Review

Mobile element biology: new possibilities with high-throughput sequencing

Jinchuan Xing[1], David J. Witherspoon[2], and Lynn B. Jorde[2]

[1] Department of Genetics, Human Genetic Institute of New Jersey, Rutgers, State University of New Jersey, Piscataway, NJ 08854, USA
[2] Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT 84112, USA

Mobile elements comprise more than half of the human genome, but until recently their large-scale detection was time consuming and challenging. With the development of new high-throughput sequencing (HTS) technologies, the complete spectrum of mobile element variation in humans can now be identified and analyzed. Thousands of new mobile element insertions (MEIs) have been discovered, yielding new insights into mobile element biology, evolution, and genomic variation. Here, we review several high-throughput methods, with an emphasis on techniques that specifically target MEIs in humans. We highlight recent applications of these methods in evolutionary studies and in the analysis of somatic alterations in human normal and tumor tissues.

Mobile DNA elements in the human genome

Mobile DNA elements, or transposable DNA elements, are discrete DNA fragments that can be moved or copied to other regions of a genome. They are present in most organisms and comprise up to two-thirds of the human genome [1,2]. Several types of mobile element, including *Alu*, long interspersed elements (LINE) 1 (L1), and SVA, have been actively transposing during human evolution history and remain active (Box 1). These elements have had profound influence on genomic architecture, and the active elements continue to change the human genomic landscape, sometimes causing human diseases in the process (see recent reviews [3–5] for more details). In addition to their genomic impact, some MEIs have occurred so recently that they are present in some individuals but not in others. These polymorphic MEIs (pMEIs) are ideal markers for studying human evolutionary history and population relations because of their known ancestral state, lack of homoplasy, and other properties [6,7]. Over the past two decades, pMEIs have been used in a large number of population genetic and phylogenetic studies in humans as well as in other species [8–11].

Since the release of the human genome reference sequence, many efforts have been made to identify pMEIs in the human genome. This, however, has been a labor-intensive task. For example, the initial release of the human Retrotransposon Insertion Polymorphisms database (dbRIPs, http://dbrip.brocku.ca) contained approximately 2100 pMEIs [12], representing all pMEIs identified in more than 50 studies over the past several decades. Hundreds of additional pMEI loci were identified when additional individual genome sequences were published [13–16]. All these efforts have been dwarfed by recent advances in high-throughput technologies, which have dramatically changed the landscape of whole-genome, population-level MEI studies. Several high-throughput methods have been developed specifically for human pMEI detection, identifying thousands of novel pMEIs that are not present in the human reference genome. In this review, we first describe these methods and then summarize some recent major findings using whole-genome MEI data. We conclude by outlining new questions that can be addressed in the near future using these technologies.

Detecting mobile elements at the whole-genome level

There are two primary types of method for profiling pMEIs at the whole-genome level: (i) targeted methods, in which DNA fragments related to a MEI are experimentally enriched before sequencing and/or genotyping; and (ii) post-sequencing bioinformatic methods, in which pMEIs are identified using whole-genome sequencing data. Although the focus of this review is on the detection of MEIs in humans, similar high-throughput MEI identification methods have recently been applied to nonhuman species (Box 2).

Targeted MEI sequencing methods

One of the most commonly used MEI selection methods is transposon display (TD). This gel-based technique was developed initially to visualize pMEIs in *Petunia* species [17] and has since been adapted for studying *Alu*s [18,19], L1s [20–22], and human endogenous retroviruses (HERVs) [23] in the human genome. Several recent techniques have adapted the TD method for HTS (Table 1) [24–26], and the basic principle of the method is illustrated in Figure 1. The key step for MEI enrichment is a PCR step using one primer complementary to the MEI sequence and another primer complementing the linker sequence (Figure 1d). Sometimes referred as a 'hemispecific PCR',

Corresponding author: Jorde, L.B. (lbj@genetics.utah.edu).
Keywords: retrotransposon; mobile DNA element; high-throughput sequencing; polymorphism; somatic insertion; cancer.

  http://dx.doi.org/10.1016/j.tig.2012.12.002

Box 1. Mobile DNA elements in the human genome

Almost half of the human genome (47.8%) is composed of readily recognizable MEIs and their fragments (Figure Ia). The most prevalent are the type I retrotransposons, including non-long terminal repeat (LTR) retrotransposon LINE families (e.g., LINE-1 and LINE-2); the SINE *Alu* family; the LTR retrotransposon endogenous retroviruses (ERV); and SVA elements, a small but still active family. The type II 'cut-and-paste' DNA transposons of the Tc1-Mariner, hAT, and piggyBac families are also present in the human genome. An additional approximately 20% of the genome comprises short remnants of ancient mobile elements and other repeats [2].

The structures of major mobile elements in the human genome are illustrated in Figure Ib. The L1 element is the most prevalent autonomous non-LTR retrotransposon (approximately half a million copies). Full-length autonomous L1 elements are approximately 6 kb in length and contain two open reading frames (ORFs) that encode the proteins necessary to catalyze L1 retrotransposition. The 5′ UTR contains an RNA Pol II promoter. Upon insertion, L1 elements generate a variable-length poly-A tail at the 3′ end and short target site duplications (TSDs) at the insertion site. Most L1s are nonautonomous due to accumulated mutations and 5′ truncations that often occur upon retrotransposition. Approximately 100 'hot' L1 elements are active today [32,62]. Present at more than one million copies, *Alu* is the most prevalent nonautonomous non-LTR retrotransposon. Canonical *Alu* elements are approximately 300 bp in length, composed of two halves (right and left monomers joined by an A-rich spacer) derived from a 7SL RNA. *Alu* elements do not encode any proteins but do carry an internal RNA Pol III promoter in the left monomer. *Alu* elements replicate by hijacking the L1 retrotransposition machinery and thus generate similar TSDs and a 3′ poly-A tail upon insertion [63]. Most *Alu* insertions are ancient and inactive, but some subfamilies (notably *Alu*Ya and *Alu*Yb) are still active. SVA elements (approximately 5000 copies) are nonautonomous and rely on L1 proteins for their retrotransposition [64]. The canonical SVA is approximately 2 kb in length and is composed of a hexameric repeat region, a stretch of *Alu*-derived sequence, a VNTR of 35–50 bp, a segment derived from the LTR of a HERV, and a polyadenylation signal followed by a poly-A tail. Autonomous ERV copies are approximately 8 kb long and encode the proteins gag, protease (pro), pol, and env, which are required for reverse transcription of their RNAs and integration into the genome [65,66]. DNA elements typically encode a transposase protein that recognizes the element by its inverted terminal repeats (longer boxed arrows in Figure I), then excises and reinserts it elsewhere in the genome. DNA elements in the human genome are ancient remnants, nonautonomous, and inactive.

