Transcendent Elements: Whole-Genome Transposon Screens and Open Evolutionary Questions

Ian Holmes[1]

Bioinformatics Group, Department of Statistics, University of Oxford, Oxford OXI 3TG, United Kingdom

Transposable elements (TEs), or transposons, form a major fraction of the eukaryotic genome (Kidwell and Lisch 2001). Dismissed for some time as junk DNA, these repetitive sequences are now recognized for their diverse evolutionary roles. In this issue of *Genome Research*, Bao and Eddy (2002) describe a software tool (RECON) for de novo annotation of transposons in genomic sequence, offering new possibilities for discovery to biologists interested in TE evolution as well as a practical tool for masking repetitive DNA from genomic annotation pipelines.

In this commentary, I begin by reviewing why transposons are relevant to studies of genome evolution. I then outline the advances of Bao and Eddy's method from previous work, highlighting certain exemplary features of the RECON method. Finally I describe some of the open questions of transposable element evolution that may be more easily addressed by large-scale bioinformatics and functional genomics approaches as RECON, and more tools like it, become available.

Why Care about Transposable Elements?

Transposable elements (Box 1) are of interest to geneticists (as an experimental tool), genome annotators (typically as junk DNA to be screened out), and structural and evolutionary biologists (for many reasons). My own bias lies toward the latter two (structural and evolutionary) aspects, and I will only briefly outline the former two (experimental and annotative) aspects before moving onto molecular evolution.

Experimental geneticists use transposons regularly as vectors for germ-line transformation, particularly gene knockouts. The use of *P* elements (class II TEs) to transform fruitflies, for example, is well-documented, and a systematic program of *P*-element-induced gene disruption is a key part of the *Drosophila melanogaster* genome project (Spradling et al. 1995). Transposons are very powerful tools in this context, and the discovery of a new transposon in a given host organism can greatly assist studies of that organism's genetics.

To many genome annotators, transposons are less exciting and more of a nuisance, because (owing to amplification of TEs following colonization of a new host) they comprise much of the repetitive content of genomes. (Other sources of repetition include signals that are locally amplified by replication error or recombinational mispairing, such as short oligonucleotide repeats.) Repetitive sequence can profoundly confuse well-intentioned statistical analysis, such as the reporting of Expectation values (*E* values) by programs like BLAST (Altschul et al. 1990) or MEME (Bailey and Elkan 1995). These *E* values are computed based on the assumption that sequences under neutral selection contain no long-range correlations, an assumption that is broken when the transposon copy number is amplified. Masking out previously characterized transposons prior to analysis (e.g., using the RepeatMasker program (A.F.A. Smit and P. Green, unpubl.) is one way around this. As for previously uncharacterized transposons, the very repetition of these sequences can be used to identify them. The approach described by Bao and Eddy is of this latter kind.

To a molecular biologist interested in protein structure and function, transposons have many interesting homologies, most notably to the replication machinery of many viruses, also to transcription factors and other specific DNA- and RNA-binding proteins. The information-processing nature of many tasks connected to transposition (specific and nonspecific sequence recognition, DNA and RNA processing, host defenses, self-regulation) and the possibility of assaying for transposition in vitro links them to an interesting variety of cellular processes.

It is, perhaps, among evolutionary theorists that transposons arouse the most interest. Since their discovery, TEs more than any other genes have highlighted the neo-Darwinian question of whether to model the gene or the organism as the fundamental unit of selection. Because TEs are a burden to the host, owing to the replicative load of the extra DNA, and (worse) because repetitive sequence content and transpositional activity are both mutagenic, TEs have in the past been regarded as purely selfish parasites (Orgel and Crick 1980). However, the mutations induced by transposons may be more structured than, say, mutations caused by irradiation or chemical toxicity, often rearranging rather than merely corrupting the host genome, and one can imagine them more readily generating a more neutral or advantageous phenotype (Kidwell and Lisch 2001); furthermore, TE activity often increases when the host is stressed (Capy et al. 2000), leading to a symbiotic or mutualistic (rather than parasitic) view of the host–TE relationship wherein TEs are agents of change or "natural genetic engineers" called in to stimulate evolution at times of stress (Shapiro 1999). If only they could talk.

