Significance [1pg]

Innovation [1/4pg]

Research strategy

AIM1: Calling the TEs on large scale **datasets.**

Rationale [1/4pg]

To understand the mechanistic regulation activity of Transposable Elements (TEs) activity we are working on *[[TESeq]]*, a framework developed by our group to unveil genomic and transcriptomic activity of Transposable Elements in whole genome and transcriptomes datasets. *[[TESeq]]* uses genomic datasets to detect, at the base level, retrotransposition of L1s, ALUs, SVAs, HERVs and retroCNVs by incorporating Paired-end and Split reads strategies. Our group is also developing TeXP, a transcriptome module of [[TESeq]] that uses fingerprints created by the mappability profile of TE subfamilies to distinguish pervasive transcription from autonomous transcription of Transposable Elements. We will apply the [[TESeq]] framework to thousands of local (aim2) and PCAWG (TCGA/ICGC) samples. In order to support these huge datasets, [[TESeq]] will be developed to leverage cloud environments. **Ultimately, these studies will deliver the most comprehensive understanding of genomic and transcriptomic somatic activity of TEs in humans and empower us to make novel biological inferences.**

Preliminary [1.5pg]

Our group has extensive experience in developing pipelines to detect structural variations in whole genome datasets. Paired-End Mapper (PEMer) is a pipeline for mapping SVs at high resolution with a confidence measure. Our group developed incorporated error models based on extensive simulations facilitated parameterization of PEMer and an evaluation of its performance.[R] Alignment with Gap Excision (AGE) is a solution that finds the optimal solution by simultaneously aligning the 5′ and 3′ ends of two given sequences and introducing a ‘large-gap jump’ between the local end alignments to maximize the total alignment score. AGE is also able to tandem duplications, inversions and complex events involving two large gaps[R].; CNVNator is a tool for CNV discovery and genotyping in a population and characterization of atypical CNVs, such as de novo and multi-allelic events.

Consortium efforts such as the 1000 Genomes Project (1000GP) estimate that a typical genome contains 2.1–2.5 thousand SVs, affecting ~20 million bases, or ~5–6 times that of SNPs.

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. As part of this effort we detected a total of 15,834 insertions of transposable elements; of which 3,048 are LINE-1 insertions and 12,786 are Alu insertions.

[[1pg - Pgenes]]

Our lab has expertise in pseudogene identification and annotation. Our work on this topic includes the creation of computational pipelines (PseudoPipe) for the automated identification of pseudogenes via sequence homology. Additionally, we have experience layering functional and comparative genomic information on top of identified pseudogenes. As a contributor to the GENCODE project, we developed the GENCODE pseudogene resource, which contains both computationally predicted and manually annotated pseudogenes as well as corresponding functional information from the ENCODE project. This includes information on transcription factor and Pol II binding sites, chromatin marks, and expression. This data can provide an indication of potential pseudogene activity.

We also apply comparative genomic methods to identify sequence conservation at both the pseudogene level as well as between the CDS and 3' UTR of a given pseudogene. These enable the identification of pseudogenes under selective pressure suggestive of a possible functional role, perhaps as noncoding RNA. This combination of functional and comparative data for identified pseudogenes has been collected in the psiDR resource (http://pseudogene.org/psidr/).

The approaches and resources described above will be leveraged in the context of cancer genomics to better understand the possible roles played by processed pseudogenes. We are well positioned to identify processed pseudogenes in cancer genomes and study their expression and activity across various cancer types and in relation to corresponding healthy tissues.

Tools for uniform processing of RNA-seq data.

We 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 long RNA-seq data for ENCODE, long and short RNA-seq data for the PsychENCODE[43] and Brainspan project as well as a custom pipeline developed for the analysis of small exRNA-seq data for the Extracellular RNA Communication Consortium (ERCC). We have already developed an efficient in-house data processing workflow for RNA-seq data that includes data organization, format conversion, and quality assessment. RSeqTools[44] is a modular tool developed for the processing of RNA-seq data and generating either transcript, gene or exon level quantifications. We also developed IQSeq[45] which calculates the relative and absolute abundance of contributing transcript isoforms to a gene from RNA-seq data using a fast algorithm based on the Fisher information matrix.

