Project title: Identification and Characterization of Transposable Elements (TEs) in cancers using PDX models

SIGNIFICANCE

Structural variations (SVs), such as deletions, duplications, insertions, inversions and translocations, are among the most significant determinants of human genetic diversity. As part of the 1000 Human Genome Project (1000GP) SV Group, we have been involved 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 datasets at high resolution and provided the research community with an unprecedented set of germline SVs from more than 2,500 normal human genomes. The 1000GP Consortium studies estimate that a typical genome contains 2.1–2.5 thousand SVs, affecting ~20 million bases, which affect far more bases than single-nucleotide polymorphisms (SNPs) combined. SVs can markedly affect phenotype in many ways, including modification of open reading frames, production of alternatively spliced mRNAs, alterations of transcription factor (TF) binding sites and structural gains or losses within the regulatory regions. A large number of studies have suggested that SVs may play important roles in human evolution, genetic diversity, and disease susceptibility.

Transposable elements (TEs) (or mobile genetic elements), including retrotransposons, DNA transposons, and endogenous retroviruses (ERVs), are one of the major mechanisms creating the structural variations in the human genomes and may contribute to both normal developmental processes and disease pathogenesis. The 1000 GP SV subgroup has developed a computational platform called Mobile Element Locator Tool (MELT) for genotyping non-reference mobile element insertions (MEI) in whole genome sequencing datasets and published a catalogue of 16,631 germline MEIs present across the 2,504 individuals. Somatic insertions of the long interspersed element-1 (LINE-1, L1) retrotransposons, the short interspersed element (SINE, Alu) and HERVs have been detected in a number of human cancers, suggesting that they may have a profound impact on tumor heterogeneity and adaptation during cancer progression. In a published study, we developed the Transposable Element Analyser (TEA) computational method to detect TE insertions in 43 cancer genomes (colorectal cancer, prostate cancer, ovarian cancer, multiple myeloma and glioblastomas) from the TCGA whole genome sequencing dataset, and identified 194 high confidence somatic TE insertions (183 L1s, 10 Alus and 1 ERV).

Cancer is a highly heterogeneous and complex disease, which is driven by genetic changes in oncogenes, tumor suppressor genes, DNA repair genes, and other regulatory elements in the human genome. Recent studies have observed 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 in cancers. While increasing evidence suggests that TEs may play important roles in tumorigenesis, the biological functions and molecular mechanisms of TE-mediated insertions in cancers remain poorly understood. The patient-derived xenograft (PDX) models are established by transplantation of patient tumors into immunodeficient mice (e.g. the NSG mice). The Jackson Laboratory and its partners have developed over 400 PDX models. Our studies have demonstrated that PDX models retain many of the key characteristics of patients' primary tumors including histology, genomic signature, cellular heterogeneity, and drug responsiveness, therefore may serve as a powerful platform for studies of cancer biology and co-clinical trials for cancer precision medicine. Since TE insertions may have profound impact on tumor heterogeneity, adaptation, and drug responsiveness, it would be important to study the activity and expression of TEs in PDX models and understand the underlying mechanisms.

While genomic profiling technologies have rapidly advanced our ability to probe and characterize TEs *in silico*, current computational platforms for novel TE discovery are limited by poor sensitivity, specificity and accuracy. We propose here to develop novel computational platforms that use an ensemble of existing callers – TEA and MELT and a novel caller – TESeq. TESeq takes a novel approach to identifying active TE elements by combining DNA and RNA sequencing data from the same patient. The ensemble will take a data mining

approach to integrate the various callers, thereby improving performance as compared to any of the three callers by themselves. The central premise of this proposal is that somatic TE elements in cancers need to be discovered, validated and characterized functionally to understand the mechanisms that lead to TE activation and its role in tumorigenesis. The objective of this study is to develop novel computational platforms to identify novel TEs in primary and PDX tumor samples, and perform functional studies of the putative TE candidates using In-silico functional analysis as well as experimental approaches.

TEAM: Scientists participating in the proposed project are leaders in SV discovery and analysis. The two PIs, Charles Lee, Ph.D. and Mark Gerstein, Ph.D. 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/Ding; 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) whole genome sequencing (WGS) datasets. The two co-investigators, Ankit Malhotra, Ph.D and Chengsheng Zhang, M.D., Ph.D. are also key participants of the 1000GP. Ankit Malhotra has extensive experience in development of computational platforms for genomic studies, whereas Chengsheng 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 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 TEs discovery, characterization and association with human cancers.

