MOBILE GENETIC ELEMENTS
2017, VOL. 7, NO. 1, e1280116 (20 pages)
http://dx.doi.org/10.1080/2159256X.2017.1280116

Taylor & Francis
Taylor & Francis Group

REVIEW

OPEN ACCESS

Population and clinical genetics of human transposable elements in the (post) genomic era

Lavanya Rishishwar[a,b,c], Lu Wang[a,b], Evan A. Clayton[a,d], Leonardo Mariño-Ramírez[b,e], John F. McDonald[a,d], and I. King Jordan[a,b,c]

[a]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA; [b]PanAmerican Bioinformatics Institute, Cali, Colombia; [c]Applied Bioinformatics Laboratory, Atlanta, GA, USA; [d]Ovarian Cancer Institute, Atlanta, GA, USA; [e]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

ABSTRACT

Recent technological developments—in genomics, bioinformatics and high-throughput experimental techniques—are providing opportunities to study ongoing human transposable element (TE) activity at an unprecedented level of detail. It is now possible to characterize genome-wide collections of TE insertion sites for multiple human individuals, within and between populations, and for a variety of tissue types. Comparison of TE insertion site profiles between individuals captures the germline activity of TEs and reveals insertion site variants that segregate as polymorphisms among human populations, whereas comparison among tissue types ascertains somatic TE activity that generates cellular heterogeneity. In this review, we provide an overview of these new technologies and explore their implications for population and clinical genetic studies of human TEs. We cover both recent published results on human TE insertion activity as well as the prospects for future TE studies related to human evolution and health.

ARTICLE HISTORY
Received 18 October 2016
Revised 3 January 2017
Accepted 4 January 2017

KEYWORDS
bioinformatics; disease; genetics; genomics; health; human; natural selection; polymorphisms; transposable elements; transposition

Human transposable element research in the (post) genomic era

Technology driven research and discovery on human transposable elements

A convergence of new technologies in three key areas –genomics, bioinformatics and high-throughput experimental techniques – is providing unprecedented opportunities for research and discovery on population and clinical genetic aspects of human transposable elements (TEs). In this review, we briefly cover these exciting technological developments and explore their implications for understanding how the activity of human TEs impacts the evolution and health of the global population. We would like to emphasize that our treatment is by no means intended as an exhaustive review of the subject, rather we are simply attempting to call the readers' attention to what we perceive to be some of the most relevant developments in this area along with the potential for future studies that these advances entail. It should also be noted that the review is focused primarily on the new bioinformatics tools that can be used to detect polymorphic TE insertions from next-generation sequence data, rather than the high-throughput experimental techniques, since we are most familiar with the computational approaches.

Developments in genomics technology, and next-generation sequencing in particular, have taken us from the analysis of a single human genome, which alone has provided profound insight into the biology of human TEs, to the population genomics era where whole genome sequences from thousands of human individuals can be compared. Concomitant developments of bioinformatics tools for genome sequence analysis have allowed for the discovery and characterization of the genetic variants that are generated via recent TE activity, *i.e.* human TE polymorphisms, via the comparative analysis of next-generation re-sequencing data from multiple human genomes. Finally, a suite of novel high-throughput experimental techniques, which also leverage next-generation sequencing data, have been developed and applied for the characterization of human

CONTACT I. King Jordan king.jordan@biology.gatech.edu School of Biology, Georgia Institute of Technology, 950 Atlantic Drive, Atlanta, GA 30332–0230, USA.

polymorphic TE insertions at the scale of whole genomes across numerous samples.

The initial analysis of the first draft of the human genome sequence was, in some sense, a watershed event for TE research. One of the most significant findings of this research was the large fraction of the human genome that was shown to be derived from TE sequences; 47% of the genome sequence was reported to be TE-derived with a single family of elements, LINE-1 (L1), making up ~17% of the genome and another family, Alu, contributing almost 11 million individual copies.[1] These remarkable results were generated using homology-based sequence analysis with the program RepeatMasker.[2] Subsequent analysis of the human genome sequence, using a more sensitive *ab initio* algorithmic approach, has revised the estimate upwards to more than two-thirds of genome being characterized as TE-derived.[3] The abundance of TE sequences found in the human genome almost surely did not come as a surprise to members of the TE research community, but this finding certainly did underscore the potentially far reaching impact of these often underappreciated genetic elements on the human condition.

The 1000 Genomes Project (1KGP) can be considered as the successor to the initial human genome project as well as the initiative that ushered human genomic research into the so-called post genomics era.[4-6] As its name implies, the 1KGP entailed the characterization of whole genome sequences from numerous human individuals, and it did so with an eye toward capturing a broad swath of world-wide human genome sequence diversity. The 1KGP resulted in the characterization of whole genome sequences for 2,504 individual donors sampled from 26 global populations, which can be organized into 5 major continental population groups. The project was executed in three phases, each of which included a substantial focus on technology development, not only with respect to sequencing methods but also for the computational techniques that are needed to call sequence variants from next-generation re-sequencing data. This focus on technology development ultimately led to the characterization of genome-wide collections of human polymorphic TE (polyTE) insertion genotypes for all individuals in the project.[7,8] Importantly, these data have been released into the public domain, thereby facilitating population and clinical genetic studies of human TE polymorphisms.

Advances in next-generation sequencing technology have also facilitated the development of high-throughput experimental techniques that can be used to detect *de novo* TE insertions, genome-wide across multiple samples. These high-throughput experimental techniques couple enrichment for sequences that are unique to active families of human TEs with subsequent next-generation sequencing and mapping techniques in order to discover the locations of novel TE insertions. Notably, these innovative experimental approaches have been successfully applied toward the characterization of somatic human TE activity in a variety of tissues, along with its potential role in cancer, as is discussed later in this review.

Active families of human TEs

As described above, a large fraction of the human genome sequence has been derived from millions of individual TE insertions. The process of TE insertion and accumulation in the genome has taken place over many millions of years along the evolutionary lineage that led to modern humans, and it turns out that the vast majority of human TE-derived sequences were generated via relatively ancient insertion events. Most ancient TE insertions have accumulated numerous mutations since the time that they inserted in the genome, and as a consequence they are no longer capable of transposition. The vast majority of TE-derived sequences in the human genome (>99%) correspond to such formerly mobile elements. The most salient aspect of these inert human TEs, with respect to population and clinical genomics, is that their insertion locations are fixed in the human genome. In other words, each individual TE sequence insertion of this kind is found at the exact same genomic location in all human individuals and for all human populations. Thus by definition, these ancient and fixed TE sequences do not contribute to human genetic variation via insertion polymorphisms.

There are, however, several families of TEs that are still active in the human genome. Elements of the HERV-K, L1, Alu and SVA families remain capable of transposition and can thereby generate insertion polymorphisms among individual human genomes. The resulting TE insertion polymorphisms have important implications for human evolution and health (disease) as detailed later in this review. HERV-K and L1 are autonomous TEs that encode all of the enzymatic machinery needed to catalyze their own transposition,

whereas Alu and SVA are non-autonomous elements that are transposed in *trans* by L1 encoded proteins.[9,10] All four active families of human TEs correspond to retrotransposons that transpose via the reverse transcription of an RNA intermediate.

Members of the HERV-K family of active human TEs are human endogenous retroviruses, which are thought to have evolved from ancient retroviral infections that made their way into the germline and eventually lost the capacity for inter-cellular infectivity via loss of coding capacity for the envelope protein. As such, HERV-K elements have genomic structures that are very similar to retroviruses, including long-terminal repeat (LTR) sequences that flank the *gag* and *pol* open reading frames, which encode structural and enzymatic (integrase and reverse transcriptase) element proteins. L1 elements are long interspersed nuclear elements (LINEs) that are classified as non-LTR containing retrotransposons. Alu and SVA elements are both classified as short interspersed nuclear elements (SINEs). Alu elements are derived from 7SL RNA and are ~300 bp in length.[11,12] SVAs are hybrid elements that are made up of SINE, VNTR (variable number tandem repeat) and Alu sequences and can vary from 100–1,500 bp in length.[13-15]

Genome-scale characterization of TE insertions

Human genome sequencing initiatives

The initial draft of the human genome sequence took more than 10 years to complete at a cost of ~2.7 billion dollars.[16] Characterization of the human genome sequence was done with Sanger sequencing technology, using essentially the same chain termination biochemistry that was invented in the mid-1970s,[17] albeit with refinements in automation. In the mid-2000s, staring with the Roche 454 pyrosequencing method, there was explosion of novel biochemical methods for DNA sequencing.[18] These so-called next-generation sequencing technologies enabled far higher throughput sequencing, at much lower cost, than the Sanger sequencing method used for the original human genome project. It is now possible to sequence an entire human genome in a single day at a cost of ~1,000 dollars using Illumina's patented sequencing by synthesis (SBS) technology. This hyper-exponential increase in sequencing capacity, and simultaneous decrease in its cost, is powering a series of human genome sequencing initiatives that have profound implications for the study of human TE genetic variation (Table 1).

The previously discussed 1KGP is the emblematic initiative for the characterization of whole human genome sequences at the population level; as such, it is difficult to overstate the impact that this project has had, and continues to have, on human population and clinical genomics. The 1KGP had the critical effect of stimulating experimental methods related to sequencing as well as numerous bioinformatics methods that are used for the analysis of genome sequence data, particularly as they relate to characterizing genetic variants. A major part of this effort was the development and refinement of methods for calling structural variants, including but not limited to TE insertion

Table 1. Large scale genome sequencing initiatives. Projects are sorted in descending order by the number of participants.

Project Name	PMID	# Participants	Description
Million Veteran Program (MVP)	26441289	1,000,000	Planned sequencing of 1 million US. Veterans (genotyping, whole genome and exome); current enrollment at 500k
SHGP	26583887	100,000	Catalog of whole genome sequences of 100k Saudis
TOPMed	N/A	62,000	Sequencing of 62k individual genomes along with a variety of data for precision medicine initiative
UK10K	26367797	10,000	Sequencing of ~10k individuals from UK to inspect the effect of rare and low-frequency variants to human traits
Human Longevity	Awaiting	10,000	Deep sequencing of 10k human genomes; Data donated to Precision FDA
Iceland Genome Project	25807286	2,636	Catalog of whole genome sequences of 2,636 Icelanders
1000 Genomes Project	26432245	2,504	International whole genome project that sampled 2,504 healthy individuals from 26 populations
EGDP	27654910	483	Catalog of whole genome sequences of 483 genomes from 148 diverse population
SGDP	27654912	300	Catalog of whole genome sequences of 300 genomes from 142 diverse population
GoNL	24974849	250	Catalog of whole genome sequences of 250 Dutch parent-offspring families
Australian Aboriginals	27654914	108	Catalog of whole genome sequences of 108 Aboriginal Australians

polymorphisms. Nevertheless, the 1KGP, which entailed the characterization of just over 2,000 whole genome sequences, has been dwarfed in scale by a number of subsequent initiatives that are currently underway (Table 1).

