NEWS & VIEWS FEATURE

Transposable elements affect maize colour.

PIXFOLIO/ALAMY

GENETICS

Junk DNA as an evolutionary force

Christian Biémont and Cristina Vieira

Transposable elements were long dismissed as useless, but they are emerging as major players in evolution. Their interactions with the genome and the environment affect how genes are translated into physical traits.

Transposable elements (TEs) — commonly called 'jumping genes' — are stretches of DNA that move around the genome of a cell, and the genomes of many higher organisms are cluttered with numerous copies of these enigmatic elements. They were discovered by Barbara McClintock in the 1950s (Box 1), but it has taken half a century to begin to understand how they act and the effects they can have. It is emerging that these elements have had a significant influence on the evolution of genomes, particularly by controlling gene activity.

The elements contain in their sequence all the instructions needed to cut themselves out of their host DNA and splice themselves into another spot. But they are not always benign 'junk' DNA — they can insert into genes or gene regulatory elements, potentially disrupting the gene's function, and they can trigger chromosome rearrangements. So, even though most copies are selectively neutral and not in themselves damaging, they have long been considered as predominantly harmful to their hosts, as they can contribute to the appearance of mutations, some of which can result in disease.

But TEs do not always have adverse effects, and their mutational activities contribute to the genetic diversity of the organism. Indeed, some TEs have been domesticated by their host genome, acting as genes or gene regulatory elements, and as a result constitute a source of genetic innovation for the organism[1,2]. Progress in understanding how these elements are regulated is bringing an appreciation of how an individual's environment can affect the expression of their genetic complement to produce their own particular characteristics (phenotypes), such as physical appearance, behaviour, susceptibility to disease and even neuronal function.

Diversity

Transposable elements are scattered throughout the genomes of many plants and animals, and can form a large proportion of the genome size (Table 1). There are two main classes of TE: DNA transposons, which act through a DNA intermediate and multiply by using the host cell's replication machinery, and retrotransposons, which act through an RNA intermediate. Retrotransposons are further subdivided into those that have 'long terminal repeats' at their ends (LTR retrotransposons) and those that do not (non-LTR retrotransposons; Box 2). In addition, TEs with composite structures are continually being discovered, illustrating the enormous flexibility of these elements. For example, DNA transposon-like elements called helitron rolling-circle elements were recently found to be responsible for copying various gene segments into new locations in the maize genome, generating a huge diversity among individual maize plants[3].

When LTR retrotransposons are excised from the genome, they leave behind an LTR sequence. Some genomes, particularly those of plants, are full of these lone LTRs. Because they

©2006 Nature Publishing Group

Box 1 | Transposable elements in gene regulation

Barbara McClintock discovered mobile genetic elements in the 1940s during her studies of maize, work for which she received the 1983 Nobel Prize in Physiology or Medicine. She observed that the patterns of colour in maize kernels changed in various breeding crosses, and she interpreted her results in terms of the regulation of gene activity by some kind of 'controlling elements' that had the ability to move from place to place on the chromosomes. As the elements move when the cells divide and proliferate, they mutate genes in only some of the cells — the changes in colour are due to their effects on pigmentation genes. McClintock's main conclusion was that these elements have a major influence on the development of organisms.

The discovery in the 1970s that bacteria, yeast and fruitflies have DNA elements able to remove themselves and insert elsewhere in the genome prompted experiments that definitively identified transposable elements as major constituents of the genome. Somatic mutations (those that are not inherited) are typical of the action of TEs, but these elements also mutate genes in the germ line, leading to heritable genetic changes.

Interest in TEs increased further as drafts of the human genome became available from 2000, revealing that around 45% of the DNA consists of TEs. But even then, their role in gene regulation was not fully recognized. Today, however, TEs are acknowledged as a main component of most genomes, and McClintock's ideas that they can control gene activity are fully accepted. **C.B. & C.V.**

can affect gene regulation, these TE leftovers can contribute to genetic diversity even though the element is no longer present.