PDX models: to our knowledge, this project may become the first study to investigate TEs in cancers using PDX models. A number of factors, including human immune system, have been suggested to affect the activity and expression of TEs. Since the PDX models lack of functional immune system and other human host factors, we may discover novel TEs in the PDX models that might be associated with tumor heterogeneity, adaptation, and drug responsiveness in the PDX models.

SPECIFIC AIMS

The overall goal of this study is to develop an integrated computational platform to identify novel TEs in primary and PDX tumor samples, and perform functional studies of the putative TE candidates using In-silico functional analysis as well as experimental approaches. There are three specific aims in the proposal.

Aim 1. Development of novel computational platform for novel TE discovery in human genomes. We are creating an ensemble of software pipelines that will help us identify TEs from whole genome sequence datasets. We propose to merge these pipelines into a novel computational platform, [name?], that offers [state advantages of new platform and how it will be applied here]]. We will apply the novel software ensemble framework to analyze [[WGS/RNAseq/WES data from]] several PDX (Aim 2) and thousands of PCAWG (TCGA/ICGC) samples. To support these huge datasets, the ensemble will be implemented on a cloud platform capable of providing the resources (both compute and storage) for such a large study. These studies will deliver the most comprehensive resource of genomic and transcriptomic somatic TEs in multiple human cancers and empower us to make novel biological inferences.

Aim 2. Identification and validation of TEs in matched primary tumors and PDX models. Here we will apply the novel computational platform developed in Aim 1 to identify and validate TEs in matched primary and PDX tumor samples. We will perform WGS and RNA-seq on primary and patient-matched PDX-derived tumor samples Novel targets will be validated experimentally by PCR/ddPCR and/or Sanger sequencing. This approach will allow us to identify common and unique TEs as a function of tumor type and to address the fidelity in TE landscapes between primary patient and PDX-derived tumors. It will also allow us to explore the possibility that PDX tumors display a different TE landscape relative to human tumors owing to the lack of a functional immune system and other human host factors.

Aim 3. Functional characterization of putative TE candidates. We will first evaluate *in silico* the oncogenic potential of TEs, focusing on those that impact 1) protein-coding genes, 2) non-coding RNAs and (3) non-coding regulatory regions. An impact score will take into account the varied ways a TE can affect genomic elements and will integrate conservation information, existing genomic annotations and epigenetic and transcriptomic datasets from sources such as ENCODE, 1000 Genomes, and GTEx. Furthermore, we will upweight the impact score of TEs overlapping elements with ubiquitous activity, high network connectivity (ie hubs) and strong allelic activity (i.e. demonstrated functional sensitivity to variants). Putative TEs will be selected for experimental validation based on whether they may affect tumor suppressor genes, gene enhancers, or network hubs.

(1) Perform In-silico functional studies of the putative TE candidates and their impact on cancers;
(2) Perform experimental studies of the potentially significant TE candidates
(3) Functional studies of the important TEs using genome-editing technologies Take top 3 in each category; do CRISPR/Cas9 mutational analysis in mutation-deficient cell line, then put into mouse and look for tumor formation/growth

APPROACH

Comment [AM1]: Lots of missing references, add references !!!

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

Rationale.
Recent literature suggests that L1 is not the only autonomous TE active in the human genome. Human endogenous retro viruses (HERVs), especially solo long terminal repeats (LTRs), were recently described as polymorphic in human populations. On the other hand, little is known about the mobilization of non-autonomous transposable elements (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. (XXX – references). To date, most of the pipelines to detect the mobilization of TE in humans focus on the mobilization of large L1Hs by using paired-end read alignments or transductions of L1Hs.

In order to understand the mechanistic regulation activity of Transposable Elements (TEs) activity we are working on creating an ensemble of software pipelines that will help us identify Transposable elements from whole genome and RNA sequence datasets. The software ensemble would include TEA (XXX) and MELT (XXX) – software developed by our group and others to specifically detect mobile elements from whole genome sequencing datasets. The ensemble would also entail the development of *TESeq*, a novel computational framework to detect genomic and transcriptomic activity of Transposable Elements from whole genome and transcriptomic datasets. We will apply the novel software ensemble framework to several of PDX (Aim 2) and thousands of PCAWG (TCGA/ICGC) samples. In order to support these huge datasets, the ensemble will be implemented on a cloud platform capable of providing the resources (both compute and storage) for such a large study. **Ultimately, these studies will deliver the most comprehensive resource of genomic and transcriptomic somatic TEs in multiple human cancers and empower us to make novel biological inferences.**
Key points to consider:

