Transposable element activity, genome regulation and human health
Lu Wang¹,² and I King Jordan¹,²

A convergence of novel genome analysis technologies is enabling population genomic studies of human transposable elements (TEs). Population surveys of human genome sequences have uncovered thousands of individual TE insertions that segregate as common genetic variants, i.e. TE polymorphisms. These recent TE insertions provide an important source of naturally occurring human genetic variation. Investigators are beginning to leverage population genomic data sets to execute genome-scale association studies for assessing the phenotypic impact of human TE polymorphisms. For example, the expression quantitative trait loci (eQTL) analytical paradigm has recently been used to uncover hundreds of associations between human TE insertion variants and gene expression levels. These include population-specific gene regulatory effects as well as coordinated changes to gene regulatory networks. In addition, analyses of linkage disequilibrium patterns with previously characterized genome-wide association study (GWAS) trait variants have uncovered TE insertion polymorphisms that are likely causal variants for a variety of common complex diseases. Gene regulatory mechanisms that underlie specific disease phenotypes have been proposed for a number of these trait associated TE polymorphisms. These new population genomic approaches hold great promise for understanding how ongoing TE activity contributes to functionally relevant genetic variation within and between human populations.

Addresses
¹ School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA
² PanAmerican Bioinformatics Institute, Cali, Colombia

Corresponding author: Jordan, I. King (king.jordan@biology.gatech.edu)

Introduction
Transposable elements (TEs) are distinguished by their ability to move, i.e. transpose, among genomic locations, often making copies of themselves as they go. TEs can replicate to extremely high copy numbers over time; at least 50% of the human genome sequence is thought to be derived from TE insertions [1,2]. The abundance of TE sequences, along with their ability to colonize a seemingly endless variety of host genomes, begs an explanation for their evolutionary success. The selfish DNA theory holds that TEs are genomic parasites, which play no functional role for their hosts and exist simply by virtue of their ability to out-replicate the genomes in which they reside [3,4]. The selfish DNA theory is still widely considered to represent the null hypothesis that best explains the presence of TEs from an evolutionary standpoint. Nevertheless, numerous studies have revealed instances of exaptation [5], also referred to as molecular domestication [6], whereby formerly selfish TE sequences have been co-opted to provide some functional utility for their host genomes. The most widely observed route of molecular domestication entails the conversion of TE sequences into host genome regulatory elements [7–9].

TE-derived sequences provide a wide variety of regulatory elements to the human genome, including promoters [10–12], enhancers [13,14,15–17], transcription terminators [18] and several classes of small RNAs [19–21]. Human TE-derived sequences can also exert higher order influences on gene regulation by shaping chromatin structure across the genome [22–26]. It is important to note that, until this time, nearly all studies on human TE regulatory elements have focused on TE-derived sequences that are remnants of relatively ancient insertion events and no longer capable of transposition. Accordingly, known human TE regulatory sequences largely correspond to so-called ‘fixed’ TE insertions, which are found at the same genomic insertion site locations within the genomes of all human individuals. This distinction is critical, since fixed TE insertions are not expected to contribute to regulatory variation among individual humans. In other words, fixed TE regulatory elements, while functionally important, do not provide a source of human population genetic variation.

Over the last several years, a convergence of genome-enabled technologies has begun to power studies that are focused squarely on structural variations generated by the ongoing activity of human TEs. There are several families of human TEs that retain the ability to transpose, primarily Alu [27,28], L1 [29,30], and SVA [31,32]. Alu and SVA elements are non-autonomous SINEs (Short Interspersed Nuclear Elements), which are mobilized in...
trans by the transposition machinery encoded from autonomous LINEs (Long Interspersed Nuclear Elements) of the L1 family. Smaller numbers of HERV-K endogenous retroviruses also remain active in the human genome [33]. When members of these TE families transpose within the human genome, they generate inter-individual variations that segregate within and between populations in the form of TE insertion site polymorphisms. Given the known regulatory properties of human TEs, it is not unreasonable to expect that segregating TE polymorphisms could have significant regulatory consequences. In particular, some human TE polymorphisms may lead