Several of the most ambitious human genome sequencing initiatives involve the characterization of cancer genome sequences. For example, the International Cancer Genome Consortium (ICGC) is collaborating with the US National Cancer Institute's The Cancer Genome Atlas (TCGA) to sequence genomes for 500 pairs of matched normal and tumor samples for 500 different tumor types, for an expected yield of 50,000 whole genome sequences.[19,20] The US National Heart, Lung, and Blood Institute's (NHLBI) TOPMed precision medicine initiative is another health-related project that aims to sequence the genomes of 62,000 individuals.[21] There are a number of other large-scale human genome sequencing initiatives that are aimed at the populations of specific countries or global sets of populations. For example, the Wellcome Trust is sponsoring the UK10K initiative to sequence the genomes of 10,000 citizens of the United Kingdom,[22] and Saudi Arabia intends to sequence 100,000 Saudi individuals for their own project.[23] The Simons Genome Diversity Project recently completed sequencing of 300 human genomes from 142 diverse populations,[24] and the Estonian Genome Diversity Project sequenced 483 genomes from 148 populations.[25] Together, these projects, along with others like them, will provide a wealth of raw sequence data that can be mined for TE insertion polymorphisms using the computational and experimental approaches described in the sections that follow.

High-throughput techniques for TE insertion detection

Bioinformatics approaches

The characterization of single nucleotide variants (SNVs) from computational analysis of next-generation re-sequencing data has proven to be relatively straightforward: sequence reads are mapped to a reference genome sequence, allowing for mismatches, and sites where the mapped reads differ in sequence from the reference are used to call variants.[26,27] The characterization of structural variants from next-generation sequence data has proven to be a far more challenging, but by no means intractable, problem.[28] Early methods for calling structural variants operated in manner that was agnostic with respect to the particular class of variant that was being characterized, whereas subsequent efforts have resulted in refined methods that are specifically tailored to individual structural variant classes.[29,30] The most widely used and reliable methods for the computational detection of human TE insertion polymorphisms fall into the latter class of more specific methods.[31] We want to be clear that these novel computational methods that we are describing are aimed at the detection of TE insertion polymorphisms, which will differ from the reference genome sequence, rather than the more mature bioinformatics methods (*e.g.* RepeatMasker) that are used to characterize the identity of the more ancient, fixed TE sequences that are included as part of a reference genome sequence.

There exist numerous computational tools that allow for the detection of TE insertions from next-generation whole genome sequence data (Table 2). While these programs may differ substantially in their details, they all tend to rely on the same two fundamental principles: discordant read pair mapping and split (or clipped) reads[32] (Fig. 1A). Discordant read pair mapping occurs when one member of a read pair maps uniquely to the reference genome sequence and the second member of the pair maps to a repetitive TE sequence that is not found in the adjacent genomic region in the reference sequence. In some cases, the second member of the pair may map partially to unique reference genome sequence and partially to the TE sequence. The presence of multiple read pairs that show this pattern, from within the same genomic interval, is taken as evidence of a TE insertion, with the specific identity of the inserted element determined by the mapping of the second member of the read pair. Typically, a TE reference library is provided to facilitate these mappings and the corresponding characterizations of insertion identities. The discordant read pair mapping technique is ideal for short read, pair end sequencing technology, such as the Illumina SBS method. The somewhat less commonly used, at least at this time, split read technology for computational detection of polymorphic TE insertions relies on longer sequence reads that map partially to unique reference genome sequence and partially to a repetitive TE sequence. This can include reads with one end in unique genome sequence and the other end in a TE sequence or reads that span an entire TE

Table 2. Computational approaches for genome-wide detection of TE insertions. Methods are sorted in order by their year of publication.

Tool Name	PMID	Year	Comments
VariationHunter	19447966	2009	Originally developed for SV detection, later refined for TE calling
HYDRA-SV	20308636	2010	General purpose SV tool; reported on mouse genome
TE-Locate	24832231	2012	Reported on 1001 Arabidopsis genomes project
Tea	22745252	2012	Specialized TE caller for cancer WGS data
ngs_te_mapper	22347367	2012	Requires TSDs; reported for Drosophila melanogaster
RetroSeq	23233656	2013	Tested on 1KGP and mouse strains
ReloaTE	23576519	2013	Requires TSDs; designed for rice genomes
Mobster	25348035	2014	Tested on 1KGP; reliable predictor for Human genome
Tangram	25228379	2014	Used in Phase II of 1KGP; no longer maintained
TEMP	24753423	2014	Reported on 1KGP and Drosophila genomes
T-lex2	25510498	2014	Reported on 1KGP and Drosophila genomes
TE-Tracker	25408240	2014	Reported on Arabidopsis genome and simulated human genome
TIGRA	24307552	2014	A breakpoint assembler and not a structural variant caller
TranspoSeq	24823667	2014	Specialized TE caller for cancer WGS data
TraFiC	25082706	2014	Specialized TE caller for cancer WGS data
MELT	26432246	2015	Used in Phase III of 1KGP; reported to work on Human, Chimp and dog.
ITIS	25887332	2015	Reported on *Medicago truncatula*; not optimized for Human genome
Jitterbug	26459856	2015	Reported on 1KGP and Arabidopsis genome
MetaSV	25861968	2015	General purpose SV tool; reported on simulated genome
DD_DETECTION	26508759	2016	Database free dispersed duplication detection approach
GRIPper	—	—	Detects non-reference gene copy insertion

insertion (*i.e.*, have a TE sequence in the middle of the read). As longer sequence read technologies – such as the Pacific Biosciences single molecule real time sequencing method (PacBio SMRT) – become more widely used for human genome sequencing, the split read approach should become increasingly useful.

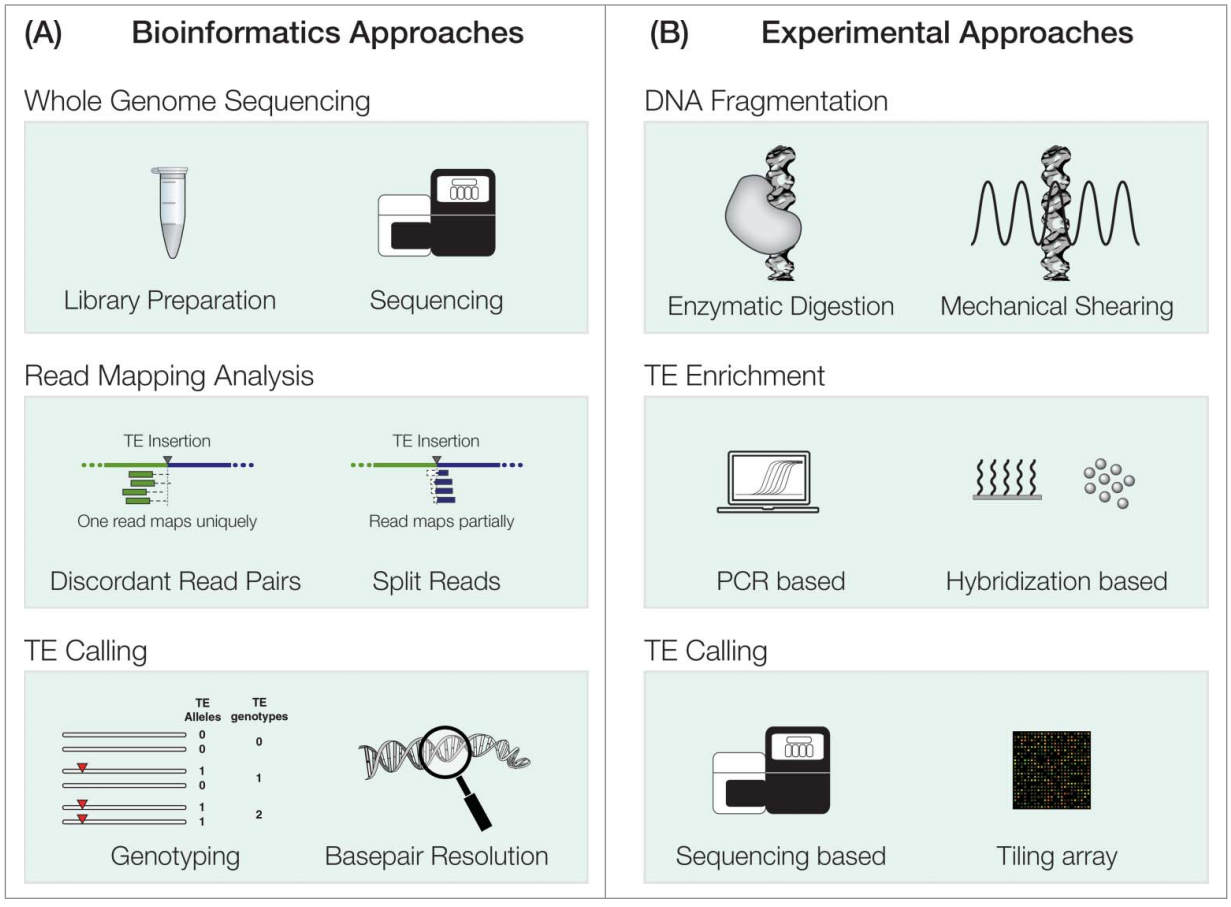

Figure 1. Schematic of the high-throughput bioinformatics (A) and experimental (B) approaches to human TE insertion discovery.

Alternatively, long reads may eventually come to be used for *ab initio* assembly of complex eukaryotic genomes, such as the human genome, thereby obviating the need for computational TE insertion detection methods altogether.

Two of the earliest computational methods developed specifically for the detection of TE insertions from next-generation sequence data are VariationHunter[33] and the program Spanner,[7] which was used for calling TE insertions in the first phase of the 1KGP. Subsequent phases of the 1KGP included additional refinement of next-generation sequence based TE insertion calling methods resulting in the Tangram[34] and MELT[8] programs, for the second and third phases of the project, respectively. RetroSeq[35] and Mobster[36] are two of the other most widely used programs for sequence based TE insertion detection. RetroSeq was implemented primarily for the detection of endogenous retrovirus insertions in the mouse genome, whereas Mobster was tested mainly on human L1 and Alu elements.