Not all TE excisions leave behind such an obvious footprint, however, so the influence of TEs on genomes has long been underestimated. But less noticeable TE remnants are being discovered, allowing the role of these elements in mutations and genome evolution to be elucidated. For instance, an ancient family of sequences that are derived from small interspersed nuclear elements (SINEs; Box 2) was discovered recently in the genomes of mammals[4]. Even though they do not encode proteins, these TE remnants seem to be under strong selective constraints, suggesting that they have some function that is being conserved during evolution.

The estimated rate of DNA mutation due to TE insertions differs between organisms. Among individual fruitflies, 50–80% of mutations are due to such insertions. This is a high value compared with that of 0.1–1% seen in the human genome, where the numerous copies of non-LTR retrotransposons are less active and mostly fixed in position. The reasons for this difference are not well understood, but global genomic characteristics, such as the rate at which genes are shuffled between generations (recombination), may be involved.

Expression

The expression of TEs — that is, the production of their encoded RNA — is tissue-specific; some elements are highly expressed during particular stages of the host organism's life, and some are even expressed differently in male and female reproductive cells (the germ lines). Such high levels and specific patterns of expression seem unexpected for apparently non-functional 'junk' DNA. This paradox had been explained by invoking either complex interactions between the TE regulatory sequences and the activity of numerous host developmental genes, or the influence of particularly highly expressed host genes on the TEs adjacent to them (the 'readthrough phenomenon'). But both of these explanations implied a puzzling waste of energy for the cells and the organisms involved, so the idea that these RNAs are deliberately expressed to fulfil a particular cellular function gained ground.

This theory received a recent boost with the observation that some retrotransposons can influence the regulation of certain host genes and affect developmental processes in mouse oocytes (egg cells) and preimplantation embryos[5]. As well as validating McClintock's concept of TEs as controlling elements (Box 1), this finding reveals that TEs could have a role in the reorganization of genome structure and the gene silencing that occur during early embryonic development.

Transposable elements could have a marked impact on the relative contributions of the two sets of parental chromosomes in early development. In many animals, certain genes are differentially expressed according to their parental origin, and in mammals some TEs act rather like these 'imprinted' genes, in a manner dependent on their type. Thus, the DNA of SINEs is methylated — and thus inactivated — in the oocyte, but unmethylated in the male germ line. By contrast, intracisternal A particles (IAPs), retrotransposons and a long interspersed nuclear element (LINE) called LINE-1 display the opposite pattern[6]. This is consistent with the observation that maternally and paternally expressed imprinted genes in the human genome tend to occupy distinct genomic regions that are activated differently in the male and female germ lines.

Table 1 | Genome size and transposable elements

		Genome size (picograms)	% TEs
Rana esculenta	Frog	5.6–8.0	77
Zea mays	Maize	5.0	60
Homo sapiens	Human	3.5	45
Mus musculus	Mouse	3.4	40
Drosophila melanogaster	Fruitfly	0.18	15–22
Caenorhabditis elegans	Worm	0.1	12
Saccharomyces cerevisiae	Yeast	0.012	3–5
Escherichia coli	Bacteria	0.0046	0.3

The silencing and activation of TEs in the male and female germ lines[7] can affect genes that are involved in the RNA interference (RNAi) pathway. This is a process by which short, double-stranded RNAs induce the sequence-specific degradation of target messenger RNAs, halting production of the encoded protein. RNAi is thought to have evolved as a form of nucleic-acid-based immunity to inactivate viruses and TEs[8].

In addition to their effects throughout embryonic development, TEs may act later in life. Indeed, LINE-1 retrotransposons seem to jump preferentially into the regulatory regions of certain neuronal genes in mice[9]. This alters the genes' expression, creating distinct populations of neuronal cells (somatic mosaicism). The impact of such TE activity on neuronal function remains to be fully elucidated, but if it occurs in humans, the activity of LINE-1 elements might be responsible for changes in the neuronal circuits in the brain, contributing to the intellectual diversity among people. This possibility opens an avenue of research that was previously unthinkable, even for the most imaginative believers in TEs as promoters of genetic and phenotypic diversity.

Environment

In many organisms, TEs are under 'epigenetic' control — where gene regulatory instructions are laid out in modifications of the DNA or its associated proteins (mainly histones) that do not alter the DNA sequence itself. These alterations include methylation of certain DNA nucleotides, and methylation or acetylation of the histone proteins around which the DNA is wound. Such epigenetic marks have a direct effect on the levels of expression of nearby genes (and TEs), and they are considered to be a second code in addition to the well-known code in the DNA sequence[10].