---

The population genomic approach for the study of TE phenotypic effects. Individuals sampled from human populations are characterized using genome (DNA-seq) and transcriptome (RNA-seq) profiling techniques. Genome-wide TE insertion genotypes are compared to tissue-specific gene expression levels to uncover TE variants implicated in gene regulation. The linkage disequilibrium patterns (LD) among TE polymorphisms and SNPs are evaluated to identify TE insertions linked to genome-wide association study (GWAS) loci. Interrogation of functional information is used to hone in on likely TE causal variants.
to differences in gene expression patterns between individuals. Furthermore, human regulatory variation generated by recent TE activity may have important implications for health and disease. This mini-review is focused on recent studies that are beginning to shed light on the ways in which ongoing TE activity can impact human health via changes in genome regulation. These studies are distinguished by their population level approach to the study of TE generated human variation (Figure 1).

**Genome-enabled approaches for characterizing TE insertion variants**

Two distinct classes of genome-enabled approaches for the characterization of TE insertion variants have emerged over the last several years [34*]: (1) bioinformatics methods that rely on the analysis of whole genome sequence data to find TE insertions that differ from a reference sequence (Figure 2A), and (2) high-throughput experimental methods that utilize next-generation sequencing to pinpoint the locations of novel TE insertions (Figure 2B).

Computational approaches for the discovery of TE insertion variants rely on one of two methods: (1) discordant read-pair mapping for short read sequencing technology, or (2) split read mapping for long read technology [35*]. Our own group recently performed a benchmarking study on 21 bioinformatics tools designed for detecting human TE insertion variants from whole genome sequence data [36**]. After an initial screen of tools that were found to be unreliable, or no longer maintained, our study focused on seven programs: ITIS [37], MELT [38*], Mobster [39], RetroSeq [40], Tangram [41], TEMP [42], and T-lex2 [43]. We found MELT to have superior performance for human TE variant detection from whole genome sequence data, but also show how a combined approach using two or more methods, including Mobster and RetroSeq, could yield superior performance. Since the publication of our paper, two new computational tools for TE insertion discovery have been published. The program STEAK [44] claims superior performance compared to existing short read methods, whereas LoRTE [45] is designed for PacBio® long read sequence technology.

At this time, given the predominance of Illumina® short read sequencing technology, discordant read-pair mapping approaches are most widely used. It should be noted that some short read methods also employ split, clipped, or insertion junction reads, in addition to discordant read-pair mapping, as part of their TE detection protocols. Nevertheless, these short read methods are still far from perfect and there is substantial room for additional development in the field. As long read sequencing technology becomes more widespread, split read approaches should

---

**Figure 2**

<table>
<thead>
<tr>
<th>(a) Bioinformatics Methods</th>
<th>(b) High-Throughput Experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Whole Genome Sequencing</strong></td>
<td><strong>DNA Fragments with TE Insertions</strong></td>
</tr>
<tr>
<td><strong>TE Insertion</strong></td>
<td><strong>PCR</strong></td>
</tr>
<tr>
<td><strong>Discordant Read Pairs</strong></td>
<td><strong>Hybridization</strong></td>
</tr>
<tr>
<td><strong>Partially Mapped Reads</strong></td>
<td><strong>Next-Gen Sequencing</strong></td>
</tr>
<tr>
<td><strong>TE Insertion Variant Calls</strong></td>
<td><strong>Tiling Array</strong></td>
</tr>
</tbody>
</table>

Genome-enabled approaches for the discovery and characterization of TE insertion variants. (A) Bioinformatics methods rely on the computational analysis of whole genome sequence read data to characterize genome-wide patterns of TE insertion alleles and genotypes. (B) High-throughput experimental methods use enrichment of genomic fragments that contain known active TE sequences followed by sequence or array based characterization of their genomic locations.
become more popular. Perhaps more importantly, we expect that split read approaches with long reads will be inherently more accurate and reliable than discordant read pair mapping, since long reads that span entire TE insertions should be mapped with much less ambiguity than shorter reads. Long reads should also help to disambiguate complex structural variants resulting from nested TE insertions.

