Absence of close-facing retrotransposons: A comparison of molecular data and theory

Alexandros Bousios, David Waxman, Stephen R. Pearce

www.elsevier.com/locate/yjtbi

Absence of close-facing retrotransposons: A comparison of molecular data and theory

Alexandros Bousios, David Waxman, Stephen R. Pearce

PII: S0022-5193(09)00556-6
DOI: doi:10.1016/j.jtbi.2009.11.014
Reference: YJTBI5779

To appear in: *Journal of Theoretical Biology*

Received date: 10 July 2009
Revised date: 16 November 2009
Accepted date: 18 November 2009

Cite this article as: Alexandros Bousios, David Waxman and Stephen R. Pearce, Absence of close-facing retrotransposons: A comparison of molecular data and theory, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2009.11.014

TITLE:

Absence of Close-facing Retrotransposons: A Comparison of Molecular Data and Theory

AUTHORS:

Alexandros Bousios, David Waxman, Stephen R. Pearce*

AFFILIATION:

School of Life Sciences

University of Sussex,

Brighton BN1 9QG

Sussex UK.

*corresponding author

ABSTRACT:

Retrotransposons occur in extremely large numbers in many eukaryotic genomes. However, little is known of the factors which affect the distribution of close proximity retrotransposons. In this work we investigate the frequency of close facing retrotransposons in a plant species with extremely high numbers of retrotransposons. Molecular observations are compared with predictions of a mathematical model that assumes a uniform probability of retrotransposon insertion into the genome. The mathematical model plays the role of a null hypothesis. We find that compared with the predictions of the model, there is a statistically significant deficit of identical copies of facing retroelements that are close to one another. This suggests that an efficient mechanism exists that removes or limits close facing retroelements.

1. Introduction

Retroelements, which include retroviruses and retrotransposons, are selfish genetic elements that are present in all eukaryotic genomes (Coffin et al., 1997). Apart from causing major illnesses, such as HIV/AIDS, retroelements comprise a substantial and changing component of eukaryotic genomes, and have a profound effect on the composition and evolution of all eukaryotic species (Feschotte et al., 2002; Kazazian, 2004; Wicker and Keller, 2007).

In this work we focus solely on the LTR (long terminal repeat) retrotransposons (Kumar and Bennetzen, 1999; Vitte and Panaud, 2005). These mobile genetic elements are a subset of retroelements that are considered to be defective retroviruses as they lack components which enable them to escape the cell and be infective. Despite this, they retain the retrovirus replicative mode of transposition, which enables them to reinfect the nuclear genome in which they reside (Willhelm and Willhelm, 2001). Successive rounds of insertion therefore lead to an increase in retrotransposon copy number, with extremely large numbers of retrotransposons being found in many eukaryotic genomes (Pearce et al., 1996a; SanMiguel et al., 1996; Meyers et al., 2001).

Although retrotransposons are dispersed throughout the genome, a direct implication of a high copy number of such elements, or a particular preference for insertion into particular regions of the genome, is that the number of *closely spaced* elements will be high. In order to investigate the distribution of closely spaced elements, we adopted a molecular technique, IRAP (Inter-Retrotransposon Amplified Polymorphism), which is able to detect two copies of the same retrotransposon when they are facing each other and reside within the range of PCR (Kalendar et al., 1999, Kalendar and Schulman, 2006). We have formulated a simple mathematical model to provide concrete predictions of the molecular data.

The system adopted as the source of empirical data was required to have high copy numbers of transpositionally active retroelements. We chose the plant, *Agave tequilana*, for this purpose, as this organism has abundant and dynamic retrotransposon populations with very high copy numbers (Bousios et al., 2007). This organism is a good subject for this study, as the total retroelement load is likely to be greater than many species studied including rice (McCarthy et al., 2002) and *Arabidopsis thaliana* (Pereira 2004). Within *A. tequilana* we restricted analysis to two examples of the major classes of retrotransposons: Ty3-*gypsy* and Ty1-*copia* elements which both occur within this organism in very large numbers (Bousios et al., 2007). This last aspect makes these elements in this organism a source of highly reliable statistics, and hence ideally suited to the present study.

In the work described here only an outward facing primer from the 5'LTR was used, so only close proximity elements in a 5' facing orientation will produce an IRAP product and be detected (figure 1).