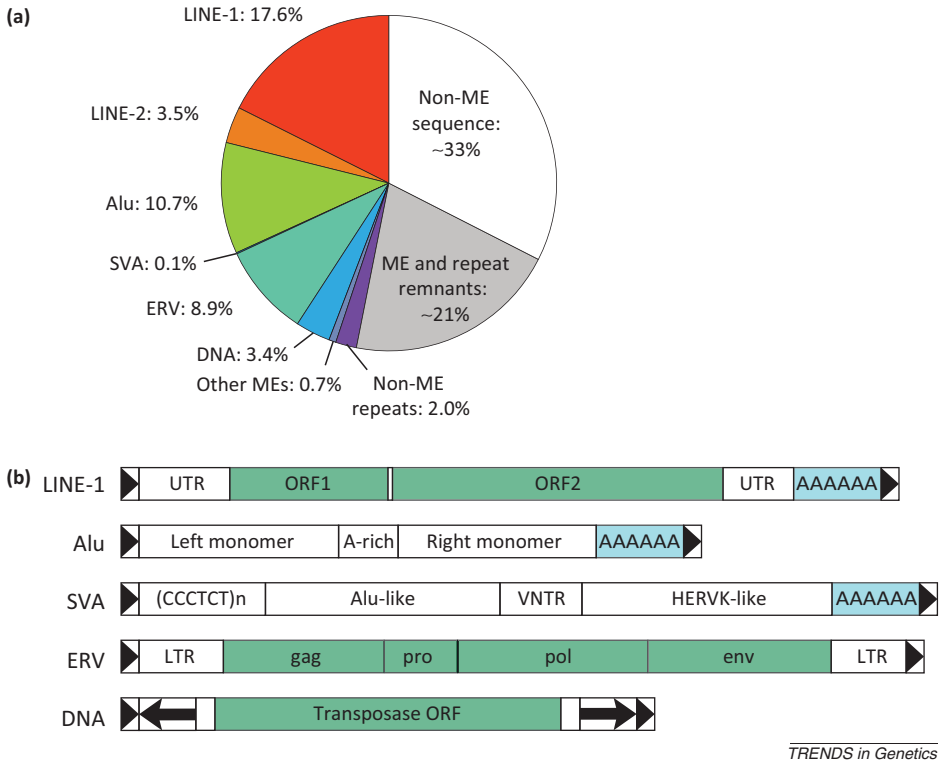

Figure I. Types and prevalence of mobile elements in the human genome. **(a)** Proportion of the human genome (reference sequence hg19) made up of different mobile element classes. **(b)** Structure of the most prevalent mobile elements in the human genome. All these elements create short target site duplications (TSDs) upon insertion into the genome, represented by boxed arrows. Abbreviations: ERV, endogenous retroviruses; LINE, long interspersed elements; LTR, long terminal repeat; ME, mobile element; ORF, open reading frame; UTR, untranscribed region.

only DNA fragments that contain the targeted MEIs will be amplified. In the 'Transposon-Seq' method [24], two rounds of nested hemispecific PCR are used to increase the specificity and to allow the sequence read to cover the junction between the MEI and the flanking genomic region (MEI–genome junction). In mobile element scanning (ME-Scan) [25], the MEI-specific primer for the first-round hemispecific PCR is biotinylated, which allows the amplification products to be easily isolated. In a slight variation of the original protocol, a single-primer extension step rather than PCR is used to amplify L1 insertions [26]. Next, anchored degenerate primers are used in separate PCR reactions to further reduce genomic complexity. Unlike methods based on TD, the retrotransposon capture sequencing (RC-Seq) method enriches MEI-containing DNA fragments by microarray hybridization [27]. Using probes that are complementary to full-length retrotransposons (*Alu*, L1, and SVA), genomic DNA fragments containing MEIs are captured by probes and subsequently sequenced using HTS.

Box 2. High-throughput pMEI identification in other organisms

The basic targeted MEI sequencing approach is completely flexible: one need only replace the primers specific to one mobile element family in one species to assay insertions of another family in another species. Indeed, the TD method, on which the sequencing approaches are based, was developed to detect *dTph1* polymorphisms in *Petunia hybrida* [17] and has since been used on many MEI families and in many species (most recently in mosquitos, maize, and rice [67–70]). Moreover, two non-human MEI families (*mPing* elements in rice [71] and *Minos* elements in wheat [72]) have been investigated by HTS-based, targeted MEI sequencing methods similar to those used in humans.

In human studies, analyses of the sequence data have been aided by the high-quality human reference genome, although the reference genome is not required. Just as MEIs were identified through TD by the positions of their corresponding products on a polyacrylamide gel, they are now more uniquely identified by the genomic sequences at their insertion sites. By either means, the presence or absence of pMEIs in an individual can be determined. This allows pMEIs to be used as genetic markers and for the study of population dynamics even in the absence of a reference genome, although analyses of the genomic contexts of MEIs would be hampered. The successful study of *Minos* elements in wheat [72] is a case in point, because the wheat genome remains largely unassembled due to its large size, hexaploid composition, and high repeat content [73].

The opportunity to apply targeted MEI sequencing to nonhuman MEIs is ripe. High-throughput library preparation methods have become more general, robust, and flexible, enabling rapid development of targeting strategies that address the peculiarities of different MEI types. The use of mechanical DNA fragmentation, as in ME-Scan [25], eliminates concerns about the distributions of random hexamers and restriction enzyme sites. The availability, capacity, and diversity of HTS platforms have increased as sequencing costs continue to drop. MEI families with lower per-genome copy numbers can be assayed on medium-throughput platforms, or assayed in many more individuals for similar cost by multiplexing samples. Longer read lengths will allow better identification and mapping of MEIs as well as insights into the internal element sequences themselves. The lack of information about family-specific sequence characteristics of nonhuman MEIs may introduce uncertainty into the MEI-specific primer design process. However, the extremely high capacity of current sequencing platforms will allow researchers to either characterize the MEI families with low-pass genome sequencing or to simply overcome the uncertainty by obtaining very high coverage using permissive or mixed primer designs. The same strategy will allow fast and reliable identification of MEIs in a family across closely related species for use as phylogenetic or species-specific markers [74].

Overall, HTS-based methods share several advantages. Hundreds of millions of sequencing reads can be generated in one experiment, which allows for indexing and pooling of multiple individuals to reduce cost. The hemispecific PCR-based methods are usually highly specific, and pMEIs from a specific MEI family or subfamily can be detected with high sensitivity. A disadvantage of HTS methods is that MEI-specific primers must be designed for each MEI family. It may also be difficult to target specifically MEI families that differ from related families or remnant copies of older families by only a few diagnostic nucleotide sites. Separate amplification experiments are needed for different MEI families and, in general, pMEI identification can only be performed at the whole-genome level (with the exception of RC-Seq).