High-throughput experimental techniques for TE variant detection also share several basic features: (1) DNA fragmentation, (2) TE enrichment, and (3) TE calling. The methods are distinguished by the approaches used for each step of the process. DNA fragmentation can be achieved via enzymatic digestion or by mechanical shearing. TE enrichment can be performed using PCR, with active TE-specific primers, or with hybridization to active TE-specific probes. Finally, TE calling is done using next-generation sequencing, for more recent methods, or with tiling arrays for the older methods. The most widely used experimental methods for TE variant detection include ME-Scan [46], L1-Seq [47], RC-Seq [48], and Transposon-Seq [49]. One area of ongoing improvement for these methods entails the refinement of algorithms used to map enriched TE fragments to genome reference sequences. For example, the TIPseqHunter algorithm was recently developed to refine and improve human TE variant calls made by the existing TIP-seq experimental method [50].

Genome-scale experimental approaches of this kind have been most widely applied to the study of somatic TE variants that characterize cancer tissues. This is one of the most promising areas of recent human TE research, and it has been extensively reviewed elsewhere [51]. This mini-review is focused instead on germline mutations that yield inter-individual differences in TE insertion patterns and manifest themselves as human population genetic variations, i.e. TE polymorphisms.

**TE polymorphisms and human genome regulation**

Our own group recently published a population-level view of the regulatory consequences of recent human TE activity [52**]. For this study, we adopted the expression quantitative trait loci (eQTL) analytical paradigm for the analysis of human TE polymorphisms. eQTL are genomic variants associated with changes in gene expression levels [53]. The eQTL approach requires multiple individual samples that have been deeply characterized at both the genomic (DNA-seq) and transcriptomic (RNA-seq) levels. Gene expression levels for individual samples are regressed against locus-specific genotypes for matched individuals to uncover eQTL associations. This approach was developed for single nucleotide polymorphism (SNP) genotypes, whereas in our case, we used locus-specific TE insertion state genotypes. TE insertion genotypes at any locus can be encoded as 0 (homozygous — insertion absent), 1 (heterozygous — one insertion present), or 2 (homozygous — two insertions present). Differences in gene expression levels across these distinct TE insertion states are indicative of TE polymorphism-to-gene expression associations (Figure 1). For the case of either SNPs or TE genotypes, the eQTL approach depends critically on the reliability of individual variant calls. Extensive benchmarking of SNP and TE variant callers has been performed, for example as part of the 1000 Genomes Project (1KGP), as previously described [54,55**]. As is standard for this kind of analysis, we only use variant calls that have been validated, including avoiding low frequency variants, for the purposes of eQTL analysis [52**].

This approach was powered by the 1KGP, phase 3 of which entailed the genome-wide characterization of TE insertion genotypes for 2504 individuals across 26 human populations [54,55**]. B-lymphocyte gene expression data, derived from EBV-transformed lymphoblastoid cell lines or LCLs, for 445 of the same 1KGP individuals, representing one African population and four European populations, were taken from the Genetic European Variation in Health and Disease (GEmUVEDIS) RNA-seq project [56]. Merging data from both projects allowed us to directly compare TE insertion site genotypes to gene expression levels from the same individuals. Furthermore, comparison of results for African and European populations allowed us to uncover population-specific regulatory effects of human TE polymorphisms.

