Published in final edited form as:

Methods. 2009 November ; 49(3): 227–235. doi:10.1016/j.ymeth.2009.05.013.

Plug and play modular strategies for synthetic retrotransposons

Wenfeng An^{a,*}, Edward S. Davis^b, Tina L. Thompson^a, Kathryn A. O'Donnell^b, Chih-Yung Lee^b, and Jef D. Boeke^b

^aSchool of Molecular Biosciences and Center for Reproductive Biology, Washington State University, Pullman, WA 99164, USA

^bThe High Throughput Biology Center and Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

Abstract

Recent progress in L1 biology highlights its role as a major driving force in the evolution of mammalian genome structure and function. This coincides with direct confirmation of the preponderance of long interspersed elements in mammalian genomes at the nucleotide level by large scale sequencing efforts. Two assay systems have been prominently featured in L1 studies over the past decade, which are used to assess L1 activities in cultured cells and transgenic mice respectively. However, constructing retrotransposon assay vectors and subsequent mapping of integration sites remain technically challenging aspects of the field. Synthetic biology approaches have changed the playing field with regard to the strategic design of retrotransposons. To streamline the construction and optimization of synthetic retrotransposons, we have implemented a highly efficient modular design for L1 vectors allowing “plug and play” swapping of individual modules as new knowledge is gained and optimization of constructs proceeds. Seven functional modules are divided by strategically placed unique restriction sites. These are utilized to facilitate module exchange and construction of L1 vectors for gene targeting, transgenesis and cell culture assays. A “double *Sfi*” strategy utilizing two non-complementary overhangs allows insert swapping to be carried out with a single, robust restriction/ligation cycle. The double-*Sfi* strategy is generic and can be applied to many other problems in synthetic biology or genetic engineering. To facilitate genomic mapping of L1 insertions, we have developed an optimized inverse PCR protocol using 4-base cutters and step-down cycling conditions. Using this protocol, *de novo* L1 insertions can be efficiently recovered after a single round of PCR. The proposed modular design also incorporates features allowing streamlined insertion mapping without repeated optimization. Furthermore, we have presented evidence that efficient L1 retrotransposition is not dependent on pCEP4 conferred autonomous replication capabilities when a shortened puromycin selection protocol is used, providing a great opportunity for further optimization of L1 cell culture assay vectors by using alternative vector backbones.

Keywords

LINE-1; retrotransposon; cell culture; transgenic mouse; modular vector; inverse PCR; EBNA-1; pCEP4

1. Introduction

Significant strides have been made in the mammalian retrotransposon field since long interspersed elements (LINEs) were first described in mammals [1]. The dominance of LINEs in mammalian genomes is now confirmed directly at the nucleotide level by recent genome

*Corresponding author. Fax: +1 509 335 9688. wenfeng_an@wsu.edu (W. An).

sequencing efforts, including representative species from placental mammals, marsupials and egg-laying monotremes. In the human genome, LINE-1s (also known as L1s) account for ~17% of the mass, with the majority of L1s being 5' truncated. Classified as non-LTR retrotransposons, full-length L1s are typically 6~7 kb, encompassing an internal promoter in the 5'UTR, two non-overlapping open reading frames (ORF1 and ORF2), and a weak polyadenylation signal in the 3'UTR. Recent progress in L1 biology highlights its role as a major driving force in mammalian genome evolution [2,3]. Two fundamental assay systems, used to assess L1 activities in cell culture and in transgenic animals, respectively [4,5], have been prominently featured in this discovery process over the past decade.

The first active L1 was isolated in 1991 [6] and its retrotransposition activity was subsequently verified in a cell-culture-based L1 functional assay [4]. The initial retrotransposition indicator cassette contains a copy of the neomycin resistance gene disrupted by an antisense intron (*neoAI*) and the level of retrotransposition is reported by colony formation after drug selection [4]. An enhanced green fluorescent protein-based reporter (*gfpAI*) has since been developed [7]. L1 functional assays in cultured cells have drastically propelled the field forward, leading to the identification of essential and non-essential L1 sequences for its retrotransposition, the identification of other active L1s, and a wide range of impacts of L1 retrotransposition on mammalian genomes [8]. These assays also make it possible to probe the effect of cellular factors on L1 retrotransposition and to provide mechanistic insights into L1 replication [9].

Another boost for L1 research came from mouse models for L1 retrotransposition, which have paved the way for a thorough understanding of L1 biology in a living organism and toward L1-based tools for random *in vivo* mutagenesis. When placed under the control of its endogenous 5'UTR promoter, a human L1 transgene is found to express exclusively in mouse testis and ovary, and its retrotransposition can be detected in the male germ line [5]. Such tissue-specific expression is consistent with previous studies on the expression of endogenous mouse and human L1 elements [10–12]. However, in a subsequent study using a similar human L1 transgene, retrotransposition was not only detected in germ cells but also in neuronal cells [13], raising a possible role of L1 somatic retrotransposition in neuronal diversity. Both human and mouse L1 transgenes can readily retrotranspose in mouse somatic cells when they are regulated by heterologous promoters [14–16]. Germ line retrotransposition frequency as high as one in every three animals has also been achieved with a synthetic mouse L1 transgene, *ORFeus* [15].

There are two primary challenges when working with L1 plasmids containing retrotransposons for either cell culture or animal experiments. The first challenge is frequently encountered during plasmid construction. The relative large size of typical retrotransposon vectors (~20 kb) makes subcloning technically demanding as DNA fragments larger than 10 kb are notoriously inefficient during almost all subcloning stages such as DNA recovery, ligation and transformation. Choice of unique 6-base cutters is limited; Eight-base cutters are prized for assembly of complex L1 constructs but frequently they are either absent from the recipient plasmid or inconveniently positioned. Although it is often desirable to swap certain functional elements in and out of an existing L1 vector, such substitution remains an inefficient and time-consuming practice unless design principles are carefully considered ahead of time. The second challenge is the lack of a standard protocol for mapping retrotransposition events once the engineered L1 is introduced into cultured cells or animals.

Here we present strategies aiming to overcome aforementioned obstacles. In section 2, we detail a blueprint for streamlining L1 vector design. Sequence components of L1 vectors are modularized, and strategically placed restriction sites are used to facilitate cassette swapping for tailored research needs. In section 3, we describe a step-by-step inverse PCR (iPCR) protocol that we have found to be useful for mapping *de novo* L1 insertions in both cultured

cells and transgenic animals, especially in DNA samples containing a complex population of individual retrotransposition events.