- Limitations of current computational platforms for the identification of TEs (such as sensitivity and specificity and/or accuracy).
- Improvement by our novel computational platforms

Preliminary results
Transposable Elements 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 normal human genomes that have been sequenced at low depth. We detected a total of 15,834 insertions of transposable elements; of which 3,048 were LINE-1 insertions and 12,786 were Alu insertions. Our group also has extensive experience in developing pipelines to detect structural variations in whole genome sequencing datasets. Meerkat, Hydra-Multi, CNVNator, AGE and Paired-End Mapper (PEMer) are pipelines developed by our group for mapping SVs at high resolution with confidence and then genotyping the discovered SVs in large populations.

Transposable Element Analyser (TEA) is a computational method to detect nucleotide resolution of somatic TE insertions into the human genome using paired end whole genome sequencing datasets. In a published study, we analyzed TE insertions using TCGA WGS dataset from 43 cancer genomes (five cancer types – colorectal, prostate, ovarian, multiple myeloma and glioblastomas). 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 - 97%) of the detected sites were validates using experimental techniques (PCR and Sanger sequencing)

Mobile Element Locator Tool (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 1000 Genomes structural variation subgroup to discover, annotate and genotype non-reference mobile element insertions (MEI) in whole

genome sequencing datasets. Using MELT the 1000 Genomes structural variation subgroup published a catalogue of 16,631 germline MEIs present across the 2,504 individuals.

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 transposable elements. We found that most of ancient and, therefore reliably mappable TEs; have read counts correlated to the most proximal genes, implying that their expression is due transcription activity from near transcriptional active regions (TAR). On the other hand, we find that read counts overlapping evolutionary young elements do not correlate with neighbor genes. Many factors including background transcription or biological noise could account for the majority of the read counts. Collectively, the low-level transcription of most of the human genome is known as pervasive transcription. For highly repetitive regions such as the TEs, pervasive transcription could be a confounding factor when estimating TE transcription level. In order to distinguish between **autonomous** transcription of and pervasive transcription of TE subfamilies we create subfamily mappability fingerprints by simulating reads from their putative mature transcripts. Using this strategy we developed *TeXP. TeXP* is a comprehensive suite that can be used to create TE subfamily **fingerprint** and also to process RNA-seq experiments in order to estimate the proportion of transcriptional signal originating from pervasive transcription and autonomous transcription of TEs. In agreement with previous works, we find that MCF-7, a cell line derived from breast cancer, shows a remarkable high level of L1Hs transcription (288 TPM), in agreement with previous works [R]. We further investigated the transcription level of L1 subfamilies in different MCF-7 cell compartments. We find that WholeCell(polyA+), WholeCell(polyA-) and Nuclear(polyA+) yield 200,000 reads mapping to L1 subfamilies (Fig 1[N]) while Cytoplasmic(polyA+) yields only 50,000. Interestingly, we find that the proportion of reads mapped to different L1 subfamilies greatly varies across different cell compartments (Figure [N]). We also find that despite the absolute difference in the number of reads mapping to each subfamily, WholeCell(polyA+) and Cytoplasmic(polyA+) have a very similar profile and, according to our TeXP methodology, indicates that approximately 50% of the reads are result of autonomous transcription of L1Hs. In contrast, less the 10% of the transcriptional signal in WholeCell(polyA-) and Nuclear(polyA+) derives from autonomous transcription of L1Hs.

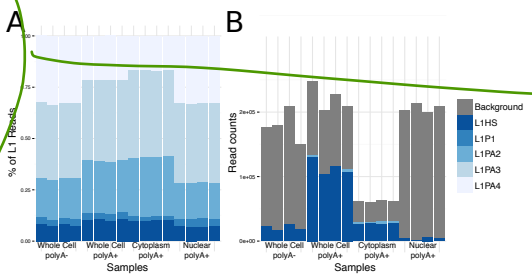


Figure 1. Alignment profile of RNA-seq reads mapping to L1 subfamilies in four MCF-7 cell compartments. Each cell compartment (Whole-Cell PolyA-, Whole Cell PolyA+, Cytoplasm polyA+ and Nuclear polyA+) has four RNA-seq replicates. A) Percentage of L1 reads mapping to L1Hs, L1P1, L1PA2, L1PA3 and L1PA4 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).

