Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Genetics & Development

Functional impact of the human mobilome

Timothy D Babatz[1,2] and Kathleen H Burns[1,2,3,4]

The human genome is replete with interspersed repetitive sequences derived from the propagation of mobile DNA elements. Three families of human retrotransposons remain active today: LINE1, *Alu*, and SVA elements. Since 1988, *de novo* insertions at previously recognized disease loci have been shown to generate highly penetrant alleles in Mendelian disorders. Only recently has the extent of germline-transmitted retrotransposon insertion polymorphism (RIP) in human populations been fully realized. Also exciting are recent studies of somatic retrotransposition in human tissues and reports of tumor-specific insertions, suggesting roles in tissue heterogeneity and tumorigenesis. Here we discuss mobile elements in human disease with an emphasis on exciting developments from the last several years.

Addresses
[1] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, United States
[2] High Throughput Biology Center, Johns Hopkins University School of Medicine, Baltimore, MD, United States
[3] Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, United States
[4] The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Corresponding author: Burns, Kathleen H (kburns@jhmi.edu)

Current Opinion in Genetics & Development 2013, **23**:264–270

This review comes from a themed issue on **Molecular and genetic bases of disease**

Edited by **Jim Lupski** and **Nancy Maizels**

For a complete overview see the Issue and the Editorial

Available online 22nd May 2013

0959-437X/$ – see front matter,

http://dx.doi.org/10.1016/j.gde.2013.02.007

Introduction

Envisioning the RNA origin of DNA may evoke thoughts of a world teeming with self-replicating ribonucleic acids with functions that blur the boundary between the inanimate and simple life. Though sequence evidence for this world is obscure, the RNA origins of our modern human genome are readily apparent; even the casual genome gazer can take in a vast genomic fossil record. Our 3 billion base pairs are overwhelmingly non-coding, and 50% or more are recognizable as repetitive sequences. Most of the repeats are interspersed, meaning that they occur discontinuously as singularly scattered copies in the genome. These include Long INterspersed Elements (LINEs) and Short INterspersed Elements (SINEs). Most abundant in mammals are the retrotransposons, which self-replicate through the reverse transcription of RNA intermediates.

Genomes are not static strings of bases, but instead are mutable, evolving, and structurally dynamic. Three families of human retrotransposons have been active in relatively recent genome evolution and continue to generate human genomic variation today. First, the autonomous LINE-1 element (L1) is 6 kb in its full-length form and contains an internal promoter, two open reading frames (ORFs), and a polyadenylation signal. After RNA polymerase II transcription, mRNA processing, and export to the cytoplasm, ORF1 and ORF2 encode two proteins (ORF1p and ORF2p) that preferentially associate with their encoding RNA [1,2]. The resulting ribonucleoprotein (RNP) particles return the RNA to the nucleus, where ORF2p initiates target-primed reverse transcription (TPRT) to insert a copied DNA sequence at a new site in the host cell genome [3–5]. There are approximately 500 000 copies of L1-derived sequence in the human genome, most of which have been deactivated by 5′-truncation on integration or accumulated mutations. The youngest human-specific L1 subfamily, L1Hs, accounts for about 1% of these copies, with a much smaller subset of 'hot L1' elements driving most of the retrotransposition events.

The remaining families of active retrotransposons are non-autonomous, relying on L1-encoded machinery for mobilization [6,7]. *Alu* elements are ~300 bp long and there are approximately 1 000 000 *Alu*-derived sequences in the human genome. Nucleotide variations within the *Alu* consensus sequence distinguish several subfamilies including the most active *Alu*Y elements (reviewed in [8]). SVA elements comprise the third family of active retrotransposons; these are a relatively heterogeneous group of sequences with several recognized subfamilies. SVAs are composite elements, named for the sources of their component parts, SINE-R, VNTR, and *Alu* sequences. The SVA consensus sequence is approximately 2 kb and there are roughly 3700 SVA-derived sequences in the human genome (reviewed in [9]).

Functional impact

Host genomes are vulnerable to disruption by retrotransposons. How a gene is affected is dependent on both where the mobile DNA insertion occurs and the internal sequence of the element. Exonic insertions can interrupt the coding region of a gene. In addition, *in vitro* cell-based assays have revealed that features of L1 sequence such as the promoter, 5′-UTR antisense promoter, and internal splice sites can affect the transcription of target regions

and transcript processing [10,11]. Concordantly, intronic L1 insertions can induce aberrant transcript initiation, premature truncation, or attenuate transcript elongation, affecting the structure and abundance of target gene transcripts [12–14]. Similarly, intronic and gene proximal *Alu* insertions can act as enhancers or promoters, introduce cryptic splice donors or acceptors, or abrogate endogenous exon splicing [15–19]. In clinical genetics, retrotransposon insertions have been implicated in rare cases of highly penetrant Mendelian disorders including hemophilia, cystic fibrosis, neurofibromatosis, cystic fibrosis, and breast and colon cancers [20,21]. In each respective example the novel insertion creates a highly damaging allele affecting the coding capacity of a known disease gene and thus is fairly definitively implicated as the cause of the disorder.

