Prediction and quality assessment of transposon insertion display data

Quang Hien Le and Thomas Bureau
McGill University, Montréal, Québec, Canada

BioTechniques 36:222-228 (February 2004)

Transposons are mobile sequences commonly found in prokaryotic and eukaryotic genomes. Their dispersal, repetitiveness, and the fact that their mobilization is a source of polymorphism make them choice candidates for use as molecular markers in mapping technologies. In recent years, variations of a technique inspired by amplified fragment length polymorphisms (AFLPs) (1) that take advantage of transposons have emerged as valuable tools for molecular analyses (2–7). These transposon-based mapping techniques, referred to as transposon insertion display (TID) after the first published report (2), have been applied to plants as well as to animals for population analysis (8,9), detection of transposition events (10), gene tagging (2), and the recovery of integration sites (3).

The common basis of all TID techniques is an adaptor-mediated multiplex PCR amplification of genomic restriction fragments that contain a transposon marker sequence, along with the variable length of the DNA sequences flanking the insertion sites (Figure 1). However, to avoid the amplification of restriction fragments that do not contain transposon marker sequences (i.e., nonspecific), different adaptor designs have been adopted (Figure 1, A and B) (2–11).

TID protocols also differ in the type of transposon chosen as a marker. However, transposon abundance, diversity, and distribution vary greatly between organisms. For example, transposon content can range from 3% in the yeast genome to over 60% in maize, and mammalian genomes primarily contain long and short interspersed nuclear elements, whereas the maize genome is mostly populated by long terminal repeat retrotransposons (12). This high variability can complicate the choice of an appropriate marker since clarity and resolution depend on the copy number of the transposon type used. Furthermore, the accuracy of TID is dependent on the design and PCR conditions of a transposon-specific primer. Transposons are characterized by structural features that may be problematic for PCR [e.g., terminal and subterminal repeats, secondary structures, A and T richness, or poly(A)/(T) tails]. Thus, the ability to predict the banding pattern generated by a specific primer in a specific genome would be useful for optimizing PCR conditions and for assessing the reliability and quality of the observed data.

A large data set of genomic sequence is currently available, including the complete sequence for eukaryotic model organisms such as *Arabidopsis*, *Caenorhabditis elegans*, *Drosophila*, mosquito, rice, and human (http://www.ncbi.nlm.nih.gov:80/PMGifs/

