whole cell (polyA-) and nuclear compartments. In contrast, cytoplasmic compartments yielded only 50,000 reads. Using a pilot version of TeXP, we estimate that approximately 50% of the reads in whole cell and cytoplasm fractions are the result of autonomous transcription of L1Hs **(Figure 1B).** The remaining 50% of reads are the result of pervasive transcription. In contrast, less than 10% of transcriptional signals from the whole cell (polyA-) and nuclear compartments is derived from autonomous transcription of L1Hs.

We further analyzed RNA-seq datasets from ENCODE and found that GM12878, a lymphoblastic cell line derived from blood from a healthy individual, had no autonomous L1Hs regardless of cell compartment and transcript selection process. In contrast, SK-MEL-5 and K562 are cancer-derived cell lines and show transcription level of respectively 20 and 30 in whole cell polyA+ experiments. We also processed over 5,000 samples from BrainSpan and GTEx projects to assess the activity of L1 elements in developmental and healthy adult tissues. **Figure 2** shows that many tissues (e.g., skin, heart, bladder, adrenal gland) have a profile of TE reads different from that expected by pervasive transcription, suggesting that these tissues could support transcriptional activity of TEs.

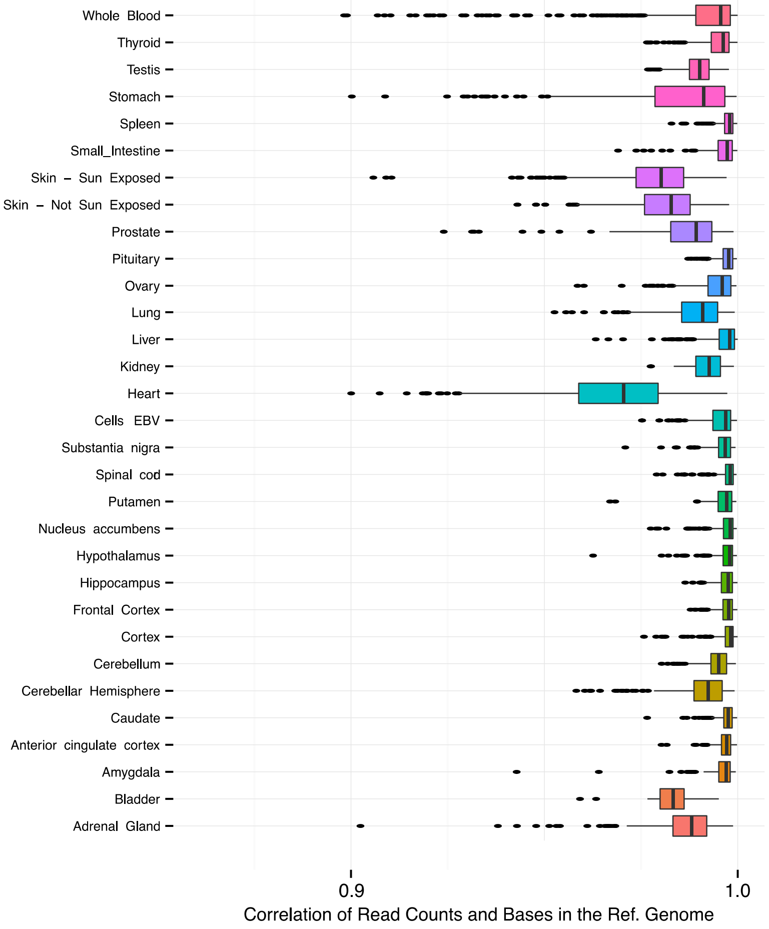

Figure 2. Distribution of the correlation coefficient between the number of RNA-seq reads mapped to L1 subfamilies and the number of bases in the reference genome annotated as that L1 subfamily. Samples are grouped by GTEx primary tissue.

Experimental Design (Aim 1)

1.1. Development of TEWG, a novel genomic TE caller: *TEWG* is being developed to detect mobilizations mediated by L1 reverse transcriptase by detecting the molecular signatures of L1 retrotransposition. This includes L1, ALUs and retroCNVs. We cluster partially aligned reads and locally assemble possible SVs by performing a sequence consensus calculation based on nucleotide frequency. Posteriorly, inferred insertions and deletions are mapped to annotated TEs (and protein coding genes to detect retroCNVs) to triage potential mobilizations of TEs. The Direct Repeat (DR) and poly(A) signals flanking L1 mediated retrotransposition are used to further support TE mobilization. At this stage, putative mobilizations of TE can be either germline or somatic. When available, we will use paired tissue information and 1000 Genomes TE polymorphism dataset to annotate germline mobilization of TEs.