Whereas some epigenetic instructions in the genome, such as gene imprinting, are usually long-lasting and heritable, the epigenetic code is more sensitive to environmental inputs than is the DNA sequence itself[11]. Epigenetic modifications even diverge gradually over the lifetimes of monozygotic twins[12]. Environmental stresses modify the genomic methylation state and thus change the structure of chromatin (the DNA and its associated proteins), inducing alterations in gene expression. Environmental stresses can thus markedly affect the way heritable changes in the DNA sequence (the genotype) are expressed as physical traits (phenotypes). Different stresses act differently on DNA methylation, histone methylation or acetylation, chromatin structure, and the production of small RNAs. Such processes interact strongly to affect genome function, and make up the core of epigenetic 'memory'[13].

The influence of the environment on

©2006 Nature Publishing Group

gene expression can also be felt through effects involving TEs, because these elements can control genes epigenetically when inserted within or very close to them[14]. TEs can act either directly by inducing the methylation (and therefore silencing) of nearby DNA, or indirectly by disrupting the normal epigenetic state of a nearby gene. The control of coat colour by the *agouti* gene in mice is a classic example of such TE–gene interaction. Overexpression of the *agouti* gene leads to yellow fur, as well as to obesity, diabetes and an increased susceptibility to tumours. When an IAP retrotransposon inserts just in front of the *agouti* gene, the expression of the gene depends on the methylation state of the inserted IAP. Differences in the IAP methylation state in different cells leads to patchy coat colour. The methylation pattern of the IAP is completely erased in the male germ line but incompletely erased in the female germ line. So the variable expression of the *agouti* gene is passed on to subsequent generations as an epigenetic inheritance, mimicking a classical maternal environmental effect on offspring[15].

The idea that environmental factors can affect how a phenotype is produced from a genotype is reinforced by the observation that early nutrition can influence the expression of various genes, including TEs, at critical developmental stages[16]. The effect of nutrition on an organism's phenotype is often viewed as a manifestation of physiological misregulation, but it seems that changes in gene expression associated with TEs are involved as well. This should make us more conscious of the impact that our environment at one developmental stage can have on further stages, even in adults — especially if changes in TE expression under new environmental conditions do turn out to modify some neuronal functions, as proposed above.

Disease

The discovery that TEs can promote mutations and chromosome rearrangements, and can be activated by changes in epigenetic state, has led to the increasing realization that these elements may be responsible for various diseases — particularly cancers[17,18]. Gene dysfunction or misregulation by TEs are considered to be responsible for around 0.5–1% of human illnesses[19]. Haemophilia, Duchenne muscular dystrophy, tumours of the oesophagus and reproductive organs and breast cancers may result from the insertion of SINEs and LINEs near or within genes. The human genome has been colonized by thousands of copies of human endogenous retroviruses (HERV), which are related to retrotransposons, and these sequences are permanently integrated in the genome. Many HERV copies have been associated with teratocarcinoma and leukaemia, and others are involved in multiple sclerosis, schizophrenia and diabetes. But HERV insertions are not necessarily detrimental, and some copies have been co-opted by the genome to carry out cellular functions. For example, the syncytin gene corresponds to part of an endogenous retrovirus and is essential to the normal development of the placenta in humans and mice[20].

As the methylation states of TEs can be modified by environmental factors, it is possible that diet or chemicals that interfere with the DNA-methylating enzymes involved may have major effects on normal physiology and the manifestation of diseases such as cancers — effects that are as yet little understood. Moreover, caution should be exercised in the use of modified retrotransposons and retroviruses for delivering gene therapy — there has been