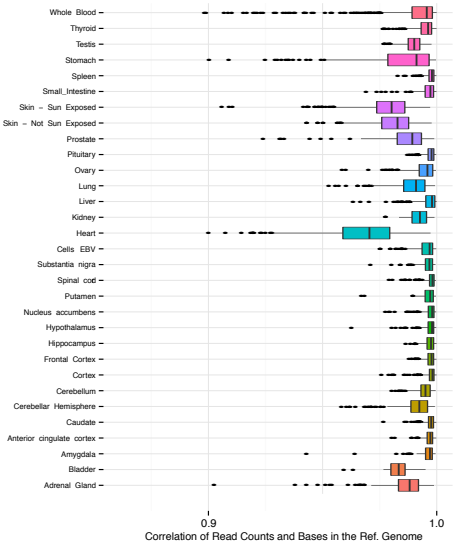


Figure 2. Distribution of the correlation coefficient between the number of RNA-seq reads mapped to L1 subfamilies and the number of bases in the reference genome annotated as that L1 subfamily. Samples are grouped by GTEx primary tissue.

We further analyzed RNA-seq datasets from ENCODE and found that GM12878, a lympoblastoid cell line derived from a healthy individual blood, had no autonomous L1Hs regardless of the cell compartment and

transcript selection process. However, 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 processed 5000 samples from BrainSpan and GTEx to assess the activity of L1 elements in developmental and adult brain tissues. Figure[2] shows that the majority of brain samples show extremely high correlation with the proportion of bases annotated as subfamilies. However, when discriminating samples per tissue we found that a number of them have lower correlation and therefore could support autonomous transcription of L1 subfamilies.

Research Plan

Development of an ensemble of existing and novel genomic and transcriptomic TE callers

We plan to develop new tools and pipelines that will work in the cloud to identify and classify structural variations (SVs) caused by the mobilization of Transposable Elements (TEs) using an ensemble of software including *TEA*, *MELT* and *TESeq*.

Our group is also developing novel pipeline (*TESeq*) to detect mobilizations mediated by L1 reverse transcriptase by detecting the signatures of L1 retrotransposition. 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 Transposable Elements (and protein coding genes to detect retroCNVs) to triage potential mobilizations of TEs. The Direct Repeat (DR) and poly(A) signal flanking L1 mediated retrotransposition is 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 germinative mobilization of TEs.

Comment [AM2]: Revisit

TESeq, will deliver 1) comprehensive identification of somatic mobilization of L1s, ALUs, SVAs, HERVs (+LTR) and retroCNVs (processed pseudogenes) in human healthy and tumoral genomes and 2) integration of RNA-seq data and TEs subfamily mappability profiles to estimate the autonomous transcriptional activity of TEs. 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. One of the insertion ALU insertions overlapped the ATR gene, known to restrict cell cycle and also thought to be a DNA damage sensor.

Ensemble Calling

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 fastq's or aligned BAM files for sequenced samples, perform quality control on the data and then process the data through the three software pipelines (TEA/MELT and TESeq) using 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 1000Genomes (XXX reference) 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 1000Genomes and PCAWG datasets). This would be the first instance of the application of a data mining approach to discovery of TE elements using an ensemble of callers and we expect to demonstrate the robustness and correctness of the approach as opposed to using single algorithms.

Cloud Computing

PCAWG is a wonderful resource for testing and validation of the pipelines developed for this project but the sheer amount of data prohibits the traditional approach to data processing. Therefore, rather than downloading and processing the files locally we've 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 analysis on large scale datasets, such as 1000 Genomes, GTEx, ENCODE and PCAWG.

Significant time and cost savings are also gained by processing a subset of the data. As described earlier, TESeq specifically targets those reads that are either [partially/poorly] aligned or unaligned, ignoring the significant portion of reads that map to the genome unambiguously.

In addition to the logistical issues involving the size of the datasets, there are other issues to consider when executing analyses of this magnitude. It has been observed that the locations and quantities of TE activities in a genome is highly polymorphic in the human population; therefore, the data generated by this analysis are particularly useful for identifying individuals. We have extensive experience analyzing potential privacy incursions and have demonstrated how such files can be de-identified without losing information.

Experimental design (Aim 1)

Expected results (Aim 1)

Pitfalls/Alternative approaches (Aim 1)

AIM 2: Identification and experimental validation of TEs in cancers

Rationale.
Identification of TEs in both primary and PDX tumor samples.
Functional immune system may affect the activity and landscape of TEs in vivo.
We may see the different TE profiles between primary tumor samples and the PDX tumor samples since NSG mice lack of functional immune system.