Research Plan [2-3pg]

Calling genomic and transcriptomic activity of Transposable Elements

a. Genomic caller - TeXP (1pg)

We plan to develop new tools and to identify and classify structural variations (SVs) caused by the mobilization of Transposable Elements (TEs). The new and improved [[*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.

Sample Selection. In order to understand the regulatory mechanisms of TEs in cancer we ought to describe the genomic and transcriptomic activity of TEs across healthy human tissue. As part of the 1000 Genomes Structural Variation our group we have access to three cohorts using the most comprehensive set of sequencing technologies. This dataset, in conjunction with 1000 Genomes Phase3, will be used to assess the populational variation of TEs across healthy individuals. In the same line of thought our group is using RNA-seq datasets extracted from healthy individuals by GTEx to assess the autonomous transcriptional activity of TEs in healthy somatic tissue. GTex-p6 is comprised of 9,798 samples from 53 human tissues. After describing the transcriptomic and genomic activity of TE in healthy individuals we will tackle the somatic activity of TEs in tumoral samples. Our group is processing datasets from TCGA+ICGC (The Pan-Cancer Analysis of Whole Genomes - PAWG) that contains [[5790 samples from 2834 donnors - http://pancancer.info/]]. A dataset of that magnitude requires special strategies to be processed (described below). We will first prioritize samples with RNA-seq [[1299]] and particularly those with paired-end RNA-seq [[162]]. [...] [[CL+AM: Add local samples from JAX?]]

Detection of somatic mobilization of Transposable Elements.
Recent literature suggests that L1 is not the only autonomous TE active in the human genome. HERVs, especially solo LTRs, were recently described as polymorphic in human populations. In 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, although only a handful of publications investigated the mobilization of these entities. 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 reads alignments or transductions of L1Hs. Non-autonomous mobilizations, such as the mobilization of SVAs, ALUs, and retroCNVs are also contains the molecular signatures of L1 reverse transcriptase, therefore, contains Direct Repeats (DR) flanking their insertion and polyA at the 3' end.

Our group is developing pipelines to detect mobilizations mediated by L1 reverse transcriptase by detecting the signatures of L1 retrotransposition and reads partially aligned to the human

reference genome. This strategy intends to save processing time by avoiding realignment of the original datasets. In a given run from TCGA, typically 30x coverage, [N%] reads are aligned to the reference genome with high quality, of these, [N%] are partially aligned. We are clustering hard and soft clipped reads to define putative structural variations supported by multiple alignments. We use sequences extracted from soft-clipped reads to perform a local assembly and infer inserted/deleted sequences. Posteriorly, inferred insertions and deletions are mapped to annotated Transposable Elements (and protein coding genes for retroCNVs) to select potential mobilizations of TEs. Due to DR, insertion points of L1 machinery retrotranspositon are never at the base pair resolution but at the DR length resolution - 7-20bp. We are using DRs and poly(A) presence/absence at the 3' extremity to further support non-autonomous TE mobilization. At this stage, putative mobilizations of TE can be germinative or somatic. When available we will use paired tissue information and 1000 Genomes TE polymorphism dataset to annotate germinative mobilization 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. [[extend?]] One of the insertion ALU insertions overlapped the ATR gene, known to restrict cell cycle and DNA damage sensor.