Figure 1

The number and size of close-proximity elements observed in the IRAP data reveal the number of adjacent, 5' facing retrotransposons and the distances between them. This allows us to test the hypothesis that retrotransposons have randomly inserted into the genome with no preference to be located near other elements of the same type, and no preference to insert in a particular genomic location. The hypothesis is introduced into this work in the form of a mathematical model which explicitly assumes a uniform probability of insertion anywhere within the genome. Such a probability of insertion does not lead to regularly spaced retrotransposons, since statistical fluctuations will automatically lead to variation of the distances between adjacent retrotransposons.

The model predicts the distribution of distances between adjacent retrotransposons, and the number of IRAP bands produced, given the number of retrotransposons and the size of the genome.

The molecular data differs significantly from the predictions of the model (which assumes a uniform probability of insertion) and indicates that close facing retrotransposon insertions are extremely rare or perhaps even absent in the genome. This may occur by either a universal mechanism to prevent close facing retrotransposon insertions or an efficient mechanism to remove retrotransposons with this orientation.

2. Results

2.1 Determining retrotransposon copy number

In order to compare observations and the predictions of the mathematical model introduced here, we require an estimate of the retrotransposon copy number, which is denoted n. We used a molecular technique to estimate n, namely quantitative Southern hybridization (Pearce et al. 1996a); see methods and Appendix A for details of this determination. This estimate is accurate within a factor of 2 so the actual copy numbers have a value, n, that lies within the range 0.5×10^8 to 2×10^8.

2.2 Modelling retrotransposon insertions

In Appendix B we present a full description of a mathematical model for the insertion of retrotransposons into the genome. The essence of the model is that retrotransposons have a uniform probability of insertion into all locations of the genome. The mathematical model plays the role of a null hypothesis in this work.

When there are n retrotransposons within a genome of length L, the distribution (probability density) of distances between adjacent retrotransposons is found to be very accurately given by the exponential distribution $f(r) = (n/L)\exp(-nr/L)$ - see Appendices B and C for details. Furthermore, when nr/L is small ($nr/L << 1$) as is relevant to this study, the distribution $f(r)$ is approximately constant:

$$f(r) \approx n/L \tag{1}$$

and hence behaves as a uniform distribution.

The retrotransposons that are actually detected by PCR are those that lie within a finite range of one another; a distance we denote D. The value of D adopted in this work is 2×10^8 bp (we use the abbreviation bp for base pairs) which is the size limit of the products of PCR (data not shown). The statistical nature of the dynamics of retrotransposons means that the number of IRAP bands observed in a particular genome is a random variable, which we denote B. When copy number n is large ($n >> 1$) and the ratio nD/L is small ($nD/L << 1$), as it is for our data sets, the number of IRAP bands is predicted from the model to have a mean or expected value (written $E[B]$) of

$$E[B] \approx n^2 D/(4L) \tag{2}$$

(see Appendix C for details). From analysis of the experimental data we arrived at the estimates in Table 1 for the number of observed IRAP bands, B_{obs}. Table 1 also contains the expected number of IRAP bands, $E[B]$, that results from Eq. (2), i.e., from the mathematical model, combined with the estimate of n and the values of D and L.

Table 1

We note that the overall number of observed IRAP bands, B_{obs}, for both types of retrotransposon, is close to the lower theoretical prediction for the expected number of bands, $E[B]$. This finding is suggestive that the overall number of observed IRAP bands is consistent with the estimated copy number although it is at the lower end of the range expected from a uniform probability of insertion. This is not surprising since a "real" population of retrotransposons (in contrast to an idealised theoretical population) will have a degree of sequence variability in their terminal sequences (through mutation) which will prevent amplification by PCR. Although this sequence variability will reduce the observed number of IRAP bands (B_{obs}) it will not influence the overall form of the size distribution of IRAP products.

In figure 2 we have plotted the theoretically predicted size distribution of IRAP fragments within the *A. tequilana* genome and compared this directly with the size profiles observed using IRAP for Teq1 and Teq17 retrotransposons.