In addition to HTS-based approaches, microarray-based methods related to transposon insertion profiling by microarray (TIP-chip) [28,29] have been developed and applied in human MEI detection [30,31]. DNA fragments containing MEIs are first enriched by hemispecific PCR and then fluorescently labeled and hybridized to a genome-tiling array to identify genomic regions adjacent to MEIs [30,31]. Compared with HTS-based methods, microarray-based methods have relatively low throughput, and only one individual can be examined on one microarray. Nevertheless, if a certain subset of the genome (e.g., a single chromosome) is the focus of a project, a customized array can be used to provide coverage for the specific genomic regions more easily than with HTS methods.

Another method for identifying full-length L1 insertions [32] is an adaptation of the fosmid-based paired-end sequencing method that was designed to detect fine-scale (kb-level) structural variation (SV) [33]. (A fosmid is a cloning vector that usually carries approximately 40 kb of genomic DNA insertion.) Candidate loci are identified by comparing the size of the fosmid inserts to the reference genome. Inserts that are approximately 6 kb larger than the reference potentially harbor a nonreference full-length L1 insertion. L1 5′ untranslated region (UTR)-specific oligonucleotide hybridization and Southern blotting are then used to identify fosmid inserts containing full-length L1 insertions. Although this method has high sensitivity for detecting full-length L1s, the size limitation of the fosmid vectors prevents it from detecting pMEIs that are smaller than 6 kb in length, including *Alu*, SVA, and approximately 99.9% of L1s.

Post-sequencing bioinformatics methods

With the advent of affordable whole-genome sequencing, large-scale detection of pMEIs can be accomplished computationally using whole-genome data. Many algorithms designed to identify SVs from whole-genome data can identify pMEIs without any knowledge of MEI characteristics. Read-pair and split-read mapping (Box 3) can both be used in this context. However, standard SV detection methods are not ideal for MEI detection because they do not explicitly take information about MEIs into account. To address this issue, methods have been developed to identify specifically MEIs from whole-genome data. For example, VariationHunter [34] discovers MEIs using a maximum-parsimony algorithm. After initial SV detection, it uses the consensus sequence of MEI families to determine whether a MEI is the mostly likely cause of a putative SV locus.

One type of commonly used algorithm, termed ‘anchored read-pair mapping’, infers MEIs using paired-end reads (Figure 2a). First, the paired-end reads are mapped to the reference genome. Read pairs in which one read maps to a unique genomic position (i.e., ‘anchored’) and the other maps to an MEI consensus sequence indicate the presence of an MEI. Because the exact MEI–genome junction is usually not retrieved in this process, multiple read pairs surrounding one region are required to support identification of an MEI. Implementations of this algorithm include RetroSeq [35] and polymorphic *Alu* insertion recognition (PAIR) [36]. Another algorithm, termed ‘split-read mapping’, is designed to determine the exact MEI position using reads that overlap the MEI–genome junction

Table 1. Targeted methods for high-throughput MEI identification

Name	Fragmentation	Enrichment method	Data generation method	Index for pooling	Advantages	Disadvantages	Refs
Transposon-Seq	Enzymatic digestion	Hemispecific PCR	HTS	Second-round PCR	High throughput, high MEI family specificity and individuals can be pooled after indexing to reduce cost	Only MEIs close to restriction digestion sites can be analyzed and different types of MEI require separate PCR experiments with different MEI-specific primers	[24]
ME-Scan	Mechanical shearing	hemispecific PCR with biotin enrichment	HTS	Ligation	High throughput, high MEI family specificity and individuals can be pooled after indexing to reduce cost	Different types of MEI require separate PCR experiments with different MEI-specific primers	[25]
L1 Display	–	Primer Extension and hemispecific PCR with random hexamers	HTS	Second-round PCR	High throughput, high L1 specificity and individuals can be pooled after indexing to reduce cost	Certain hexamers only amplify a subset of genome	[26]
RC-Seq	Mechanical shearing	Hybridization	HTS	Ligation	High throughput and individuals can be pooled after indexing to reduce cost	Different types of MEI require separate PCR experiments with different MEI-specific primers	[27]
TIP-Chip	Enzymatic digestion	Hemispecific PCR	Tiling array	N/A	High specificity, different MEI families can be examined on the same chip and a subset of the genome can be examined	Relatively low throughput, only one individual can be analyzed per chip and only MEI close to restriction digestion sites can be analyzed	[30]
AIP	Enzymatic digestion	Primer extension with biotin enrichment and hemispecific PCR	Tiling array	N/A	High specificity and a subset of the genome can be examined	Relatively low throughput, only one individual can be analyzed per chip and only MEI close to restriction digestion sites can be analyzed	[31]
Fosmid-based paired-end sequencing	Fosmid library	Allele-specific oligonucleotide hybridization and Southern blot	Fosmid sequencing and/or ATLAS	N/A	Specifically identify full-length L1 insertions	Relatively low throughput. Can only apply to full-length L1s or MEIs larger than 6 kb	[32]

(Figure 2b). If one part of a read can be mapped to a unique position of the genome and the rest of the read is mapped to the MEI sequence, the position of an MEI can be determined. This type of algorithm generally requires longer read lengths than the anchored read-pair methods because it relies on unambiguous mapping of a portion of a read. With long sequence reads, both ends of the MEI–genome junction, as well as the whole MEI sequence, can be recovered from a single read. This greatly increases the accuracy of MEI inference. In the 1000 Genomes Project, for example, the split-read method identified more than 4000 nonreference MEIs in 24 individuals using relatively long reads from the Roche 454 sequencer [37].

With paired-end whole-genome sequencing, reads that can provide information for either anchored read-pair mapping or split-read mapping are usually present in a single data set. Therefore, some MEI discovery methods attempt to integrate these two algorithms, along with other characteristics of MEIs. One example is the TE analyzer (Tea) pipeline [38], in which anchored read-pair mapping is used to infer the MEI candidate and split-reading mapping is used to identify the exact MEI-genome junction. In addition, the presence of an MEI is inferred from signatures of a typical retrotransposition event, such as target site duplications (TSDs) and poly(A) stretches. Several recent analyses, including the 1000 Genomes Project, have employed similar combinations of algorithms in their MEI discovery pipelines [37,39].