1.2. Ensemble calling genomic mobilizations of TE: We will develop new tools and pipelines that will work in the cloud to identify and classify SVs caused by the mobilization of TEs using TEA, MELT and TEWG. The three TE callers would be unified into an ensemble within a dockerized Virtual Machine (VM). The virtual machine would include complete functionality to take unaligned fastqs or aligned BAM files for sequenced samples, perform quality control on the data and then process the data through the three software pipelines using pre-optimized parameter sets. The resulting TE callsets would then be unified using an integration data model that will be built using an expectation maximization algorithm on the results from the 1000GP[3] and PCAWG consortium (preliminary and unpublished). The data model would be built by parsing callsets into subsets based on a few discriminating features (type, length, GC content) and then determining the best combination of the three callers that maximizes overlap with the truth call set (constructed from 1000GP and PCAWG datasets). This would be the first application of a data mining approach to discovery of TE elements using an ensemble of callers. We expect to demonstrate significantly greater robustness and correctness of this approach versus single algorithms.

1.3. Quantification of TE transcriptomic activity: From the pilot analysis, we learned that ancient TEs have a number of reads correlated with the closest gene, implying that their expression is associated with nearby transcriptional active regions (TARs). However, the same analysis suggests that read counts associated with evolutionarily young elements (e.g., L1Hs, Alus and SVAs) do not correlate with their nearby genes. Many factors could contribute to this, including pervasive transcription and autonomous transcription. Pervasive transcription is defined as the low-level transcription of most of the genome. For highly repetitive regions such as TEs, pervasive transcription is a major confounding factor. To distinguish between autonomous transcription and pervasive transcription of TE subfamilies, we will create TE subfamily fingerprints by simulating reads from their putative mature transcripts. TeXP is a computational suite that will be used to create TE subfamily mapping fingerprints and process RNA-seq experiments in order to regress the proportion of autonomous and pervasive transcription of TEs. It will enable the use of RNA-seq datasets to assess the transcriptional activity of TE subfamilies.

1.4. Cloud computing for 1000 genomes and PCAWG cohorts: We will use the ensemble somatic TE detection pipeline to call mobile elements from two important data cohorts. To understand how mobile elements are active and segregating within normal individuals, we would process data from 2,504 individuals from the 1000GP consortium. To understand how mobile elements become activated in cancers, we will identify TE insertions in tumor genomes from 2,834 individuals of the PCAWG consortium. Both of these cohorts are wonderful resources for testing and validating the computational pipelines developed for this project, but the sheer amount of data, in the order of Petabytes, prohibits the traditional approach to data processing. Therefore, rather than downloading and processing the files locally, we designed our ensemble caller as a virtual machine that can be easily deployed in the cloud. We believe this would give us a unique advantage as we will be ideally positioned to be able to execute this and other similar analysis on large-scale datasets quickly and efficiently.

Expected results (Aim 1): These studies will yield a cloud-enabled ensemble pipeline that utilizes both whole genome and RNA sequencing data from individuals to identify somatic mobile element insertions in genomes, as well as a comprehensive catalog of activated mobile elements identified from healthy (1000 Genomes) and diseased (PCAWG cohort) subjects that will provide a baseline for subsequent discovery and functional interpretation from our PDX patient cohort (Aims 2 and 3). They will also help answer questions about population-level extent of mobile element insertions across thousands of individuals. By making the ensemble pipeline available as a community resource, we expect this work to propel future genome-level mobile element analyses for the research community.

Pitfalls/Alternative approaches (Aim 1): One criticism is that we lack numbers on the performance of the ensemble pipeline. We acknowledge and accept this criticism; however, because of our extensive knowledge in building SV and TE detection algorithms, we expect that the ensemble pipeline would, at the worst, be as good as TEA and MELT algorithms, two of the published callers in the ensemble. In the TEA study, we were able to validate 38/39 somatic L1 insertions identified in colorectal and ovarian tumors. In the 1000Genomes Phase 3 structural variation study, MELT identified 16,631 mobile element insertions with a site FDR of 4% and a sensitivity estimate of 96%. These results indicate that the expected sensitivity rates for our ensemble pipeline would be similarly high. In the worst case, we would use TEA and MELT as algorithms for detection of the TE and find overlap between their callsets as our callset for further investigation.

AIM 2: Identification and experimental validation of TEs in cancers

Rationale. Recent studies have identified somatic TE insertions in a number of human cancers, but the frequency and extent of TEs in cancer genomes varies by cancer types and individual patients. Somatic TE insertions may inactivate tumor suppressor genes and/or activate oncogenes as well as induce genomic instability. However, the biological functions and the molecular mechanisms of TE-mediated insertions in cancers remain poorly understood.

Studies of human cancers have long been a challenge because of lack of good disease models. In this study, we propose to investigate the role of TEs in cancer using PDX models because these models very closely mimic patients' primary tumors characteristics, including histology, genomic signature, cellular heterogeneity, and drug responsiveness, and therefore serve as a renewable 3D *in vivo* resource for studies of cancer biology and co-clinical trials for cancer precision medicine. Through a collaboration with Seoul National University (SNU)(see letter: Kim), we have access to a comprehensive tissue bio-banking system that preserves the patient primary tumor tissues, normal matched tissue, and the PDX tumors derived from the same patient. To our knowledge, a lot of previous studies that were mainly dependent on the TCGA database often have extremely limited access to the original patient samples for additional validation or functional studies. On the other hand, the bio-banked primary tumor samples and matched tissue samples as well as the PDX tumor samples from SNU, can provide adequate resources that allow us to compare the landscapes of TEs in the primary and PDX tumors and perform additional functional studies *in vitro* and *in vivo*.