Summarize at a high level what you will do here and how. Explain how these experiments will contribute to the overarching goals of the proposal.

Preliminary results (Aim 2)

Identification of TEs from TNBC samples (Ankit is working on this)

Experimental design (Aim 2)
2.1 Sample Preparation: In this study, we proposed to investigate 5 cancer types, including breast, gastric, colorectal, lung, and bladder cancers based on the sample availability at our lab. Ten primary and 10 PDX tumor samples from each cancer type will be used for this project. Snap frozen tumor tissue samples from the patients and PDX mice will be obtained from the tissue banks at the Jackson Laboratory and the EWHA University in Korea. This study was reviewed and approved by the Institutional Review Board (IRB) of the Jackson Laboratory and the EWHA University in Korea (JAX IRB Number 121200011 for PDX studies).

2.2 DNA isolation: Genomic DNA will be isolated from the frozen tissues according to the protocol provided by the manufacture (QIAGEN). Briefly, the tissue will be dissected into small pieces and treated with 3ml of cell lysis solution at room temperature for one hour (samples in cell lysis are stable for up to three years). Add 15µl of Proteinase K to tissue sample and incubate at 55°C for 3 hours or overnight at room temperature (no pieces of tissue should be seen if completely lysed). Add 15µl of RNAse A to the sample and incubate in 37ºC water bath for 15 minutes. Place the sample tube on ice for 3 minutes and the add 1ml of the Protein Precipitation Solution to sample, vortex for 1 minute at high speed, and spin sample tube at 2000xg for 5 minutes. Pour the supernatant into 3ml of 100% isopropanol and invert the tube until the DNA is visible. Spin down and aspirate solution leaving the DNA in the tube and washed the DNA pellet with 70% ethanol. Dry the DNA pellet and use low sale TE buffer (10mM Tris.HCl and 1mM EDTA, pH. 8.0) to dissolve the DNA. Check the DNA quantity and quality using agarose gel electrophoresis. The DNA samples are stored at -20°C freezer for future applications.

2.3 DNA library construction and sequencing: The DNA sequencing library will be constructed using a paired-end Illumina TruSeq Kit according to the protocol provided by the manufacture. Briefly, 1 µg of genomic DNA will be sonicated using a Covaris S220 and selected for an insert size of 200–250 bp for library construction. The concentration and insert size of the libraries will be measured using an Agilent Bioanalyzer 2100. The Libraries will be sequenced on an Illumina HiSeq 2000 at the Genome Technology Center of the Jackson Laboratory for Genomic Medicine.

Identification of TEs by computational analysis

Identification of TEs in the primary and PDX tumor samples by computational analysis
Add the novel computational platforms from Yale and JAX in this section.

Find the common and unique TEs in the primary and PDX tumor samples

2.5 Experimental validation of the putative TEs

2.5.1 Selection of TE candidates for the validation: A panel of TE candidates (50-100) will be selected for validation to confirm the false discovery rate of the computational analysis pipelines. We will choose the candidates that hit important genes and/or the regulatory elements of the gene targets (e.g. oncogenes, tumor suppressor genes, splicing sites, microRNA, lncRNA, etc.)

Validation approaches

2.5.2.1 PCR/ddPCR validation: We will perform PCR validations on the putative TEs Primers flanking the predicted breakpoint will be designed to detect the pre-insertion allele. A primer within the L1 and a primer in the flanking genomic sequence will be designed to detect the somatic L1 insertions. Additional primers will be designed to detect L1 insertions that failed the PCR amplification described above. 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. Generally, the PCR assays will be performed using Invitrogen Platinum Taq polymerase in a 25 µL reaction volume for 35 cycles with an annealing temperature of 60°C. In addition, we will use the Bio-Rad Dropt Digital PCR (ddPCR) to detect the copy numbers of the TE candidates.

2.5.2 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.

Expected results (Aim 2)

Identification of novel TEs in these tumor samples.
We may find novel TEs in the PDX samples
We may also find unique TEs in the primary tumor samples that were selected by human immune system and/or the microenvironment.

How would these expected outcomes advance the field? Bring about the goals of the project?

Pitfalls/Alternative approaches (Aim 2)

What pitfalls might you expect and how would you deal with them?