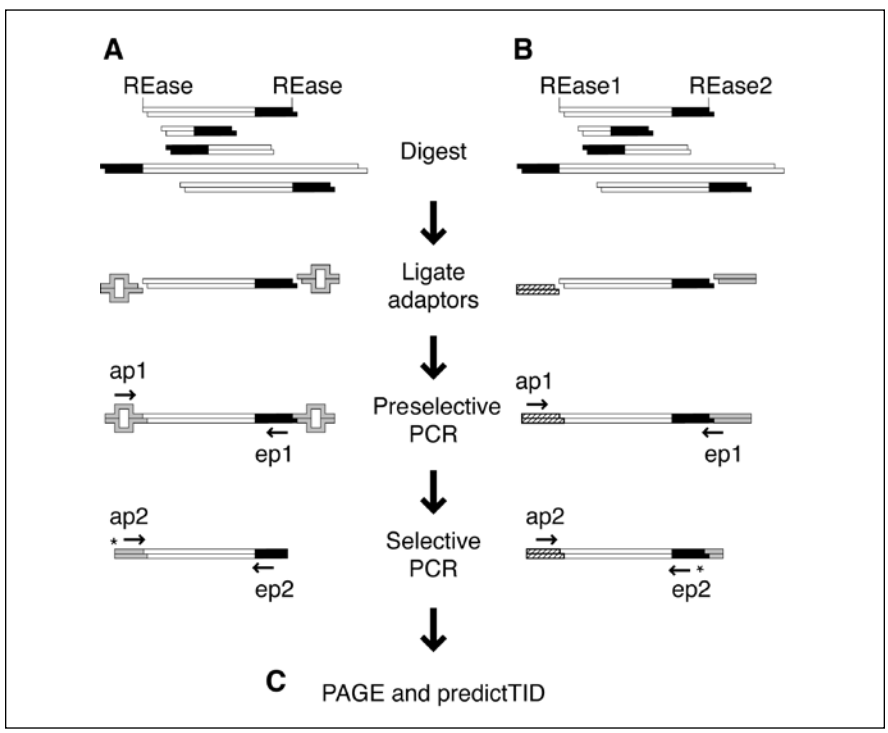

Figure 1. Transposon insertion display (TID) strategies. Transposon (black), genomic (white), and adaptor sequences (gray and hatched boxes) are represented. Genomic DNA is digested with restriction endonucleases (REases), yielding transposon fragments with varying lengths of adjacent sequence. Adaptors are ligated, and a preselective PCR amplifies transposon termini with flanking DNA sequences using primers directed against the adaptor (ap1) and transposon family (ep1). A selective amplification using nested or more specific primers enrich for transposon-specific products (optional in some protocols) (10). (A) The vectorette (11) is a specially designed double-stranded adaptor that is not completely complementary in sequence (represented by a bulge). Elongation from ap1 can only occur after first-strand synthesis from ep1, selecting for transposon-specific products. (B) TID strategy using different restriction enzymes (REase1 and REase2) and adaptors (gray-shaded and hatched boxes), each specific to either the transposon side or the flanking side. In this manner, amplification can be initiated from different adaptor-specific primers. Biotinylation of the transposon-specific adaptor can also serve as an enrichment step prior to the selective PCR (6). (C) Radioactive or fluorescently labeled primers (indicated by asterisks) allow for detection after size fractionation using polyacrylamide gel electrophoresis (PAGE), and predictTID uses ep2 sequence to calculate the size of the expected fragments. PredictTID serves to optimize experimental design, allows polymorphic bands to be distinguished from artifacts and, in experiments, can provide support for the observed data.

Table 1. cac1 Elements in *Arabidopsis*[a]

Element	Accession No.[b]	Position	Size on TID (bp)
CAC1	AC005897	52296–60774	735
CAC2	AC069160	34626–38790	478
CAC3	AC006429	85404–76949	836
CAC4	AC027135	72903–64858	849

TID, transposon insertion display.
[a]See Reference 15. [b]GenBank.

Genomes/euk_g.html). In addition, many other genomes are currently in the process of being sequenced. Available sequence information has been exploited in applications to calculate the expected sizes of PCR and AFLP products (http://www.in-silico.com/ and http://elanor.sci.muni.cz/cgi-bin/vpcr2.cgi; unpublished data) (13,l4), but these are site-specific or do not deal with the amplification of transposon-specific fragments. In this report, we describe a method to predict the TID banding pattern of a given primer from the available genomic sequence information.

We compared the observed against the predicted TID profile of a family of terminal inverted repeat (TIR) transposons called cac1. These elements were first identified from *Arabidopsis*, and the location of the four members (CAC1, CAC2, CAC3, and CAC4) within the completely sequenced genome is known (Table 1) (15). Exploiting conserved regions near the cac1 TIRs, preselective [(cac1-1; 5′-(T/C)TTTCGTAATGCTATGGTTGAAACACCTAAC-3′) and selective (cac1-2; 5′-CATACAATTCTGACGCTATC-3′)] primers for TID were designed by visual examination. The nucleotide sequences were retrieved from GenBank® (http://ncbi.nlm.nih.gov/Entrez/) and aligned using the PileUp function from the GCG® suite of programs (version 10; Accelrys, Burlington, MA, USA).

We wrote a Perl (version 5.6.0) script, predictTID, to calculate the sizes of bands expected from TID. PredictTID first uses the cac1-2 selective primer sequence as a query in a Basic Local Alignment Search Tool (BLAST®) search (version 2.2.3; ftp://ftp.ncbi.nih.gov/blast/) (16) against the *Arabidopsis* genome sequence (downloaded from The Institute for Genome Research; ftp://ftp.tigr.org/pub/data/a_thaliana). BLASTN parameters are set to default values, and the repeat filter is set to false. The BLAST search result is then parsed, and high scoring pairs (HSP) (16) that contain more mismatches than allowed by a user-defined variable (threshold limit) are discarded. We observed no differences in the predicted results when 0%–20% mismatches with the primer sequence were allowed. Genomic sequence (1000 bp) flanking the regions of similarity are retrieved and examined for the first *Bfa*I restriction pattern found upstream or downstream, depending on whether the BLAST subject was in the same or in reverse orientation, respectively, relative to the BLAST query on HSP.