Preliminary results (Aim 2)

Identification and validation of putative mobile elements We have used conventional PCR, RT-PCR, droplet digital PCR (ddPCR), microarray, Sanger sequencing and next-generation sequencing (NGS) to study the structure, function and evolution of the human genome[7]. For example, one of our previous studies identified 194 somatic TE insertions by computational analysis of 43 high-coverage WGS datasets from five cancer types (glioblastoma, multiple myeloma, colorectal, prostate, and ovarian cancers) and their matched normal genomes. We then conducted experimental validations using PCR and Sanger sequencing **(Figure 3),** and found that a majority of the TE candidates (97%) were validated[7]. This study also showed that the extent of somatic mobile element insertions can vary by cancer types. All of the somatic L1 and Alu insertions were observed in epithelial cancers (colorectal, prostate, and ovarian cancers), with colorectal tumors showing the highest frequency of somatic L1 insertions, but not in blood or brain cancers. Our preliminary data, profiling WGS datasets from the PCAWG pilot project, indicated that breast, kidney, liver, ovarian, uterine, cervical and lung cancers had higher than expected Alu insertions per genome. Other studies have indicated that a subset of gastric cancers have a high degree micro-satellite instability and could have an increased rate of mobile element insertions in the genomes[25].

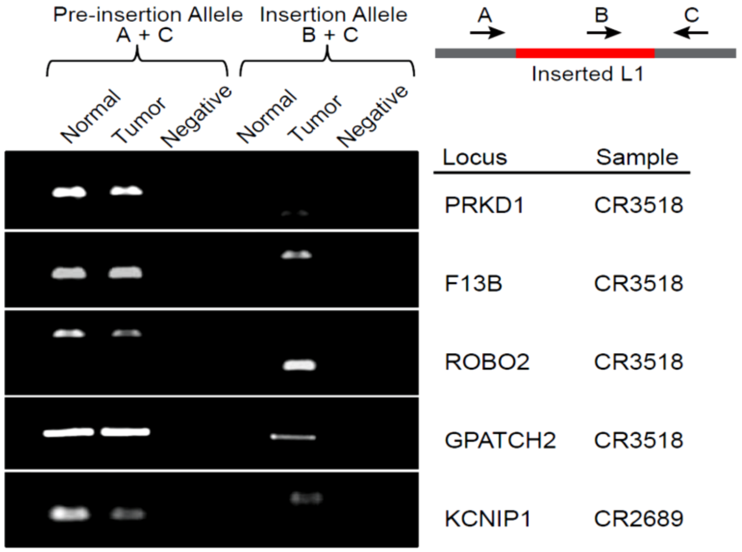

Figure 3. A representative image showing the results of PCR validation of somatic L1 insertions. Comparison of the products amplified using primers flanking the insertion (A+C) with the products amplified using a primer internal to the L1 and a downstream primer (B+C) shows that the insertion alleles are present in the tumor sample and absent in the matched normal sample (Lee et al, 2012).

In collaboration with Dr. Jong Il Kim at SNU and other scientists, we have been developing PDX models from primary tumor samples for over 400 patients across seven different tumor types. We established a comprehensive tissue bio-banking system at SNU that preserves the patient primary tumor tissues, the matched normal tissues, and the PDX tumors, all derived from the same patients. In addition, whole genome, whole exome and RNA sequencing data is being generated from each patient's tumor, its matched normal tissue and engrafted tumor specimens **(Figure 4**; see letter: Kim). For all the cancer types, as part of our collaboration with SNU, we already have sequencing data for at least 10 patient PDX models from five different cancer types (breast, gastric, colorectal, lung, and bladder cancers).

In this study, we propose to study five cancer types (breast, gastric, colorectal, lung, and bladder), and 10 patients from each cancer type, by utilizing the PDX models developed under this collaboration.