Whole-genome versus targeted sequencing

Although whole-genome analyses have provided an overview of all types of pMEI, targeted sequencing has a powerful advantage over whole-genome sequencing when the goal is to identify rare pMEIs across many DNA samples. To detect heterozygous MEIs with high sensitivity using whole-genome sequencing, each human individual must be sequenced at high coverage, which requires approximately 90×10^9 bp of sequencing at 30× coverage

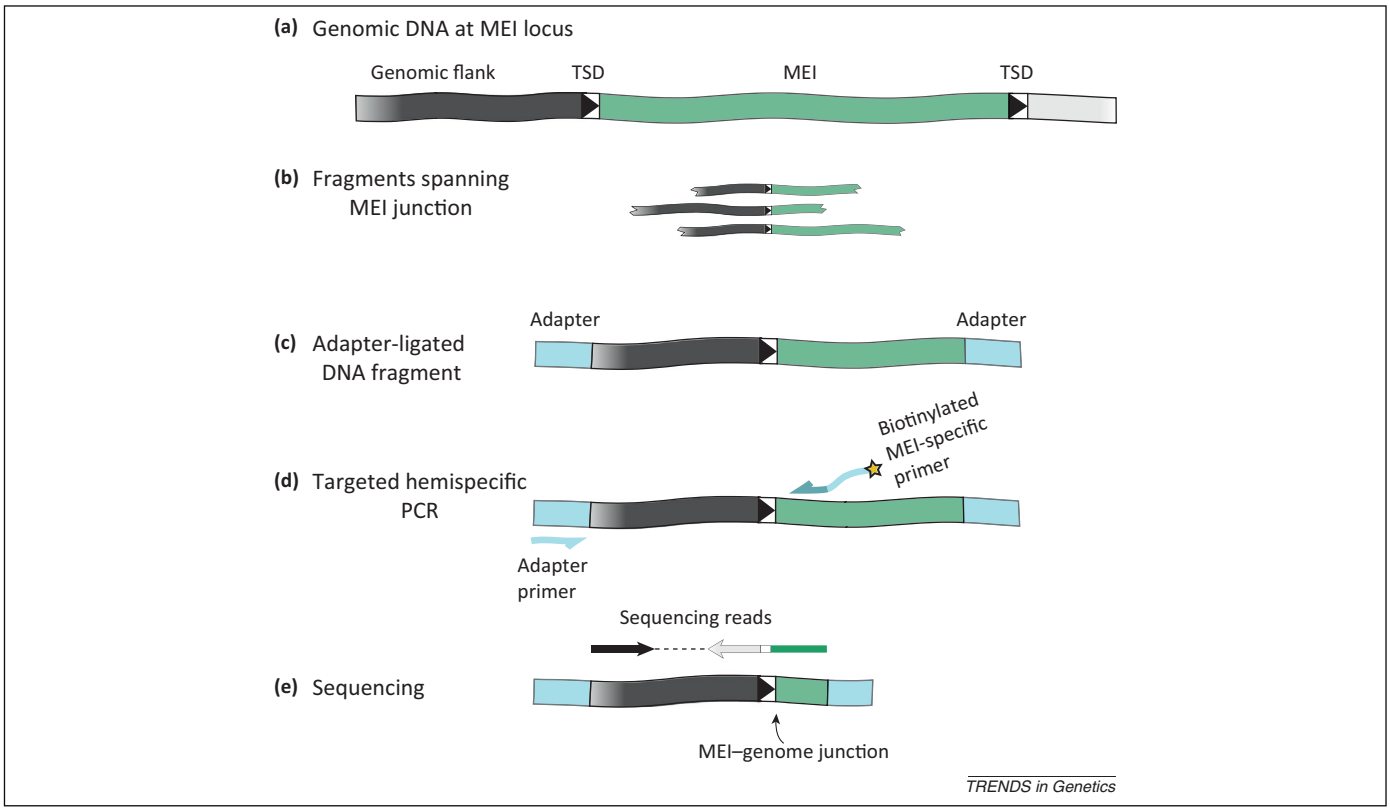

Figure 1. Identification of mobile element insertions (MEIs) using targeted high-throughput sequencing. **(a)** Genomic DNA (gray) containing a MEI (green). Targeted site duplications (TSDs) are shown as boxed arrows. **(b)** DNA is fragmented by mechanical shearing or enzymatic digestion, generating multiple fragments that span the junction of the MEI and the flanking unique genomic sequence. **(c)** If necessary, adapter oligonucleotides (blue) are ligated onto them. **(d)** A primer that anneals to the mobile element family of interest and an adapter primer are used to carry out hemispecific MEI PCR. Here, the targeting primer is biotinylated (star) to facilitate the specific capture and enrichment of fragments carrying the targeted MEIs. This is the method used in mobile element scanning (ME-Scan) [25]. Other methods use hybridization to enrich insertion-carrying fragments, whereas some use carefully designed PCR strategies that amplify only the fragments of interest. **(e)** High-throughput sequencing is used to generate reads from unique genomic sequences immediately adjacent to the insertions in the enriched library or spanning their junctions; both types of read are depicted.

($30 \times 3 \times 10^9$ bp). By contrast, accurate MEI detection using targeted sequencing requires far less sequencing per individual. For example, the Yb8/9 Alu subfamily, which has approximately 6500 copies per individual, accounts for approximately 30% of active mobile elements in the human genome. Assuming paired 100 bp reads and aiming for 120× coverage per MEI copy to allow for variation of coverage across loci, targeted sequencing can detect these elements with high sensitivity using only approximately 0.2% of the reads required for whole-genome sequencing ($120 \times 6500 \times 2 \times 100 = 0.16 \times 10^9$ bp). To exploit this advantage fully requires indexing and pooling samples in groups of approximately 200. This is no longer a limiting factor with current technology, because efficient pooling of more than 500 samples has been demonstrated recently [40]. Furthermore, because most library preparation steps are carried out after pooling, the per-sample cost remains low. The throughput advantage of targeted methods (approximately two orders of magnitude) may be used to process hundreds or thousands of samples quickly, or to assay fewer samples with greatly increased sensitivity, a critical factor for detecting low-prevalence insertions in somatic tissues or tumors.

Human population history studies using genome-wide, population-level MEI data

High-throughput MEI-profiling methods have opened the door to a new dimension in the study of mobile element biology and human population genetics. Although the sequencing of a single genome provided insight into the activity of mobile elements in our evolutionary past (e.g., [1,16]), it could not illuminate the patterns of genetic variation that mobile elements create across individuals and populations. Limited numbers of polymorphic *Alu* and L1 insertion loci, ascertained first in European samples, helped to shed light on the pattern of MEI-generated genetic diversity across populations [8,41]. Compared with these earlier efforts, high-throughput MEI-profiling approaches can identify thousands of pMEIs of all frequency classes across hundreds of individuals without ascertainment bias, with a much greater potential to infer evolutionary history [24–26,39,42]. These studies have provided an unprecedented view of the breadth of variation generated by pMEIs.