The expected sizes reported by predictTID were then compared with the banding pattern of *Arabidopsis thaliana* (Columbia) cac1 elements using a vectorette-mediated TID strategy (2,11) with primers cac1-1 and cac1-2 (Figure 2). Genomic DNA was extracted from one or two rosette leaves of *A. thaliana* (Columbia-0) using a DNeasy® Plant Mini Kit (Qiagen, Mississauga, ON, Canada), following the manufacturer's instructions. TID as described by Korswagen et al. (2) was modified for the DNA 4200 fluorescence system (Li-Cor, Lincoln, NE, USA) (8). Approximately 100 ng of genomic DNA

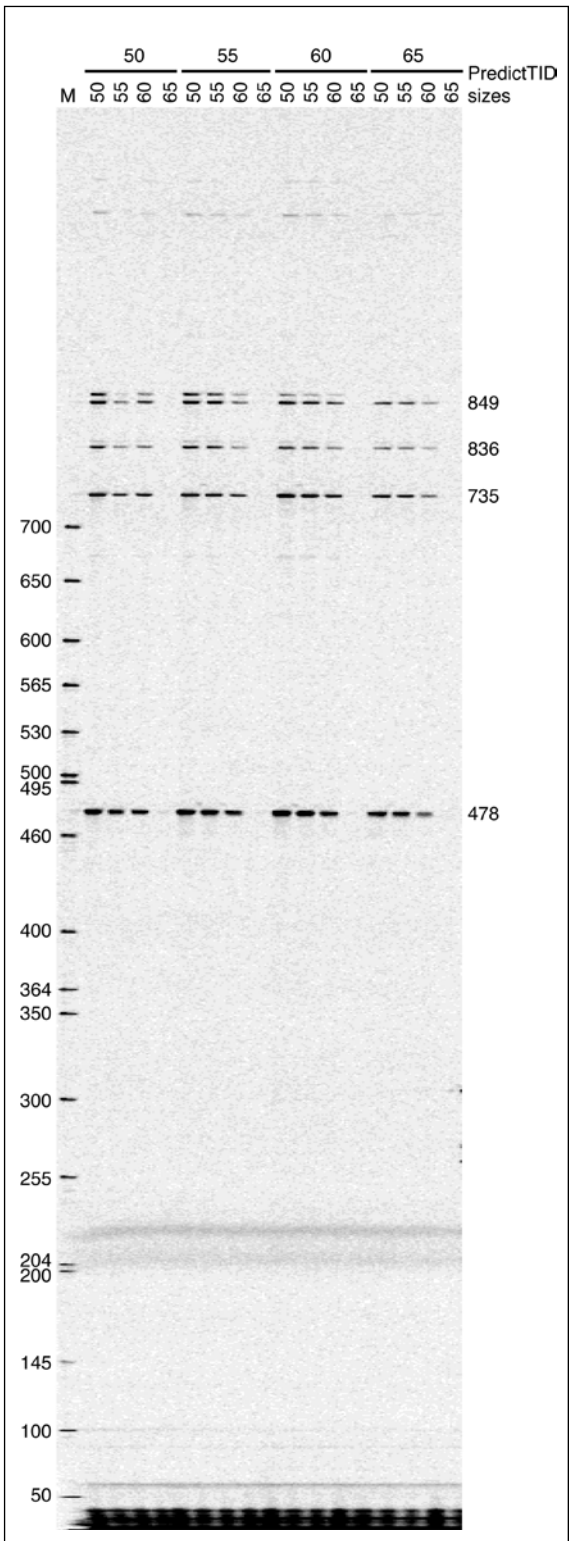

Figure 2. A transposon insertion display (TID) using cac1-1 and cac1-2 primers. Annealing temperatures for preselective (top, horizontal) and selective (bottom, vertical) PCRs are indicated above each lane. Lane (M), 50–700 bp DNA molecular weight marker. The sizes of the bands observed on the gel are in agreement with the sizes calculated by predict TID (indicated on the right) and correspond to the CAC1, CAC2, CAC3, and CAC4 elements listed in Table 1.