Despite the seeming anthropomorphism of the mutualism/parasitism debate, the underlying question—how TE and host evolution are linked—is a rational and important one to ask. One way to address this question is to assemble a picture of the various processes, molecular and evolutionary, that are linked to transposition in one way or another, using sequence analysis and direct experiment to collect data on the observed interactions between TEs and their hosts (Kidwell and Lisch 2001). Such interactions include potential TE-related mutations in the host (Table 1), host defenses against TEs (Table 2), and evolution and self-regulation of TEs (Table 3). We can, in principle, amass many of these data by detailed sequence analysis (Kidwell and Lisch 2001). This requires that we know, first of all, where in the genome the transposons actually are.

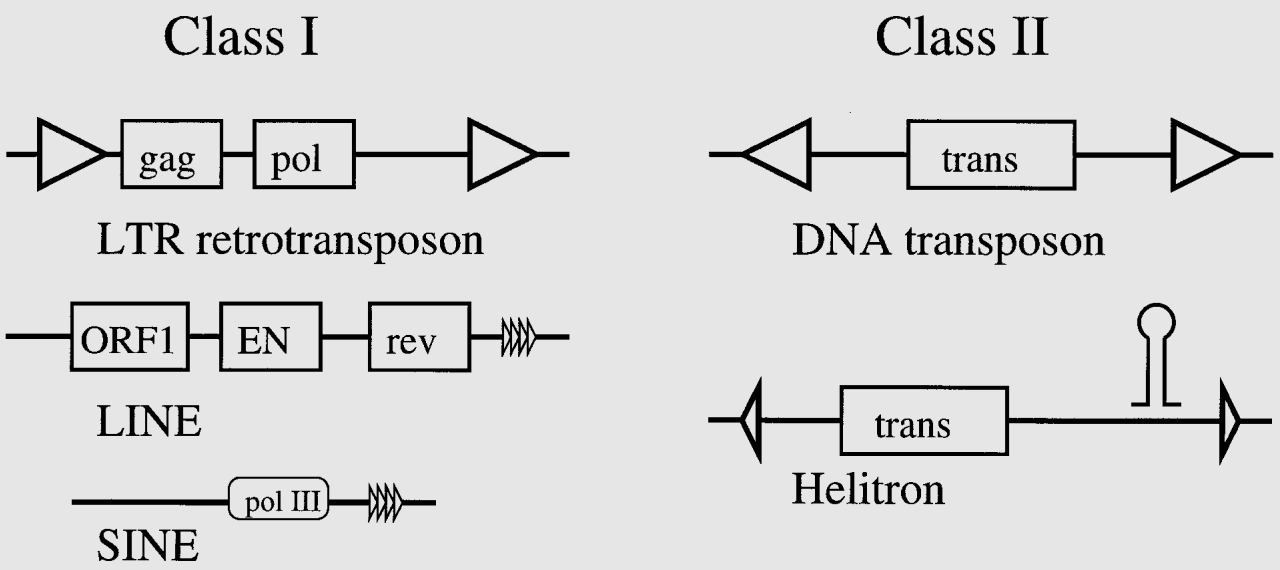
Automated Annotation: Hunting for Repetition

The task of de novo, automated annotation of all TEs in a genome is a difficult one, as explained by Bao and Eddy (2002). The principle of identifying repeated sequences, by

[1]E-MAIL holmes@stats.ox.ac.uk; FAX 44 1865 272595.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.453102.

12:1152–1155 ©2002 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

Box 1: Some Known Types of Transposable Elements



Transposable elements are DNA sequences that move or are copied from one genomic location to another (Feschotte et al. 2002). They can be classified according to their transposition intermediate, RNA (class I) or DNA (class II), and whether they code for genes that catalyze transposition (autonomous TEs) or require these genes to be provided, usually by other TEs in the host (nonautonomous TEs). The genes required for autonomy are different for class I and class II transposons.

Class I TEs are transcribed to mRNA and then reverse-transcribed into a new locus. These include long terminal repeat (LTR) retrotransposons, close relatives of retroviruses with LTRs, requiring *gag* (capsid) and *pol* (protease, reverse transcriptase, integrase) genes for autonomy. The usual difference between LTR retrotransposons and viruses is the absence of an *env* gene, which allows viruses to breach host cell membranes and survive in the extracellular matrix. Having said this, some class I TEs are active retroviruses, such as *Drosophila*'s *gypsy* (Mejlumian et al. 2002). Other kinds of class I TEs include long and short interspersed nuclear elements (LINEs and SINEs), the former of which contain *gag*-like, endonuclease and reverse transcriptase genes, the latter a pol III promoter; both end with a short repeat.