To evaluate the performance of our genomic caller we took advantage of deep-coverage, PCR-Free WGS data from 27 samples sequenced by the 1000 Genomes Project. Using the annotated SVs, specifically the L1 and Alu mobilizations, from the 1000GP Phase 3, we performed k=3 cross fold validation on this cohort, wherein we built a model using 18 samples and applied the model to the other 9 samples for SV discovery *ab initio*. This step was repeated 1000 times with random selection for the learning samples and the test samples.

b. Retrodup Caller [STL/AA]

c. Transcriptome caller - TeXP (1.5pg)

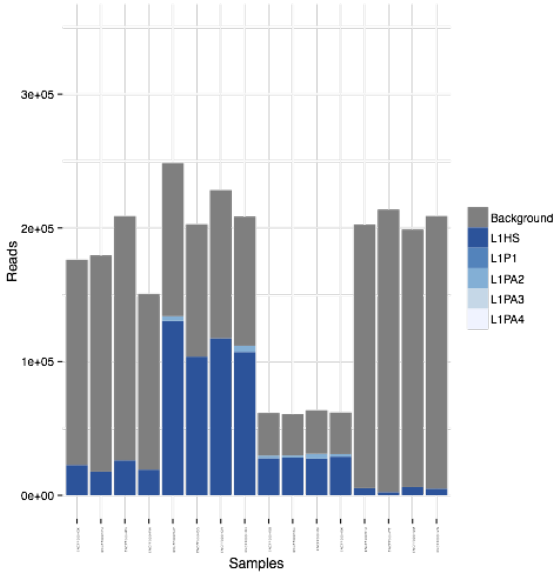
We analyzed more than 5,000 [[How much is enough for a pilot analysis?]] RNA-seq experiments from available datasets ([N]Table S1) from [N] human organs and explored the expression of repetitive elements across the human samples. We found that most of ancient and, therefore highly expressed and reliably mappable TEs - such as LTRs, DNA transposons, L2s; correlate with the most proximal genes, implying that its expression is due to background transcription near transcriptional active regions (TAR) ([N]Supplementary Figure 1).

In other hand, for most RNA-seq experiments, read counts overlapping evolutionary young elements correlate with their cumulative subfamily bases count in the genome (Figure[[N]]). We hypothesize that most of this signal is generated by pervasive transcription of regions annotated as repetitive elements. In order to distinguish between **autonomous** transcription of L1, LTR and SVAs subfamilies and passive transcription of L1 subfamilies we simulated reads originating from their respective putative subfamily transcripts. The simulated reads were aligned to the human reference genome and the TE subfamily mappability fingerprint was created (Figure[N]). We found that younger TEs tends to have more reads mapped to other closely related subfamilies. For example, we find that only ~30% of reads from L1Hs (the most

recent – and supposedly active. L1) maps back to loci annotated as L1Hs. While older subfamilies such as L1PA4, have a higher proportion of reads mapping back to its instances. We rationalized that the number of reads mapped to each subfamily must be a combination of signals generated by autonomous transcription of L1 Subfamilies and Pervasive transcription (Supplementary figure [N]).

We developed a method called TeXP. TeXP is a comprehensive suite that creates TE subfamily **fingerprint** and also process RNA-seq experiments to estimate the proportion of transcriptional signal originating from pervasive transcription and autonomous transcription of TEs. Using TeXP, we first estimated the transcription level of L1 subfamilies across RNA-seq experiments generated by ENCODE (Table [N]).

We find that MCF-7, a cell line derived from breast cancer, shows a remarkable high level of L1Hs transcription (288 TPM) as in agreement with previous works [R]. We further investigated the transcription level of L1 subfamilies in different cell compartments. Using RNA-seq from different cell compartments publicly available from ENCODE, we find that WholeCell(polyA+), WholeCell(polyA-) and Nuclear(polyA+) yield 20 thousand reads mapping to L1 subfamilies (Fig Sup[N]) while Cytoplasmic(polyA+) yields a [N]. Interestingly, the number of reads mapped to different L1 subfamilies varies across different cell compartments (Figure [N]). We find that despite the absolute difference, WholeCell(polyA+) and Cytoplasmic(polyA+) have a similar profile that according to TeXP 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. Conversely, most of the Nuclear(polyA+) signal is originated by pervasive transcription, suggesting that, at least in MCF-7, the signal of autonomous L1Hs transcription is mostly represented by potentially functional cytoplasmatic mature L1Hs RNA.