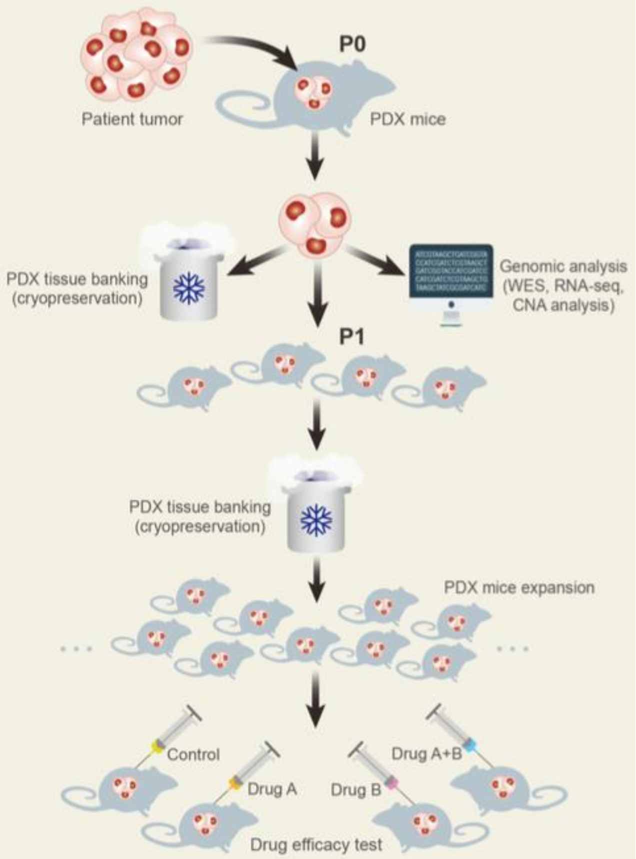

Figure 4. Generation of PDX models. Surgical specimens from cancer patients are divided into small pieces and transplanted into immunodeficient mice (P0). When tumors are grown in P0 mice, xenografts are used for genomic analysis including whole genome sequencing (WGS), whole exome sequencing (WES), RNA sequencing (RNA-seq), and copy number alteration (CNA) analysis, and then maintained in cryo-banks for preservation. After expanding tumor xenografts in immunodeficient mice (P1 and more), *in vivo* drug responsiveness is screened in these models.

Experimental designs (Aim 2)

2.1. Identification of TEs: In this study, we will employ the integrated computational platforms developed in Aim 1 to identify TE insertions in the primary tumor, matched normal tissue and PDX tumor from the same patient, by computational analysis of WGS datasets. In addition, we will investigate the transcriptional activity of TE in these PDX tumors. The TE subfamily transcriptional level will be estimated for normal, primary and P0 (first engraftment) tumors **(Figure 4)[26].** As TE transcriptional activity can be used as proxy for potential genomic activity, we will also use the TeXP TE transcriptional level to investigate the activation of TE subfamilies in the normal control, primary tumor and PDX tumors.

2.2. Experimental validation of the putative TEs: To examine the false discovery rate of our computational analysis pipelines, we will perform experimental validation on the TE calls identified in Aim 2.1 using the methods described below. Based on our previous studies and other published data, we expect to identify 200-300 TE calls from the 100 tumor samples (50 primary and 50 PDX tumor samples). While we expect to observe some unique TE calls between the primary (A) and PDX tumors (B), respectively, we do expect that a majority of the TE calls would be common ones (C) that are shared by the primary and PDX tumors **(Figure 5).**

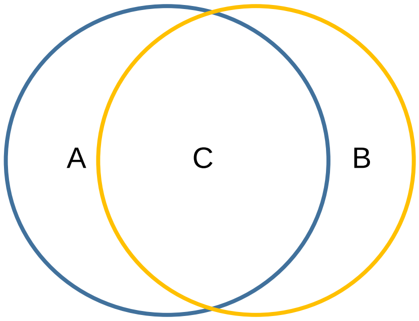

Figure 5. Unique and overlap TE calls identified in the primary and PDX tumors. A indicates unique calls in the primary tumors; B indicates the unique calls in the PDX tumors; C indicates the common calls shared by the primary and PDX tumors.

2.2.1. Normal control, the primary and PDX tumor samples: In this study, we proposed to investigate five cancer types, including breast, gastric, colorectal, lung, and bladder cancers. Ten primary and 10 PDX tumor samples from each cancer type (along with matched normal tissues as controls) will be used for this project. Snap-frozen tissue samples from the cancer patients and PDX mice will be obtained as part of an ongoing research study for which Dr. Lee is a co-Investigator ("Development of Personalized Medicine Platform with Patient-derived Xenograft Models", and approved by the Institutional Review Board (IRB) of SNU, IRB Number #1402-054-555**.**

2.2.2. DNA isolation and PCR validation: Genomic DNA will be isolated from the frozen tissues according to the protocol provided by the manufacturer (QIAGEN). We will then perform PCR validations on the putative TEs chosen for validation. The PCR primers flanking the predicted breakpoints will be designed to detect the pre-insertion allele according to the following pipeline: First, the genomic sequence of a 500 bp region adjacent to the breakpoints of each TE will be extracted from the UCSC Genome Browser, GRCh37/hg19 assembly. Second, Primer3 Plus[27] will be used to compute a set of primer pairs flanking the breakpoint for these regions. Third, the quality score of the primers will be checked using Netprimer (PREMIER Biosoft International, Palo Alto, CA) software. The primers would not be used if the quality score is less than 80%. Fourth, all primer pairs will be tested for their uniqueness across the human genome using In Silico PCR from the UCSC Genome Browser. Finally, the NCBI 1000 Genome Browser will be used to check if there are any SNPs in the primer-binding region. In this study, a primer within the L1 and a primer in the flanking genomic sequence will be designed to detect the somatic L1 insertions **(Figure 3).** Primers flanking the insertion site will be designed to detect Alu insertions. The pre-insertion allele will produce a PCR fragment that is consistent with the predicted size of the reference genome, whereas the allele with Alu insertion will generate a PCR band that is ~300 bp larger that the wild type allele. PCR validation of the HERV insertions will be performed using the primers flanking the insertion sites. Both normal and tumor tissue samples will be used for PCR validation to confirm the germline and/or somatic insertions. The PCR amplifications will be performed in 25µl reaction using BioRad c1000 Touch Thermal Cycler. PCR reactions will be conducted as described[7] and the PCR products will be analyzed by gel electrophoresis and visualized using the Genebox instrument.