Box 2 | Structure and nomenclature of transposable elements

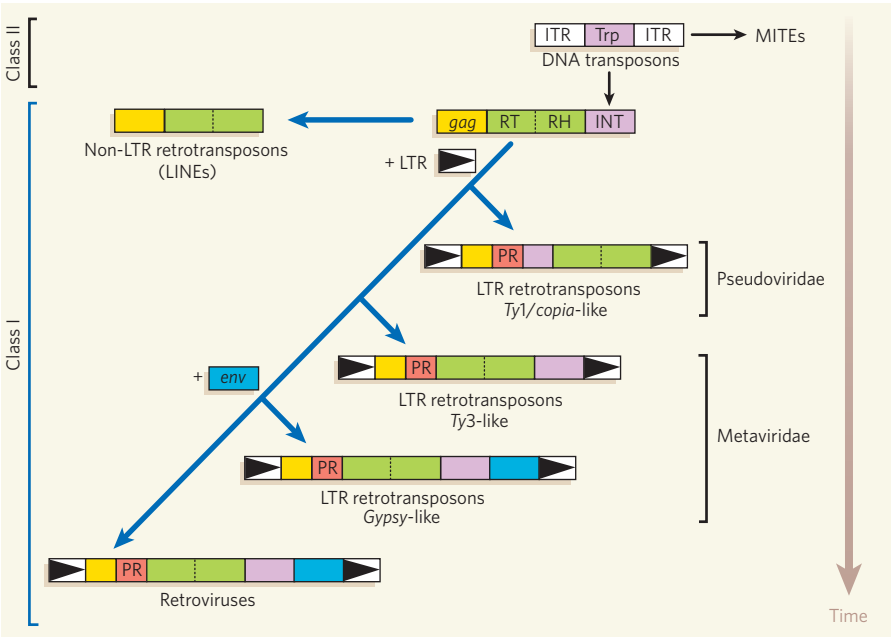
There is no officially agreed system for classifying transposable elements, but here is a simple system based on their evolution (phylogeny) and the genetic modules they contain (modified from ref. 30). The term transposon is often used as a generic term instead of transposable element, but it was originally coined to name the first characterized TE. This first TE moves about the genome through a DNA intermediate, using the transposase (Trp) enzyme to splice itself in and out of the DNA. The term DNA transposon (class II elements) is often used to distinguish this kind of TE from LTR (long terminal repeat) and non-LTR retrotransposons (class I elements) that move by means of an RNA intermediate and use the reverse transcriptase (RT) enzyme.

Inverted terminal repeats (ITRs) are needed for the movement of DNA transposons. The *gag* gene specifies the components of the molecular complex that is associated with the RNA transposition intermediate of retrotransposons. Retrotransposons also encode the RT enzyme, which synthesizes a complementary single-stranded RNA from the inserted DNA of the TE, and converts it to a double-stranded DNA that will be integrated in the genome elsewhere. They also encode the enzyme ribonuclease H (RH), which degrades the DNA–RNA hybrids obtained during transposition. Retrotransposons have genes encoding the integrase enzyme (INT), which splices the double-stranded DNA into a new spot in the host genome, and an enzyme (a protease; PR) that cuts up precursor proteins and is involved in particle assembly. Some retrotransposons have gained the envelope gene (*env*), which encodes surface proteins that interact with the host cell membrane, conferring infectious characteristics on the elements.

The human genome has about half a million copies of long interspersed nuclear elements (LINEs), of which 50–100 are still active. And there are more than a million copies of a short interspersed nuclear element (SINE) called *Alu* in the human genome. *Alu* elements are also a major source of genomic diversity in dogs. These SINEs have a particular structure and depend on LINEs for their transposition.

Miniature inverted-repeat transposable elements (MITEs) are an ancient TE family that is characterized by short-sequence, terminal or subterminal inverted repeats flanked by short direct repeats, and they have no coding potential. They are distributed ubiquitously and seem to originate from DNA transposons.

Other classifications have been proposed on the basis of the transposition mechanism involved, for example, and a system similar to that used by virologists for viruses exists for LTR retrotransposons, which are then classified as Metaviridae and Pseudoviridae[30]. But these classifications are not generally accepted. **C.B. & C.V.**