Figure 2

2.3 Size distribution of IRAP products

The frequencies of IRAP products in the 0-1500bp size range that were observed for the two *A. tequilana* retrotransposons Teq1 and Teq17 were compared with the size distribution predicted in the 0-1500bp size range from random insertions of an element with similar copy numbers (denoted U in Figure 2).

The mathematical model predicted a size distribution of IRAP bands that is effectively uniform over the size range detected by IRAP (20-1500bp) – see frequency distribution U in Figure 2. The IRAP experiments for both types of retroelements show a slightly greater number of fragments in the middle of the size range (800bp) with lower levels than those predicted by the model at both the higher and lower ends of the size range. We make the assumption that the number of fragments in the middle of the size range (800bp) as being indicative of a "true" uniform distribution of fragment sizes. The mismatch of the observed frequency of 800bp fragments and the frequency of the theoretical distribution, labelled U, in Figure 2, can be taken as a measure of the accuracy of the estimate of copy number, *n*. The reduction in observed numbers of upper size fragments (>800bp) is not unexpected, given that longer PCR products are more difficult to amplify under the experimental conditions. The completely unanticipated observation is that IRAP products in the lower size range of the experiment are effectively absent. These 'effectively absent' fragments lie within the optimal size range for PCR amplification. This constitutes strong evidence that facing, closely spaced adjacent retroelements (i.e. those separated by less than 300bp) are an extreme rarity in this genome. The deficit of retroelements with small separations is statistically significant within the framework of the null hypothesis (i.e., under the assumption of a uniform probability of insertion). Obtaining the observed numbers of closely adjacent retroelements, given a uniform probability of insertion, occurs with probability $p < 10^{-4}$.

3. Discussion

Eukaryotic genome structure is complex and dynamic and no part of these genomes is as dynamic as the repetitive component (Morgante, 2006). The distribution and copy number of

retrotransposons observed in eukaryotes is the outcome of evolution, i.e., a complex history of amplification, mutation and deletion, resulting in profiles of retrotransposon insertion which are unique to particular lineages (Vitte and Bennetzen, 2006). The high number of repetitive sequences in eukaryotic genomes indicates that these make a significant contribution to genome size. In humans an expansion in genomic size was an ancient event, with transposition declining over the past 40 million years (Lander et al., 2001). By contrast, in many plants a genomic expansion has occurred in the last few million years (SanMiguel et al., 1998; Bennetzen, 2000; Vicient and Schulman, 2002).

For species with larger genomes, gene densities appear to be especially non-random, with the general picture emerging that genomes are organised into gene-dense regions separated by heterochromatic gene-poor blocks (Moore et al., 1995). Although it would be expected that retrotransposon insertion would be tolerated in non-coding regions of the genome (Bennetzen, 2000; Wicker et al., 2001), in-situ hybridisations show that retroelements are present (with minor exceptions) throughout the entire genome (Pearce et al., 1996b; Waugh et al., 1997). On a finer scale, and in particular where the distribution of individual retroelements is studied in more detail, the distribution of elements is less uniform with extremely high densities of retrotransposons existing in particular genomic locations (Vitte and Panaud, 2005). A number of studies have shown that insertion is common in coding sequences. In particular, in human and mouse genomes, many transposable elements are found in protein-coding genes (Nekrutenko and Li, 2001) with up to 25% of promoter regions of human genes containing transposable element sequences (Jordan et al., 2003). The discovery of retroelements influencing the expression of genes in wheat (Kashkush et al., 2003) and the preference of *Tos17* retrotransposon for gene-rich regions of rice (Miyao et al., 2003) show that the interaction of retroelements with genes is common. We can therefore see that although retroelements are present in almost every part of

the genome, each particular retroelement may have its own preferred location and although may not be abundant in the whole genome may be abundant locally (Vitte and Panaud 2005).