2.2.4. Detection of copy number by Droplet Digital PCR (ddPCR): We will use the Bio-Rad QX200 ddPCR platform to quantify copy number of the TE candidates. The ddPCR reactions will be performed according to the manufacturer's protocols. All ddPCR experiments will include at least one normal control (NA12878 and/or NA10851) and a no-template control (NTC) with water. All samples and controls will be run in duplicates. The ddPCR data will be analyzed with the QuantaSoft software provided by the manufacturer (Bio-Rad).

2.2.5 Sanger sequencing: To further confirm the breakpoints and specificity of the TEs, PCR products will be purified using the Qiagen MinElute kit and sequenced using the Sanger method covering both of 5' and 3' insertion junctions. Briefly, the sequencing PCR will be performed in a 50ul reaction and then PCR product will be run in a 1% Agarose gel to separate DNA. The target band will be cut from the gel and purified using the Gel Extraction and PCR Clean-Up Kit (Clontech Laboratories). The DNA concentration will be measured using the NanoDrop 2000 to get an optimal DNA concentration for Sanger sequencing (10-20ng/µl). DNA sequencing will be done by Eton Bioscience. The MEGA program[28] will be used for the analysis of the Sanger sequencing data.

Expected results (Aim 2): Based on our previous studies and other published data, we expect to identify 200-300 TE calls from the 100 tumor samples (50 primary and 50 PDX tumor samples). While the majority of the TE calls will be common ones that are shared by the primary and PDX tumors, we also expect to observe some unique TE calls in the primary and PDX tumors, respectively **(Figure 4).** In particular, due to lack of functional immune system and different microenvironments, the PDX tumors may present some novel TEs that would have been suppressed in the primary tumors by the human immune system and/or the host restriction factors. Our study will be the first to investigate the landscape of somatic TEs in PDX environments.

Pitfalls/Alternative approaches (Aim 2): One concern is the possible contamination of NSG mouse cells or tissues with the human xenograft tumor tissues in the PDX mice, which may complicate the sequencing analysis and PCR amplification. To overcome these problems, our lab has developed a number of computational pipelines to filter and clean mouse sequence data from human data. In addition, JAX has established a mouse genome sequencing consortium and multiple computational pipelines to separate human from mouse sequencing data in PDX mice. Moreover, we will pay special attention to design primers/probes that are specific to human genomic sequences. Finally, we may need to design additional PCR primers to detect L1 or other TE insertions that failed the PCR amplification in the first round of reactions described above.

AIM 3: Functional characterization of validated TE candidates

Rationale. TEs have been discovered in a variety of cancer genomes. For example, disruption of ABC, RB1 and PTEN tumor suppressor genes by L-1 insertions were detected in colorectal cancer, retinoblastoma, and endometrial cancer, respectively[14,29,30]. Activation of *Alu* sequences may induce transposon-mediated genomic instability and contribute to tumor heterogeneity[31]. Recurrent *Alu*-mediated non-allelic homologous recombinations have been identified in a number of cancer-associated genes, such as MLH1, MSH2 etc. MLH1 and MSH2 are two of the important genes involved in the mismatch repair system and are associated with hereditary non-polyposis colorectal cancer (Lynch syndrome). Interestingly, MLH1 has 20% Alus and MSH2 has 40% Alus within their intronic sequence, which are significantly higher than the 10% average *Alu* density in genome. Moreover, *Alu* Y has been suggested to play a potential role in the development of drug resistance to Irinotecan (SN38) or oxaliplatin in colorectal cancer[32].

Despite increasing evidence suggested that TEs may play important roles in the pathogenesis of a variety of human diseases, including cancers, the functional impact of their mobilization in a somatic genome-wide fashion and the molecular mechanisms by which they mediate oncogenesis remain largely unknown. In this study, we propose to develop a framework to perform a comprehensive *in silico* characterization of TEs over three contexts: (1) those impacting protein coding genes; (2) those impacting non-coding RNAs; and (3) those impacting non-coding regulatory regions such as transcription factor binding sites. The impact analysis will also integrate conservation information, allelic activity, existing genomic annotations and epigenetic and transcriptomic datasets from sources such as ENCODE, 1000 Genomes, and GTEx. Furthermore, we will perform a series of experimental studies to understand the biological functions and the molecular mechanisms of TE-mediated insertions in cancers.

Preliminary results (Aim 3)

Preliminary results from the computational analysis: We have developed a number of computational tools to perform *in silico* functional studies of the validated TE candidates and their impact on cancers. We developed Variant Annotation Tool (VAT) to annotate protein sequence changes of mutations. VAT provides transcript-specific annotations and annotates mutations as synonymous, missense, nonsense or splice-site disrupting changes[33]. We used VAT to systematically survey loss-of-function (LoF) variants in a cohort of 185 healthy people as part of the Pilot Phase of the 1000GP[34]. We also participated in the 1000GP Phase 3 studies on LoF variants and the functional impact of SVs, and found that a typical genome contains ~150 LoF variants. Our ncVar pipeline further analyzes genetic variants across biotypes and subregions of ncRNAs, e.g. showing that miRNAs with more predicted targets show higher sensitivity to mutation in the human population[35]. We have extensive experience performing annotation of non-coding regulatory regions and expertise in developing tools to analyze ChIP-Seq data to identify genomic elements and interpret their regulatory potential. For ChIP-Seq, we developed two tools—PeakSeq and MUSIC—that identify regions bound by transcription factors and chemically modified histones[36,37]. PeakSeq has been widely used in consortium projects such as ENCODE[36,22]. MUSIC is a newly developed tool that uses multiscale decomposition to help identify enriched regions in cases where strict peaks are not apparent.