Until very recently, all of these individual programs had only been benchmarked and validated individually by the same groups that developed each one. In other words, there was no independent and controlled comparison of the accuracy, runtime performance and usability of these tools. We recently performed just such a benchmarking and validation comparison of 21 different programs for sequence based TE insertion detection in an effort to provide researchers with an unbiased assessment of their utility.[31] Our benchmarking study was focused solely on human TE detection, owing both to the importance of human TE detection for population and clinical genetic studies as well as the availability of an experimentally validated set of TE insertions for an entire human genome.

The first phase of our benchmarking study entailed an effort to select tools that would be of the most potential use to the human TE research community. This included eliminating all programs from consideration that: 1) performed general structural variant detection (these typically have worse performance of TE insertion detection), 2) were specifically designed for cancer and required matched normal-cancer genome pairs, or 3) perform breakpoint assembly for TE insertion identification and are not able to detect insertion site locations without prior information. In this phase, we also eliminated all programs that were no longer supported and/or could not be used due to non-user generated errors such as previously reported bugs. This process resulted in reducing the original set of 21 programs down to 7 programs, which we then evaluated using simulated and actual human genome sequences.

The final 7 programs that we evaluated were: MELT,[8] Mobster,[36] RetroSeq,[35] Tangram,[34] TEMP,[37] ITIS[38] and T-lex2[39] (Table 2). For each of these programs, we provided a detailed set of notes in support of their installation and use, including the exact commands and parameters that are required for their optimal performance. We compared all of the programs with respect to a set of qualitative and quantitative benchmarks. The qualitative benchmarks were ease of installation, ease of use, level of detail in the user manual and source code availability (*i.e.,* open or closed source). The quantitative benchmarks were precision and recall accuracy measures along with the runtime parameters: CPUtime, walltime, RAM and the number of CPUs used. The simulated data that we used consisted of artificial genome sequences with randomly generated TE insertions and sequence read pairs simulated based on the Illumina sequencing profile. The empirical data was taken from a single individual from the 1KGP whose genome was extensively characterized, including with PacBio long read sequence technology, resulting in an experimentally validated set of 893 TE insertions genome-wide.

When all of these factors were taken into consideration, the program MELT showed the best overall performance followed by the programs Mobster and RetroSeq. The superior performance of MELT on these particular data should be taken with some caution given the fact that it was developed and refined on the exact same human data set. Indeed, the programs that were designed to perform more broadly, such as TEMP, or for different species, such as ITIS and T-lex2, did not perform as well, consistent with the possibility that they were at an inherent disadvantage when benchmarked on human genome sequence data from the 1KGP. Nevertheless, our benchmarking analysis clearly supports the use of MELT, and to a lesser extent Mobster and RetroSeq, for the computational detection of human TE insertions from next-generation sequence data.

There remain a number of caveats and open issues that should be considered when using these kinds of programs to predict TE insertions from whole genome sequence data. The first thing to consider is that no

single method can produce optimal overall results. The best strategy is to use two or more of the top 3 methods – MELT, Mobster and RetroSeq – and then to combine the methods by looking for consensus TE calls that are supported by multiple methods. This approach has the potential effect of increasing precision at only a minor cost to recall, *i.e.,* it is simultaneously conservative but can also increase the number of total TE calls by using multiple methods. Of course, a combined approach of this kind can be quite user intensive and could exceed the ability of some labs to readily implement. Perhaps the most pressing open issue regarding computational methods for TE insertion detection relates to the level of resolution at which the insertion sites can be located in the genome sequence. In our experience, TE insertions can only be accurately localized within approximately ± 100 bp. This lack of resolution makes it particularly difficult to combine results from multiple methods, as suggested above, since the same predictions will most often not be located at exactly the same genomic location. This limitation can be overcome by considering TE insertions detected within ± 100 bp windows to represent the same calls. Nevertheless, further algorithm development aimed at more precise TE insertion location should prove to be an important future development in the field.

High-throughput experimental approaches

In addition to driving the bioinformatics based efforts at TE insertion detection, next-generation sequencing techniques have also enabled a number of high-throughput experimental approaches for the detection of novel TE insertions (Table 3). Much like the computational approaches for TE detection, these high-throughput experimental techniques also share a core set of design principles[40] (Fig. 1B). The first phase of these experiments consists of fragmentation of genomic DNA followed by enrichment for sequence elements that uniquely correspond to active human TE subfamilies, mainly Alu and L1. Different methods are distinguished by the approaches that they take to genomic fragmentation as well as whether they use PCR or hybridization for the enrichment step. Enrichment of sequence fragments from active TE subfamilies is followed by next-generation sequencing, for the most recently developed methods, or hybridization to tiling arrays for some of the older methods.

Table 3. High-throughput experimental approaches for TE insertion detection. Next-generation sequence based methods are presented separately from methods that used tiling arrays or Sanger sequencing. Methods are sorted in descending order by their year of publication.

Next-generation sequence based			Tiling arrays/Sanger based		
Method	PMID	Year	Method	PMID	Year
L1-Seq	20488934	2010	TIP-Chip	20602999	2010
Transposon-Seq	20603005	2010	Fosmid-based	20602998	2010
ME-Scan	20591181	2010	AIP	22495107	2012
RC-Seq	22037309	2011			

The first attempt at the systematic and unbiased characterization of novel human TE insertions was based on tiling array technology and was relatively low throughput.[41] A number of next-generation sequence based techniques for TE insertion, which allowed for a substantial increase in the numbers of TE insertions that could be detected, were independently developed right around the same time in 2010 and 2011. Three such methods were published in 2010: ME-Scan,[42] L1-Seq[43] and Transposon-Seq.[44] ME-Scan was used to characterize polymorphic Alu insertions, L1-Seq was applied to L1 insertions, and Transposon-Seq was used for TE insertion discovery with both families of elements. A fourth early sequence based method for TE insertion detection employed a lower-throughput approach that utilized fosmid sequences, characterized via Sanger sequencing, to characterize L1 insertions.[45] The RC-Seq method was developed in 2011 and is the only method of its kind to be applied to all three families of active human TEs: Alu, L1 and SVA. RC-Seq combines tiling array based hybridization with next-generation sequencing for TE insertion discovery.

Evolutionary genetics of active human TEs

The high-throughput approaches to TE insertion detection described in the previous section, particularly the computational genome sequence based methods, have the potential to yield genome-wide catalogs of human TE insertion polymorphisms across numerous individuals from multiple populations. The realization of this possibility is exemplified by the 1KGP, phase 3 of which includes the public release of 16,192 TE insertion genotype calls for 2,504 individuals from 26 global populations. data sets of this kind have the potential to yield unprecedented insight into the

nature of the evolutionary forces that act on TE polymorphisms.

Human genetic variation from TE activity

The first step in any genome-scale evolutionary analysis of human TE insertional polymorphisms involves a basic description of the nature of the genetic variation that is generated by TE activity. This includes descriptive statistics regarding the levels of TE insertion variation within and between populations along with a sense of how polymorphic TEs are distributed across the genome, particularly with respect to the location of functionally relevant genomic features such as genes and regulatory elements.

Levels and patterns of TE genetic variation

TE insertion detection programs yield presence absence genotype calls for individual loci – homozygous absent (0), heterozygous (1) and homozygous present (2) – across the entire genome, when applied to whole genome next-generation sequence data. For large scale human genome sequence initiatives, such as the 1KGP, this yields the kind of data that can be used to calculate polyTE insertion allele frequencies within and between populations. PolyTE allele frequencies (p_{TE}) can be calculated from site-specific genotype data as the total number of TE insertions observed at any given genomic site (TE_i) normalized by the total number of chromosomes in the population under consideration ($2n$): $TE_i/2n$. This can be done for individual populations or for groups of related populations, such as the 5 major continental population groups characterized as part of the 1KGP.

Population level polyTE allele frequencies can in turn be used in turn to calculate a variety of population genetic parameters that measure how genetic variation is apportioned among populations, such as heterozygosity (H) and related fixation index (F_{ST}) statistics.

$$H = 1 - \left(p_{TE}^2 + (1 - p_{TE})^2\right) \quad (1)$$

$$F_{ST} = \frac{(H_T - H_S)}{H_T} \quad (2)$$

where H_S is the sample (within population) polyTE heterozygosity and H_T is the total (between population) polyTE heterozygosity. These kinds of statistics are ideal for measuring the effects of natural selection on TE insertion polymorphisms as described later in this review.

Genomic landscape of TE insertions

Genome-wide catalogs of polyTE genotypes can also be used to systematically evaluate the landscape TE insertions and to compare their locations to the locations of functionally important genomic features such as genes, regulatory elements and epigenetic chromatin marks. The overall human TE genomic landscape is already very well defined, dating to the initial analysis of the draft human genome sequence[1] and even earlier,[46-49] but the extent to which polyTE distributions resemble those of the more ancient, fixed TEs that predominate in the human genome remains an open question. When all TE-derived sequences are considered, there a number of anomalous genomic regions that are particularly enriched or depleted for human TE sequences, and these are thought to be related to gene density and tight regulatory requirements. The rate of recombination is also an important factor influencing the genomic distribution of human TEs. It is thought that TE density should be reduced in regions of high recombination owing to the deleterious effects of ectopic recombination among dispersed element insertions.[50] However, the situation is more complicated in the human genome where marked differences in TE genomic distributions can be seen for different families of TEs and even for different age classes within the same TE family.[1,51,52] Across the entire genome, LINE elements (L1) tend to be enriched in AT-rich DNA and are primarily found in low recombining intergenic regions, whereas SINE elements (Alu) are enriched in GC-rich DNA regions in and around gene sequences. These TE distribution patterns correspond very well to previously defined isochores,[53] which are large regions of DNA with uniform GC-content patterns.[54]

One particularly interesting finding from the initial analysis of the human genome sequence was that the distribution patterns of Alus change drastically for different age classes. Older subfamilies of Alus, *i.e.,* those that inserted in the genome long ago, show the most skewed genomic distributions and the highest enrichment in GC-rich DNA. As the Alu subfamilies under consideration become progressively younger, they are progressively less enriched in GC-rich DNA; in fact, the very youngest AluY subfamily shows a preference

for AT-rich DNA. These results were taken to indicate that Alus are preferentially retained in GC-rich DNA, and conversely more frequently lost from AT-rich DNA, since Alus are known to insert into the AT-rich target sequences favored by L1 encoded endonucleases. This was initially thought to be due to some positive selective force acting on Alus in GC-rich DNA,[1,55] but was later shown to be more likely related to the relative ease with which Alu deletions were tolerated in gene poor AT-rich regions, compared to gene rich GC-regions where Alu deletions via ectopic recombination between nearby insertions would be far more deleterious.[52,56-60] This issue has received substantial attention in the ensuing years and remains controversial. Now that there is a complete catalog of very recent Alu insertions, it will be very interest to see if this same patterns holds up.