We further analyzed ENCODE RNA-seq datasets and found that GM12878, a lympoblastoid cell line derived from a healthy individual blood, have no autonomous L1Hs regardless of the cell compartment and transcript selection process. As well as the most of the compartments from K562 and SKMEL5. However, in contrast to GM12878, SK-MEL-5 and K562 are cancer derived cell lines and show transcription level of respectively [N] and [N] in whole-cell polyA+ datasets. If

pervasive transcription is the dominant mechanism L1s in a given experiment, we expect the number of reads overlapping each subfamily to correlate with the number of bases annotated as each subfamily. Indeed, We find that all experiments using GM12878 have correlations higher than .9[N]. With that result in mind we reprocessed [N] samples from BrainSpan and GTEx to assess the activity of L1 elements in developmental and adult brain tissues. Figure[N] 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.

[[+Tumoral brain samples?What sort of result I should add at this point?]]

d. Ankit/Charles Caller

e. Cloud Computing [1 pg]

PCAWG is a wonderful resource for this project but the file sizes prohibit standard processing pipelines. For example, for the lung cancer PCAWG data there are 180 BAM files with a mean size of 145 GB. Standard processing involves downloading the data and running analyses locally, however, assuming a 10 Mbps internet connection, it would take approximately one month to simply download the lung cancer data. Scaling this to all of the tissues available in PCAWG brings the download time to over two years.

Rather than downloading and processing the files locally we've adapted TESeq to run in the cloud. While conceptually this process seems straightforward, there are significant hurdles to creating the algorithms in this fashion; TESeq remain the only transposable elements caller that are fully cloud compatible [is this true???]. This makes us uniquely positioned to be able to execute this analysis on the scale of the entire PCAWG dataset.

The callers are run by … doing something in the cloud.

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. ????

Expected Results + Pitfalls/Alternative approaches [1/2 pg]

AIM2: **Validation.**

Rationale [1/4pg]

Preliminary [1.25pg]

Research Plan [3pg]

Expected Results + Pitfalls/Alternative approaches [1/2 pg]

AIM3: **Characterization of TEs activity.**

Develop tools to analyze the functional impact of TEs. 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 score 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).

Rationale [1/4pg]

Complex SVs are frequently associated to genetic diseases and are responsible for more nucleotides variation than single nucleotide polymorphism in the human genome. Despite their relevance, little is known about their functional impact in a genome-wide fashion. These events are disproportionately observed in the noncoding part of the genome and we anticipate that comprehensive assessment of TEs functional impact will require the integration of large-scale data resources such as ENCODE, 1000

Genomes and GTEx. We also anticipate that this proposal will catalogue the largest number of TEs so far; therefore, new methods to functionally prioritize TEs and select a smaller number for the association studies will be necessary.

Preliminary Results [1.5pg]

Mechanism Classification (NAHR v NHR)

We have intensively studied the distinct features of SVs originated from different mechanisms. This indicates specific creation processes and potentially divergent functional impacts[24092746][26028266]. The most notable type, NAHR, is associated with activating enhancers and open chromatin environment. Our analysis also showed that micro-insertions flanking NH type breakpoints are templated from late replicating genome sited with characteristic distances from breakpoints. These results not only shed light on SV forming processes but also indicate differences in functional impacts of different SVs types. We also performed SV mechanism annotations for the 1000 Genomes Phase 3 deletions using BreakSeq [20037582]; and categorized 29,774 deletions into NAHR, NHR, TEI and VNTR by their origination mechanisms. Among these, NHR is the most prevalent mechanism (~73% of all categorized deletions) [1000G Phase3 SV reference].

Functional Impact in Genes and Pseudogenes

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 developed PseudoPipe, the first large-scale pipeline for genome wide human pseudogene annotation\cite{16574694}, and then obtained the “high confidence” pseudogenes by combining computational predictions with extensive manual curation\cite{22951037,25157146}. We identified parent gene sequence from which the pseudogene arises based on their sequence comparisons\cite{22951037}. We have also studied the mechanisms of pseudogene formation by relating pseudogenes to segmental duplications\cite{20615899} and retroduplication events\cite{24026178}. Through integration of functional genomics data generated by the ENCODE Project, we identified a broad spectrum of biological activity for pseudogenes, and in particular, revealed ~15% of pseudogenes are transcribed\cite{25157146}.