A powerful way to integrate diverse genomic data is through networks representations. We have great experience studying regulatory network and relating variants to networks. In particular, we have integrated multiple biological networks to investigated gene functions. We found that functionally significant and highly conserved genes tend to be more central in various networks[38]. Furthermore, we have extensive experience with allelic activity analysis. We recently applied this pipeline on a population scale to RNA-Seq data from the 1000GP, and used this analysis to create AlleleDB, a database of genomic regions with high allelic activity[39].

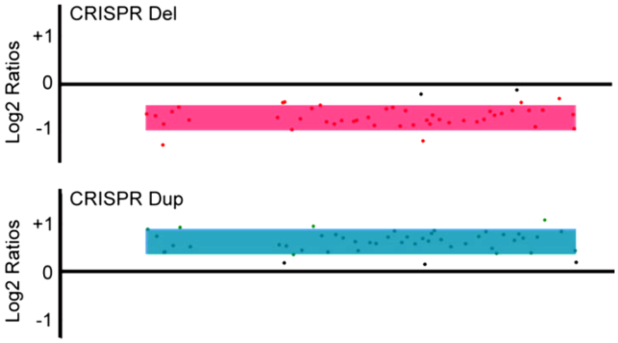

Figure 6. Microarray analyses showing deletion (CRISPR Del) and duplication (CRISPR Dup) of the 16p11.2 region in CRISPR-treated lines is shown. Gains or losses of 16p11.2 region were determined by normalized log2 ratios.

Preliminary results from experimental studies: We have developed a variety of assay systems for functional studies, including cell culture, PCR, real-time PCR, western blot and immunohistochemistry staining, microarrays, and animal models. In addition, we have developed two CRISPR/Cas-9-based genome-editing platforms. One uses the lentiviral vectors (Addgene, MA), and another employed a CRISPR system consisting of two guide RNA vectors in pCas-Guide and donor vector with predesigned homologous arms with Knockin GFP-Puro for double selections (OriGene, MD). For example, we developed a platform to mimic non-allelic homologous recombination *in vivo* and generate microdeletions and microduplications in human genomes by targeting segmental duplications with the lentiviral vector-based CRISPR technology **(Figure 6)**[40]. Recently, we initiated a project to investigate the biological functions of the tumor suppressor gene PTEN using the homology-directed repair (HDR)-based CRISPR technology, and have successfully generated a panel of cell lines with homozygous or heterozygous PTEN deletions **(Figure 7).** In addition, as described above, we have extensive experience in developing mouse models for studies of cancer biology, genomics, and drug responses **(Figure 4).**

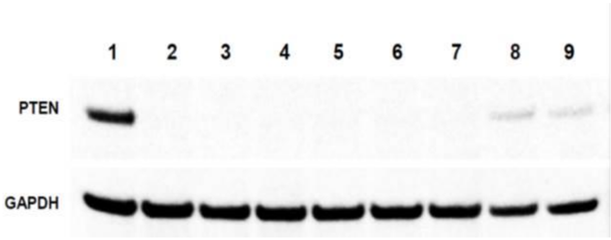

Figure 7. Detection of PTEN protein expression in 293T cell lines by Western blot. Lane 1, untreated control cells; Lane 2-7, CRISPR targeted cell lines with homozygous deletion; Lane 8-9, CRISPR targeted cell lines with heterozygous deletion.

Experimental design (Aim 3)

3.1 Prioritization of somatic TE insertion candidates:

Our somatic TE insertion prioritization pipeline will integrate many features into a single priority measure per mobile element insertion. We will first identify the functional impact of TE insertions identified and validated in aim 2 based on the annotation of the insertion point. Slightly different strategies will be used if the insertion overlaps: 1) protein-coding genes; 2) non-coding transcripts 3) non-coding regulatory elements. Also, we will further prioritize somatic TE insertions by investigating insertions overlapping conserved elements; allelic elements; the network connectivity of the target gene. Each functional overlap increases the score measure and hence the significance of the TE insertion.

3.1.1. Identifying the functional impact of TE insertion on protein coding genes. We will investigate how somatic TE insertions creates LoF of the target protein coding gene. We will first identify putative LoF-causing TE insertions as those that induce: 1) premature stop codons; 2) frame-shifted open reading frames; or 3) truncated proteins due to changes of RNA splice sites or either predicted or verified changes in splicing pattern from RNA-Seq data. L1s and Alus for example are frequently truncated, however, they frequently include a strong poly(A) signal, potentially truncating target genes with intron insertions. We will also use RNA-seq data to measure the influence of insertions into protein coding genes. Target genes will also be analyzed in regard of their differential expression and, when available, we will use paired (normal/tumor) sample RNA-seq to evaluate changes in target gene expression.