Polymorphic TE insertions as ancestry informative markers

Ancestry informative markers (AIMs) are genetic variants that distinguish evolutionary lineages, different species or distinct populations within the same species, and can thereby be used to reconstruct evolutionary histories.[61,62] For a number of reasons, TE insertions have proven to be extremely useful as AIMs, both within and between species.[63] Most critically, locus-specific TE insertions nearly always represent synapomorphies, *i.e.,* shared derived character states that are free from homoplasies where identical states do not result from shared ancestry.[57,64,65] TE insertions also have the advantage that the ancestral state can be assumed to be absence of the insertion, and TE insertions are ideal AIMs for the very practical reason that they can be rapidly and accurately typed via PCR based assays.

A number of studies from the pre-genomic era used polyTE insertions to study human evolution and ancestry.[66-73] Most of these studies have focused on Alu elements, owing both to their relative abundance and the ease with which their shorter sequences can be PCR amplified. Far fewer studies have used L1s as AIMs, and to our knowledge, SVAs have yet to be used as markers in human evolutionary studies. Our own lab recently published the first evolutionary analysis of human populations using the genome-wide collection of human polyTE insertions characterized as part of the 1KGP.[74] These data confirmed that human polyTE insertions are substantially geographically differentiated with many population-specific insertions. Furthermore, the patterns of polyTE insertion divergence within and between populations recapitulate known patterns of human evolution. African populations show both the highest numbers of polyTE insertions and the highest levels of polyTE sequence diversity, consistent with their ancestral status. Evolutionary relationships among human populations computed from the analysis of polyTE genotypes were entirely consistent with those that have been derived from single nucleotide polymorphisms (SNPs). In addition, when select subsets of population differentiated polyTEs are used as AIMs, they were able to accurately predict patterns of human ancestry and admixture.

It is becoming increasingly apparent that patterns of human genetic ancestry and admixture are relevant to the study of human health and disease. In particular, there are numerous health disparities between human populations, and many of these are likely to be genetically based.[75,76] Thus, the utility of polyTEs as AIMs could prove to be of clinical relevance, in applications such as admixture mapping for instance,[77,78] in addition to their applications to population genetic studies.

Effects of natural selection on polymorphic TE insertions

The ability to calculate polyTE allele frequencies genome-wide, as detailed in the previous section of this review, should prove to be critical for measuring the effects of natural selection on TE insertions. One aspect of natural selection on polyTE insertions is already abundantly clear: the role that negative (purifying) selection plays in eliminating deleterious insertions from the population. The fact that TE insertions are deleterious is underscored by the numerous studies that have linked TE insertions to human disease.[79-83] We describe a number of such clinically related human TE studies in subsequent sections of this review. The deleterious nature of mutations generated by TE activity is not at all surprising when you consider that TE insertions can be hundreds to thousands of base pairs long. Such large-scale mutations are clearly far more substantial mutational changes than the more commonly considered SNPs. In addition, the simple fact that TE mutations are insertions of

DNA sequence, rather than duplications or other rearrangements, also attests to their potentially disruptive nature.[74]

Our own previous genome-wide study of polyTE insertions turned up several lines of evidence consistent with the action of negative selection on human TEs. First of all, human polyTE insertions tend to be found at very low allele frequencies within and between human populations. Indeed, the allele frequency spectrum of polyTE insertions is highly skewed toward the lower end, and even more so than seen for SNPs, consistent with purifying selection. In addition, polyTE insertions were found to be severely under-represented in functional genomic regions including genes and exons.

Despite the documented deleterious effects of TE insertions, the majority of TE insertion events are likely to be neutral or nearly so. This can be attributed in part to humans' relatively low effective population size, which renders natural selection less able to eliminate individual TE insertions that have only moderate fitness effects.[84,85] Indeed, most recently integrated Alu elements have been shown to evolve as neutral alleles,[86] and relatively short L1 insertions, which result from 5′ truncations that occur frequently during L1 retrotransposition,[87,88] have also been shown to evolve neutrally in the human genome.[89-91] The findings on the neutrality of short L1 elements underscore the importance of distinguishing among different types of TE insertions, short versus long insertions in particular, when considering the potential effects of selection on TE polymorphisms.

As alluded to previously, the results demonstrating the action of negative selection on human polyTE insertions are not surprising considering the disruptive nature of genic TE insertions and their known link to diseases. In addition to the deleterious effects of TE insertions on gene function, TEs can also affect fitness post-insertionally by mediating ectopic recombination and by causing cellular toxicity via DNA damage.[92] Selection against longer TE insertions, along with their relative paucity in high recombining regions, is consistent with deleterious effects of ectopic recombination among dispersed TE copies.[93] It has been suggested that TEs represent such a potent mutational threat that host genomes were forced to evolve global regulatory mechanisms to repress their activity. For example, a number of epigenetic regulatory systems may have originally evolved to defend against TE activity and were only subsequently coopted to serve as host gene regulators.[94,95] Nevertheless, for us it is also particularly interesting to speculate as to a possible role for positive (adaptive) selection in sweeping polyTE insertions to (relatively) high frequencies along specific population lineages. If positive selection on polyTE insertions was to be detected, it would suggest that such sequences can somehow encode functional utility for the human genome.

The possibility that TE sequences can provide functional utility for their host genomes is well supported by numerous studies on the phenomenon of exaptation,[96,97] or molecular domestication,[98] whereby formerly selfish TE sequences come to encode essential cellular functions. This has been seen most often in the context of regulatory sequences.[99] Human TE sequences have been shown to provide a wide variety of gene regulatory sequences including promoters,[100-102] enhancers,[103-107] transcription terminators[108] and several classes of small RNAs.[109-111] Human TE sequences can also affect host gene regulation via changes in the local chromatin environment.[1,112-116] However, all of the human TE-derived regulatory sequences studied to date correspond to relatively ancient TE insertions that are no longer capable of transposition and are consequently fixed with respect to their genomic locations. Accordingly, it is not known whether exaptation of TE sequences can occur on the far shorter time scale that would be needed in order for polyTE insertions to show evidence of evolving by positive selection.

At this time, there are some tentative lines of evidence that are consistent with a role for positive selection in shaping the evolution of human polyTE insertions. Closer inspection of the polyTE insertion allele frequency spectrum mentioned above revealed a shift at the higher end of the spectrum, suggesting that some TE insertions may have increased in frequency owing to the effects of positive selection. This pattern was seen for Asian and European populations but not for African populations. Thus, it is possible that this shift could reflect genetic drift, and accordingly less efficacious selection, in human populations that have historically lower effective population sizes. Additional work is needed to distinguish between these two possibilities. There is also data from a more narrowly focused study on polymorphic L1 insertions showing patterns of linkage disequilibrium and

extended haplotypes that are consistent with positive selection on human polyTE insertions.[117]

More detailed studies on human TE genetic variation will be needed to fully assess the role that positive selection has played in the evolution of polyTEs. The flood of whole genome sequence data coming from human genome initiatives around the world, coupled with the maturing computational techniques for characterizing polyTE insertions from those data, should provide ample opportunities for studies of this kind. In addition, the analytical framework for detecting positive selection at the genomic level is already well established[118-120] and should be readily portable to genome-wide studies of TE genetic variation.

Clinical genetics of polymorphic TE insertions

TE insertions in Mendelian disease

Human TE insertions are relatively large scale mutations that are considered to be both rare and deleterious, particularly if they occur in genes or other functionally important genomic elements. In other words, TE insertions often correspond to highly penetrant mutations, and accordingly they have been linked to many Mendelian diseases.[121,122] Indeed, the ability of L1 sequences to transpose was first confirmed by a study showing that a novel L1 insertion into the *F8* (Coagulation Factor VIII) gene causes hemophilia A.[123] Subsequent studies have implicated Alu insertions in a number of Mendelian diseases, including hemophilia B,[124] cystic fibrosis[125] and Apert syndrome.[126] An SVA insertion in the *BTK* (*Bruton Tyrosine Kinase*) gene causes X-linked agammaglobulinaemia.[127]

Despite their known disease causing properties, TE insertion mutations are often not considered in screens for disease causing variants. For example, widely used exome based methods for disease variant discovery will necessarily overlook the contribution of TE insertions to human disease. Computational and experimental approaches to TE insertion discovery provide a number of potential advantages with respect to the discovery of TE mutations that can cause Mendelian diseases. As we have discussed previously, these kinds of approaches allow for the systematic and unbiased characterization of deleterious TE insertions genome-wide, a critical dimension of genomic approaches to the diagnosis of disease. In addition, characterization of the genomic landscape of TE insertions for large scale population based genome initiatives (Table 1) will provide an important reference panel of TE mutations that are found in healthy individuals for the purpose of screening for rare potential disease causing variants.

TE activity and cancer

There are a number of lines of evidence that indicate a relationship between the activity of human TEs and the etiology of cancer, particularly for the active subfamily of L1 elements. The initial studies that uncovered a potential connection between TEs and cancer focused on expression, both transcript and protein, of L1 elements in tumor tissue samples. While it was previously thought that L1 expression was largely repressed in somatic tissue, it has been shown that numerous L1 elements are also expressed in a wide variety of tumor types including testicular cancer,[128] germ cell tumors[129] and breast cancer.[130,131] More recently, nearly half of all human cancers were found to be exclusively immunoreactive to L1 *ORF1* encoded proteins compared to matched normal tissue samples, suggesting that the ORF1 proteins could serve as cancer diagnostic biomarkers.[132]

In addition to the aforementioned L1 expression analysis, numerous studies have employed next-generation sequence analysis based techniques, followed by validation with PCR and Sanger sequencing, in order to characterize the TE insertion landscape of human cancers. Tumor genome sequences from a wide variety of cancer types have been found to be enriched for L1 insertions; these include colorectal tumors,[133] esophageal carcinoma,[134135] and gastrointestinal tumors.[136] In one particularly broad survey, 53% of 244 cancer genomes were found to have L1 insertions, many of which included 3′ transduced sequences that are introduced as copying errors from run-on transcripts during the reverse transcription process.[137] As was the case for TE cancer expression research, these surveys of TE insertion in cancer genomes were suggestive and interesting but did not necessarily establish a causal relationship for TE activity in the etiology of cancer (*i.e.*, tumorigenesis).