Functional Impact in non-coding RNAs

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}.

Functional Impact in non-coding regulatory regions

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, such as diversity and fraction of rare variants, to characterize selection pressure 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}.

Preliminary results related to networks and allelic expression

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}.
We have also developed a tool, AlleleSeq\cite{21811232}, for the detection of candidate variants associated with allele specific binding (ASB) and allele specific expression (ASE) based on the construction of a personal diploid genome sequence (and corresponding personalized gene annotation) using genomic sequence variants (SNPs, indels, and SVs).

Research Plan [1pg]

1. 2-3pg Charles

2. (1) JZ/DL K562 & GM12878 - building regulatory network, relate to epigenome, how TE gets turned off

By comparing the L1Hs from K562 and GM12878, we aim to answer two questions: 1) how and what regulatory signal changes around the full-length L1Hs open the transcriptional machineries; 2) how will the newly transcribed copies affect the downstream gene expressions.

While roughly 17% of the human genome is derived from L1 elements, only a small fraction, which is about 7,000 in average human genome (needs to confirm, https://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-9-r100), of elements are full-length and capable of retrotransposition. 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 it is possible to locate most of the parental L1Hs.

We first aim to characterize both proximal and distal regulatory changes to identify the active one (autonomous?) from the potentially mappable full-length 1,644 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. Besides, we have also developed a match-filter based enhancer discovery algorithm to discover enhancers in these cell lines. Another algorithm called ENGINE will try to utilize Hi-C or ChIA-PET experiment to find target regions of the discovered enhancers. These distal regulatory elements could help to uncover the underlying mechanism of L1Hs transcription. 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 and histone ChIP-seq data. Based on

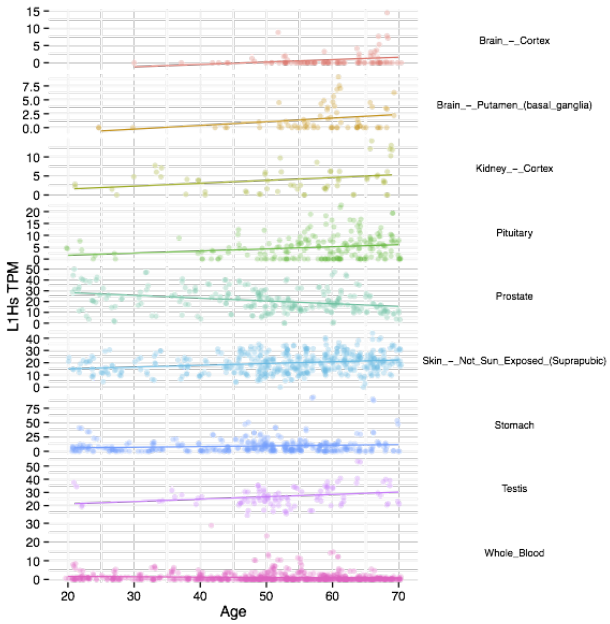
these profiles, we will identify parental L1Hs that were responsible for insertions specific to K562 and cancer etiology.

In addition , we will try to uncover the functional impact of the newly transcribed L1Hs (is this call newly transcribed?). 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) SK functional impact of TEs

4. (1) FN GTEx+Aging

Tissue	Correlation	p-value
Brain - Cortex	0.23	1.8x10-2
Brain - Putamen	0.25	3.1x10-2
Kidney	0.27	2.9x10-2
Pituitary gland	0.17	9x10-3
Prostate	-0.31	5.4x10-8
Skin (not exposed)	0.22	1.38x10-5
Stomach	0.12	1.9x10-2
Testis	0.24	1.5x10-2
Whole blood	-0.1	4x10-3



Expected Results + Pitfalls/Alternative approaches [1/2 pg]

PROJECT SUMMARY (0.5pg)

SPECIFIC AIMS (1pg)