3.1.2. Prioritizing non-coding transcripts from structural variant data. To prioritize the effects of somatic TE insertions in ncRNAs, we will investigate insertions in regulatory elements. To perform this analysis, we will define categories of RNA regions that display human population-level conservation, and combine these features to generate RNA element scores. We have defined annotations of biochemical interactions and regulatory motifs that are enriched for rare variants in the human population and will use these sensitive RNA regions to score and prioritize potential deleterious TE insertions in ncRNA.

3.1.3. Prioritizing non-coding insertions on regulatory elements. Unlike protein-coding genes and non-coding RNAs, TF binding motifs are relatively small in size. Thus, we are going to analyze insertions that occur close to TF binding motifs and analyze where these insertions lead to the breakage of existing or the creation of new motifs. In the prioritization scheme, we will also penalize changes in distance between motifs and newly created motifs if they occur close to an existing transcription factor motif. For example, an insertion of a full L1H (6Kb) should be more harmful than an Alu (360 base pairs).

3.1.4. Additional prioritization features: Insertion point conservation. For evolutionary properties, we will quantify the conservation of insertion points using intra-human variation data (from The 1000GP) by comparing the ratio of low-frequency variants and high-frequency variation. This index will be used to define regions under selection at the population time scale. We will also use cross-species evolutionary conservation (using classical measures such as the GERP score[41] to prioritize variants disrupting evolutionarily relevant regions.

Network connectivity. We will examine the network topological properties of the genomic elements affected by somatic TE insertions. Insertions disrupting regulatory elements with high connectivity—network hubs and bottlenecks—will be up-weighted based on their scaled centrality scores, as we know that disruption of highly connected genes or their regulatory elements is more likely to be deleterious[38]. For network features, we will comprehensively define associations between somatic TE insertion and their target protein-coding and non-coding genes. For each target (host) gene, we will use a variety of networks—e.g., regulatory network, metabolic pathways, etc.—to assess the impact of the insertion.

Allelic activity. We will prioritize TE insertions that overlap highly allelic regions throughout the genome, based on AlleleDB[42] – our resource of such regions has been identified through allele-specific RNA-Seq analysis from over 300 individuals generated by the GEUVADIS consortium[43]. The impact analysis will also integrate conservation information, allelic activity, existing genomic annotations, and epigenetic and transcriptomic datasets from sources such as ENCODE, 1000GP, and GTEx.

3.2. Experimental studies of the putative TE candidates: We will select a panel of putative TE candidates predicted to significantly impact gene expression and have important functional consequences based on the impact score assigned by our *in silico* characterization pipeline described in Aim 3.1. We will focus on TE candidates that may disrupt: 1) protein-coding genes (e.g., oncogenes, tumor suppressor genes); 2) non-coding transcripts (e.g., microRNA, IncRNA); or 3) non-coding regulatory elements (e.g., promoters, enhancers, super enhancers, splicing sites, or other non-coding regulatory elements). We will select 10 top TE candidates (two from each cancer type if possible) from each of the three candidate groups for *in vitro* studies.

3.2.1 Detection of gene expression levels by RT-PCR: We will perform RT-PCR to evaluate whether the TE insertion will affect target gene expression. Briefly, total RNA will be isolated from the normal control and tumor samples using RNeasy Kit (QIAGEN) and quantified by spectrometry. The quality of the RNA samples will be assessed on an Agilent 2100 Bioanalyzer. RT-PCR will be performed according to the protocol provided by the manufacturer (QIAGEN).

3.2.2 Detection of protein expression levels by Western blot and immunohistochemistry: We will perform western blot and immunohistochemistry to quantify protein expression of the target genes. For western blot, tissues or cell pellets will be lysed in lysis buffer (50mM HEPES, 1%Triton X-100, 50mM NaCl, and protease inhibitor cocktail). Western blot will be performed using antibodies specific to the respective target protein. For example, we can use anti-PTEN monoclonal antibody to detect PTEN protein expression level as shown in **Figure 7.** For immunohistochemistry, we will use 5µm-thick sections of the formalin-fixed, paraffin-embedded (FFPE) tissue blocks to detect the target protein expression in the tissues. Although we do not know the exact target proteins we will work on at this time, we can expect to purchase the respective antibodies from various companies, such as Santa Cruz Biotechnology and BD Biosciences. Both Western blot and immunohistochemistry will be performed according to the protocol provided by the manufacturer.

3.2.3 Microsatellite instability (MSI) assays: Previous studies indicate that TE insertions affect DNA mismatch repair genes (e.g., MLH1, MSH2) and induce genomic instability in colorectal cancers. In this regard, we will perform MSI assays according to protocols provided by manufacturers (ThermoFisher Scientific). The five markers recommended by the National Cancer Institute will be used to assess MSI: BAT25 and BAT26 to assess mononucleotide repeats (A)n and D2S123, D5S346, and D17S250 to assess dinucleotide repeats (CA)n. MSI status will be determined using protocols as described[44].