A smaller number of studies have shown even more direct evidence that specific TE insertions play a causal role in the etiology of cancer. The application of the RC-Seq technique[138] to 19 hepatocellular carcinoma genome sequences uncovered two different L1

insertions, each of which initiated tumorigenesis via a different oncogenic pathway.[139] Independent L1 insertions were found in the *MCC* (*Mutated in Colorectal Cancers*) and *ST18* (*Suppression of Tumorigenicity*) tumor suppressor genes in this study. Perhaps the strongest evidence for an L1 insertion that is an actual driver mutation for tumorigenesis was recently reported for colorectal cancer.[140] Investigators in this study found a somatic L1 insertion in one allele of the *APC* (Adenomatous Polyposis Coli) tumor suppressor gene, and they showed that this L1 insertion coupled with a point mutation in the second allele of the same gene to initiate tumorigenesis via the so-called two hit colorectal cancer pathway.

TE activity and aging

A number of recent studies have also suggested a link between human TE activity and aging.[141] The connection between TE activity and aging is tangentially supported by TEs' involvement in a number of age related diseases including cancer,[142] as described in the previous section, and intriguingly the rate of L1 transposition in cancer does appear to increase with age.[133] More direct evidence for a link between TE activity and aging comes from several studies that have uncovered evidence that the epigenetic silencing of human TEs declines with age. For example, the methylation levels of Alu were shown to decline with age[143] as was heterochromatin based silencing of L1 elements.[144] In senescent fibroblast cells, the chromatin environment for relatively young members of the Alu, L1 and SVA families becomes progressively more open leading to an increase in both their levels of transcription and rates of transposition.[145] Finally, transcriptional derepression of Alu elements has been implicated in aging via an indirect route linked to nuclear cytotoxicity in senescent stem cells that is caused by DNA damage.[146] Upregulated Alu transcription in senescent stem cells inhibits their ability to efficiently repair DNA damage, and suppression of Alu transcription was shown to reverse this effect. Research on the connection between TE activity and aging is a particularly new and promising area of investigation.

Polymorphic TE insertion associations with common diseases

The association of TE insertions with both Mendelian disease and cancer, discussed in previous sections, rests on the assumptions that TE mutations are rare, deleterious and penetrant. However, recent results from analysis of the 1KGP sequences indicate that numerous TE insertions can be found in the genomes of healthy individuals.[8] Population genetic analysis of these data shown that TE polymorphisms segregate within and between human populations and can, albeit relatively rarely, increase to high allele frequencies.[74] In other words, TE polymorphisms can in some cases come to represent common genetic variants. Common genetic variants of this kind, also referred to common mutations, have been widely used over the last decade or so in association studies that aim to characterize the genetic architecture of common human diseases or conditions. Genomic characterization of TE insertion genotypes, for hundreds of thousands of individuals among various human populations, can provide an ideal source of data for genome wide association studies (GWAS), which to date have almost exclusively been conducted using SNPs.

GWAS require hundreds or thousands of cases and controls in order to have sufficient statistical power to detect associations between common genetic variants and disease. Despite the drastic decreases in the cost of whole genome sequencing over the last several years, it is still not practical to use this approach for most GWAS. Accordingly, these studies rely on the use of array technology to characterize variant alleles for hundreds of thousands of known SNPs genome-wide. This approach yields disease associations with SNP alleles that do not necessarily represent causal mutations. In other words, an associated SNP may simply tag a genomic region that contains a nearby disease causing variant that is in linkage disequilibrium (LD) with the associated SNP.

The existence of LD structure provides an important opportunity for the association of TE insertion polymorphisms with common diseases. As more and more whole genome sequences accumulate from the various genome sequencing initiatives around the world, the genomic landscape of TE insertions should become increasingly well characterized, assuming computational methods for TE insertion detection are accurately applied to these data. The accumulation of thousands of whole genome sequences, from diverse human populations, that include genome-wide catalogs of TE insertion genotypes provides the opportunity for imputation of TE insertion genotypes via

comparison with SNP array data. In this way, TE insertion polymorphisms could be associated with disease via thousands of existing GWAS studies along with untold numbers of future GWAS. The potential of this approach to TE GWAS is supported by a recent genome-wide survey of human L1 insertions that found abundant evidence of LD between these TE polymorphisms and nearby SNPs.[117]

TE insertion associations with quantitative traits

The same logic that applies to the association of TE insertion polymorphisms with common diseases via GWAS can be used to associate TE polymorphisms with a number of different quantitative traits. These include anthropometric phenotypes, measures of human performance and a wide variety of so-called endophenotypes, which are considered as intermediate physiological traits that underlie higher order, observable phenotypes.[147] Gene expression levels are perhaps the most widely studied class of endophenotype. Expression quantitative trait loci (eQTL) analysis correlates levels of gene expression with genetic variant genotypes in order to characterize the influence of genetic variants on gene regulation. As is the case with GWAS, the vast majority of eQTL studies compare SNP genotypes with gene expression levels. However, more recent studies have begun to analyze different classes of genetic variants using the eQTL framework. For example, copy number genotypes for short tandem repeat sequences at >2,000 loci were recently shown to be associated with the expression of numerous human genes using an eQTL approach.[148] In addition, the Structural Variation Group[8] of the 1KGP used the eQTL approach to quantify the influence of structural variants on human gene expression using RNA-seq data characterized for 1KGP samples from European and African populations by the GUEVEDIS project.[149]

Many of the large scale genome initiatives listed in Table 1 will include abundant donor meta-data along with their genome sequences. For example, the NHLBI TOPMed precision medicine initiative will collect molecular, behavioral, imaging, environmental, and patient clinical data along with a variety of omics data sources, including DNA methylation, metabolite and RNA expression profiles. These kinds of quantitative data can all be compared to the genetic variants that will be characterized by whole genome sequencing, including TE insertion polymorphisms, in order to characterize the genetic architecture of a variety of quantitative human traits.

Imputation of TE genotypes onto SNP array data, as described previously in the context of GWAS, could also provide abundant opportunities to characterize TE-eQTLs in particular. The GTEx eQTL project, for instance, has compared genome-wide SNP genotypes from hundreds of individuals to their RNA-seq gene expression data for 53 human tissue types.[150] Imputation of TE insertion genotypes onto the SNP arrays used for this study could lead the discovery of TE influences on human gene expression related to a wide variety of phenotypes.

Conclusions and prospects

Human TE research has been profoundly influenced by the ongoing revolution in genomic technology. There are a number of new computational and experimental approaches that allow for the genome-wide characterization of TE insertions across numerous samples. These kinds of techniques are continually being refined and improved, and this process often goes hand-in-hand with large scale genome sequencing initiatives, such as was the case for the 1KGP. These new approaches are making it possible to study the population and clinical genetics of human TEs at the genome-scale for the first time.

The explosion of genome sequencing initiatives, which are often explicitly motivated by evolutionary or clinical considerations, will provide abundant opportunities for the application of these novel genomic techniques for TE discovery and research. Nevertheless, the sheer abundance of the data that is being generated by such initiatives will provide substantial challenges to the research community. The temptation could exist to focus on the most easily accessible sequence variants, *i.e.,* SNPs, and disregard the more difficult to characterize structural variants. We feel that this would be a mistake, as it is simply not possible to appreciate the full scope of human genetic variation without considering TE insertion polymorphisms. Hopefully the new genomic technologies for TE discovery and characterization will come to be even more widely used and applied for future genome powered studies of human genetics.

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

Funding

L.R. and I.K.J. were supported by the IHRC-Georgia Tech Applied Bioinformatics Laboratory (ABiL). L.W. and E.A.C. were supported by the Georgia Tech Bioinformatics Graduate Program. This work was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and National Center for Biotechnology Information (NIH, NLM, NCBI).

References

[1] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. Nature 2001; 409(6822):860-921; PMID:11237011; http://dx.doi.org/10.1038/35057062

[2] Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0 2015 <http://www.repeatmasker.org>

[3] de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 2011; 7(12): e1002384; PMID:22144907; http://dx.doi.org/10.1371/journal.pgen.1002384

[4] Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature 2010; 467(7319):1061-73; PMID:20981092; http://dx.doi.org/10.1038/nature09534

[5] Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature 2012; 491(7422):56-65; PMID:23128226;http://dx.doi.org/10.1038/nature11632

[6] Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. A global reference for human genetic variation. Nature 2015; 526(7571):68-74; PMID:26432245; http://dx.doi.org/10.1038/nature15393

[7] Stewart C, Kural D, Stromberg MP, Walker JA, Konkel MK, Stutz AM, Urban AE, Grubert F, Lam HY, Lee WP, et al. A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet 2011; 7(8):e1002236; PMID:21876680; http://dx.doi.org/10.1371/journal.pgen.1002236

[8] Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. An integrated map of structural variation in 2,504 human genomes. Nature 2015; 526 (7571):75-81; PMID:26432246; http://dx.doi.org/10.1038/nature15394

[9] Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 2003; 35(1):41-8; PMID:12897783; http://dx.doi.org/10.1038/ng1223

[10] Ostertag EM, Goodier JL, Zhang Y, Kazazian HH, Jr. SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet 2003; 73(6):1444-51; PMID:14628287; http://dx.doi.org/10.1086/380207

[11] Ullu E, Tschudi C. Alu sequences are processed 7SL RNA genes. Nature 1984; 312(5990):171-2; PMID:6209580; http://dx.doi.org/10.1038/312171a0

[12] Schmid CW, Deininger PL. Sequence organization of the human genome. Cell 1975; 6(3):345-58; PMID:1052772; http://dx.doi.org/10.1016/0092-8674(75)90184-1

[13] Shen L, Wu LC, Sanlioglu S, Chen R, Mendoza AR, Dangel AW, Carroll MC, Zipf WB, Yu CY. Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. J Biol Chem 1994; 269 (11):8466-76; PMID:8132574

[14] Ono M, Kawakami M, Takezawa T. A novel human nonviral retroposon derived from an endogenous retrovirus. Nucleic Acids Res 1987; 15(21):8725-37; PMID:2825118; http://dx.doi.org/10.1093/nar/15.21.8725

[15] Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA. SVA elements: a hominid-specific retroposon family. J Mol Biol 2005; 354(4):994-1007; PMID:16288912; http://dx.doi.org/10.1016/j.jmb.2005.09.085

[16] NHGRI. 2003 October 3, 2016. The human genome project completion: frequently asked questions. <https://www.genome.gov/11006943/>. October 3, 2016.