AIM 3: Functional characterization of putative TE candidates

Rationale
Transposable elements have been frequently associated with genetic diseases and are responsible for considerable amount of polymorphism in the human genome and somatic variation in cancer genomes. Despite their relevance, little is known about the mechanisms contributing to their role in oncogenesis or the functional impact of their mobilization in a somatic genome-wide fashion. These events are disproportionately observed in the noncoding part of the genome and we anticipate that comprehensive assessment of TEs the regulatory mechanisms of L1 and the investigation of functional impact will require the integration of large-scale data resources such as ENCODE, 1000 Genomes and GTEx. We anticipate that most of TEs discovered in the human genome will not impact coding regions; thus, methods to evaluate the functional impact of TEs need to be genome-wide, including non-coding regions. We propose to develop a framework to evaluate TEs over three contexts: (1) Impacting protein coding genes; (2) Impacting non-coding RNAs; (3) Impacting non-coding regulatory regions such as Transcription Factor Binding Sites (TFBS). The impact analysis will take into account the varied ways a TE can affect genomic elements (e.g. partial overlap or engulf) and will integrate conservation information, existing genomic annotations, and epigenetic and transcriptomic datasets from sources such as ENCODE, 1000 Genomes, and GTEx. Furthermore, we will upweight the impact score of SVs overlapping elements with ubiquitous activity, high network connectivity (ie hubs) and strong allelic activity (i.e. demonstrated functional sensitivity to variants).

HOW REG

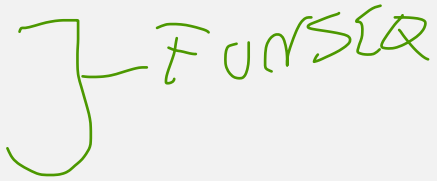
Preliminary results
We have developed a number of computational tools to perform *in-silico* functional studies of the putative TE candidates and their impact on cancers. We have extensive experience in functional interpretation of coding mutations. To this end, we develop Variant Annotation Tool (VAT, vat.gersteinlab.org) to annotate protein sequence changes of mutations. VAT provides transcript-specific annotations and annotates mutations as synonymous, missense, nonsense or splice-site disrupting changes\cite{22743228}. We have 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 1000 Genomes Project\cite{22344438}, distinguishing deleterious LoF alleles from common LoF variants in nonessential genes. We have done an integrative annotation of variants from 1092 humans from the 1000 Genomes Project Phase 1 study\cite{24092746}. By using enrichment of rare nonsynonymous SNPs as an estimate of purifying selection, we showed that genes tolerant of LoF mutations are under the weakest selection, whereas cancer-causal genes are under the strongest. We have also participated in the 1000 Genomes Project Phase 3 studies on LoF variants and functional impact of SVs and found that a typical genome contains ~150 LoF variants. Furthermore, we discovered a significant depletion of SVs (including deletions, duplications, inversions and multiallelic copy number variants) in CDS, UTRs and introns of genes, compared to a random background model, which implies strong purifying selection.
We have also developed RSEQtools and IQseq, tools that build gene models and determine gene- and isoform-level RNA-Seq quantifications \cite{21134889, 22238592}. Beyond quantification of RNA in gene regions, we have also been interested in identifying transcription in unannotated regions, and have developed specific tools to help quantify specific types of transcripts that require special processing, particularly pseudogenes and fusion transcripts \cite{17567993,25157146, 22951037, 20964841}. We have applied our expertise in RNA-Seq analysis to analyze and compare the transcriptomes of human, worm, and fly, using ENCODE and modENCODE datasets. We found a finding striking similarity between the processes regulating

transcription in these three distant organisms \cite{21177976, 25164755, 22955620}. We have also developed tools that specifically analyze features of ncRNAs. Our incRNA pipeline combines sequence, structural and expression features to classify newly discovered transcriptionally active regions into RNA biotypes such as miRNA, snRNA, tRNA and rRNA\cite{21177971}. 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 \cite{21596777}.

We have extensive experience performing annotation of non-coding regulatory regions, with expertise in developing tools to analyze ChIP-Seq data to identify genomic elements and interpret their regulatory potential. For ChIP-Seq, we have developed two tools - PeakSeq and MUSIC - that identify regions bound by transcription factors and chemically modified histones \cite{19122651, 25292436}. PeakSeq has been widely used in consortium projects such as ENCODE \cite{19122651, ENCODE main paper}. MUSIC is a newly developed tool that uses multiscale decomposition to help identify enriched regions in cases where strict peaks are not apparent. This tool has the advantage that it robustly calls both broad and punctate peaks\cite{25292436}. We have further developed methods to use ChIP-Seq signals to identify regulatory regions such as enhancers and to predict gene expression, using both supervised and unsupervised machine learning techniques \cite{21324173, 22039215, 22955978, 25164755, 22950945}. We developed method called Target Identification from Profile (TIP) to predict a TF's target genes\cite{22039215}. Furthermore, we have analyzed the patterns of variation within functional noncoding regions, along with their coding targets\cite{21596777,22950945,22955619}. We used metrics, inter species conservation and the ratio between common and rare variants in different human populations, to characterize selection pressure and, therefore, ultra sensitive regions on various classes and subclasses of functional annotations\cite{21596777}. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region\cite{22955616}.