were digested with 2.5 U *Bfa*I (New England Biolabs, Beverly, MA, USA) and ligated to 15 pmol adaptor cassettes (5′-TAGCAAGGAGAGGACGCT-GTCTGTCGAAGGTAAGGAACG-GACGAGAGAAGGGAGA-3′ and 5′-TCTTCCCTTCTCGAATCGTA-ACCGTTCGTACGAGAATCGCT-GTCTCTCCTTGC-3′) with T4 DNA ligase (Invitrogen, Burlington, ON, Canada). The ligation product was diluted 4-fold before a preselective amplification using cac1-1 and ap1 (5′-CGAATCGTAACCGTTCGTA-CGAGAATCGCT-3′). Preselective amplification products were diluted 100-fold and reamplified using cac1-2 and ap2 (5′-GTACGAGAATCGCT-GTCCTC-3′), the latter being labeled with a IRDye™ 700 fluorescent dye (Li-Cor). We used the AmpliTaq® PCR system (Perkin-Elmer, Boston, MA, USA) in a PTC-225 DNA Engine Tetrad® Thermal Cycler (MJ Research, Waltham, MA, USA) for both amplifications, which consisted of 94°C for 10 min; 20 cycles of 94°C for 1 min, 50°, 55°, 60°, or 65°C for 1 min, and 72°C for 1 min; and 72°C for 10 min. Six microliters of loading dye (95% formamide, 10 mM EDTA, 0.1% bromophenol blue) were then added to the final amplification products, which were separated by size and visualized on a 5.5% denaturing polyacrylamide gel (BioShop, Burlington, ON, Canada). Fluorescently labeled DNA (50–700 bp) (50–700 sizing standard; Li-Cor) served as molecular weight markers.

Combinations of preselective and selective annealing temperatures between 50°–65°C, with 5°C increments, were tested. Four bands were expected from predictTID, and these could be reliably matched on TID. Annealing temperatures between 55°–60°C yield bands that are in best agreement with sizes calculated from predictTID. The estimated annealing temperatures for the selective amplification primers are between 50.3°–58.0°C (17). As a control, we manually retrieved and visually examined the sequences flanking CAC1, CAC2, CAC3, and CAC4 elements to confirm the sizes of fragments calculated using predictTID. Fragment sizes for the four cac1 members that were determined by manual examination are indicated in Table 1 and Figure 2. An unpredicted band (approximately 850 bp) could be seen but was no longer observed when higher annealing temperatures were used for the preselective amplification, which suggests that this may be a misannealing product.

Although not observed in this study, there are potential limitations to the reliability of predictTID. First, the maximum length of sequences flanking an insertion that is examined by predictTID was set to 1000 bp because, in practice, longer fragments cannot be reliably resolved by TID. Here we used *Bfa*I, and analysis of the restriction pattern of the complete *Arabidopsis* genome indicated that the majority (89.2%) of *Bfa*I fragments generated were shorter than 1000 bp.

Second, the initial step of pattern matching the primer sequence is handled by the BLASTN program (16) and may not exactly reflect PCR annealing conditions in terms of mismatches and gaps. The fact that predictTID gives the same weight to mismatches on the 5′ and 3′ ends of the primer sequence may be a source of difference.

Third, the applicability of our program is of course dependent on the amount of available sequence information. Even with completely sequenced genomes, there are gaps, especially within the repetitive sequence-laden centromeric and telomeric regions. Sequencing projects also focus on a specific genotype, and in other lines or strains, sequence polymorphisms may be a source of error. Nevertheless, as long as a large quantity of genome sequence information is available, predictTID can be useful to assess the validity of TID-generated bands.