Regression of gene expression against TE insertion site genotypes revealed hundreds of eQTL associations, and TE-eQTL were found both within and between the African and European populations. A number of TE polymorphisms were shown to be associated with expression differences between population groups. One advantage of using TE insertion site genotypes for eQTL analysis is that the relatively low number of common TE genotypes across the genome (~16,000) allows for both cis and trans eQTL analysis. This is because the number of possible eQTL associations is the product of the number of genes and the number of variants being compared; accordingly, the analysis of millions of SNPs times thousands of genes presents a combinatorically daunting bioinformatics analysis challenge. For this reason, most SNP eQTL studies focus exclusively on cis SNPs that are found within or in close proximity to individual genes. Since our study was not limited in this way, we were able to discover many trans associations of TE polymorphisms with human gene regulation. In fact, we were surprised to find that trans regulatory effects for TE polymorphisms were even more common than cis effects.

For one particular example, the B cell specific transcription factor PAX5, we uncovered a potential mechanism
The impact of TE polymorphisms on gene regulatory networks. The eQTL approach is used to discover associations between TE insertion variants and tissue-specific gene expression levels (i.e. TE-eQTLs). A TE insertion variant found in cis to a transcription factor (TF) can lead to coordinated changes across a gene regulatory network via transitive effects on downstream targets of the TF. An example is shown, similar to what has been observed for the TF gene PAX5, where TE associated increase in the expression of a TF leads in turn to increased expression of the TF target genes. This will reveal itself as multiple trans TE-eQTL associations for the same TE insertion variant.
that could explain the numerous trans TE-eQTL that we observed (Figure 3). This example also underscores how individual TE loci can participate in the rewiring of entire regulatory networks. The PAX5 gene has a cis Alu eQTL that is associated with increased expression in B lymphocytes. This same Alu insertion is associated with increased expression of numerous PAX5 target genes, presumably by virtue of a transitive effect whereby increased PAX5 expression in turn increases the expression of downstream targets in its regulatory network.

To our knowledge, this is the first and only study of its kind in humans. However, analogous genome-scale approaches have been used to discover TE associations with gene expression in the model organisms Arabidopsis [57] and maize [58]. It is important to point out that the eQTL results summarized here are very much cell-type dependent. Expansion of this kind of eQTL analysis to multiple cell and tissue types is expected to reveal distinct TE-gene regulatory effects. The recently completed Genotype-Tissue Expression (GTEx; https://www.gtexportal.org/) project provides eQTL data for more than 50 cell/tissue types, providing a tremendous opportunity for further work of this kind.

**TE polymorphisms and complex common disease**

Two studies published in 2017 have taken a similar population-level view of the phenotypic effects of human TE polymorphisms [59,60]. For each of these studies, associations between TE insertion site genotypes and complex common diseases were explored. Both studies relied on the analysis of linkage disequilibrium (LD) patterns to discover TE polymorphisms linked to SNPs that were previously associated with health or disease related phenotypes via genome-wide association studies (GWAS). An implicit rationale for genome-scale surveys of this kind is the notion that TE insertions are expected to be more disruptive than SNP variations given the larger scale genomic changes that they entail. Interestingly, both studies report that TE polymorphisms are enriched at GWAS loci, highlighting their potential impact. The first study of this kind, from the group of Kathleen Burns, found 44 Alu insertions in tight LD with previously discovered GWAS trait associated SNPs [59]. The authors pointed out that this represents a >20-fold increase over the number of polymorphic Alu insertions that were previously known to be associated with human phenotypes, thereby underscoring the power of population genomic approaches for studies on the phenotypic impact of TE polymorphisms. Furthermore, the implicated Alu polymorphisms were found to be associated with a very broad range of health and disease related phenotypes.

Our own study on the impact of TE polymorphisms on complex common disease was designed to explore the connection between TE-mediated genome regulation and disease related phenotypic effects [60]. To achieve this aim, we used a progressive set of genome-wide bioinformatics screens that searched for polymorphic TE insertions that are: (1) found in LD with known GWAS SNPs, (2) located within tissue-specific enhancers, and (3) associated with tissue-specific gene expression levels. We further narrowed our search for candidate
TE polymorphisms to those associated with genes with blood or immune related functions, consistent with the fact that the gene expression data we analyzed is from B lymphocytes. This progressive and stringent genomic screen uncovered six TE polymorphisms that are likely to be associated with disease phenotypes by virtue of their gene regulatory effects. These included both Alu elements, as previously reported, as well as SVA elements. For example, we discovered an SVA insertion in the cell-type specific enhancer of the B4GALT1 gene (Figure 4). B4GALT1 acts to convert the Immunoglobulin G (IgG) antibody from a pro-inflammatory to an anti-inflammatory form. The SVA insertion is associated with both down-regulation of the B4GALT1 gene, thereby potentially leading to increased inflammation, and is linked to a genomic region implicated by GWAS in both inflammatory conditions (Crohn’s disease) and autoimmune disease (systemic lupus erythematosus).