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\cite{23505346} and positioned on the top level of regulatory networks \cite{22955619}. Further studies showed relationships between selection and protein network topology (for instance, quantifying selection in hubs relative to proteins on the network periphery\cite{18077332,23505346}). Incorporating multiple network and evolutionary properties, we have developed a computational method - NetSNP\cite{23505346} to quantify the indispensability of each gene. This method shows strong potential for interpretation of variants involved in Mendelian diseases and in complex disorders probed by genome-wide association studies.

We have also developed a wide range of analyses on biological networks, with a particular focus on regulatory networks. We constructed regulatory networks for data from the ENCODE and modENCODE projects, identifying functional modules and analyzing network hierarchy\cite{22955619}. To quantify the degree of hierarchy for a given hierarchical network, we defined a metric called hierarchical score maximization (HSM) to infer the hierarchy of a directed network\cite{25880651}. We also developed Loregic to integrating gene expression and regulatory network data and characterize the cooperativity of regulatory factors and interrelate gate logic with other aspects of regulation, such as indirect binding via protein-protein interactions, feed-forward loop motifs and global regulatory hierarchy\cite{25884877}. We have also introduced several software tools for network analysis, including Topnet, tYNA and PubNet\cite{14724320, 17021160,16168087}.



Experimental design (Aim 3)

3.1 Perform **In-silico** functional studies of the putative TE candidates and their impact on cancers (Mark Gerstein's team from Yale).

3.1.1 Building regulatory network, relate to epigenome, of activation of parental TE

In order to understand how and what regulatory signal changes around potentially active TE and how somatic TE insertions affect the downstream gene expressions we will compare the regulatory landscape of TE from two cell-lines GM12878 and K562. These cell lines respectively derived from a healthy and a tumoral cell line. The changes in regulatory context during the transition from healthy to tumoral context will gives insights on the activation of TE.

While roughly 17% of the human genome is derived from L1s, only a small fraction of these are full-length, therefore, potentially capable of retrotransposition. L1Hs is the only L1 subfamily active in the human genome. There are 1,644 L1Hs in the human genome of which 304 are full-length. Moreover, it is believed that approximately 100 full-length L1Hs are capable of translating the L1Hs reverse transcriptase machinery (Hot-L1). These are the elements that are called parental, since they are the template for new L1 mobilizations. Our preliminary analysis revealed that only a small fraction of L1Hs retrotransposon insertions falls into the human genome blacklist regions. Of 1,653 potential L1Hs regions, only 9 full-length L1Hs were overlapped with these regions. Hence, the regulatory context of Hot-L1s are analyzable through epigenomic, histone Chip-seq and DNase-seq experiments.

We first aim to characterize both proximal and distal regulatory changes to identify the active L1 from the potentially mappable 304 full-length L1Hs. In specific, we will set up models to quantify the TF binding events using ChIP-seq experiments as TF scores to search for promoter like regions. Such TF scores will represent the potential of L1Hs 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 L1Hs transcription via distal regulatory elements.

Finally, the differences in epigenomic landscape within and flanking regions of these distal and proximal regulatory events will be characterized between K562 and GM12878 using the ENCODE DNase-seq, histone ChIP-seq, and WGBS methylation data. Based on these profiles, we plan to identify parental L1Hs that were responsible for insertions specific to K562 and cancer etiology.

In addition, we will try to uncover the functional impact of the newly retrotransposed L1Hs. A full category of cell-line-specific functional elements for K562 and GM12878 will be extracted from the ENCODE project, and we will investigate the effect of newly inserted regions to these functional elements. In particular, we will divide the potential functions into two categories, oncogenic and tumor-suppressive. For example, we will systematically search for disruptive functions of tumor suppressor gene transcription in K562 (compared to GM12878) through their distal or proximal regulatory elements as cancer suppressive events. We will also investigate the TF regulatory networks to search for the alternation of key TFs that are regulating either oncogene or tumor suppressor gene.