Class II TEs, in contrast, are excised and reinserted as DNA. Characterized by terminal inverted repeats (TIRs), these elements can be autonomous with but a single transposase gene, which must specifically recognize the TIRs and catalyze the *cut and paste* transposition reaction (van Luenen et al. 1994).

A new type of eukaryotic TE called a *Helitron*, which was tentatively characterized as a class II (DNA) transposon but predicted to have a distinctive transpositional intermediate, has recently been discovered in the genomes of *C. elegans*, *Arabidopsis*, and rice (Kapitonov and Jurka 2001). Helitrons show sequence and structural homology to bacterial rolling-circle transposons, which transpose by a three-step mechanism: nuclease cut, strand transfer, and repair (Mendiola et al. 1994). They have short terminal repeats and specific sequences at each terminus (including a short palindromic signal at the 3′ end that may form a DNA stem–loop) direct the targeted cleavage reactions of transposition.

Nonautonomous TEs can be any DNA sequences containing transpositional-activating signals specifically recognized by proteins from autonomous TEs. Typically, nonautonomous elements are derived from autonomous elements by deletions or other mutations.

Nontransposable repetitive DNA includes local repeats and microsatellites caused by replicative errors such as polymerase stutter, as well as local duplications and larger satellite runs caused by chromosome mispairing during meiotic recombination.

The above diagram is not to scale. Exon structure and promoters are not shown (except the pol III promoter in SINEs). TEs may contain additional genes and regulatory sequences.

clustering hits from a BLAST self-comparison or similar search, is straightforward enough. The problem is that certain TEs are often found to be associated, for example, if one TE jumps next to or into another and the resulting chimera TE is then amplified, so that automated programs can easily conflate adjacent or nested TEs (Holmes 1998).

Bao and Eddy have developed an elegant solution to this problem. As with previously described methods, their RECON algorithm starts by doing a BLAST-versus-self of the input genome. In place of the single-linkage clustering of prior methods, however, their algorithm examines the BLAST coordinates in detail. Wherever the density of BLAST hits to a region changes sharply, RECON places a boundary between elements. The sensitivity of the algorithm to changes in hit density, that is, the willingness of RECON to split up elements, is a parameter that can be fine-tuned. RECON also uses a simple length-based heuristic to distinguish major insertion and deletion variants from close familial relatives.

The approach of Bao and Eddy appears to work well. In a test on 3 Mb of human sequence data, RECON identified 6 out of 10 known repeat families and one new family (f179); in most of the known families, the reported consensus matched the canonical sequence closely. Several of the larger families are broken up, but this is probably inevitable to a certain extent and is certainly better than lumping distinct families together.

The problem addressed by RECON, of separating sequence motifs that may frequently be found adjacent to or nested within each other, is one that crops up throughout bioinformatics. RECON's solution is a clustering approach that is mindful of the nature of the data and the statistical issues involved. As pointed out by the authors, a similar approach may be useful for identifying the boundaries of protein domains; indeed, a clustering approach was used to build the protein hidden Markov model database Pfam (Sonnhammer et al. 1997). In protein sequences, domain-adjacency is more common than domain-nesting, but the domain boundaries should still be apparent from changes in the density of hits.

Table 1. Transposon-Related Host Mutations

Host mutation	Example	Reference
Insertion and excision	*Ds*-like TE in fungus *A. immersus*	Colot et al. 1998
Duplications, inversions, translocations (due to homozygous recombination or DNA repair)	Many examples in *C. elegans Ty* elements in *S. cerevisiae*	Holmes 1998 Kim et al. 1998
Transport of short signals (exons, promoters) during transposition	*Tpn1* in Japanese morning glory LINE-1 in *H. sapiens*	Takahashi et al. 1999 Pickeral et al. 2000
Horizontal gene transfer by viral elements	*env* genes in invertebrate retroviruses	Malik et al. 2000
Expression of host genes under TE's own promoters	Pol II promoters in *R. rattus* TEs	Tomilin 1999
Cooption of TE domains by host	V(D)J recombination Telomerase	Agrawal et al. 1998 Eickbush 1997
TF-derived pseudogenes	Vertebrate genomes	Brosius 1999

An open and ongoing algorithmic challenge is to develop clustering algorithms akin to RECON, reflecting the underlying biology as much as possible. Examples of patterns that could conceivably be modeled are proximity effects, substitutions and indels in the TE (and the rates of such mutations), patterns of TE aggregation and nesting, subfamilies, deletion variants, and prediction of TE class (Box 1). Probabilistic models seem a natural choice for this task, as many of the rules will be stochastic rather than deterministic.