[17] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977; 74(12):5463-7; PMID:271968; http://dx.doi.org/10.1073/pnas.74.12.5463

[18] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016; 17(6):333-51; PMID:27184599; http://dx.doi.org/10.1038/nrg.2016.49

[19] Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer analysis project. Nat Genet 2013; 45(10):1113-20; PMID:24071849; http://dx.doi.org/10.1038/ng.2764

[20] Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang J, Whitty B, et al. International cancer genome consortium data portal–a one-stop shop for cancer genomics data. Database (Oxford) 2011; 2011:bar026; PMID:21930502

[21] Initiative NPM. Mar 15. Trans-Omics for Precision Medicine (TOPMed) program. <http://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed>. Accessed 2016 Mar 15.

[22] Consortium UK, Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M, et al. The UK10K project identifies rare variants in health and disease. Nature 2015; 526(7571):82-90; PMID:26367797; http://dx.doi.org/10.1038/nature14962

[23] Project Team SG. The Saudi Human Genome Program: An oasis in the desert of Arab medicine is providing

clues to genetic disease. IEEE Pulse 2015; 6(6):22-6; http://dx.doi.org/10.1109/MPUL.2015.2476541

[24] Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 2016; 538(7624):201-206; http://dx.doi.org/10.1038/nature18964

[25] Pagani L, Lawson DJ, Jagoda E, Morseburg A, Eriksson A, Mitt M, Clemente F, Hudjashov G, DeGiorgio M, Saag L, et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature 2016; 538 (7624):238-242; http://dx.doi.org/10.1038/nature19792

[26] Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 2014; 32(3):246-51; PMID:24531798; http://dx.doi.org/10.1038/nbt.2835

[27] Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 2016; 3:160025; PMID:27271295; http://dx.doi.org/10.1038/sdata.2016.25

[28] Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet 2011; 12(5):363-76; PMID:21358748; http://dx.doi.org/10.1038/nrg2958

[29] Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 2009; 6(9):677-81; PMID:19668202; http://dx.doi.org/10.1038/nmeth.1363

[30] Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res 2009; 19(7):1270-8; PMID:19447966; http://dx.doi.org/10.1101/gr.088633.108

[31] Rishishwar L, Marino-Ramirez L, Jordan IK. Benchmarking computational tools for polymorphic transposable element detection. Brief Bioinform 2016; PMID:27524380

[32] Ewing AD. Transposable element detection from whole genome sequence data. Mob DNA 2015; 6:24; PMID:26719777; http://dx.doi.org/10.1186/s13100-015-0055-3

[33] Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, Sahinalp SC. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics 2010; 26 (12):i350-7; PMID:20529927; http://dx.doi.org/10.1093/bioinformatics/btq216

[34] Wu J, Lee WP, Ward A, Walker JA, Konkel MK, Batzer MA, Marth GT. Tangram: a comprehensive toolbox for mobile element insertion detection. BMC Genomics 2014; 15:795; PMID:25228379; http://dx.doi.org/10.1186/1471-2164-15-795

[35] Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics 2013; 29(3):389-90; PMID:23233656; http://dx.doi.org/10.1093/bioinformatics/bts697

[36] Thung DT, de Ligt J, Vissers LE, Steehouwer M, Kroon M, de Vries P, Slagboom EP, Ye K, Veltman JA, Hehir-Kwa JY. Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biol 2014; 15(10):488; PMID:25348035; http://dx.doi.org/10.1186/s13059-014-0488-x

[37] Zhuang J, Wang J, Theurkauf W, Weng Z. TEMP: a computational method for analyzing transposable element polymorphism in populations. Nucleic Acids Res 2014; 42(11):6826-38; PMID:24753423; http://dx.doi.org/10.1093/nar/gku323

[38] Jiang C, Chen C, Huang Z, Liu R, Verdier J. ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. BMC Bioinformatics 2015; 16:72; PMID:25887332; http://dx.doi.org/10.1186/s12859-015-0507-2

[39] Fiston-Lavier AS, Barron MG, Petrov DA, Gonzalez J. T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. Nucleic Acids Res 2015; 43(4):e22; PMID:25510498; http://dx.doi.org/10.1093/nar/gku1250

[40] Xing J, Witherspoon DJ, Jorde LB. Mobile element biology: new possibilities with high-throughput sequencing. Trends Genet 2013; 29(5):280-9; PMID:23312846; http://dx.doi.org/10.1016/j.tig.2012.12.002

[41] Wheelan SJ, Scheifele LZ, Martinez-Murillo F, Irizarry RA, Boeke JD. Transposon insertion site profiling chip (TIP-chip). Proc Natl Acad Sci U S A 2006; 103 (47):17632-7; PMID:17101968; http://dx.doi.org/10.1073/pnas.0605450103

[42] Witherspoon DJ, Xing J, Zhang Y, Watkins WS, Batzer MA, Jorde LB. Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. BMC Genomics 2010; 11:410; PMID:20591181; http://dx.doi.org/10.1186/1471-2164-11-410

[43] Ewing AD, Kazazian HH, Jr. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res 2010; 20(9):1262-70; PMID:20488934; http://dx.doi.org/10.1101/gr.106419.110

[44] Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, Devine SE. Natural mutagenesis of human genomes by endogenous retrotransposons. Cell 2010; 141(7):1253-61; PMID:20603005; http://dx.doi.org/10.1016/j.cell.2010.05.020

[45] Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran JV. LINE-1 retrotransposition activity in human genomes. Cell 2010; 141(7):1159-70; PMID:20602998;http://dx.doi.org/10.1016/j.cell.2010.05.021

[46] Prak ET, Kazazian HH, Jr. Mobile elements and the human genome. Nat Rev Genet 2000; 1(2):134-44; PMID:11253653; http://dx.doi.org/10.1038/35038572

[47] Goldman MA, Holmquist GP, Gray MC, Caston LA, Nag A. Replication timing of genes and middle repetitive sequences. Science 1984; 224(4650):686-92; PMID:6719109; http://dx.doi.org/10.1126/science.6719109

[48] Manuelidis L, Ward DC. Chromosomal and nuclear distribution of the HindIII 1.9-kb human DNA repeat segment. Chromosoma 1984; 91(1):28-38; PMID:6098426; http://dx.doi.org/10.1007/BF00286482

[49] Soriano P, Meunier-Rotival M, Bernardi G. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. Proc Natl Acad Sci U S A 1983; 80(7):1816-20; PMID:6572942; http://dx.doi.org/10.1073/pnas.80.7.1816

[50] Langley CH, Montgomery E, Hudson R, Kaplan N, Charlesworth B. On the role of unequal exchange in the containment of transposable element copy number. Genet Res 1988; 52(3):223-35; PMID:2854088; http://dx.doi.org/10.1017/S0016672300027695

[51] Graham T, Boissinot S. The genomic distribution of L1 elements: the role of insertion bias and natural selection. J Biomed Biotechnol 2006; 2006(1):75327; PMID:16877820

[52] Medstrand P, van de Lagemaat LN, Mager DL. Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res 2002; 12(10):1483-95; PMID:12368240; http://dx.doi.org/10.1101/gr.388902

[53] Meunier-Rotival M, Soriano P, Cuny G, Strauss F, Bernardi G. Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. Proc Natl Acad Sci U S A 1982; 79(2):355-9; PMID:6281768; http://dx.doi.org/10.1073/pnas.79.2.355

[54] Cuny G, Soriano P, Macaya G, Bernardi G. The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity. Eur J Biochem 1981; 115(2):227-33; PMID:7238506; http://dx.doi.org/10.1111/j.1432-1033.1981.tb05227.x

[55] Schmid CW. Does SINE evolution preclude Alu function? Nucleic Acids Res 1998; 26(20):4541-50; PMID:9753719; http://dx.doi.org/10.1093/nar/26.20.4541

[56] Hackenberg M, Bernaola-Galvan P, Carpena P, Oliver JL. The biased distribution of Alus in human isochores might be driven by recombination. J Mol Evol 2005; 60 (3):365-77; PMID:15871047; http://dx.doi.org/10.1007/s00239-004-0197-2

[57] Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet 2002; 3(5):370-9; PMID:11988762; http://dx.doi.org/10.1038/nrg798

[58] Brookfield JF. Selection on Alu sequences? Curr Biol 2001; 11(22):R900-1; PMID:11719231; http://dx.doi.org/10.1016/S0960-9822(01)00547-4

[59] Stenger JE, Lobachev KS, Gordenin D, Darden TA, Jurka J, Resnick MA. Biased distribution of inverted and direct Alus in the human genome: implications for insertion, exclusion, and genome stability. Genome Res 2001; 11(1):12-27; PMID:11156612; http://dx.doi.org/10.1101/gr.158801

[60] Lobachev KS, Stenger JE, Kozyreva OG, Jurka J, Gordenin DA, Resnick MA. Inverted Alu repeats unstable in yeast are excluded from the human genome. EMBO J 2000; 19(14):3822-30; PMID:10899135; http://dx.doi.org/10.1093/emboj/19.14.3822

[61] Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 2003; 73(6):1402-22; PMID:14631557; http://dx.doi.org/10.1086/380416

[62] Ding L, Wiener H, Abebe T, Altaye M, Go RC, Kercsmar C, Grabowski G, Martin LJ, Khurana Hershey GK, Chakorborty R, et al. Comparison of measures of marker informativeness for ancestry and admixture mapping. BMC Genomics 2011; 12:622; PMID:22185208; http://dx.doi.org/10.1186/1471-2164-12-622

[63] Shedlock AM, Takahashi K, Okada N. SINEs of speciation: tracking lineages with retroposons. Trends Ecol Evol 2004; 19(10):545-53; PMID:16701320; http://dx.doi.org/10.1016/j.tree.2004.08.002

[64] Ray DA, Xing J, Salem AH, Batzer MA. SINEs of a nearly perfect character. Syst Biol 2006; 55(6):928-35; PMID:17345674; http://dx.doi.org/10.1080/10635150600865419