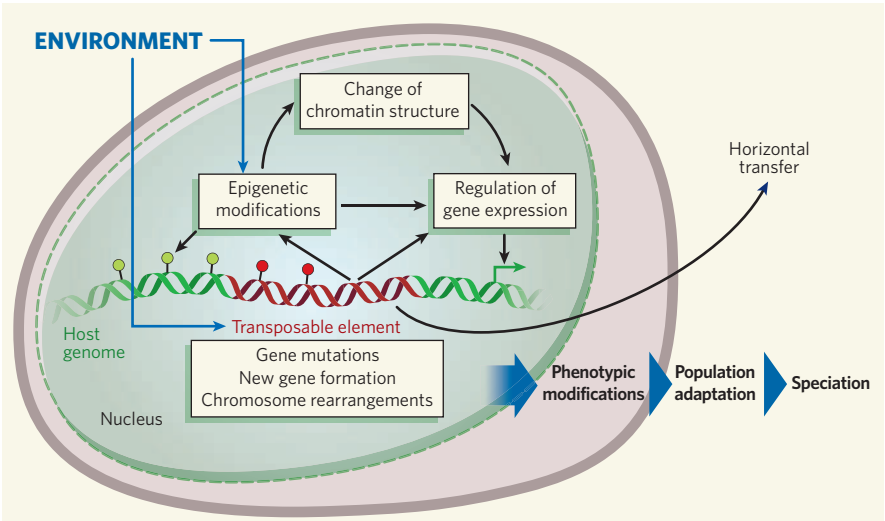


©2006 Nature Publishing Group

Box 3 | Transposable elements and genome–environment interactions

Transposable elements are increasingly seen as major originators of genetic change, allowing populations to adapt to change and species to evolve, as shown in the figure. They can also move between genomes of different species. Such horizontal transfer allows these elements to escape the various regulatory mechanisms imposed on them by their host genome, and to invade new genomes where they increase their copy number until new mechanisms evolve there to limit their spread.

Although many TEs are regulated in copy number and expression by various molecular mechanisms, some limiting forces are also at work at the population level. These forces suggest that there is selection against the direct deleterious effects of insertions, even if these effects are small, and against the chromosomal rearrangements that frequently occur when TEs of the same family are present. The impact of such forces depends on the TE involved, the structure of the population and its reproductive system. It results in a slight tendency towards either the elimination of TE copies or their accumulation in genomes, making time an important factor in shaping the aspects of the genomes affected by TEs. As a result of these controlling forces, genomes contain a mixture of TEs, some of which are still active, whereas others are ancient relics that have degenerated. **C.B. & C.V.**



an instance of retroviral activation of a cancer-promoting gene in gene-therapy trials[21].

Populations

Because TEs can transpose at high frequency, with a rate ranging from 10^{-3} to 10^{-5} per element per generation, depending on the element, they are more powerful producers of the raw material of evolution than is the classical nucleotide-base substitution rate, which is around 10^{-8}–10^{-9} per nucleotide per generation. So, waves of mobilization or loss through evolution may have had a major effect on the formation of new species, as has been suggested occurred in rodents[22].

Species and populations differ in the structure and copy numbers of their TE sequences[23] (Table 1). Indeed, although the number of genes in a genome increases from bacteria to higher organisms, it is the proportion of TEs and other kinds of repeated sequence that accounts for the main differences in genome size among species. Genome size is thus not correlated with the complexity of the organism. From a population-genetics point of view, genome size may increase in a passive way as a consequence of small population size[24]. This is because, in small populations, the efficiency of selection against the detrimental effects of TE insertions is reduced and the effect of random processes is increased, leading to the accumulation of TEs. But it is an open question whether the variation in genome size is indirectly associated with host population size, or whether it is directly promoted by environmental stress or by the novel environmental conditions that populations encounter when they invade a new habitat. The answer will bear on our understanding of, for example, how ancestral humans adapted after they migrated out of Africa.

Although analysis of the genomes of various species is necessary to understand their variation in size, composition and structure, such an approach will give little information on the historical mechanisms that changed them. This knowledge will come from comparison of the genomes of many populations from a given species, together with their environmental conditions. For example, in the fruitfly *Drosophila melanogaster*, such work has shown that TE copy number varies greatly among populations, which have sometimes been invaded by novel TEs from other species (horizontal transfers)[25]. Furthermore, waves of invasion or loss of TEs during evolution have been suggested for various elements, including the numerous LINEs in the human genome, with some TEs being in the process of invading while others are in the process of being eliminated[26]. Population surveys should be able to test this theory, and to furnish material to analyse the forces that control TE and gene expression in nature and that shape our genome (Box 3).