In contrast, less deleterious insertions that are inherited and constitutive in an individual are major sources of structural variation in the human genome, genetic variation in populations, and occasional causes of genetic disease. Several studies in 2010 demonstrated that retrotransposon insertion polymorphisms (RIPs) are significantly more polymorphic among individual human genomes than previously recognized.

Germline insertions

Because *de novo* insertions are templated from the relatively few active subfamilies of repeats, each new insertion bears high internal sequence homology to the subfamily consensus sequence. This feature can be leveraged in targeted approaches for finding mobile DNA variants. These methods can be broadly categorized as either using laboratory-based genomic assays to map repeats or using computational means to recognize insertions in unselected reads in sequencing studies. The first category has included variations on hemi-specific PCRs using one primer within the repeat family of interest and a reaction schema that amplifies adjacent DNA [22•,23–25] (Figure 1). The latter has had the advantage of using data from more large-scale collaborative efforts such as the 1000 Genomes project, to identify large numbers of new insertions and phase these with neighboring SNPs [26,27]. At the same time, Beck *et al.* reported polymorphism of full-length 'hot' L1Hs elements among populations, suggesting possible inter-individual variation in endogenous L1 activity [28]. Collectively, new perspectives resulting from these discovery efforts have upheld earlier assertions that mobile DNAs are a major source of human genetic diversity. We currently estimate that between 1/100 and 1/200 individuals harbor a *de novo* LINE-1 insertion, and that about 1/20 individuals have a *de novo Alu* element insertion. This degree of activity is expected to result in approximately 11 000 L1Hs and 65 000 *Alu*Y common variable insertion alleles (MAF > 0.05) segregating in human populations and additional increasingly rare alleles.

Recently, two studies elucidated the mutational mechanism of inherited insertions in recessive diseases, one of which was a novel insertion discovery. A third study reported a novel mechanism of accumulated *Alu* RNA cytotoxicity.

Taniguchi-Ikeda *et al.* functionally characterized the polymorphic SVA insertion in the 3′ UTR of the *FUKUTIN* gene, which causes Fukuyama muscular dystrophy (FCMD). FCMD is a rare autosomal recessive disease mainly described in the Japanese population [29] and currently has no effective treatments. This ancient SVA is on a haplotype shared by >80% of FCMD chromosomes and likely arose as a single ancestral allele [30]. Long-range RT-PCR and sequence analysis in FCMD lymphoblast mRNA and controls revealed a novel splicing event between a weak alternative donor site in the final *FUKUTIN* coding exon and an acceptor site within the SVA insertion. This aberrant splicing excises the normal stop codon, deletes coding sequences for the 38 C-terminal amino acids of FUKUTIN, and adds 129 amino acids encoded by the SVA sequence. Normal FUKUTIN localizes to the Golgi, while FUKUTIN with SVA-derived C-terminus mislocalizes to the ER.

Remarkably, the authors were able to restore normal mRNA splicing of FCMD alleles to 40% of control levels using a combination of antisense oligonucleotides (AONs) targeted to the alternative splice donor, acceptor, and an exonic splicing enhancer. This strategy increased normal *FUKUTIN* mRNA expression significantly and rescued functional Fukutin protein *in vivo* in knock-in mice homozygous for humanized SVA-containing *FUKUTIN*. This study elucidated the mutational mechanism of a polymorphic disease-causing retrotransposon and demonstrated the possibility of splicing modulation therapy as a treatment for FCMD.

Tucker *et al.* [31•] performed exome sequencing on a single proband with retinitis pigmentosa (RP), a typically autosomal recessive ciliopathy characterized by loss of photoreceptor cells and a significant amount of genetic heterogeneity. After stringent filtering of sequence data from both ABI and Illumina sequencing instruments and prioritization of identified variants, there was a surprising lack of plausible disease-causing mutations verified by Sanger sequencing and present in both datasets. The ABI data showed a compound heterozygous pair of mutations in exon 9 of male germ cell-associated kinase (*MAK*) that were not in the Illumina data, and Sanger sequencing of the exon surprisingly revealed a homozygous 353-bp *Alu* insertion. The authors investigated why this important variant was missed by one method and misidentified by the other. The ABI protocol completely removed the *Alu*-MAK junction sequences and therefore sequenced chimeric DNA molecules and identified two heterozygous variants, and the Illumina protocol sequenced the