2. Modular design of L1 vectors for cell culture and animal studies

2.1 General considerations

Several synthetic biology standards for assembling complex series of “standardized parts” such as BioBricks [17] have been proposed, and some have been adopted by large segments of the synthetic biology community (e.g. the Registry of Standard Biological Parts; see http://partsregistry.org/Main_Page). The main drawback to “BioBricking” the various components of retrotransposons is that the retrotransposons assemblies can constitute combinations of ten or more “parts” and thus it is advantageous to be able to swap out individual parts one at a time. Therefore, we have adopted a strategy that uses a series of relatively rare cutting and well-behaved restriction enzyme sites located at strategic positions. Current practices and conventions have been carefully reviewed.

For testing L1 retrotransposition in cultured cells, a marked L1 element is typically subcloned into pCEP4 or its derivatives (10 kb backbone; [4]). pCEP4, marketed by Invitrogen, was initially chosen because it carries the Epstein-Barr Virus replication origin and nuclear antigen EBNA-1 that jointly permit its extrachromosomal replication in primate cell lines [18]. Additionally, this vector encodes the hygromycin B resistance gene that can be used to enrich transfected cells by drug selection. To facilitate this subcloning process, the first L1 element tested was engineered to have a unique *NotI* site upstream of its 5'UTR and a unique *BamHI* site downstream of the reporter cassette, and was subsequently inserted between *NotI* and *BamHI* sites of pCEP4 ([4]; Figure 1A). The *BamHI* site at the 3' end echoes our original Ty1 design [19]. In this configuration, pCEP4 provides a heterologous CMV promoter and a SV40 polyadenylation signal for L1 in addition to its native promoter, 5'UTR. The promoter activity of 5'UTR can be tested independently if a *NotI*-*BamHI* fragment of L1 is subcloned into a modified pCEP4 vector from which the CMV promoter upstream of *NotI* site has been previously deleted [4]. Currently, similar strategies have been adopted to test L1 isolates from at least three different species, including human, mouse, and rat [4,20,21], and zebrafish L2 isolates [22]. However, the subcloning of other L1 sequences or reporters into a recipient pCEP4 plasmid remains an exceedingly involved process requiring numerous intermediate steps due to the lack of convenient or unique restriction sites. It is precisely these clumsy and time consuming steps that our strategies aim to eliminate or reduce.

The initial mouse models of L1 utilized a human L1 sequence in pCEP4 derivatives, and the SV40 polyadenylation signal downstream of the *BamHI* site from pCEP4 was used as the transcriptional terminator ([5]; Figure 1A). In one scheme, the transgene is subcloned into a pCEP4 derivative in which the CMV promoter in pCEP4 has been replaced by a mouse RNA polymerase II large subunit promoter (pPolII); thus, its expression is regulated by both pPolII and 5'UTR. In another scheme, the transgene is subcloned into a pCEP4 derivative lacking the CMV promoter so that its expression is solely regulated by 5'UTR. Both these transgenes lack internal *SalI* site, and are flanked by two *SalI* sites from the pCEP4 backbone. Thus, both can be released from the vector as a *SalI* fragment for pronuclear microinjection ([5]; Figure 1A). Later our group employed a synthetic mouse L1 element ([23]; *ORFeus*) as transgenes, which were subcloned into pBluescript II (Stratagene) for reduced backbone size (3kb; [14,15]). One such *ORFeus* line features a constitutive, heterologous CAG promoter [15]; the other includes a transcriptional stop cassette between the constitutive promoter and L1 coding sequences, allowing tempopsatial control of L1 activity in transgenic mice by a Cre-loxP system ([15]; *ORFeusLSL* in Figure 1C). An additional *NotI* site was engineered to the end of these *ORFeus*-based transgenes, so they can be linearized by a single *NotI* digestion reaction.

To streamline L1 subcloning, we are implementing a modular vector design containing seven functional modules (Figure 1B). Strategically placed restriction sites are used to demarcate individual modules and to facilitate vector construction and module swapping. The repertoire for each of these functional modules is rapidly expanding. With suitable recipient plasmid backbones, the same set of modules can be used for multiple testing conditions (Figure 1B): as a gene targeting vector (e.g. modules 1–7 in pBluescript II), as a transgene (e.g. modules 2–6 in pBluescript II) or as a cell culture vector (e.g. modules 2,4,5 in pCEP4 or derivatives). Thus, once an L1 vector is made conforming to this design, it can be easily subcloned into different recipient vectors for different purposes. New versions of any functional module can also be efficiently “plugged” into an existing vector by using flanking restriction sites. In addition, we have constructed and tested several vectors that can be used as the starting point for constructing specific L1 vectors (Figure 1C).

2.2. Strategically positioned restriction sites

Each module is flanked by a group of restriction sites that should allow selective exchange of any functional elements using pairwise combinations of sites, all of which have noncompatible overhangs. To increase its probability of each site’s uniqueness in the final vector, we place a minimum of one >6-base restriction site. Such restriction sites include *RsrII* (7-base), *AscI*, *AsiSI*, *FseI*, *NotI*, *PacI*, *PmeI*, *SbfI*, *SfiI*, *SgrAI*, *SwaI* (all 8-base). These rare cutters are supplemented by several 6-base cutters that generate either sticky or blunt ends to maximize subcloning flexibility. To enhance compatibility with existing L1 plasmids from various labs, we preserve the positioning of two landmark restriction sites: One *NotI* site located immediately upstream of the L1 promoter, and a *BamHI* site positioned downstream of the reporter but upstream of the L1 polyadenylation signal. It should be noted that several rare cutters (e.g. *NotI*, *PacI*, *SfiI* and *SwaI*) are intentionally represented in the modularized vector more than once. The utility of these four rare cutters will be discussed in the following sections.