An analysis of the data generated by the 1000 Genomes Project from 185 individuals identified 7380 polymorphic MEIs, including 5370 insertions that are not present in the human reference genome [37]. Approximately one-third of these (2649) had not been observed in previous studies. Approximately 85% were due to *Alu* insertions, 12% were L1 elements, and 3% were SVAs. These proportions mirror the relative retrotransposition rates estimated for the three element classes (0.039, 0.0056, and 0.002 insertions per genome per generation, respectively). MEI frequency spectra for African, Asian, and European samples show more pMEIs in African samples, especially at the lower-frequency

Box 3. Read-pair and split-read methods for SV detection

The read-pair algorithm uses paired-end reads that are sequenced from the two ends of a DNA fragment. The size of the DNA fragment is usually larger than the length of the two sequencing reads; therefore, the middle of the DNA fragment is usually not sequenced. After the reads are mapped to a reference genome, the reads that have a nonstandard mapping pattern (e.g., too far apart, not in the correct direction, etc.) are used to infer SV in the genome. In the split-read algorithm, the goal is to identify sequence reads that contain the exact breakpoints of SVs. The reads are generally split into two sections and mapped separately to a reference genome. In contrast to the read-pair algorithm, the split-read algorithm does not require paired-end reads. However, long sequencing reads are needed to map accurately a section of the read to the reference. Therefore, the appropriate use of these algorithms partly depends on the format of the HTS data. For current HTS systems, reads from the 454 GS FLX+ of Roche and the PacBio RS of Pacific Biosciences are more suitable for the split-read algorithm, whereas reads from the HiSeq 2000 of Illumina and the SOLiD 5500 of Life Technology are more suitable for the read-pair algorithm.

end. Principal components analyses grouped individuals according to their continents of origin, indicating stratification of MEI allele frequencies across populations. The 1000 Genomes Project relied mainly on pooled low-coverage sequencing data (1–3× per individual) from many individuals for MEI identification. This approach is biased towards common insertions because an insertion allele present in multiple individuals will effectively receive high coverage across the pooled data set. Despite this detection bias, most MEIs detected in this study are rare (<10% allele frequency).

In another HTS-based analysis, 4342 non-reference *Alu* insertions were identified in eight individuals [42]. The majority (79%) of these insertions had not been previously observed, suggesting that they are usually rare and could only be detected with the higher sensitivity allowed by high sequencing coverage. Similar to [37], more novel insertions were observed in African individuals than in nonAfricans. A complementary study [39] extended the computational search for L1 insertions in the 1000 Genomes Project data set to 310 individuals. Of the 1016 L1 insertions resulting from this and other studies, 104 were present only in African samples. Such a large number of private alleles (more than detected in any other population) is consistent with the 'Out of Africa' hypothesis, which posits that modern humans originated in Africa and then migrated to populate the rest of the world. The migrating populations experienced a population size bottleneck and therefore lost genetic diversity, leading to the observed pattern of higher genetic diversity in African versus nonAfrican populations [43].

Whole-genome sequence data allow simultaneous analysis of all mobile element types and are unaffected by modest element truncations and mutations that can interfere with targeted approaches. However, cost imposes sharp limits on this approach, especially if high sensitivity for rare insertions and, thus, high coverage sequencing are required. The lower cost, efficiency, and sensitivity of targeted MEI sequencing approaches has allowed researchers to scan rapidly samples of their choosing and to identify thousands of additional pMEIs.

As an example of this approach, ME-Scan was used in a proof-of-principle analysis of two active *Alu* element subfamilies (*Alu*Yb8 and *Alu*Yb9) in four Asian individuals. The scan identified 5053 *Alu* insertions, 487 of which had not been previously observed [25]. In Table 2, data from a subsequent application of ME-Scan (3228 *Alu*Yb8/9 insertions identified in 102 individuals) shows that *Alu*Yb8/9 insertions in Europeans are largely a subset of those observed in African samples. Although rare insertions are more likely to be population specific than are common ones, rare African insertions are more likely to be observed only in African populations. A study targeting active L1 subfamilies in 25 individuals from Africa, Europe, and Japan identified 797 L1 insertion loci [26]. The between-population frequency differences of these pMEIs enabled a parsimony-based analysis to group the samples correctly according to their geographic origin.

A consistent overall pattern emerges from these studies: in each one, hundreds to thousands of novel pMEIs are detected. Most novel pMEIs are rare, as expected for derived alleles under a neutral model of drift in a population. Many MEIs are stratified by population and some appear to be population specific, particularly those in African populations. African populations have more pMEI loci than do nonAfrican populations, and pMEIs tend to have higher frequencies in Africans, resulting in higher heterozygosity in African populations. The higher genetic diversity among Africans is expected under the 'Out of Africa' hypothesis, as described above.

Among the thousands of novel MEIs detected, almost no insertions interrupt protein-coding exons. The 1000 Genomes Project confirmed the presence of just two *Alu* elements inside coding exons out of 5370 nonreference MEIs detected [37]. This implies a rate that is 46 times lower than expected for randomly distributed insertions. In another study, the TIP-Chip method was used to search for novel L1 insertions on the X chromosome in 69 males with presumptively X-linked intellectual disability [30]. No insertions in coding exons were found, although two novel insertions were found in introns of genes related to intellectual disability. The absence of even rare pMEIs from protein-coding exons implies that MEIs are highly disruptive of gene function and eliminated quickly by natural selection.

Another unique property of MEIs is that they have a very low probability of inserting at any particular position in the genome per generation (although certain insertion 'hotspots' have been reported; e.g., [44,45]). Thus, the presence of a pMEI marks the surrounding DNA sequence as a reservoir of ancient genetic diversity. Analysis of many such regions, flagged by polymorphic *Alu* insertions, demonstrated that the effective population size of human ancestors living approximately 2 million years ago was approximately 20 000 [46] [age modified according to improved single nucleotide polymorphism (SNP) mutation rate estimates [47]]. This is a remarkably small number for a species that left remains and artifacts across the entire Old World.

Somatic MEIs in cancers

Retrotransposons have long been known to play a role in cancer etiology (reviewed in [3]). With high-throughput methods, we can now compare tumor and normal tissue