3.1.2 Prioritization of somatic TE insertions

Our somatic TE insertion prioritization pipeline will integrate many features. We will first identify functional impact of TE insertions on base on the annotation of the insertion point. Slightly different strategies will be used if the insertions overlaps: 1) protein-coding genes; 2) non-coding transcripts 3) non-coding regulatory elements. Also, will will further prioritize somatic TE insertions by investigating insertions overlapping conserved elements; allelic elements; the network connectivity of the target gene.

Identifying the functional impact of TE insertion on protein coding genes.

We will investigate how somatic TE insertions creates loss of function (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) frameshifted open reading frames; or 3) truncate 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 quantify the confidence of these LoFs using features such as whether they are in highly duplicated regions and the number of paralog genes.

Prioritizing non-coding transcripts from structural variant data.

To prioritize the effects of somatic TE insertions in ncRNAs, we will investigate insertions in transcripts 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. For example, we will mine RNA interactions between proteins (e.g., CLIP-Seq) and miRNAs (e.g., TargetScan) to create a compendium of biochemical active regions on protein coding transcripts50-54. We have found annotations of all of the above types—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.

Prioritizing non-coding insertions on regulatory elements.
Unlike protein-coding genes and ncRNAs, 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 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 TF motif. For example, an insertion of a full L1Hs (6Kb) should be more harmful than and ALU (360 base pairs). We will use TF binding nc elements by leveraging better enhancer definitions provided by the Epigenome Roadmap56-58 and ENCODE to create a compendium of regions the human regulatory elements.

We will further prioritize TE insertions based on conservation, networks, and allelic activity. After performing annotation-based assessment of identified TE insertions, the following functional features will be used for prioritization.

i) *Conservation of the insertion point.* For evolutionary properties, we will quantify the conservation of insertions points using intra-human variation data (from The 1000 Genomes Project) 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 as cross-species evolutionary conservation (using classical measures such as the GERP score93) to prioritize variants disrupting evolutionary relevant regions.

ii) Network connectivity. We will examine the network topological properties of the genomic elements affected by somatic TE insertions. Variants disrupting regulatory elements with high connectivity—network hubs and bottlenecks—will be upweighted based on their scaled centrality scores. 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 variety of networks -- e.g., regulatory network, metabolic pathways, etc; to assess the impact of the insertion. We will examine their network centralities (eg hubs, bottlenecks and hierarchy tops), as we know that disruption of highly connected genes or their regulatory elements is more likely to be deleterious73,75.

iii) Allelic activity. We will prioritize TE insertions that overlap our database of strongly allelic regions throughout the genome, based on AlleleDB\cite{}, our resource of such regions identified through allele-specific RNA-Seq analysis from over 300 individuals generated by the GEUVADIS consortium48.
RNA-seq of the primary and PDX tumor samples: RNA-seq will be performed according to protocols described previously.

3.1.3 DNA Methylation Analyses: The L1 promoter methylation level will be performed according to protocols described previously.
Functional impact of TEs by building network analysis
Expected Results + Pitfalls/Alternative approaches [1/2 pg]

Perform experimental studies of the potentially significant TE candidates

Detection of gene expression levels by RT-PCR (Charles Lee's team from JAX): Total RNA will be isolated from the tumor samples using RNeasy Kit (QIAGEN) and RNA 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 protocols described previously.

Detection of protein expression levels by Western blot and/or immune histochemistry (Charles Lee's team from JAX)

Western blot: Tissues or cell pellets will be lysed in the lysis buffer (50mM HEPES, 1%Triton X-100, 50mM NaCl, and protease inhibitor cocktail. Western blot will be performed using the specific antibodies against the target proteins according to the protocols described previously.

Immunohistochemistry (IHC): To detect the target protein expression level in the tissue sections, 5μm-thick sections of formalin-fixed, paraffin-embedded (FFPE) tissue samples will be used for IHC according to the protocols described previously.

3.2.3 Microsatellite instability (MSI) assays: The Five markers recommended by the National Cancer Institute will be sued to assess the 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 described previously (Ashktorab et al. 2003; Muller et al. 2004).

Functional studies of the important TEs using genome-editing technologies (Charles Lee's team from JAX): We will employ the CRISPR/Cas9 genome-editing technologies to study the functions of important TEs using cell-based assays.

Expected results (Aim 3)

What do you expect to find? How would these expected outcomes advance the field? Bring about the goals of the project?

Pitfalls/Alternative approaches (Aim 3)
What pitfalls might you expect and how would you deal with them?