2.3. Individual functional modules

In the blueprint, seven functional modules are presented in a 5’→3’ linear order according to their predicted positions in all possible L1 vectors (Figure 1B). Module 1 (5’ targeting homology arm) and module 7 (3’ arm) are present in gene targeting vectors only (see 2.4.1). Module 2 is reserved for an L1 promoter, which can be either a heterologous promoter or the native 5’UTR promoter. Module 3 (optional transcriptional control and/or selection module; see 2.3.1 for details) is typically used in gene targeting vectors and some transgenes. Module 4 contains L1 coding sequences ORF1 and ORF2. Module 5 (“reporter”) is designed to accommodate different retrotransposition indicator cassettes (*neoAI*, *gfpAI*, etc) and other utility elements, such as gene trapping cassettes (see 2.3.2 for details). Module 6 encodes a polyadenylation signal, which is usually derived from pCEP4 for cell culture vectors. Two of these modules are highlighted below.

2.3.1. Control module (module 3)—Module 3 is configured to serve multiple functions. A typical application of module 3 is to include an exogenous positive selectable/screenable marker for gene targeting applications, such as *neo* or β -*geo*, which allows effective selection of desired mouse ES clones. Such a selectable marker usually carries a strong polyadenylation signal(s) and is flanked by a pair of loxP sites. The presence of a strong polyadenylation signal(s) between the L1 promoter (module 2) and coding sequences (module 4) prevents L1 expression until module 3 has been removed by Cre recombinase. These features make it possible to achieve conditional control of L1 expression by using a Cre-loxP system in transgenic animals [14]. The Cre-loxP system allows spatiotemporal control of L1 expression by simply breeding an L1 transgenic line to various Cre-expressing mouse lines. An alternative approach for conditional control of L1 expression is to use a conditional promoter element at module 2 in conjunction with chemical inducers.

2.3.2. Reporter module (module 5) and double *SfiI* sites—A salient feature of the reporter module (which itself can be the subject of intense multimodular engineering) is its two non-complementary flanking *SfiI* sites. The recognition sequence of *SfiI* is GGCCNNNN/NGGCC (where N is any base and "/" is the point of cleavage). The modular vector incorporates two *SfiI* sites that differ in the spacer region (Left "*SfiI*_L" = GGCCAAAA/TGGCC and Right "*SfiI*_R" = GGCCTGTC/AGGCC). This design allows directional cloning of the reporter module by a single *SfiI* digestion exploiting the noncompatible 3-base overhangs (Figure 2). pESD202 is one such *SfiI*-ready recipient vector (Figure 1C).

2.4. Mix-and-match for different applications

2.4.1. Gene targeting cassette (modules 1–7) and double *PacI* sites—The inclusion of 5' and 3' homology arms in the modular design allows targeting of an enclosed L1 element into a specific genomic locus through homologous recombination (Figure 1B). Although often studied for convenience on episomes, L1 is naturally a chromosome-resident element. Targeting of a marked L1 into a specific chromosomal locus allows its activity and regulation to be studied in a more native state. Two applications of L1 gene targeting can be conceived. The first is to study a targeted L1 in somatic cell lines, which is greatly facilitated by targeting L1 to an X-linked genetic locus like *HPRT* in a male cell line. A second application is to study a targeted L1 in live animals by homologous recombination in mouse ES cells. In the modular design, gene-specific 5' and 3' arms are flanked externally by two *PacI* sites. Thus, a single *PacI* digestion can release an L1 gene targeting cassette from the pBluescript II backbone. An intermediate vector carrying 5' and 3' homology arms for the human *HPRT* gene has been constructed (Figure 1C), which was successfully used to target an HIV-1 provirus into the *HPRT* locus of a human cell line [24].

2.4.2. Transgene cassette (modules 2–6) and double *NotI* sites—The transgene cassette consists of a promoter, an optional control module, L1 ORFs, a reporter, and a polyadenylation signal (Figure 1B). It is flanked by a pair of *NotI* sites. Thus, the transgene can be linearized by a single *NotI* digestion for pronuclear microinjection. pBluescript II is the preferred backbone because of its smaller size.

2.4.3. Cell culture assay cassette (modules 2, 4 and 5)—For cell culture experiments, an L1 element is typically subcloned as a *NotI*-*Bam*HI fragment into pCEP4 [8] or its puromycin-resistant derivative pCEP-Puro ([7]; Symer D.E. and Boeke J.D., unpublished data). The latter features hastened killing of non-transfected cells in just three days by puromycin selection and thus has become the vector of choice. It should be noted that this subcloning strategy renders the SV40 polyadenylation signal from pCEP4 as the default polyadenylation signal for the L1. Alternatively, it is also possible to use a non-replicating plasmid backbone if puromycin selection is employed to enrich transfected cells or if a transient assay procedure is followed (see section 2.6). The control module (module 3) is usually not used in cell culture assays but it can be inserted into an existing cell culture vector between modules 2 and 4 to derive Cre-loxP ready gene targeting vector or transgenes.

2.5. Other important features

2.5.1. Embedded iPCR priming sites—iPCR can be used to efficiently map L1 integration events (see section 3). To avoid repeated optimization efforts that are typically required if new primers and/or restriction enzymes are to be used for each new L1 vector, two iPCR priming sites are incorporated in the modular vector (Figure 1B). JB8822 is a reverse primer placed upstream of the *SfiI*_L site, and JB8897 is a forward primer that is positioned downstream of the *SfiI*_R site. These two primers have been extensively tested in our optimized iPCR-based protocol ([14, 15]; see section 3). In addition, when designing reporter/utility cassettes, we routinely screen the 3' end of the L1 vector sequence for several 4-base cutters

that are featured in our iPCR protocol, such as *MspI* and *TaqI*. Such restriction sites will be eliminated, if at all possible, from the 3' end of the L1 vector between JB8822 and the polyadenylation cleavage site. In doing so, the same iPCR protocol can be used for different L1 vectors without testing additional primers and restriction enzymes.

2.5.2. Tandem 5x *SwaI* sites—We incorporate five copies of the rare cutting *SwaI* sites in tandem between modules 6 and 7 as an additional measure to ensure iPCR specificity. These *SwaI* sites are located downstream to the polyadenylation signal cleavage site and, therefore, should not be present in *de novo* L1 insertions. After the ligation step in iPCR (see section 3.3), a single *SwaI* digestion selectively re-linearizes DNA molecules derived from donor L1 elements, and renders them unsuitable templates for a subsequent iPCR reaction.