[65] Ray DA, Batzer MA. Reading TE leaves: new approaches to the identification of transposable element insertions. Genome Res 2011; 21(6):813-20; PMID:21632748; http://dx.doi.org/10.1101/gr.110528.110

[66] Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad BV, Reddy PG, Rasanayagam A, et al. Genetic evidence on the origins of Indian caste populations. Genome Res 2001; 11(6):994-1004; PMID:11381027; http://dx.doi.org/10.1101/gr.GR-1733RR

[67] Witherspoon DJ, Marchani EE, Watkins WS, Ostler CT, Wooding SP, Anders BA, Fowlkes JD, Boissinot S, Furano AV, Ray DA, et al. Human population genetic structure and diversity inferred from polymorphic L1 (LINE-1) and Alu insertions. Hum Hered 2006; 62 (1):30-46; PMID:17003565; http://dx.doi.org/10.1159/000095851

[68] Watkins WS, Rogers AR, Ostler CT, Wooding S, Bamshad MJ, Brassington AM, Carroll ML, Nguyen SV, Walker JA, Prasad BV, et al. Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. Genome Res 2003; 13(7):1607-18; PMID:12805277; http://dx.doi.org/10.1101/gr.894603

[69] Ray DA, Walker JA, Hall A, Llewellyn B, Ballantyne J, Christian AT, Turteltaub K, Batzer MA. Inference of human geographic origins using Alu insertion polymorphisms. Forensic Sci Int 2005; 153(2-3):117-24; PMID:16139099; http://dx.doi.org/10.1016/j.forsciint.2004.10.017

[70] Novick GE, Novick CC, Yunis J, Yunis E, Antunez de Mayolo P, Scheer WD, Deininger PL, Stoneking M, York DS, Batzer MA, et al. Polymorphic Alu insertions and the Asian origin of Native American populations. Hum Biol 1998; 70(1):23-39; PMID:9489232

[71] Nasidze I, Risch GM, Robichaux M, Sherry ST, Batzer MA, Stoneking M. Alu insertion polymorphisms and the genetic structure of human populations from the Caucasus. Eur J Hum Genet 2001; 9(4):267-72; PMID:11313770; http://dx.doi.org/10.1038/sj.ejhg.5200615

[72] Batzer MA, Stoneking M, Alegria-Hartman M, Bazan H, Kass DH, Shaikh TH, Novick GE, Ioannou PA, Scheer WD, Herrera RJ, et al. African origin of human-specific polymorphic Alu insertions. Proc Natl Acad Sci U S A 1994; 91(25):12288-92; PMID:7991620; http://dx.doi.org/10.1073/pnas.91.25.12288

[73] Batzer MA, Arcot SS, Phinney JW, Alegria-Hartman M, Kass DH, Milligan SM, Kimpton C, Gill P, Hochmeister M, Ioannou PA, et al. Genetic variation of recent Alu insertions in human populations. J Mol Evol 1996; 42 (1):22-9; PMID:8576959; http://dx.doi.org/10.1007/BF00163207

[74] Rishishwar L, Tellez Villa CE, Jordan IK. Transposable element polymorphisms recapitulate human evolution. Mob DNA 2015; 6:21; PMID:26579215; http://dx.doi.org/10.1186/s13100-015-0052-6

[75] Risch N, Burchard E, Ziv E, Tang H. Categorization of humans in biomedical research: genes, race and disease. Genome Biol 2002; 3(7):comment2007; PMID:12184798; http://dx.doi.org/10.1186/gb-2002-3-7-comment2007

[76] Collins FS, Green ED, Guttmacher AE, Guyer MS, Institute USNHGR. A vision for the future of genomics research. Nature 2003; 422(6934):835-47; PMID:12695777; http://dx.doi.org/10.1038/nature01626

[77] Smith MW, O'Brien SJ. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat Rev Genet 2005; 6(8):623-32; PMID:16012528; http://dx.doi.org/10.1038/nrg1657

[78] Winkler CA, Nelson GW, Smith MW. Admixture mapping comes of age. Annu Rev Genomics Hum Genet 2010; 11:65-89; PMID:20594047; http://dx.doi.org/10.1146/annurev-genom-082509-141523

[79] Reilly MT, Faulkner GJ, Dubnau J, Ponomarev I, Gage FH. The role of transposable elements in health and diseases of the central nervous system. J Neurosci 2013; 33 (45):17577-86; PMID:24198348; http://dx.doi.org/10.1523/JNEUROSCI.3369-13.2013

[80] Hancks DC, Kazazian HH, Jr. Roles for retrotransposon insertions in human disease. Mob DNA 2016; 7:9; PMID:27158268; http://dx.doi.org/10.1186/s13100-016-0065-9

[81] Chenais B. Transposable elements in cancer and other human diseases. Curr Cancer Drug Targets 2015; 15 (3):227-42; PMID:25808076; http://dx.doi.org/10.2174/1568009615666150317122506

[82] Burns KH, Boeke JD. Human transposon tectonics. Cell 2012; 149(4):740-52; PMID:22579280; http://dx.doi.org/10.1016/j.cell.2012.04.019

[83] Beck CR, Garcia-Perez JL, Badge RM, Moran JV. LINE-1 elements in structural variation and disease. Annu Rev Genomics Hum Genet 2011; 12:187-215; PMID:21801021; http://dx.doi.org/10.1146/annurev-genom-082509-141802

[84] Koonin EV. Splendor and misery of adaptation, or the importance of neutral null for understanding evolution. BMC Biol 2016; 14(1):114; PMID:28010725; http://dx.doi.org/10.1186/s12915-016-0338-2

[85] Lynch M. The origins of genome architecture. Sunderland, MA: Sinauer Associates; 2007.

[86] Cordaux R, Lee J, Dinoso L, Batzer MA. Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. Gene 2006; 373:138-44; PMID:16527433; http://dx.doi.org/10.1016/j.gene.2006.01.020

[87] Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH, Jr. Many human L1 elements are capable of retrotransposition. Nat Genet 1997; 16(1):37-43; PMID:9140393; http://dx.doi.org/10.1038/ng0597-37

[88] Pavlicek A, Paces J, Zika R, Hejnar J. Length distribution of long interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection. Gene 2002; 300(1-2):189-94; PMID:12468100; http://dx.doi.org/10.1016/S0378-1119(02)01047-8

[89] Boissinot S, Davis J, Entezam A, Petrov D, Furano AV. Fitness cost of LINE-1 (L1) activity in humans. Proc Natl Acad Sci U S A 2006; 103(25):9590-4; PMID:16766655; http://dx.doi.org/10.1073/pnas.0603334103

[90] Boissinot S, Chevret P, Furano AV. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol 2000; 17(6):915-28; PMID:10833198; http://dx.doi.org/10.1093/oxfordjournals.molbev.a026372

[91] Pascale E, Liu C, Valle E, Usdin K, Furano AV. The evolution of long interspersed repeated DNA (L1, LINE 1) as revealed by the analysis of an ancient rodent L1 DNA family. J Mol Evol 1993; 36(1):9-20; PMID:8433380; http://dx.doi.org/10.1007/BF02407302

[92] Hedges DJ, Deininger PL. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutat Res 2007; 616(1-2):46-59; PMID:17157332; http://dx.doi.org/10.1016/j.mrfmmm.2006.11.021

[93] Dolgin ES, Charlesworth B. The effects of recombination rate on the distribution and abundance of transposable elements. Genetics 2008; 178(4):2169-77; PMID:18430942; http://dx.doi.org/10.1534/genetics.107.082743

[94] Matyunina LV, Bowen NJ, McDonald JF. LTR retrotransposons and the evolution of dosage compensation in Drosophila. BMC Mol Biol 2008; 9:55; PMID:18533037; http://dx.doi.org/10.1186/1471-2199-9-55

[95] McDonald JF, Matzke MA, Matzke AJ. Host defenses to transposable elements and the evolution of genomic imprinting. Cytogenet Genome Res 2005; 110(1-

4):242-9; PMID:16093678; http://dx.doi.org/10.1159/000084958

[96] Bowen NJ, Jordan IK. Exaptation of protein coding sequences from transposable elements. Genome Dyn 2007; 3:147-62; PMID:18753790

[97] Jordan IK. Evolutionary tinkering with transposable elements. Proc Natl Acad Sci U S A 2006; 103 (21):7941-2; PMID:16705033; http://dx.doi.org/10.1073/pnas.0602656103

[98] Miller WJ, Hagemann S, Reiter E, Pinsker W. P-element homologous sequences are tandemly repeated in the genome of Drosophila guanche. Proc Natl Acad Sci U S A 1992; 89(9):4018-22; PMID:1315047; http://dx.doi.org/10.1073/pnas.89.9.4018

[99] Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet 2008; 9(5):397-405; PMID:18368054; http://dx.doi.org/10.1038/nrg2337

[100] Conley AB, Piriyapongsa J, Jordan IK. Retroviral promoters in the human genome. Bioinformatics 2008; 24 (14):1563-7; PMID:18535086; http://dx.doi.org/10.1093/bioinformatics/btn243

[101] Jordan IK, Rogozin IB, Glazko GV, Koonin EV. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends in Genetics 2003; 19(2):68-72; PMID:12547512; http://dx.doi.org/10.1016/S0168-9525(02)00006-9

[102] Marino-Ramirez L, Lewis KC, Landsman D, Jordan IK. Transposable elements donate lineage-specific regulatory sequences to host genomes. Cytogen Genome Res 2005; 110(1-4):333-341; PMID:16093685; http://dx.doi.org/10.1159/000084965

[103] Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 2006; 441(7089):87-90; PMID:16625209; http://dx.doi.org/10.1038/nature04696

[104] Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 2016; 351(6277):1083-7; PMID:26941318; http://dx.doi.org/10.1126/science.aad5497

[105] Chuong EB, Rumi MA, Soares MJ, Baker JC. Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat Genet 2013; 45(3):325-9; PMID:23396136; http://dx.doi.org/10.1038/ng.2553

[106] Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet 2010; 42(7):631-4; PMID:20526341; http://dx.doi.org/10.1038/ng.600

[107] Notwell JH, Chung T, Heavner W, Bejerano G. A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. Nat Commun 2015; 6:6644; PMID:25806706; http://dx.doi.org/10.1038/ncomms7644