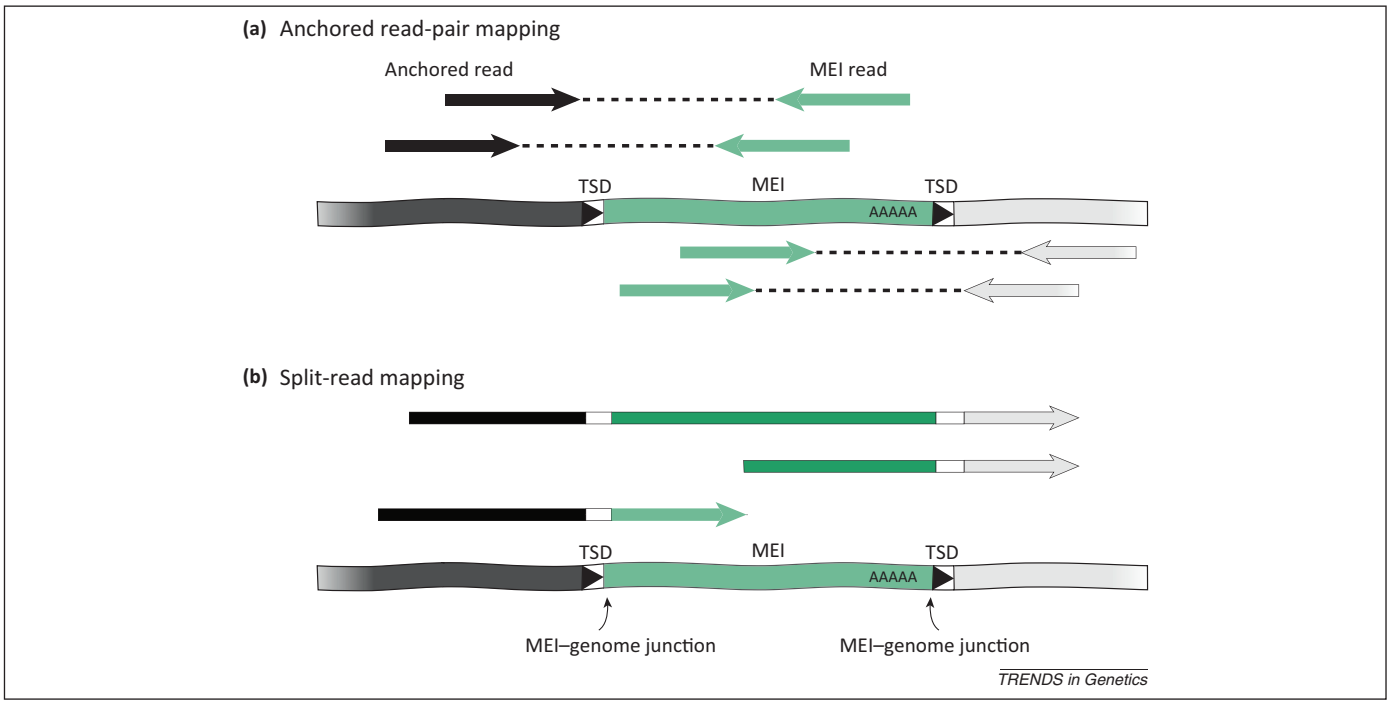

Figure 2. Bioinformatic mobile element insertion (MEI) identification methods. **(a)** Anchored read-pair mapping. Targeted MEIs are shown as green boxes. The flanking genomic sequence is shown as a black bar (5′) or a gray bar (3′), respectively. The paired-end reads are shown in arrows with patterns matching the corresponding genomic regions. The arrows indicate the direction of the read. The middle of the DNA fragment that is not sequenced is shown as broken lines. **(b)** Split-read mapping. Reads are shown in arrows with patterns matching the corresponding genomic regions. The top read illustrates a long sequence read that covers the whole MEI and both targeted site duplications (TSDs). This type of read provides the highest confidence in MEI inference.

pairs from the same individual at the whole-genome level. This allows us to identify tumor-specific MEIs and assess their roles in cancer initiation and progression. Several types of tumor have been examined using high-throughput methods, and the results have revealed important insights into retrotransposon activity in tumors [24,27,38,48].

For example, one study examined *Alu* and L1 insertions in 20 non-small cell lung tumors and their matched normal adjacent tissues as well as ten brain tumors with matched blood leukocytes [24]. Nine tumor-specific insertions were detected in lung tumors, but none were found in brain tumors. Interestingly, the tolerance of the tumors for somatic MEIs was correlated with their methylation status: all six tumors that contained somatic L1 insertions were hypomethylated compared with the tumor samples without L1 insertions.

In another study [38], 194 putative tumor-specific MEIs (mostly L1s) were identified using whole-genome sequences from 43 tumors and blood samples from the same individuals, including samples from multiple myeloma, glioblastoma, and colorectal, prostate, and ovarian cancer. Strikingly, all 194 insertions were found in epithelial tumors (colorectal, prostate, and ovarian cancers), whereas no insertions were found in blood or brain tumors. Almost all of the newly identified L1 insertions were heavily truncated, with an average length of 545 bp, omitting their coding regions. Somatic L1 insertions were identified in 62 annotated genes, including genes with potential tumor suppressor function and genes frequently mutated in colorectal tumors. Although no somatic L1s were identified in the coding region, genes containing L1 insertions in the sense direction showed a significant reduction in expression level, suggesting that the new insertions have a functional impact. Consistent with [24], the insertions appeared to be enriched in regions of the genome that are hypomethylated in tumors. Because MEI insertion rates may vary among tissues, a comparison of tumor tissue with blood cells could overestimate the number of tumor-specific MEIs. Ideally, such studies should compare tumor tissue with normal surrounding tissue to control for potential tissue-specific variation.

In a recent study of colorectal tumors [48], the L1-Seq [26] and RC-Seq methods [27] were used to identify tumor-specific L1 insertions among 16 pairs of tumor/normal matched colorectal samples. Among the L1 insertions that were present in the tumor but not in normal tissues, 73 were validated by locus-specific PCR. Similar to the other studies, these tumor-specific L1s were found in many genes that are frequently mutated in cancers and were heavily truncated (mean size 585 bp). This study also showed that the new L1 insertions likely occurred after cancer initiation (single-cell stage) and were heterozygous among cancer tissues.

These studies have revealed several important features common to tumor-specific somatic MEIs: (i) most tumor-specific somatic L1 insertions are heavily truncated; (ii) hypomethylation is correlated with the increase in somatic MEI insertions; and (iii) different cancer types and/or samples vary greatly in their receptiveness to retrotransposon insertions. Somatic insertions might play a role in tumor progression by affecting the expression of genes that have tumor repressor function. However, tumor initiation and the global changes that allow

Table 2. Allele sharing and allele frequency for 3228 *Alu* insertion polymorphisms[a]

Allele frequency (a.f.)	0 < a.f. < 0.05	0.05 < a.f. < 0.10	a.f. > 0.10
Probability of observing an *Alu* outside Africa, given ascertainment in Africa	0.09	0.25	0.80
Probability of observing an *Alu* in Africa, given ascertainment outside Africa	0.41	0.76	0.97

[a]Samples: African: 53 Bantu and Pygmy; nonAfrican: 49 HapMap TSI and Indian Brahmin.

somatic retrotransposition might have happened before these somatic MEIs appeared.

Somatic MEIs in normal tissues

Before high-throughput methods became available, somatic retrotransposition activity was studied primarily using engineered retrotransposon constructs in cell culture systems or transgenic mice and/or rats. This research suggested somatic retrotransposon activity in early embryogenesis [49–52], tumor cell lines [53,54], and neural progenitor cells [55–57], but little is known about retrotransposon activity in normal somatic tissues. More recently, HTS data have been used to examine several tissue types for somatic MEIs, including brain [27,48,58], liver, and testis [48]. Thus far, somatic insertions have not been identified in most tissues, with the exception of the brain [27,58]. For example, a recent report based on low-coverage sequencing suggested substantial somatic L1 retrotransposition activity in the hippocampus [27]. However, most of the somatic insertion candidates are singletons, and low-coverage candidates from HTS have high false-positive rates. Site-specific validation PCRs of approximately 30 somatic insertion candidates were carried out using a modified two-round PCR protocol on the enriched library, not the original DNA samples. The additional rounds of PCR, in the presence of large numbers of DNA templates, could introduce artifacts that inflate the validation rate.