2.5.3. Cryptic splice sites and polyadenylation signals—Aberrant splicing and premature polyadenylation are frequently observed for L1 assay vectors, resulting in attenuated retrotransposition rates [25,26]. Such negative effects on L1 activity can be largely prevented by screening and eliminating cryptic splice donor/acceptor sites and polyadenylation signals from synthetic L1 vectors, including both coding sequences and reporter/utility cassettes.

2.6 Efficient retrotransposition by modular L1 assay vectors

We previously reported high retrotransposition rates from synthetic mouse L1, *ORFeus* using transient L1 assays with pCEP4-based vectors in both human and mouse cells [23]. Some of these vectors have since been converted into a pCEP-Puro backbone (pWA121 and pWA122; Figure 3); both of these are driven by tandem CMV and 5'UTR promoters, and marked by a conventionally subcloned *neoAI* reporter. We have now constructed several new *ORFeus* assay vectors that conform to the proposed modular design strategies (pTT011, pTT012 and pTT013; Figure 3); all of them carry identical double *SfiI* flanked *neoAI* reporters. When these modular L1 vectors were tested in HeLa cells, a consistent, high level of retrotransposition was observed (Figure 3).

2.7 Use of non-replicating vector for L1 cell culture assays

The use of episomal vectors, such as pCEP4 and derivatives, comes with some limitations. First of all, the sequence elements responsible for episomal maintenance in human cells (i.e., EBNA-1 gene and EBV replication origin oriP) dramatically increase the plasmid size (a minimum contribution of 4.6 kb). Second, if episomal replication is critical for L1 cell culture assays, such vectors may not fare well in non-primate cell lines because oriP vectors replicate most efficiently in primate cells [18]. However, recent developments on L1 assay protocols indicate that the capability of replicating episomally may not be needed for L1 cell culture assays. The first evidence is the successful establishment of a transient L1 assay with oriP-based vectors [27], which proceeds directly (3 days post-transfection) to the second round of antibiotic selection without the typical first round selection against untransfected cells. Retrotransposition rates from different samples are normalized by transfection efficiency inferred from cotransfected GFP plasmid. The second development is the substitution of puromycin resistance for hygromycin resistance in pCEP4, resulting in pCEP-Puro ([7]; Symer D.E. and Boeke J.D., unpublished data). With pCEP-Puro, the first round of selection can be shortened to 3 days. Although these newer assay schemes have yet to be tested on a wide range of non-primate cells, both lead to comparable retrotransposition rates in the prototypic HeLa cell line when using the *neoAI* reporter ([27] and this report). To directly assess the requirement of EBNA-1 for L1 assays, we inactivated EBNA-1 by a frameshift mutation, and compared retrotransposition rates between L1 assay vectors carrying wild type EBNA-1 (pWA121) and EBNA-1 mutant (pTT014; Figure 3). We followed the pCEP-Puro assay protocol using a 3-day puromycin selection before plating cells for G418 selection, and obtained only a modestly higher retrotransposition rate for pTT014 than pWA121 (Figure 3). Since functional EBNA-1

protein is essential for pCEP4 replication [18], our data strongly suggest that autonomous replication capability is not required for achieving high rates of retrotransposition when a shortened puromycin selection protocol is used.

3. Inverse PCR mapping of *de novo* L1 insertions

3.1 General considerations

Mapping the integration site of a *de novo* L1 insertion and interrogating the inserted L1 sequence and its surrounding genomic DNA sequences have provided a rich source of information about preferences and consequences of L1 integration in both cultured cells and transgenic mice [13,15,16,28–32]. Unlike DNA transposons and LTR-retrotransposons, L1 insertions have two unique features that complicate global mapping/sequencing schemes: (1) the 5' end of L1 varies from insertion to insertion due to 5' truncation; (2) a poly(A) tail of variable length separates the 3' end of an L1 insertion from the flanking genomic DNA (Figure 4). Several approaches have been employed to clone L1/genomic DNA junctions, including plasmid rescue [29,32], inverse PCR [13,15,31], and thermal asymmetric interlaced (TAIL) PCR [16]. Here, we describe an optimized iPCR protocol that amplifies the element's 3' junction using 4-base cutters and step-down cycling conditions. This iPCR strategy allows efficient amplification of multiple insertions in a single round of PCR reaction of 35 cycles [15]. It can be used not only to recover individual events but also to assess the relative level of retrotransposition in cells and tissues [14]. The following factors should be considered when adapting the described protocol to other L1 transgenes or applications.

3.2 Critical parameters

3.2.1 Restriction enzymes—Four-base cutters producing sticky ends, such as *MspI* and *TaqI*, are preferred over 6-base cutters in this protocol for the following reasons. First, restriction with 4-base cutters produces shorter genomic DNA fragments that can be more efficiently amplified by PCR. Second, the cyclization of a linear DNA fragment is most efficient if DNA length falls in the range of 200 to 4000 bp [33]. The choice of a CG containing 4-base cutter puts the fragment sizes squarely into the sweet spot of this range [34]. Third, the ligation efficiency is usually higher for DNA with sticky ends than those with blunt ends. Fourth, CpG methylation sensitive enzymes should be avoided.

In selecting 4-base cutters, an L1 vector should be analyzed to ensure sufficient room at the 3' end of the L1 vector for designing specific iPCR primers (see 3.2.2.). For the specific implementation we use, any enzyme leaving a 5'CG overhang can be employed, alone or in combination, provided it does not cut between the reverse primer site and the polyadenylation signal cleavage site. Using a blend of two or more digests (performed separately) helps avoid the problem of those sites that happen to lie close to the element end in the 3' flanking gDNA, which can give rise to sequence tags too short to accurately map the insertion.

3.2.2. Primer design—Inverse PCR requires two diverging primers relative to the template DNA [35,36]. The suitability of primers for iPCR must be empirically determined. We use primer3 (<http://frodo.wi.mit.edu/>) to design iPCR primers. As a general guideline, the melting temperature (T_m) of each primer should be ~65°C. The positioning of iPCR primers should be carefully considered. Ideally, the reverse primer (Rvs) should be positioned ~50 bp downstream of the desired restriction site (M), and the forward primer (Fwr) should be positioned ~50 bp upstream of the polyadenylation signal cleavage site (Figure 4). An interval of 50 bp is preferred for two reasons. First, a larger distance will unnecessarily increase the size of every iPCR amplicon. Second, each iPCR amplicon should retain sufficient L1-specific sequence, which can later be used to verify the authenticity of the cloned DNA after sequencing.