Considering the apparent non-random distribution of retroelements in genomes, it is interesting to speculate whether high copy number retroelements are directed to, or away from, particular areas of the genome or whether post-transposition mechanisms have operated to remove inserted elements in particular areas. Miyao et al. (2003) suggest that aggregation of retrotransposon insertions in host genomes is common and although mechanisms for retroelement removal were disputed for some time, these are now firmly established (Bennetzen, 2005). In rice the estimated half-life of a retroelement is 790,000 years (Wicker and Keller, 2007). Recombination and deletions between the long terminal repeats of the same retroelement has been proposed to occur through unequal homologous recombination (Bennetzen, 2005). The efficiency of removal through this mechanism is expected to increase with the size of the LTR (Bennetzen, 2005) and also with an inverse of the length of the intervening sequence, and could well be the mechanism behind the removal of close facing retroelement insertions. In humans, Alu repeats constitute 10% of the genome and even though they have a different mechanism of integration to the copia-like retroelements, closely spaced and inverted Alu elements are extremely unstable (Lobachev et al., 2000; Rowold and Herrera, 2000; Stenger et al., 2001). Similarly the findings of this study suggest that close proximity retroelement insertion is a similarly unstable event.

Although a growing amount of information is available on the distribution of repetitive sequences in eukaryotic genomes, this data is concentrated on a few species of high economic significance, and even with continued expansion of this activity, information on the distribution of repetitive elements in wild or uneconomic species will not become readily available. As IRAP banding numbers are dependant on retroelement copy number and genomic distribution, we have

demonstrated how this technique, in combination with an essentially null theoretical model, can be used to investigate the insertional dynamics of a particular retroelement. By comparing IRAP profiles and copy numbers with a random insertion model we observe that the insertion of the two high copy number elements studied here is broadly in line with the model but that close retroelement facing insertions are extremely uncommon. Although close opposed insertions may represent a small fraction of total insertion events for randomly inserting high copy number elements, this effect will be much greater for any retroelements which specifically target particular genomic locations and may be a significant mechanism for reducing their overall copy number.

4. Materials and methods

***4.1 Plant materials*:** *Agave tequilana* (var. Weber azul) DNA was prepared from a leaf sample by DNeasy plant mini kit (Qiagen).

4.2 Inter-Retrotransposon amplified Polymorphism (IRAP) method.

Primers were designed to complement the 5'LTR of Teq1 and Teq17 retrotransposons (Bousios et al., 2007). In principle, retrotransposons can integrate into the genome in either orientation, giving head to head, tail to tail, or head to tail orientations. In this work single outward facing primers were used, thus only adjacent retrotransposons with 5'LTRs facing were detected (figure 1).

IRAP was performed in a 20µl reaction mixture containing 20ng DNA, 1 x PCR buffer, 5pmol primer, 200nM dNTP, 1u Taq polymerase (NEB). The PCR programme consisted of 1 cycle at 94˚C for 5 minutes followed by 30 cycles of 1 minute at 94˚C, 1 minute at 64˚C and 2 minutes at

72˚C. ^{33}P dATP was included in the reaction to enable products to be visualized by autoradiography after resolution by denaturing acrylamide gel electrophoresis.

Size determination of IRAP products was achieved by excising fragments from the dried gel followed by re-amplification, subcloning into the TOPO TA vector (Invitrogen) and sequencing.

4.3 Slot Blotting

Agave tequilana genomic DNA was quantified spectrophotometrically, serially diluted, denatured in 0.2M NaOH/2M NaCl for 5 min, and transferred to a Hybond N^+ membrane (Amersham) using a vacuum slot blotter. Control double stranded plasmid DNAs containing the individual agave RNaseH fragments specific to each probe (Bousios et al., 2007) were quantified by ethidium bromide agarose gel electrophoresis spectrophotometrically, serially diluted and transferred as above. Cloned RNaseH PCR fragments were direct labelled with alkaline phosphatase, using AlkPhos direct labelling (Amersham). These were hybridized to the filters and washed with varying levels of stringency with the highest being equivalent to 0.1 x SSC at 60˚C. Hybridisation was detected by chemiluminescence using CDP-Star detection reagent (Amersham) using Kodak MXB film.

Figure legends

Figure 1: Detection of close facing retroelement insertions by IRAP.

Retrotransposon structure is illustrated, with elements in different orientations represented as arrows (a-d). The intervening and flanking sequence are indicated with wavy lines. As the single primer is homologous to the reverse complement of the 5' LTR terminal sequence (black arrow), (d) is the only orientation which produces IRAP products (indicated by a bold wavy line and grey arrows).

Figure 2: Size distribution of IRAP products.