3.3 Functional studies of the selected TEs using cell-based assays: Based on the data from Aim 3.1 and 3.2, we will select nine TE candidates (three from each category described in Aim 3.2) and perform functional studies using specific cell-based assays with CRISPR genome-editing technologies.

3.3.1 Development of TE-inserted cancer cell lines using the CRISPR/Cas-9 Technology: Since human primary cells are difficult to culture and much less efficient for gene-transfer studies, we will select the appropriate cell lines based on the tumor types. For example, to study a TE insertion associated with breast cancer, we will select a human breast cancer cell line. After we select the appropriate cell lines, we will generate the cell lines with a specific TE insertion in the genome using CRISPR/Cas-9 technology according to methods described previously[45,46]. We will divide the experiments into the following three groups:

Group one:

1) Three newly established cell lines with the specific TE targeting the protein-coding gene.
2) Each parental cell line without the specific TE in the genome

Group two:

1) Three newly established cell line with the specific TE targeting the non-coding transcript region.
2) Each parental cell line without the specific TE in the genome

Group three:

1) Three newly established cell line with the specific TE targeting the non-coding regulatory elements.
2) Each parental cell line without the specific TE in the genome

We will perform PCR, ddPCR, microarray and Sanger sequencing to confirm the location of the TE insertion in the genome and check the potential off target effects caused by the CRISPR/Cas-9 technology.

3.3.2 Functional characterization of the TE-inserted cancer cell lines: We will perform functional studies of these cell lines and understand the mechanisms of TE activation.

3.3.2.1 Functional characterization of the TE-inserted cancer cell lines by the cell-based assays: To examine the effects of the inserted TE on the host cells, we will perform the functional studies described in Aim 3.2.1-3.2.3 to examine the gene and protein expression levels as well as the genomic instability. In addition, we will carry on the following functional analysis: 1) to examine the cell growth rate using the cell proliferation assay (Promega) to examine cell cycle progression by flow cytometric analysis with propidium iodine (Abcam) to examine cell colony formation using the soft agar-based clonogenic assay as described previously[47]

3.3.2.2. Understanding the mechanisms of TE activation: To understand how the regulation of TE changes in transition from normal to cancerous, we will investigate the regulatory context around potentially active parental

TEs and how somatic TE insertions affect the nearby regulatory context. First, we will investigate the transitions in the regulatory context of parental TEs. We will investigate the Sanger sequences (Aim 2.2.5) to learn which instances of parental TEs are recurrently activated in normal/cancer transformation. As a model we will focus on well-characterized paired normal/cancer cell lines (e.g. GM12878/K562). We envision that new paired cell lines will be characterized in the near future by consortia such as ENCODE. We will further set up models to quantify the TF binding events using ChIP-seq experiments as TF scores to search for promoter-like regions from potentially active parental TEs. Such transcription factor scores will represent the potential of TE proximal regions to initiate the transcription process. Using our expertise in enhancer discovery and target prediction, we are planning to uncover the underlying mechanism of TE activation via proximal and distal regulatory elements. We will also investigate the regulatory disruption caused by the somatic insertion of TEs. The selected TE somatic insertions candidates will be compared with respect to changes in regulatory landscape in well-established cell lines before and after CRISPR/Cas9-based introduction of the TE element of interest. We will generate RNA-seq, DNA methylation and H3K27ac ChIP-seq from these cell lines, as well as use publicly available data for wildtype cell lines when available. Similar to the parental TE activation analysis, we will use our models to determine the changes in promoter-like loci close to the impactful insertion to understand how the TE insertions affect nearby regulatory elements.

Expected results (Aim 3): We expect to generate a panel of functionally important TE candidates in primary and PDX tumors by our *in silico* functional studies. In particular, we expect to identify some novel and functionally important TEs in the PDX tumors. We expect to develop a panel of stable cell lines bearing the specific TE using the genome-editing technologies and the respective mouse models. These cell lines and mouse models will become valuable resources for the scientific communities to study the biological functions of TEs and their contributions to cancer development, evolution, and drug responses.

Pitfalls/Alternative approaches (Aim 3): We may have potential off-target effects from the genome-editing studies using CRISPR technology. Therefore, we will perform genome-wide microarray to detect non-specific deletions/duplications in the genome and identify the cell lines without the off-target changes for the functional analysis proposed in this study.

	Task	Year 1				Year 2				Year 3				Year 4				Year 5			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Aim 1	Develop ensemble of genomic and transcriptomic TE callers	■	■	■																	
	Ensemble calling		■	■	■	■	■														
	Quantification of TE transcriptomic activity		■	■	■	■															
Aim 2	Identification of TEs using novel computational platform				■	■	■	■													
	Experimental validation of putative TEs			■	■	■	■	■	■	■											
Aim 3	*In silico* functional studies of the putative TE candidates								■	■	■	■	■	■	■	■	■	■	■	■	■
	Experimental studies of putative TE candidates											■	■	■	■	■	■	■	■	■	■
	Functional studies of important TEs													■	■	■	■	■	■	■	■