Two primers, JB8822 and JB8897, have been extensively tested in our iPCR protocol. These priming sites have subsequently been incorporated into the modular design (see 2.5.1).

Under some conditions, the synthetic L1 donor element can be inadvertently amplified by the iPCR reaction due to digestion and subsequent re-circularization of the donor transgene concatemer. In particular, this phenomenon occurs in transgenic lines in which the copy number of the transgene concatemer is high. To avoid amplification of ligated fragments from L1 donor transgene, the iPCR reaction can be modified by designing a new reverse primer that spans the intron splicing junction in the reporter module. This new reverse primer will only anneal to L1 insertions (and not donor elements) due to intron removal during retrotransposition, allowing for efficient amplification of new L1 insertions without amplification of the donor transgene. In designing splice-site-spanning iPCR primers, the relative position of the splice site within the primer needs to be considered. In general, the iPCR specificity is optimal if the splice junction is positioned in the middle of the primer (after the first 10 nt in a typical 22–23 nt primer) but decreased if the splice junction is moved more towards the 5' end.

3.2.3. Hot-start Taq DNA polymerase and “step-down” PCR cycling conditions

—Borrowing from both “hot-start” and “step-down” PCR methods [37], this iPCR protocol requires the use of hot-start Taq DNA polymerase and ten initial cycles of amplification with incrementally decreasing annealing temperatures from cycle to cycle, both of which are critical for ensuring amplification specificity and efficiency. Typically, nonspecific amplification from control samples is minimal but specific insertions can be amplified and visualized on an agarose gel simply after a single round of 35-cycle PCR reaction [14,15].

3.3. Step-by-step iPCR protocol

3.3.1. Genomic DNA preparation—Mouse genomic DNA (gDNA) is extracted from various mouse tissues using a Gentra Puregene tissue kit (Qiagen) per manufacturer's instructions. DNA is resuspended in 50 µl of supplied DNA hydration solution (also known as TE; 10 mM Tris-HCl, 1 mM EDTA, pH 8.0). This method generates high-quality high-molecular-weight gDNA that can be used for initial PCR genotyping and for other downstream applications, including iPCR and Southern blotting analysis. The typical yield from a tail snip of 0.5 cm length ranges from 5 to 10 µg.

3.3.2. Restriction enzyme digestion—Restriction digestion of mouse genomic DNA is performed in a 50 µl volume. An *MspI* digestion reaction contains the following: 5 µl 10x buffer 2, 2 µl *MspI* (20 units/µl, New England Biolabs), 0.5 µg DNA, and dH₂O to 50 µl. Incubate at 37°C for 2 hours. Inactivate *MspI* at 65°C for 20 min. It should be noted that no effect on iPCR efficiency was observed when we varied the incubation time from 2 hours to overnight. No DNA purification step is needed before setting up the ligation reaction.

3.3.3. Ligation reaction—DNA ligation reaction is performed at a DNA concentration of 1 µg/ml in a 500 µl volume. Set up the reaction by adding the following components to the heat-inactivated restriction reaction: 397.5 µl dH₂O, 50 µl 10x T4 DNA ligase buffer, and 2.5 µl T4 DNA ligase (400 cohesive end units/µl, New England Biolabs). Incubate at 16°C overnight. Inactivate T4 DNA ligase at 65°C for 10 min.

3.3.4. Pre-PCR sample processing—Before setting up iPCR amplification, ligated genomic DNA needs to be concentrated by using Microcon YM-100 columns (Millipore). Load the 500 µl ligation reaction to a YM-100 column. Spin at 500 g for 15 min. Check the remaining volume in each column, and spin at 1000 g for an additional 5 min if the sample volume is > 50 µl. Invert the column and place onto a new collection tube, and spin at 1000 g for 3 min to recover the concentrate. These centrifugation steps consistently produce a final sample volume

≤ 50 µl. If necessary, adjust the volume of the concentrate to 50 µl with dH₂O. Such adjustment is required if the experiment is intended to compare the retrotransposition activity among multiple samples. Note: The molecular weight (MW) cutoff for YM-100 columns is 100,000, which should effectively retain double-stranded DNA molecules >125 bp in length. This step may also remove some protein components from the concentrate since the nominal MW of T4 DNA ligase (55 kDa) and *MspI* (29 kDa) are well below the MW cutoff.

3.3.5. PCR reaction—A standard iPCR reaction volume is 50 µl (Table 1). The reaction volume can be scaled down to 25 µl proportionally with no adverse effect on PCR performance. If multiple samples are to be compared for retrotransposition activity, PCR setup in a 96-well plate is strongly recommended. A minimum of two independent PCR reactions should be set up for each ligation reaction. However, the exact number of replicates should be tailored to match the objective of the experiment. In our experience, duplicate reactions provide sufficient starting material for recovering multiple unique insertions. To achieve maximal coverage of potential insertions in a sample, six or more replicates may be used.

Upon completion, 5 µl of PCR reaction can be inspected on a 1% agarose gel. During electrophoresis, replicate reactions for each sample should be loaded in adjacent lanes, allowing direct comparison of band pattern differences. A typical result features multiple bands ranging from 0.1 to 3 kb in each reaction, with a completely or largely distinct band pattern for replicate reactions from the same sample [15]. Such band patterns indicate the presence of a large number of unique insertions in the starting gDNA sample and the random sampling nature of the iPCR reaction. If a band is consistently amplified in all replicate reactions, it likely represents a germ line insertion or a somatic insertion that arose early during embryogenesis. Weakly amplified bands are more likely to be somatic insertions that occurred late during development.

3.3.6. PCR product cleanup—The PCR product can be purified either directly as a pool of bands without gel electrophoresis or as individual bands after gel electrophoresis. (1) If the objective is to recover a large number of insertion events, especially if working with multiple samples, maximal efficiency can be achieved by recovering iPCR products as a pool. We use QIAquick PCR purification columns (Qiagen) to clean up PCR reactions. Replicate reactions for each genomic DNA sample can be loaded to a single column at this step, and DNA is eluted into 30 µl supplied elution buffer (10 mM Tris-HCl, pH 8.5). (2) If specific bands are to be recovered, the remaining iPCR reaction (~45 µl) should be resolved on a 1% agarose gel. We use the QIAquick gel purification kit (Qiagen) to purify DNA fragments from agarose gels. DNA is eluted into 30 µl supplied elution buffer, and can be used for subsequent cloning steps.