Frequencies of IRAP products in the 0-1500bp size range observed for the two *A. tequilana* retrotransposons Teq1 and Teq17 are presented, along with the size distribution predicted in the 0-1500bp size range from a model with a uniform probability of insertion (labelled U in the Figure), using an estimated value of the element copy numbers, *n*.

Figure 3: Copy number of Teq1 and Teq17 Ty1-*copia* retrotransposons in *Agave tequilana*.

(For each panel) Row A contains cloned target DNA with 2.207×10^9 molecules (left slot), 2.207×10^{10} molecules (middle slot) and 2.207×10^{11} molecules (right slot). Row B contains *A. tequilana* genomic DNA with 2.207×10^4 genomes (middle slot) and 2.207×10^5 genomes (right slot).

Table 1: This table contains the *A. tequilana* genome size *L* (in base pairs), the maximum size *D* (in base pairs) of the products of PCR, the copy number *n*, of Teq1 and Teq17 retroelements (see Appendix A for the method of copy number estimation), the expected number of IRAP bands

that were predicted from the mathematical model $E[B]$, and the number of IRAP bands determined from Molecular data B_{obs}. Generally, the number and size of IRAP bands varies from genome to genome because each retrotransposon has a unique history of activity and insertion profile.

References

Bennett, M. D., Leitch, I. J., 2004. Angiosperm DNA C-values database (release 5.0, Dec. 2004). http://www.kew.org/cvalues/homepage.html.

Bennetzen, J. L., 2000. Transposable element contributions to plant gene and genome evolution. Plant Molecular Biology, 42, 251-269.

Bennetzen, J. L., 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. Current Opin Genet & Dev 15, 621-627.

Bousios, A., Saldana-Oyarzabal, I., Valenzuela-Zapata, A. G., Wood, C., Pearce, S. R., 2007. Isolation and characterization of Ty1-*copia* retrotransposon sequences in the blue agave (*Agave tequilana* Weber var. azul) and their development as SSAP markers for phylogenetic analysis. Plant Science 172, 291-298.

Coffin, J. M., Hughes, S., Varmus, H. E., 1997. Retroviruses. Cold Spring Harbor Press (USA).

Feschotte, C., Jiang, N., Wessler, S. R., 2002. Plant transposable elements: Where genetics meets genomics. Nature Reviews Genetics 3, 329-341.

Hogg, R. V. and Craig, A. T., 1970. *Introduction to Mathematical Statistics*. Macmillan. London

Jordan, I. K., Rogozin, I. B., Glazko, G. V., Koonin, E. V., 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19, 68-72.

Kalendar R., Grob, T., Regina, M., Suoniemi, A., Schulman, A. H., 1999. IRAP and REMAP: Two new retrotransposon-based DNA fingerprinting techniques. Theor Appl Genet 98, 704-711.

Kalendar, R., Schulman, A. H., 2006. IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. Nature protocols 1, 2478-2484.

Kashkush, K., Feldman, M., Levy, A. A., 2003. Transpositional insertion of retrotransposons alters the expression of adjacent genes in wheat. Nature Genetics 33, 102-106.

Kazazian, H. H., 2004. Mobile elements: drivers of genome evolution. Science 303, 1626-1632.

Kumar, A., Bennetzen, J. L., 1999. Plant retrotransposons. Ann Rev Genet 33, 479–532.

Lander, E. S., Linton, L. M., Birren, B., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Lobachev, K. S., Stenger, J. E., Kozyreva, O. G., Jurka, J., Gordenin, D. A., Resnick, M. A., 2000. Inverted Alu repeats unstable in yeast are excluded from the human genome. EMBO J 19, 3822-3830.

McCarthy, E. M., Liu, J., Lizhi, G., McDonald, J. F., 2002. Long terminal repeat retrotransposons of *Oryza sativa.* Genome Biol 3, R 0053.1-0053.11.

Meyers, B. C., Tingley, S. V., Morgante, M., 2001. Abundance, distribution and transcriptional activity of repetitive elements in the maize genome. Genome Research 11, 1660-1676.

Miyao, A., Tanaka, K., Murata, K., Sawaki, H., Takeda, S., Abe, K., Shinozuka, Y., Onosato, K., Hirochika, H., 2003. Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. Plant Cell 15, 1771-1780.