Another recent study examined L1 insertions in neuronal cells from the cerebral cortex and caudate nucleus [58]. Individual neuronal cells were first isolated from postmortem brains of three neurologically normal individuals. A combination of single-cell multiple-displacement amplification and an L1 profiling procedure similar to L1-Seq [26] identified L1 insertions in 300 neuronal cells (50 cells from the cerebral cortex and 50 from the caudate nucleus for each individual). After excluding known L1 polymorphisms and insertions that were present in other tissues from the same individual, the remaining somatic insertion candidates were subjected to PCR validation. The validation results suggest that most somatic insertion candidates are artifacts, and only five insertions were validated in the 300 cells. After correcting for sensitivity, this result extrapolates to an L1 neuronal somatic insertion rate of one insertion per 25 cells, consistent with the rate of insertions per cell division in the germline. The low somatic insertion rate argues against an essential role for L1s in generating neuronal diversity in these brain regions.

Future directions for high-throughput mobile element biology

The development of high-throughput technologies, along with new bioinformatics tools, has transformed the study of mobile elements in the human genome. As sequencing read lengths continue to increase with improved technology, we can expect further improvement in the accuracy of MEI identification. One pressing question is the sensitivity of current methods to detect insertions that are present in only a small number of cells. Several methods show the potential to identify MEIs present in a single cell or only a small proportion of cells, but carefully designed experiments with different proportions of MEI fragments and/or MEI-containing cells will be needed to address this question fully.

Results from recent studies also highlight the necessity of validation for pMEI candidates, especially for those that are supported by only a few or even one sequence read. In our experience, previously unknown (novel) MEIs observed in only one individual or a sample (singletons) and supported by only by a few sequencing reads (ten or less) have a validation rate of approximately 20% (by locus-specific PCR). A similarly poor replication rate was reported by a study of somatic pMEIs in single cells [58]. Novel MEIs supported by just one sequencing read in a very high-coverage experiment are likely to be false positives, presumably reflecting rare chimeras generated during library preparation. For germline insertions that are present in all somatic tissues of an individual, validation and high-throughput replication studies will determine thresholds that clearly distinguish between true and false positives. Such validation will be more difficult for somatic insertions that are present in only portions of a tissue or even in a single cell.

With population-level, genome-wide pMEI profiles, several long-standing questions related to mobile element biology will soon be answered. One of these is the *de novo* insertion rate of mobile elements, which has been estimated indirectly using phylogenetic and population methods [16,26,59]. However, these approaches will not detect MEIs that were lost soon after integration, and such insertions may have important functional consequences. The *de novo* insertion rate can be directly obtained using trio data, counting new insertions that are present in the offspring but absent in both parents. Because the rate of retrotransposon insertion is likely to be low, a large number of trios must be analyzed to obtain a reliable rate. Methods described here allow simultaneous analysis of hundreds of individuals in a cost-effective fashion, enabling this question to be answered.

Another important issue is the role of somatic MEIs in cancer etiology and brain development. The cancer MEI studies reviewed here provide critical information about the characteristics of cancer-specific MEIs and suggest high retrotransposition activity in certain regions of the brain. Nevertheless, only a few types of cancer and several brain samples were examined, and the numbers of documented MEIs remain relatively small. Do cancer-specific MEIs play a pivotal role in cancer initiation and/or progression in certain types of cancer? If so, what is the mechanism behind it? What are the frequency and impact of brain-specific MEIs? Do these insertions happen in

certain stage(s) of development or certain areas of the brain? These questions may be answered with surveys of large panels of different samples, in combination with *in vitro* analysis of the genes affected by somatic MEIs.

Lastly, whole-genome MEI profiling will provide the groundwork for a better understanding of the genomic control of mobile elements. It is known that host factors, methylation, and small RNAs [small interfering RNAs and Piwi-interacting RNA (piRNAs)] all play important roles in governing retrotransposon activities in the genome (reviewed in [5,60,61]). Recent studies have demonstrated great variability of pMEIs among human individuals as well as populations [37,39]. Therefore, it is likely that retrotransposon regulation activity also varies among human individuals and/or populations. With population-level MEI profiling, we can begin to explore the correlation between retrotransposon activity and host controlling mechanisms, eventually elucidating how mobile elements are regulated in the human genome.

Acknowledgments

The authors thank the three anonymous reviewers for their constructive and valuable comments. The authors declare no competing financial interests. This study was supported by a grant from the National Institutes of Health (GM059290). J.X. is supported by the National Human Genome Research Institute (R00HG005846).