3.3.7. PCR reamplification of individual bands—This step is optional and normally not needed. It is a procedure that we initially used in verifying the identity of individual PCR fragments after gel electrophoresis [15]. One may find it useful if having technical difficulty with low DNA yield. To reamplify, the gel slice containing the desired band (excised at step 3.3.6) can be quickly frozen and thawed in the presence of 100 µl 10 mM Tris. Smash the gel slice with a pipette tip after thawing. Spin briefly. Take 2 µl as templates for reamplification PCR (Table 2). It is expected that in some reactions a few minor bands coexist with a major band that you intend to amplify. At this point, another round of gel purification may be necessary. For the same reason, we generally discourage direct sequencing of purified PCR products.

3.3.8. PCR product cloning and sequencing—In this step, purified PCR product, either as a pool or from individual bands (see step 3.3.6), will be subcloned into a suitable vector for subsequent sequencing. PCR amplicons generated with Hot start ExTaq DNA Polymerase (Takara) contain a mixture of blunt ends and 3' A overhangs, which allow efficient cloning with vectors with 3' T overhangs. Both TOPO TA cloning kit (Invitrogen) and StrataClone

PCR cloning kit (Stratagene) produce satisfactory results. Competent cells, which are supplied as part of either kit, are transformed per manufacturer's instructions and plated on LB-carbenicillin plates for blue-white color screening. Pick white colonies for DNA sequencing analysis. Bidirectional sequencing is required if universal primers from the cloning vector are to be used. We use M13-forward and M13-reverse primers for pCR2.1 (Invitrogen), T3 and T7 primers for cloning vector pCR4 (Invitrogen) and pSC-A (Stratagene). Alternatively, the forward and reverse iPCR primers may be used.

De novo L1 insertions end in homopolymeric poly(A) tails averaging 60–103 nt [15,16,29, 32]. In most cases, sequence quality rapidly deteriorates beyond a run of 20 A's. Therefore, for L1 insertions that possess poly(A) tails longer than 20 nt, the forward sequencing read (i.e. from sequencing reaction with the forward iPCR primer) can not provide information about the 3' flanking genomic DNA. If the iPCR amplicon is less than 800 bp, the sequencing reaction with the reverse primer should be able to read across the gDNA sequence, and reach the junction between 3' gDNA and the poly(A) tail. If the iPCR amplicon is over 800 bp long, the precise junction between 3' gDNA and poly(A) tail needs to be determined by primer walking. It is important, in our opinion, always to sequence from both directions, as the forward read can provide assurance of the cloned sequences being authentic L1 insertions.

3.3.9. Genomic mapping of 3' integration site—The genomic location of individual insertion can be determined by extracting the 3' flanking genomic DNA from sequencing data and aligning to assembled human or mouse genome.

3.3.10. Determination of 5' junction—This iPCR protocol is designed primarily to capture 3' L1/gDNA junctions. Occasionally, the 5' gDNA/L1 junction can be obtained simultaneously from the same reverse read if an insertion is 5' truncated between the restriction site and the reverse primer (refer to Figure 4). In all other cases, mapping 5' gDNA/L1 junction necessitates the use of 5' junction PCR (Table 3). The aim of such PCR is to amplify the 5' gDNA/L1 junction with a genomic primer and an L1 primer. Considering the unpredictability of an L1 insertion's 5' end and the frequent occurrence of sequence deletion and inversion at the 5' junction, it is recommended to design a minimum of two forward primers specific to the genomic DNA sequence upstream to the mapped 3' integration site, and a series of reverse primers at an average interval of 1 kb over the length of the L1 vector. Multiple 5' junction PCR reactions can then be set up by pairing each forward genomic primer with each reverse L1 primer. A subset of these PCR reactions will produce discrete single bands, which are indicative of successful amplification of the 5' gDNA/L1 junction. The 5' integration site of an insertion can be determined by cloning and sequencing of such PCR fragments.

Concluding remarks

We have proposed a modular design for L1 vectors that aims to streamline functional studies of L1 elements in both cell culture and transgenic animals. The eventual success of such a designing scheme depends on active yet voluntary participation of individual L1 investigators and beyond. Thus, we urge our colleagues to implement the proposed modular design whenever a new L1 plasmid or functional modules are made. Such a concerted effort will ultimately benefit the overall L1 research community as a result of rapid dissemination of this key research reagent and minimal downtime in constructing new vectors to suit each lab's specific research needs. We view this as version 1.0 of a synthetic retrotransposon design standard, and hope that this paper will stimulate the development of an improved standard in the future. For example, one area that awaits further improvement is to use alternative vector backbones for L1 assays. We have presented evidence here that efficient L1 retrotransposition is not dependent on pCEP4 conferred autonomous replication capabilities. One can envision the use of non-pCEP4 backbones that will facilitate the use of many sites naturally located in the pCEP4

backbone. Also presented here is a thorough description of an optimized method to map L1 integration sites from samples containing a multitude of retrotransposition events. This protocol is especially suited for high-throughput recovery of L1 insertions from mouse tissues. It is also semi-quantitative and can be used to rapidly access the overall retrotransposition activity of a given L1 transgene in mouse samples.

Acknowledgements

This work was supported in part by start-up funds from Washington State University (W.A.) and by National Institutes of Health Grant CA16519 (J.D.B.).