Challenges for Bioinformatics: Smoking Icicles

A primary task for postgenomic bioinformatics studies of transposon evolution is the characterization of the full transposon complement of sequenced genomes. This includes the annotation of all transposons, using both de novo and homology-based methods, including classification in the vocabulary of Box 1 and more stringent familial groupings. Clearly, the work of Bao and Eddy is a significant step in this direction.

With the transposon complement annotated, it will be possible to investigate more systematically the evolutionary relevance of the various processes outlined in Tables 1, 2, and 3. Some of these stories can be expected to feature more smoking guns than others. For example, it is possible that relatively few class II TE-induced mutations directly implicate transposons simply because the evidence (the TE itself) disappears quickly. Rather than a smoking gun, this is reminiscent of Agatha Christie's perfect murder weapon: an icicle. We can, however, assemble a statistical picture of the rates and patterns of these mutations, for example, by looking at duplications of noncoding DNA within a genome (Holmes 1998), and see how this fits with other measurable rates of TE evolution.

The host responses of Table 2 should be more amenable to direct investigation in the lab. The TE evolutionary and self-regulatory processes of Table 3 may similarly yield to a combination of sequence analysis and experimental methods. An example is the coexistence of multiple related TE subfamilies within a single host. Previous informatics-led work revealed six previously undescribed families in the *Caenorhabditis elegans* genome related to the class I element *Tc1/mariner*. Sequence analysis suggests that these families have rapidly evolved distinct transposase-DNA specificity, thus avoiding crossmobilization (Holmes 1998). Hypotheses like this can be developed and tested by in vitro footprint-

Table 2. Host Responses to Transposons

Host response	Example	Reference
DNA methylation, chromatin remodeling	SINEs in *H. sapiens* *Athila*-like TEs in *A. thaliana*	Greally 2002 Okamoto and Hirochika 2001
dsRNA-mediated silencing	*Tc1* in *C. elegans*	Ketting et al 1999
Stress response of host defenses	*Tntl* in *N. tabacum*	Capy 2000

Table 3. Theorized Modes of TE "Self"-Regulation and Evolution

Type of regulation/evolution	Example	Reference
Targeted insertion	Gene-proximal regions in *C. elegans*	Holmes 1998
Heterochromatin vs. euchromatin	*A. thaliana* centromeric regions	Okamoto and Hirochika 2001
Interactions between subfamilies	*Tc1* variants in *C. elegans*	Holmes 1998
Interactions between autonomous and nonautonomous elements	*Tc1* hitchhikers in *C. elegans*	Holmes 1998
Changes in transposase-DNA specificity	*Tc1/mariner* and relatives	Holmes 1998; Lampe et al. 2001
Mutations destroying autonomy	Deletion products of *D. melanogaster* P elements	Engles 1996
Alternative splicing	ERV-3 retrovirus in *H. sapiens*	Larsson et al. 1997
Conversions between TEs and viruses	Invertebrate retroviruses	Malik et al. 2000
Colonization of new host species	Spread of *gypsy* in *Drosophila*	Mejlumian et al. 2002

ing (Colloms et al. 1994) and transposition assays (Lampe et al. 1996), making this fertile ground for collaborations between computational and experimental biologists.

Of central interest are questions of evolutionary timing. How fast, and in what ways, do transposable elements evolve at the sequence level? How does this compare to the rates of transposition and recolonization? What can we learn about colonization from comparing the transposon complements of closely related species? Can we link transposon invasions to bursts of evolutionary activity? How, if at all, do transposons and their hosts coevolve?

The fundamental evolutionary issue at hand is the role and raison d'être for what was once called junk DNA. This fascinating junk now forms an important part of our evolutionary picture of genomes as information repositories in flux. Computational whole-genome transposon screens like RECON offer the possibility of moving beyond anthropomorphic debates of selfishness-versus-altruism to pragmatic questions about the organization of genomes and the mutational/selective forces that shape their history.