Despite these potential limitations, we have shown that available sequence information can be used to test the reliability of primers in TID experiments. In a number of model organisms and for a wide range of transposon families, TID techniques are being adopted for a variety of applications. In these systems, predictTID can provide theoretical support for observed data, thereby facilitating the optimization of PCR conditions and the determination of the quality of designed primers. In addition, and based on the principle used by predictTID, two programs, predictAFLP and predictRAPD, are also available for calculating the sizes of bands expected from AFLP and other PCR-based mapping techniques. All programs are freely available by request and are also accessible online (http://bailly.biol.mcgill.ca/predictTID.html).

ACKNOWLEDGMENTS

We thank Dr. M.-A. Grandbastien, Dr. S.M. Tam, Nikoleta Juretic, and Fabienne Saadé for critical comments on our manuscript. We are grateful to Newton Agrawal for providing computer-programming advice. This work was funded by a National Science and Engineering Research Council (NSERC) grant to T.B. and a McGill Major Fellowship to Q.-H.L.

REFERENCES

1. **Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, et al.** 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. *23*: 4407-4414.
2. **Korswagen, H.C., R.M. Durbin, M.T. Smits, and R.H.A. Plasterk.** 1996. Transposon *Tc1*-derived, sequence-tagged sites in *Caenorhabditis elegans* as markers for gene mapping. Proc. Natl. Acad. Sci. USA *93*:14680-14685.
3. **Kohli, A., J. Xiong, R. Greco, P. Christou, and A. Pereira.** 2001. Tagged Transcriptome Display (TTD) in indica rice using Ac transposition. Mol. Genet. Genomics *266*:1-11.
4. **Yephremov, A. and H. Saedler.** 2000. Technical advance: display and isolation of transposon-flanking sequences starting from genomic DNA or RNA. Plant J. *21*:495-505.
5. **Waugh, R., K. McLean, A.J. Flavell, S.R. Pearce, A. Kumar, B.B. Thomas, and W. Powell.** 1997. Genetic distribution of BARE-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Mol. Gen. Genet. *253*:687-694.
6. **Van den Broeck, D., T. Maes, M. Sauer, J. Zethof, P. De Keukeleire, M. D'hauw, M. Van Montagu, and T. Gerats.** 1998. Transposon Display identifies individual transposable elements in high copy number lines. Plant J. *13*:121-129.
7. **Ayyadevara, S., J.J. Thaden, and R.J. Shmookler Reis.** 2000. Anchor polymerase chain reaction display: a high-throughput method to resolve, score, and isolate dimorphic genetic markers based on interspersed repetitive DNA elements. Anal. Biochem. *284*:19-28.
8. **Wright, S.I., Q.H. Le, D.J. Schoen, and T.E. Bureau.** 2001. Population dynamics of an *Ac*-like transposable element in self- and cross-pollinating *Arabidopsis*. Genetics *158*:

1279-1288.
9. **Ellis, T.H., S.J. Poyser, M.R. Knox, A.V. Vershinin, and M.J. Ambrose.** 1998. Polymorphism of insertion sites of Ty1-copia class retrotransposons and its use for linkage and diversity analysis in pea. Mol. Gen. Genet. *260*:9-19.
10. **Melayah, D., E. Bonnivard, B. Chalhoub, C. Audeon, and M.-A. Grandbastien.** 2001. The mobility of the tobacco Tnt1 retrotransposon correlates with its transcriptional activation by fungal factors. Plant J. *28*:159-168.
11. **Arnold, C. and I.J. Hodgson.** 1991. Vectorette PCR: a novel approach to genomic walking. PCR Methods Appl. *1*:39-42.
12. **Kidwell, M.G.** 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica *115*:49-63.
13. **Schuler, G.D.** 1997. Sequence mapping by electronic PCR. Genome Res. *7*:541-550.
14. **Rombauts, S.,Y. Van De Peer, and P. Rouzé.** 2003. AFLPinSilico, simulating AFLP fingerprints. Bioinformatics *19*:776-777.
15. **Miura A., S. Yonebayashi, K. Watanabe, T. Toyama, H. Shimada, and T. Kakutani.** 2001. Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. Nature *411*:212-214.
16. **Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*: 3389-3402.
17. **Breslauer K.J., R. Frank, H. Blocker, and L.A. Marky.** 1986. Predicting DNA duplex stability from the base sequence. Proc. Natl. Acad. Sci. USA *83*:3746-3750.