References

1. Singer MF. Cell 1982;28:433–434. [PubMed: 6280868]
2. Han JS, Boeke JD. Bioessays 2005;27:775–784. [PubMed: 16015595]
3. Kazazian HH Jr. Science 2004;303:1626–1632. [PubMed: 15016989]
4. Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr. Cell 1996;87:917–927. [PubMed: 8945518]
5. Ostertag EM, DeBerardinis RJ, Goodier JL, Zhang Y, Yang N, Gerton GL, Kazazian HH Jr. Nat Genet 2002;32:655–660. [PubMed: 12415270]
6. Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian HH Jr. Science 1991;254:1805–1808. [PubMed: 1662412]
7. Ostertag EM, Prak ET, DeBerardinis RJ, Moran JV, Kazazian HH Jr. Nucleic Acids Res 2000;28:1418–1423. [PubMed: 10684937]
8. Moran JV. Genetica 1999;107:39–51. [PubMed: 10952196]
9. Goodier JL, Kazazian HH Jr. Cell 2008;135:23–35. [PubMed: 18854152]
10. Branciforte D, Martin SL. Mol Cell Biol 1994;14:2584–2592. [PubMed: 8139560]
11. Ergun S, Buschmann C, Heukeshoven J, Dammann K, Schnieders F, Lauke H, Chalajour F, Kilic N, Stratling WH, Schumann GG. J Biol Chem 2004;279:27753–27763. [PubMed: 15056671]
12. Trelogan SA, Martin SL. Proc Natl Acad Sci U S A 1995;92:1520–1524. [PubMed: 7878012]
13. Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH. Nature 2005;435:903–910. [PubMed: 15959507]
14. An W, Han JS, Schrum CM, Maitra A, Koentgen F, Boeke JD. Genesis 2008;46:373–383. [PubMed: 18615728]
15. An W, Han JS, Wheelan SJ, Davis ES, Coombes CE, Ye P, Triplett C, Boeke JD. Proc Natl Acad Sci U S A 2006;103:18662–18667. [PubMed: 17124176]
16. Babushok DV, Ostertag EM, Courtney CE, Choi JM, Kazazian HH Jr. Genome Res 2006;16:240–250. [PubMed: 16365384]
17. Shetty RP, Endy D, Knight TF Jr. 2008 J Biol Eng;2:5.
18. Yates JL, Warren N, Sugden B. Nature 1985;313:812–815. [PubMed: 2983224]
19. Boeke JD, Garfinkel DJ, Styles CA, Fink GR. Cell 1985;40:491–500. [PubMed: 2982495]
20. Kirilyuk A, Tolstonog GV, Damert A, Held U, Hahn S, Lower R, Buschmann C, Horn AV, Traub P, Schumann GG. Nucleic Acids Res 2008;36:648–665. [PubMed: 18073200]
21. Naas TP, DeBerardinis RJ, Moran JV, Ostertag EM, Kingsmore SF, Seldin MF, Hayashizaki Y, Martin SL, Kazazian HH. EMBO J 1998;17:590–597. [PubMed: 9430649]
22. Ichiyanagi K, Nakajima R, Kajikawa M, Okada N. Genome Res 2007;17:33–41. [PubMed: 17151346]
23. Han JS, Boeke JD. Nature 2004;429:314–318. [PubMed: 15152256]
24. Han Y, Lin YB, An W, Xu J, Yang HC, O'Connell K, Dordai D, Boeke JD, Siliciano JD, Siliciano RF. Cell Host Microbe 2008;4:134–146. [PubMed: 18692773]
25. Belancio VP, Hedges DJ, Deininger P. Nucleic Acids Res 2006;34:1512–1521. [PubMed: 16554555]
26. Perepelitsa-Belancio V, Deininger P. Nat Genet 2003;35:363–366. [PubMed: 14625551]
27. Wei W, Morrish TA, Alisch RS, Moran JV. Anal Biochem 2000;284:435–438. [PubMed: 10964437]

28. Gilbert N, Lutz S, Morrish TA, Moran JV. Mol Cell Biol 2005;25:7780–7795. [PubMed: 16107723]
29. Gilbert N, Lutz-Prigge S, Moran JV. Cell 2002;110:315–325. [PubMed: 12176319]
30. Morrish TA, Garcia-Perez JL, Stamato TD, Taccioli GE, Sekiguchi J, Moran JV. Nature 2007;446:208–212. [PubMed: 17344853]
31. Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, Batzer MA, Moran JV. Nat Genet 2002;31:159–165. [PubMed: 12006980]
32. Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, Boeke JD. Cell 2002;110:327–338. [PubMed: 12176320]
33. Shore D, Langowski J, Baldwin RL. Proc Natl Acad Sci U S A 1981;78:4833–4837. [PubMed: 6272277]
34. Bishop DT, Williamson JA, Skolnick MH. Am J Hum Genet 1983;35:795–815. [PubMed: 6310992]
35. Ochman H, Gerber AS, Hartl DL. Genetics 1988;120:621–623. [PubMed: 2852134]
36. Triglia T, Peterson MG, Kemp DJ. Nucleic Acids Res 1988;16:8186. [PubMed: 3047679]
37. Zhang Z, Gurr SJ. Gene 2000;253:145–150. [PubMed: 10940551]

Figure 1.

Modular design of L1 vectors. (A) A schematic of the first generation of L1 vectors. Both the first cell culture assay vector and L1 transgene are housed in a pCEP4-based backbone as a *NotI*-*BamHI* fragment. The L1 element is under the control of either 5'UTR alone or two tandem promoters (a heterologous promoter "P" upstream of the *NotI* site, plus 5'UTR). (B) A blueprint for modularized L1 vectors. A hypothetical L1 vector can contain seven modules from 5' to 3' end: 5' targeting homology arm, promoter, conditional control element, L1 coding sequence (ORF1 and ORF2), reporter, polyadenylation signal, and 3' targeting homology arm. These modules are delimited by strategically placed restriction sites (indicated at the top of the vector) so that new vectors can be derived by swapping individual functional elements using flanking restriction sites. The same vector can also be used to generate vectors for different applications when subcloned into desired vector backbones. Selected restriction sites are denoted as "universal sites", which should be uniformly preserved at specified positions in all modular vectors to facilitate module swapping. (C) Exemplary modular L1 vectors. pWA203 is an intermediate vector for targeting L1 into intron 3 of the human *HPRT* gene; *ORFeusLSL* is a conditional L1 transgene using Cre-loxP system; and pESD202 is our latest base vector in which the 3' end fully conforms to the modular design shown in panel B.

Figure 2.