Conclusions

One important caveat regarding the surveys of the effects of TE polymorphisms on human gene regulation and disease reviewed here relates to the fact that they rely on association studies. While this class of approaches has great potential to reveal connections between TE generated variation and health related phenotypes, association-based methods do not necessarily uncover causal variants (i.e., correlation ≠ causation). In this sense, the TE-phenotype associations uncovered by these studies are perhaps best considered as hypotheses, which will need to be further interrogated by experimental studies in order to provide deeper insight into causality and mechanism. Accordingly, one expectation is that these kinds of large-scale association studies can substantially narrow the experimental search space, with respect to possible TE-phenotype interactions, and thereby serve as a valuable point of departure for subsequent work.

The population genomics view of TE s exemplified by the recent studies reviewed here has the potential to expand our understanding of the phenotypic impact of human TEs. While ongoing human TE activity has widely considered to be deleterious, the presence of TE insertion variants that segregate as common polymorphisms among human populations indicates that many novel TE insertions must have escaped the action of purifying selection. Accordingly, polymorphic human TE insertion variants comprise an important source of naturally occurring genetic variation with subtle effects on genome regulation and human health. Functionally relevant TE polymorphisms of this kind are likely to provide crucial source material for ongoing human evolution.

Conflict of interest statement

Nothing declared.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest


A detailed account of endogenous retrovirus (ERV) sequences that have been exapted to provide interferon inducible enhancers to a diverse group of mammalian genomes. This study is distinguished by its use of the CRISPR-Cas9 system to experimentally confirm the regulatory role of candidate ERV enhancers.


A recent review that provides a comprehensive overview of the prospects for population genomic studies of human TEs. Both evolutionary and clinical aspects of human TE biology are explored.


A review paper that covers the algorithmic approaches used to discover novel TE insertion variants via the analysis of whole genome sequence data.


A comprehensive benchmarking and validation study of 21 programs designed to discover and characterize TE insertion variants from whole genome sequence data. The focus of the study is on the discovery of TE insertion variants in human genome sequences.


The bioinformatics tool MELT was developed by members of the 1000 Genomes Project structural variation group for the sequence discovery and characterization of polymorphic human TE insertions. The superior performance of MELT has been independently validated [56].


A comprehensive and in depth review that covers the very latest developments on the role of TE activity in cancer, which is one of the most promising areas of human TE research.


The first study of its kind wherein the regulatory potential of polyomorphous human TE families was explored using the expression quantitative trait loci (eQTL) analytical paradigm. Numerous associations between TE polymorphisms and gene expression were uncovered including...
population-specific gene regulatory effects as well as coordinated changes to gene regulatory networks.


A comprehensive catalog of human genome structural variants produced by the structural variation working group of the 1000 Genomes Project consortium. This catalog of variants includes a genome-wide collection of polymorphic TE insertions for Alu, L1, and SVA families and serves as an invaluable resource for population genomic studies of human TEs.


The first study linking TE insertion variants with disease risk alleles from genome-wide association studies (GWAS). The findings represent a >20-fold increase in the number of polymorphic Alu insertions associated with human traits, underscoring the power of the population genomic approach for the study of TE phenotypic impacts.


Another recent report linking polymorphic human TEs with disease phenotype loci. This study is distinguished by the revealed connections between the regulatory efforts of TE polymorphisms and the molecular mechanisms that underlie particular disease phenotypes.