Moore, G., Devos, K., Wang, Z., Gale M., 1995. Cereal genome evolution: Grasses, line up and form a circle. Curr Biol 5, 737-739.

Morgante, M., 2006. Plant genome organization and diversity: the year of the junk! Curr Opin in Biotech 17, 168-173.

Nekrutenko, H., Li, W., 2001. Transposable elements are found in a large number of human protein-coding genes. Trends Genet 17, 619-621.

Pearce, S. R., Harrison, G., Li, D., Heslop-Harrison, J. S., Kumar, A., Flavell, A. J., 1996a. The Ty1-*copia* group retrotransposons in *Vicia* species: Copy number, sequence heterogeneity and chromosomal localisation. Mol Gen Genet 250, 305-315.

Pearce, S. R., Harrison, G., Pich, U., Heslop-Harrison, J. S., Kumar, A., Flavell, A. J., 1996b. The Ty1-*copia* group retrotransposons of *Allium cepa* are distributed throughout the chromosomes but are enriched in the terminal heterochromatin. Chromosome Res 4, 357-364.

Pereira, V., 2004. Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. Genome Biol 5, R79.

Rowold, D. J., Herrera, R. J., (2000). Alu elements and the human genome. Genetica, 108, 57-72.

SanMiguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z.,. Bennetzen, J. L., 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274, 765–768.

SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., Bennetzen, J. L. 1998. The Paleontology of intergene retrotransposons of maize. Nature Genetics 20, 43-45.

Stenger, J. E., Lobachev, K. S., Gordenin, D., Darden, T. A., Jurka, J., Resnick, M. A., 2001. Biased distribution of inverted and direct Alus in the human genome: Implications for insertion, exclusion and genome stability. Genome Research 11, 12-27.

Vicient, C. M., Schulman, A. H., 2002. Copia-like retrotransposons in the rice genome: few and assorted. Genome Lett 1, 35-47.

Vitte, C., Bennetzen, J. L., 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci (USA) 103, 17638-17643.

Vitte, C., Panaud, O., 2005. LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet Genome Res 110, 91-107.

Waugh, R., McLean, K., Flavell, A. J., Pearce, S. R., Kumar, A., Thomas, W. T. B., Powell, W., 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.

Wicker, T., Stein, N., Albar, L., Feuillet, C., Schlagenhauf, E., Keller, B., 2001. Analysis of a contiguous 211kb sequence in diploid wheat (*Triticum monococcum L.*) reveals multiple mechanisms of genome evolution. Plant J 26, 307-316.

Wicker, T., Keller, B., 2007. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. Genome Research 17, 1072-1081.

Willhelm, M., Willhelm F. X., 2001. Reverse transcription of retroviruses and LTR retrotransposons. Cell Mol Life Sci 58, 1246-1262.

Appendix A: Retroelement copy number determination

In this Appendix, we provide details of the estimation of copy number of elements within the genome.

Estimation of total copy number was carried out by quantitative Southern blotting (see Figure 3).

Figure 3

Specific retroelement probes, corresponding to 164bp sections of the RNaseH gene, were generated from clones containing larger fragments of the Teq1 and Teq17 retroelements (Bousios et al., 2007). A series of concentrations of control target DNA (row A) and *A. tequilana* genomic DNA (row B) were analysed. As the genome size of *A. tequilana* is known, namely 2C = 8.4 x 10^9bp (Bennett and Leitch, 2004) the exact number of genomes loaded can be calculated (1G = 1μg = 2.207 x 10^5 genomes). The amount of target was chosen to correspond to different numbers of 1G genomes, ranging from 10^4 to 10^6. Direct comparison of the intensity of genome hybridization and that of the control sequences provide a good estimate of the number of elements in the genome. For both Teq1 and Teq17, the intensity of the genome hybridization (1G) approximately coincides with the 10^5 control lane, thereby indicating this number of element copies, per genome, in each case. This result is accurate within an overall factor of 2, hence the retrotransposon copy number has a value in the range 0.5×10^5 to 2×10^5.

Appendix B*:* Mathematical model

In this Appendix, we define the mathematical model on which we base our analysis of the distribution of distances between adjacent retrotransposons. This Appendix also contains a summary of the key results of the model. All calculational details leading to the results in this Appendix are given in Appendix C.