References

1 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
2 de Koning, A.P. *et al.* (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* 7, e1002384
3 Hancks, D.C. and Kazazian, H.H., Jr (2012) Active human retrotransposons: variation and disease. *Curr. Opin. Genet. Dev.* 22, 191–203
4 Cordaux, R. and Batzer, M.A. (2009) The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* 10, 691–703
5 Burns, K.H. and Boeke, J.D. (2012) Human transposon tectonics. *Cell* 149, 740–752
6 Nishihara, H. and Okada, N. (2008) Retroposons: genetic footprints on the evolutionary paths of life. *Methods Mol. Biol.* 422, 201–225
7 Ray, D.A. *et al.* (2006) SINEs of a nearly perfect character. *Syst. Biol.* 55, 928–935
8 Watkins, W.S. *et al.* (2003) Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. *Genome Res.* 13, 1607–1618
9 Xing, J. *et al.* (2007) Mobile DNA elements in primate and human evolution. *Am. J. Phys. Anthropol.* 134 (Suppl. 45), 2–19
10 Kriegs, J.O. *et al.* (2006) Retroposed elements as archives for the evolutionary history of placental mammals. *PLoS Biol.* 4, e91
11 Shimamura, M. *et al.* (1997) Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature* 388, 666–670
12 Wang, J. *et al.* (2006) dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum. Mutat.* 27, 323–329
13 Levy, S. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.* 5, e254
14 Wang, J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature* 456, 60–65
15 Bentley, D.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59
16 Xing, J. *et al.* (2009) Mobile elements create structural variation: analysis of a complete human genome. *Genome Res.* 19, 1516–1526
17 Van den Broeck, D. *et al.* (1998) Transposon Display identifies individual transposable elements in high copy number lines. *Plant J.* 13, 121–129
18 Roy, A.M. *et al.* (1999) Recently integrated human Alu repeats: finding needles in the haystack. *Genetica* 107, 149–161
19 Mamedov, I.Z. *et al.* (2005) Whole-genome experimental identification of insertion/deletion polymorphisms of interspersed repeats by a new general approach. *Nucleic Acids Res.* 33, e16
20 Sheen, F.M. *et al.* (2000) Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. *Genome Res.* 10, 1496–1508
21 Ovchinnikov, I. *et al.* (2001) Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. *Genome Res.* 11, 2050–2058
22 Badge, R.M. *et al.* (2003) ATLAS: a system to selectively identify human-specific L1 insertions. *Am. J. Hum. Genet.* 72, 823–838
23 Buzdin, A. *et al.* (2002) A technique for genome-wide identification of differences in the interspersed repeats integrations between closely related genomes and its application to detection of human-specific integrations of HERV-K LTRs. *Genomics* 79, 413–422
24 Iskow, R.C. *et al.* (2010) Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* 141, 1253–1261
25 Witherspoon, D.J. *et al.* (2010) Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. *BMC Genomics* 11, 410
26 Ewing, A.D. and Kazazian, H.H., Jr (2010) High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res.* 20, 1262–1270
27 Baillie, J.K. *et al.* (2011) Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 479, 534–537
28 Wheelan, S.J. *et al.* (2006) Transposon insertion site profiling chip (TIP-chip). *Proc. Natl. Acad. Sci. U.S.A.* 103, 17632–17637
29 Gabriel, A. *et al.* (2006) Global mapping of transposon location. *PLoS Genet.* 2, e212
30 Huang, C.R. *et al.* (2010) Mobile interspersed repeats are major structural variants in the human genome. *Cell* 141, 1171–1182
31 Cardelli, M. *et al.* (2012) Alu insertion profiling: array-based methods to detect Alu insertions in the human genome. *Genomics* 99, 340–346
32 Beck, C.R. *et al.* (2010) LINE-1 retrotransposition activity in human genomes. *Cell* 141, 1159–1170
33 Tuzun, E. *et al.* (2005) Fine-scale structural variation of the human genome. *Nat. Genet.* 37, 727–732
34 Hormozdiari, F. *et al.* (2010) Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 26, i350–i357
35 Wong, K. *et al.* (2010) Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol.* 11, R128
36 Sveinbjornsson, J.I. and Halldorsson, B.V. (2012) PAIR: polymorphic Alu insertion recognition. *BMC Bioinformatics* 13 (Suppl. 6), S7
37 Stewart, C. *et al.* (2011) A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.* 7, e1002236
38 Lee, E. *et al.* (2012) Landscape of somatic retrotransposition in human cancers. *Science* 337, 967–971
39 Ewing, A.D. and Kazazian, H.H., Jr (2011) Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res.* 21, 985–990
40 Reyes, A. *et al.* (2012) IS-seq: a novel high throughput survey of *in vivo* IS6110 transposition in multiple *Mycobacterium tuberculosis* genomes. *BMC Genomics* 13, 249
41 Witherspoon, D.J. *et al.* (2006) Human population genetic structure and diversity inferred from polymorphic L1(LINE-1) and Alu insertions. *Hum. Hered.* 62, 30–46
42 Hormozdiari, F. *et al.* (2011) Alu repeat discovery and characterization within human genomes. *Genome Res.* 21, 840–849
43 Harpending, H. and Rogers, A. (2000) Genetic perspectives on human origins and differentiation. *Annu. Rev. Genomics Hum. Genet.* 1, 361–385
44 Wimmer, K. *et al.* (2011) The NF1 gene contains hotspots for L1 endonuclease-dependent *de novo* insertion. *PLoS Genet.* 7, e1002371
45 Conley, M.E. *et al.* (2005) Two independent retrotransposon insertions at the same site within the coding region of BTK. *Hum. Mutat.* 25, 324–325
46 Huff, C.D. *et al.* (2010) Mobile elements reveal small population size in the ancient ancestors of *Homo sapiens*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2147–2152
47 Roach, J.C. *et al.* (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328, 636–639
48 Solyom, S. *et al.* (2012) Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res.* 22, 2328–2338

49 Ostertag, E.M. *et al.* (2002) A mouse model of human L1 retrotransposition. *Nat. Genet.* 32, 655–660
50 Prak, E.T. *et al.* (2003) Tracking an embryonic L1 retrotransposition event. *Proc. Natl. Acad. Sci. U.S.A.* 100, 1832–1837
51 Kano, H. *et al.* (2009) L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes Dev.* 23, 1303–1312
52 Garcia-Perez, J.L. *et al.* (2007) LINE-1 retrotransposition in human embryonic stem cells. *Hum. Mol. Genet.* 16, 1569–1577
53 Garcia-Perez, J.L. *et al.* (2010) Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. *Nature* 466, 769–773
54 Moran, J.V. *et al.* (1996) High frequency retrotransposition in cultured mammalian cells. *Cell* 87, 917–927
55 Coufal, N.G. *et al.* (2009) L1 retrotransposition in human neural progenitor cells. *Nature* 460, 1127–1131
56 Muotri, A.R. *et al.* (2010) L1 retrotransposition in neurons is modulated by MeCP2. *Nature* 468, 443–446
57 Muotri, A.R. *et al.* (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* 435, 903–910
58 Evrony, G.D. *et al.* (2012) Single-neuron sequencing analysis of l1 retrotransposition and somatic mutation in the human brain. *Cell* 151, 483–496
59 Cordaux, R. *et al.* (2006) Estimating the retrotransposition rate of human Alu elements. *Gene* 373, 134–137
60 Levin, H.L. and Moran, J.V. (2011) Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.* 12, 615–627
61 Castaneda, J. *et al.* (2011) piRNAs, transposon silencing, and germline genome integrity. *Mutat. Res.* 714, 95–104
62 Brouha, B. *et al.* (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. U.S.A.* 100, 5280–5285
63 Dewannieux, M. *et al.* (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* 35, 41–48
64 Hancks, D.C. *et al.* (2011) Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum. Mol. Genet.* 20, 3386–3400
65 Dewannieux, M. *et al.* (2006) Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res.* 16, 1548–1556
66 Lee, Y.N. and Bieniasz, P.D. (2007) Reconstitution of an infectious human endogenous retrovirus. *PLoS Pathog.* 3, e10
67 Yang, G. *et al.* (2012) ATon, abundant novel nonautonomous mobile genetic elements in yellow fever mosquito (*Aedes aegypti*). *BMC Genomics* 13, 283
68 Smith, A.M. *et al.* (2012) TCUP: a novel hAT transposon active in maize tissue culture. *Front. Plant Sci.* 3, 6
69 Zerjal, T. *et al.* (2012) Maize genetic diversity and association mapping using transposable element insertion polymorphisms. *Theor. Appl. Genet.* 124, 1521–1537
70 Dong, H.T. *et al.* (2012) A Gaijin-like miniature inverted repeat transposable element is mobilized in rice during cell differentiation. *BMC Genomics* 13, 135
71 Naito, K. *et al.* (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* 461, 1130–1134
72 Yaakov, B. and Kashkush, K. (2012) Mobilization of Stowaway-like MITEs in newly formed allohexaploid wheat species. *Plant Mol. Biol.* 80, 419–427
73 Winfield, M.O. *et al.* (2012) Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol. J.* 10, 733–742
74 Grzebelus, D. *et al.* (2011) Dynamics of Vulmar/VulMITE group of transposable elements in Chenopodiaceae subfamily Betoideae. *Genetica* 139, 1209–1216