Finding the links between changes in epigenetic gene regulation during the early and late stages of development on the one hand, and changes in the environment, genome size, and population size and structure on the other, is the next task for this field. Possible systems in which to study these effects have been uncovered recently. For example, variations in gene expression have been observed between genetically identical mice[27], together with variation in tissue-specific gene expression among natural populations of the teleost fish *Fundulus heteroclitus*[28]. Also, strain-specific epigenetic variation that is associated with RNA abundance of a non-LTR retrotransposon has been found in the flowering plant *Arabidopsis*[29]. Research into understanding these mechanisms and how they might apply to disease will then be crucial. Such research promises to shed light on environment–genotype interactions and on the link between genotype and phenotype. What was once dismissed as junk DNA must now be regarded as a major player in many of the processes that shape the genome and control the activity of its genes. ■

Christian Biémont and Cristina Vieira are in the Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, CNRS, Université Claude Bernard Lyon 1, 69622 Villeurbanne Cedex, France.
e-mail: biemont@biomserv.univ-lyon1.fr

1. Brandt, J. *et al. Gene* **345,** 101–111 (2005).
2. Medstrand, P. *et al. Cytogenet. Genome Res.* **110,** 342–352 (2005).
3. Messing, J. & Dooner, H. K. *Curr. Opin. Plant Biol.* **9,** 157–163 (2006).
4. Nishihara, H. *et al. Genome Res.* **16,** 864–874 (2006).
5. Peaston, A. *et al. Dev. Cell* **7,** 597–606 (2004).
6. Walter, J., Hutter, B., Khare, T. & Paulsen, M. *Cytogenet. Genome Res.* **113,** 109–115 (2006).
7. Kalmykova, A. I., Klenov, M. S. & Gvozdev, V. A. *Nucleic Acids Res.* **33,** 2052–2059 (2005).
8. Buchon, N. & Vaury, C. *Heredity* **96,** 195–202 (2006).
9. Muotri, A. R. *et al. Nature* **435,** 903–910 (2005).
10. Martienssen, R. A., Doerge, R. W. & Colot, V. *Chromosome Res.* **13,** 299–308 (2005).
11. McDonald, J. F., Matzke, M. A. & Matzke, A. J. *Cytogenet. Genome Res.* **110,** 242–249 (2005).
12. Fraga, M. F. *et al. Proc. Natl Acad. Sci. USA* **102,** 10604–10609 (2005).
13. Kim, S. Y. *et al. Plant Cell* **17,** 3301–3310 (2005).
14. Lippman, Z. *et al. Nature* **430,** 471–476 (2004).
15. Morgan, H. D. *et al. Nature Genet.* **23,** 314–318 (1999).
16. Waterland, R. A. & Jirtle, R. L. *Nutrition* **20,** 63–68 (2004).
17. Hughes, J. F. & Coffin, J. M. *Genetics* **171,** 1183–1194 (2005).
18. Feinberg, A. P. *et al. Nature Rev. Genet.* **7,** 21–33 (2006).
19. Kazazian, H. H. *Curr. Opin. Genet. Dev.* **8,** 343–350 (1998).
20. Frendo, J. L. *et al. Mol. Cell Biol.* **23,** 3566–3574 (2003).
21. Kaiser, J. *Science* **310,** 1894–1896 (2005).
22. Grahn, R. A. *et al. Cytogenet. Genome Res.* **110,** 407–415 (2005).
23. Biémont, C. & Vieira, C. *Cytogenet. Genome Res.* **110,** 25–34 (2005).
24. Lynch, M. & Conery, J. S. *Science* **302,** 1401–1404 (2003).
25. Kidwell, M. G. & Lisch, D. R. *Evolution* **55,** 1–24 (2001).
26. Fablet, M. *et al. Gene* **375,** 54–62 (2006).
27. Pritchard, C. *et al. Genome Biol.* **7,** R6 (2006).
28. Whitehead, A. & Crawford, D. L. *Genome Biol.* **6,** R13 (2005).
29. Rangwala, S. H. *et al. PLoS Genet.* **2,** 271–281 (2006).
30. Capy, P. *Cytogenet. Genome Res.* **110,** 457–461 (2005).

©2006 Nature Publishing Group