[108] Conley AB, Jordan IK. Cell type-specific termination of transcription by transposable element sequences. Mob DNA 2012; 3(1):15; PMID:23020800; http://dx.doi.org/10.1186/1759-8753-3-15

[109] Kapusta A, Kronenberg Z, Lynch VJ, Zhuo XY, Ramsay L, Bourque G, Yandell M, Feschotte C. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. Plos Genetics 2013; 9(4); PMID:23637635; http://dx.doi.org/10.1371/journal.pgen.1003470

[110] Piriyapongsa J, Marino-Ramirez L, Jordan IK. Origin and evolution of human microRNAs from transposable elements. Genetics 2007; 176(2):1323-37; PMID:17435244; http://dx.doi.org/10.1534/genetics.107.072553

[111] Weber MJ. Mammalian small nucleolar RNAs are mobile genetic elements. Plos Genetics 2006; 2(12):1984-1997; http://dx.doi.org/10.1371/journal.pgen.0020205

[112] Jacques PE, Jeyakani J, Bourque G. The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet 2013; 9(5): e1003504; PMID:23675311; http://dx.doi.org/10.1371/journal.pgen.1003504

[113] Pavlicek A, Jabbari K, Paces J, Paces V, Hejnar JV, Bernardi G. Similar integration but different stability of Alus and LINEs in the human genome. Gene 2001; 276 (1-2):39-45; PMID:11591470; http://dx.doi.org/10.1016/S0378-1119(01)00645-X

[114] Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell 2012; 148(1-2):335-48; PMID:22244452; http://dx.doi.org/10.1016/j.cell.2011.11.058

[115] Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res 2014; 24(12):1963-76; PMID:25319995; http://dx.doi.org/10.1101/gr.168872.113

[116] Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015; 518 (7539):317-30; PMID:25693563; http://dx.doi.org/10.1038/nature14248

[117] Kuhn A, Ong YM, Cheng CY, Wong TY, Quake SR, Burkholder WF. Linkage disequilibrium and signatures of positive selection around LINE-1 retrotransposons in the human genome. Proc Natl Acad Sci U S A 2014; 111(22):8131-6; PMID:24847061; http://dx.doi.org/10.1073/pnas.1401532111

[118] Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet 2013; 47:97-120; PMID:24274750; http://dx.doi.org/10.1146/annurev-genet-111212-133526

[119] Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, et al. Identifying recent adaptations in large-scale genomic data. Cell 2013; 152(4):703-13; PMID:23415221; http://dx.doi.org/10.1016/j.cell.2013.01.035

[120] Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. Positive natural selection in the human lineage. Science 2006; 312(5780):1614-20; PMID:16778047; http://dx.doi.org/10.1126/science.1124309

[121] Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet 2009; 10 (10):691-703; PMID:19763152; http://dx.doi.org/10.1038/nrg2640

[122] Deininger PL, Batzer MA. Alu repeats and human disease. Mol Genet Metab 1999; 67(3):183-93; PMID:10381326; http://dx.doi.org/10.1006/mgme.1999.2864

[123] Kazazian HH, Jr., Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 1988; 332(6160):164-6; PMID:2831458; http://dx.doi.org/10.1038/332164a0

[124] Vidaud D, Vidaud M, Bahnak BR, Siguret V, Gispert Sanchez S, Laurian Y, Meyer D, Goossens M, Lavergne JM. Haemophilia B due to a de novo insertion of a human-specific Alu subfamily member within the coding region of the factor IX gene. Eur J Hum Genet 1993; 1(1):30-6; PMID:8069649

[125] Chen JM, Masson E, Macek M, Jr., Raguenes O, Piskackova T, Fercot B, Fila L, Cooper DN, Audrezet MP, Ferec C. Detection of two Alu insertions in the CFTR gene. J Cyst Fibros 2008; 7(1):37-43; PMID:17531547; http://dx.doi.org/10.1016/j.jcf.2007.04.001

[126] Oldridge M, Zackai EH, McDonald-McGinn DM, Iseki S, Morriss-Kay GM, Twigg SR, Johnson D, Wall SA, Jiang W, Theda C, et al. De novo alu-element insertions in FGFR2 identify a distinct pathological basis for Apert syndrome. Am J Hum Genet 1999; 64(2):446-61; PMID:9973282; http://dx.doi.org/10.1086/302245

[127] Lester T, McMahon C, VanRegemorter N, Jones A, Genet S. X-linked immunodeficiency caused by insertion of Alu repeat sequences. J Med Gen 1997; 34:1417-1417.

[128] Bratthauer GL, Fanning TG. Active LINE-1 retrotransposons in human testicular cancer. Oncogene 1992; 7 (3):507-10; PMID:1312702

[129] Bratthauer GL, Fanning TG. LINE-1 retrotransposon expression in pediatric germ cell tumors. Cancer 1993; 71(7):2383-6; PMID:8384068; http://dx.doi.org/10.1002/1097-0142(19930401)71:7%3c2383::AID-CNCR2820710733%3e3.0.CO;2-P

[130] Bratthauer GL, Cardiff RD, Fanning TG. Expression of LINE-1 retrotransposons in human breast cancer. Cancer 1994; 73(9):2333-6; PMID:8168038; http://dx.doi.org/10.1002/1097-0142(19940501)73:9%3c2333::AID-CNCR2820730915%3e3.0.CO;2-4

[131] Asch HL, Eliacin E, Fanning TG, Connolly JL, Bratthauer G, Asch BB. Comparative expression of the LINE-1 p40 protein in human breast carcinomas and normal breast tissues. Oncol Res 1996; 8(6):239-47; PMID:8895199

[132] Rodic N, Sharma R, Sharma R, Zampella J, Dai L, Taylor MS, Hruban RH, Iacobuzio-Donahue CA, Maitra A, Torbenson MS, et al. Long interspersed element-1 protein expression is a hallmark of many human cancers. Am J Pathol 2014; 184(5):1280-6; PMID:24607009; http://dx.doi.org/10.1016/j.ajpath.2014.01.007

[133] Solyom S, Ewing AD, Rahrmann EP, Doucet T, Nelson HH, Burns MB, Harris RS, Sigmon DF, Casella A, Erlanger B, et al. Extensive somatic L1 retrotransposition in colorectal tumors. Genome Res 2012; 22(12):2328-38; PMID:22968929; http://dx.doi.org/10.1101/gr.145235.112

[134] Doucet-O'Hare TT, Rodic N, Sharma R, Darbari I, Abril G, Choi JA, Young Ahn J, Cheng Y, Anders RA, Burns KH, et al. LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. Proc Natl Acad Sci U S A 2015; 112(35): E4894-900; PMID:26283398; http://dx.doi.org/10.1073/pnas.1502474112

[135] Doucet-O'Hare TT, Sharma R, Rodic N, Anders RA, Burns KH, Kazazian HH, Jr. Somatically Acquired LINE-1 Insertions in Normal Esophagus Undergo Clonal Expansion in Esophageal Squamous Cell Carcinoma. Hum Mutat 2016; 37(9):942-54; PMID:27319353; http://dx.doi.org/10.1002/humu.23027

[136] Ewing AD, Gacita A, Wood LD, Ma F, Xing D, Kim MS, Manda SS, Abril G, Pereira G, Makohon-Moore A, et al. Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. Genome Res 2015; 25(10):1536-45; PMID:26260970; http://dx.doi.org/10.1101/gr.196238.115

[137] Tubio JM, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, Gundem G, Pipinikas CP, Zamora J, Raine K, et al. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science 2014; 345(6196):1251343; PMID:25082706; http://dx.doi.org/10.1126/science.1251343

[138] Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature 2011; 479 (7374):534-7; PMID:22037309; http://dx.doi.org/10.1038/nature10531

[139] Shukla R, Upton KR, Munoz-Lopez M, Gerhardt DJ, Fisher ME, Nguyen T, Brennan PM, Baillie JK, Collino A, Ghisletti S, et al. Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. Cell 2013; 153(1):101-11; PMID:23540693; http://dx.doi.org/10.1016/j.cell.2013.02.032

[140] Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res 2016; 26(6):745-55; PMID:27197217; http://dx.doi.org/10.1101/gr.201814.115

[141] Gorbunova V, Boeke JD, Helfand SL, Sedivy JM. Human Genomics. Sleeping dogs of the genome. Science 2014; 346(6214):1187-8; PMID:25477445; http://dx.doi.org/10.1126/science.aaa3177

[142] St Laurent G, 3rd, Hammell N, McCaffrey TA. A LINE-1 component to human aging: do LINE elements exact a longevity cost for evolutionary advantage? Mech Ageing Dev 2010; 131(5):299-305; PMID:20346965; http://dx.doi.org/10.1016/j.mad.2010.03.008

[143] Jintaridth P, Mutirangura A. Distinctive patterns of age-dependent hypomethylation in interspersed repetitive sequences. Physiol Genomics 2010; 41(2):194-200; PMID:20145203; http://dx.doi.org/10.1152/physiolgenomics.00146.2009

[144] Van Meter M, Kashyap M, Rezazadeh S, Geneva AJ, Morello TD, Seluanov A, Gorbunova V. SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age. Nat Commun 2014; 5:5011; PMID:25247314; http://dx.doi.org/10.1038/ncomms6011

[145] De Cecco M, Criscione SW, Peckham EJ, Hillenmeyer S, Hamm EA, Manivannan J, Peterson AL, Kreiling JA, Neretti N, Sedivy JM. Genomes of replicatively senescent cells undergo global epigenetic changes leading to gene silencing and activation of transposable elements. Aging Cell 2013; 12(2):247-56; PMID:23360310; http://dx.doi.org/10.1111/acel.12047

[146] Wang J, Geesman GJ, Hostikka SL, Atallah M, Blackwell B, Lee E, Cook PJ, Pasaniuc B, Shariat G, Halperin E, et al. Inhibition of activated pericentromeric SINE/Alu repeat transcription in senescent human adult stem cells reinstates self-renewal. Cell Cycle 2011; 10(17):3016-30; PMID:21862875; http://dx.doi.org/10.4161/cc.10.17.17543

[147] Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet 2011; 13(2):135-45; http://dx.doi.org/10.1038/nrg3118

[148] Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet 2016; 48(1):22-9; PMID:26642241; http://dx.doi.org/10.1038/ng.3461

[149] Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 2013; 501(7468):506-11; PMID:24037378; http://dx.doi.org/10.1038/nature12531

[150] Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015; 348(6235):648-60; PMID:25954001; http://dx.doi.org/10.1126/science.1262110