Received 22 August 2003; accepted 19 November 2003.

Address correspondence to Quang Hien Le, Laboratoire de Biologie Cellulaire, INRA Centre de Versailles, 78026 Versailles, France. e-mail: hien.le@versailles.inra.fr

High cell density induces expression from the carbonic anhydrase 9 promoter

Milota Kaluzová[1,2], Stefan Kaluz[1,2], and Eric J. Stanbridge[1]
[1]University of California at Irvine, Irvine, CA, USA and [2]Slovak Academy of Sciences, Bratislava, Slovak Republic

BioTechniques 36:228-234 (February 2004)

Efficient ways to control the level and timing of the expression of specific genes in cultured cells and tissues, including tumors, have received considerable attention (1,2). Inducible expression systems should have a minimal activity in the basal state but should allow rapid accumulation of the heterologous protein upon stimulation. The most advanced inducible systems employ combinations of functional domains from prokaryotic, eukaryotic, or viral proteins to create chimeric transactivators capable of modulating gene expression in a drug-dependent manner (1). The second component in these systems is an inducible promoter, which consists of a multimerized transactivator binding sequence linked upstream of a minimal promoter. In the presence of inducer, the chimeric activator binds specifically to its DNA recognition sequence and activates the transcription of the target gene (1). Although very specific and effective, these chimeric systems usually require specialized cell lines or have to be prepared in several relatively time-consuming steps. Therefore, it may be advantageous to use systems that are easier to generate and yet retain significant inducibility. Here we describe the cell density-dependent activity of the carbonic anhydrase 9 (*CA9*) promoter and propose its utility as an inducible expression system.

The expression of carbonic anhydrase IX (CAIX, previously known as MN) has been detected in a large number of carcinomas and carcinoma-derived cell lines but not in the corresponding normal tissues (References 3, 4, and references therein). The mechanism of CAIX induction in dense cultures was the subject of our previous study (5). Earlier, oxygen levels in sparse (10^6 cells in 100-mm plates) and dense (10^6 cells in 34.8-mm plates) LNCaP human prostate cells were established as 13% (96 mmHg) and 9% (70 mmHg), respectively (6). Reoxygenation by stirring abrogated CAIX expression, suggesting that CAIX expression in cultured cells is indeed triggered by an intermediate decrease of O_2 tension due to increased O_2 consumption and not by cell contacts per se. This decreased O_2 tension, also termed pericellular hypoxia (6), is too high for an appreciable stabilization of hypoxia-inducible factor 1α (HIF-1α), but it is sufficient for the activation of a phosphatidylinositol 3′-kinase (PI3-K)-dependent pathway (5). Earlier studies defined the *CA9* promoter in the (-173; +31) region (the numbers in parentheses indicate each position relative to the transcription start), which appears to contain the critical regulatory elements for *CA9* transcriptional activation (7). Among these, the hypoxia-response element (HRE) (8) and SP1/SP3 binding protected region 1 (PR1) (9) are crucial for *CA9* transcriptional activity.

The striking effect of cell density on CAIX expression prompted us to investigate the utility of the *CA9* promoter as a cell density-inducible expression system. The (-173; +31) *CA9* promoter fragment was cloned in the pGL2-Basic (Promega, Madison, WI, USA) and pEGFP-1 (BD Biosciences Clontech, Palo Alto, CA, USA) vectors. The (-2361; +298) vascular endothelial growth factor (VEGF) gene fragment was also cloned in the pGL2-Basic vector. The simian virus 40 (SV40) early promoter-driven pGl2 control vector was obtained from Promega. To prevent the possible modulation of *CA9* promoter activity by the SV40 promoter/enhancer sequence present in pEGFP-1 (GenBank® accession no. U55761; positions 1694–1925),