SfiI-mediated swapping of reporter modules. (A) Efficient cassette exchange mediated by noncomplementary double *SfiI* sites. A double *SfiI* reporter cassette and an *SfiI*-ready recipient L1 plasmid were prepared by *SfiI* digestion, gel purified, ligated at a 1:5 vector:insert molar ratio, and used to transform *E. coli* strain TOP10 (Invitrogen). Plasmid DNA was prepared from each of twenty-eight colonies, digested with *SfiI*, and resolved on a 0.8% agarose gel in the presence of ethidium bromide. Colonies carrying a successful *SfiI* mediated cassette exchange event are predicted to produce two bands: 1.1 kb for the reporter cassette and 9.8 kb for the vector. M: 1 kb DNA ladder (New England Biolabs). (B) Directionality of *SfiI* mediated cassette exchange. The orientation of the double *SfiI* reporter cassette is checked by *EcoRV* digestion of all 19 recombinant clones (lanes 1–19). In all 19 cases, the band pattern is consistent with the orientation being correct. *EcoRV* is predicted to cut twice in a recombinant clone: once in the reporter insert and once in the backbone. If the orientation is correct, it should

give a banding pattern of 9.6 kb and 1.3 kb. If the orientation is reversed, the bands should be 10.0 kb and 0.9 kb. A *Sfi*I digested miniprep DNA (predicted to release 2 fragments of 9.8 kb and 1.1 kb, irrespective of orientation; lane 20) and a sample of uncut DNA (lane 21) serve as size controls. M: 2-log DNA ladder (New England Biolabs).

Figure 3.

Retrotransposition assays with modular L1 vectors and an EBNA-1 mutant. (A) Schematic representation of assay vectors and results. All vectors shown are subcloned into pCEP-Puro backbone. pWA121, pWA122 and pTT014 are structurally identical except the following mutations: pWA122 carries D212G and D709Y point mutations that abolish ORF2 endonuclease and reverse transcriptase activities (marked by “x”); pTT014 has a frameshift mutation in the EBNA-1 gene on the vector backbone. pWA121 and pWA122 are pCEP-Puro version of previously reported pCEPsmL1 and pCEPsmL1mut. pTT014 is derived from pWA121 by introducing a 4 bp insertion at amino acid position 15 through *AvrII* restriction and filling in reactions. pTT011, pTT012, and pTT013 are modular *ORFeus* vectors in which the *neoAI* reporter was introduced by using the double *SfiI* scheme (dotted box). L1 expression is controlled by one or two tandem promoters, including CMV promoter (hatched box), 5’UTR of mouse L1 (open box) and/or CAG promoter (filled box). Cells are transfected in triplicate, selected with puromycin for 3 days, and for each transfection, three 10 cm dishes are seeded with 1×10^4 cells each under G418 selection for 9 days. Retrotransposition activity shown is the average number of colonies per 10^4 cells seeded. (B) Predicted open reading frames for the wt EBNA-1 gene (as in pWA121) and its frameshifted counterpart (as in pTT014). Note the frameshift mutation may produce a truncated 38 aa protein with the first 14 aa being in frame. The frameshift mutation is marked by a downward arrow. Drawn to scale. (C) Representative images of G418 resistant colonies from above assays.

Figure 4.

Schematic of iPCR mapping for L1 insertions. Multiple L1 insertions in the mouse genome are illustrated as solid green rods. One insertion is magnified to display its sequence features, including target site duplications (filled triangles) flanking the insertion and a poly(A) stretch at the 3' end of the insertion. It should be noted that the majority of L1 insertions are variably 5' truncated. Thus, the 3' junction between an L1 insertion and its flanking genomic DNA is chosen to be the iPCR target. The relative position of both forward (Fwr, in red) and reverse (Rvs, in black) iPCR primers and restriction sites (M) are shown. Genomic DNA is represented by thin black lines and the PCR cloning vector as thicker blue line. Refer to text for details.

Table 1

iPCR reaction conditions.

Ingredient	Volume (μl)
dH ₂ O	34.95
10x ExTaq buffer	5.0
dNTPs, 2.5 mM each	5.0
Forward primer, 10 μM	1.9
Reverse primer, 10 μM	1.9
Hot start ExTaq DNA polymerase (Takara), 5 units/μl	0.25
DNA from concentrated ligation reaction	1.0
Total	50.0

PCR cycling parameters: denature at 94°C for 2 min; 10 cycles of {94°C for 15 sec, 70°C (−0.5°C/cycle) for 30 sec, 72°C for 60 sec}; 25 cycles of {94°C for 15 sec, 65°C for 30 sec, 72°C for 60 sec}; extend at 72°C for 7 min.

Note: A PCR cyclor capable of programming temperature decrement from cycle to cycle must be used. Two different cyclor models have been tested. Both Mastercycler Gradient (Eppendorf) and GeneAmp PCR System 9700 (Applied Biosystems) cyclors can reliably generate the characteristic band pattern as described in section 3.3.5. The bands from GeneAmp 9700 are noticeably lighter though, which can be compensated by adding 2 or 3 more PCR cycles to aforementioned conditions if necessary.

Table 2
PCR conditions for reamplification of individual bands

Ingredient	Volume (μl)
dH ₂ O	36.75
10x ExTaq buffer	5.0
dNTPs, 2.5 mM each	4.0
Forward primer, 10 μM	1.0
Reverse primer, 10 μM	1.0
ExTaq DNA polymerase (Takara), 5 units/μl	0.25
DNA from gel slice	2.0
Total	50.0

PCR cycling parameters: denature at 94°C for 2 min; 10 cycles of {94°C for 15 sec, 67.5°C (−0.5°C/cycle) for 30 sec, 72°C for 60 sec}; 20 cycles of {94°C for 15 sec, 62.5°C for 30 sec, 72°C for 60 sec}; extend at 72°C for 7 min.

Table 3

PCR conditions for 5' junction PCR.

Ingredient	Volume (μl)
dH ₂ O	37.75
10x buffer	5.0
dNTPs, 2.5 mM	4.0
Forward genomic primer, 10 μM	1.0
Reverse L1 primer, 10 μM	1.0
ExTaq DNA polymerase (Takara), 5 units/μl	0.25
gDNA, 100 ng/μl	1.0
Total	50.0

PCR cycling parameters: denature at 94°C for 2 min; 10 cycles of {94°C for 15 sec, 67.5°C (−0.5°C/cycle) for 30 sec, 72°C for 60 sec}; 35 cycles of {94°C for 15 sec, 62.5°C for 30 sec, 72°C for 60 sec}; extend at 72°C for 7 min.