REFERENCES

Agrawal, A., Eastman, Q.M., and Schatz, D.G. 1998. *Nature* **394:** 744–751.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. *J. Mol. Biol.* **215:** 403–410.

Bailey, T.L. and Elkan, C. 1995. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology* (eds. C. Rawlings et al.), pp. 21–29. AAAI Press, Menlo Park, CA.

Bao, Z. and Eddy, S. 2002. *Genome Res.* (this issue).

Brosius, J. 1999. *Genetica* **107:** 209–238.

Capy, P., Gasperi, G., Biémont, C., and Bazin, C. 2000. *Heredity* **85:** 101–106.

Colloms, S.D., van Luenen, H.G., and Plasterk, R.H. 1994. *Nucleic Acids Res.* **22:** 5548–5554.

Colot, V., Haedens, V., and Rossignol, J.L. 1998. *Mol. Cell. Biol.* **18:** 4337–4346.

Eickbush, T.H. 1997. *Science* **277:** 911–912.

Engels, W.R. 1996. In *Transposable elements* (eds. E. Saedler and A. Gierl), pp. 103–123. Springer-Verlag, Berlin.

Feschotte, C., Jiang, N., and Wessler, S.R. 2002. *Nat. Rev. Genet.* **3:** 329–336.

Greally, J.M. 2002. *Proc. Natl. Acad. Sci.* **99:** 327–332.

Holmes, I. 1998. "Studies in probabilistic sequence alignment and evolution." Ph.D thesis, Chapters 5, 7.The Sanger Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom. Available at http://www.stats.ox.ac.uk/~holmes/thesis/.

Kapitonov, V.V. and Jurka, J. 2001. *Proc. Natl. Acad. Sci.* **98:** 8714–8719.

Ketting, R.F., Haverkamp, T.H., van Luenen, H.G., and Plasterk, R.H. 1999. *Cell* **99:** 133–141.

Kidwell, M.G. and Lisch, D.R. 2001. *Int. J. Org. Evol.* **55:** 1–24.

Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A., and Voytas, D.F. 1998. *Genome Res.* **8:** 464–478.

Lampe, D., Churchill, M., and Robertson, H. 1996. *EMBO J.* **15:** 5470–5479.

Lampe, D.J., Walden, K.K.O., and Robertson, H.M. 2001. *Mol. Biol. Evol.* **18:** 954–961.

Larsson, E., Venables, P., Andersson, A.C., Fan, W., Rigby, S., Botling, J., Oberg, F., Cohen, M., and Nilsson, K. 1997. *Leukemia* **3:** 142–144.

Malik, H.S., Henikoff, S., and Eickbush, T.H. 2000. *Genome Res.* **10:** 1307–1318.

Mejlumian, L., Pelisson, A., Bucheton, A., and Terzian, C. 2002. *Genetics* **160:** 201–209.

Mendiola, M.V., Bernales, I., and de la Cruz, F. 1994. *Proc. Natl. Acad. Sci.* **91:** 1922–1926.

Okamoto, H. and Hirochika, H. 2001. *Trends Plant Sci.* **6:** 527–534.

Orgel, L.E. and Crick, F.H. 1980. *Nature* **284:** 604–607.

Pickeral, O.K., Makalowski, W., Boguski, M.S., and Boeke, J.D. 2000. *Genome Res.* **10:** 411–415.

Shapiro, J.A. 1999. *Genetica* **107:** 171–179.

Sonnhammer, E.L.L., Eddy, S.R., and Durbin, R. 1997. *Proteins* **28:** 405–420.

Spradling, A.C., Stern, D., Kiss, I., Roote, J., Laverty, T., and Rubin, G.M. 1995. *Proc. Natl. Acad. Sci.* **92:** 10824–10830.

Takahashi, S., Inagaki, Y., Satoh, H., Hoshino, A., and Iida, S. 1999. *Mol. Genet. Genet.* **261:** 447–451.

Tomilin, N.V. 1999. *Int. Rev. Cytol.* **186:** 1–48.

van Luenen, H., Colloms, S., and Plasterk, R. 1994. *Cell* **79:** 293–301.