The retrotransposons under consideration have a moderate size - approximately 8000bp (bp = base pair), and we shall focus on the 20bp IRAP primer site in the forward (5'LTR) terminal sequence of a retrotransposon. We term this 20bp region the "tip" of the element. Then the assumptions of the model are as follows.

Assumptions

1. The size of the retrotransposon tip is negligible compared with the length of the genome (the length of the genome is 8.4×10^9bp)
2. The possible locations of the tip of a transposable element, within the genome, can take continuous values between 0 and L, where L is the length of the genome in bp. That is, we neglect discreteness of the genome.
3. The tip of a retrotransposons has a uniform probability of insertion at all locations of the genome.
4. Only 1/4 of all adjacent retrotransposons that are within the PCR distance D of one another are detected. The factor 1/4 arises since only the tips of adjacent retrotransposons with 5'LTRs facing are detected (see Figure 1) and each of the four possible orientations of adjacent elements are assumed equally likely.
5. The fraction of the genome occupied by tips of transposable element is sufficiently small that there is negligible chance of tips of different retrotransposons inserting within one another.

Under these assumptions it is possible to predict the *distribution of distances* between the tips of adjacent retrotransposons and make an explicit prediction about the *mean number* of IRAP bands observed (see Appendix C for details).

In particular, when there are n retrotransposons within the genome, the distribution (probability density) of distances between adjacent retrotransposons, neglecting the length of the tips, is found (see Appendix C) to be given by $f(r) = (n/L)(1 - r/L)^{n-1}$ and when n is large (n>>1) this distribution is very accurately given by the exponential distribution

$$f(r) = (n/L)\exp(-nr/L). \quad (1)$$

The distribution labelled U, in Figure 2, is based on Eq. (1).

The number of observed IRAP bands in a particular genome is a random variable that we denote B. The definition of B follows primarily from Assumption (4) above, namely an IRAP band is observed when adjacent retrotransposons have 5'LTRs facing and their tips are less than D base pairs apart (D is the maximum size of PCR products). We have determined the expected value B, written $E[B]$, which represents the average of B over all possible genomes with n transposable elements. When copy number n is large ($n >> 1$) and nD/L is small ($nD/L << 1$), as it is for our data set, $E[B]$ is found (see Appendix C) to be given by

$$E[B] \approx n^2 D/(4L). \quad (2)$$

Appendix C: Calculations of IRAP banding from the model

In this Appendix, we give details of the calculations for the mathematical model adopted in this paper.

We assume there are n transposable elements in the genome and that these may be treated as point entities which have positions arising from a uniform probability of insertion into the genome. Thus their positions lie in the continuous range 0 to L (L = length of the genome).

The positions of the elements are X_j (j =1, 2, ..., n) and these are independent and identically distributed random numbers that are drawn from a uniform distribution on [0, L]. It is convenient to go to an ordered description of the positions. We thus set Y_n to be the largest of the X_j, Y_{n-1} to be the next largest, ..., and ultimately Y_1 to be the smallest. The Y_j are termed order statistics (Hogg and Craig, 1970). The probability density of distances between adjacent elements is given as a function of distance, r, by

$$f(r) = \frac{1}{n-1}\sum_{j=1}^{n-1} E\left[\delta\left(r - (Y_{j+1} - Y_j)\right)\right] . \qquad (3)$$

Here $\delta(r)$ denotes a Dirac delta function (it is a spike of infinite height, zero width, and an area of unity that is located at $r = 0$) and E[...] denotes a mathematical expectation corresponding to an average over all possible genomes with n transposable elements. The joint probability density of the order statistics Y_j and Y_{j+1} is written $g(y_j, y_{j+1})$ and is non zero only for the range $0 \le y_j < y_{j+1} \le L$. In this range $g(y_j, y_{j+1})$ is given by (Hogg and Craig, 1970)

$g(y_j, y_{j+1}) = \frac{n!}{(j-1)!\,(n-j-1)!}[P(y_j)]^{j-1}[1 - P(y_{j+1})]^{n-j-1} p(y_j) p(y_{j+1})$ where $p(y) = \frac{1}{L}$ and $P(y) = \frac{y}{L}$. It follows that for $0 \le r \le L$ we have

$E\left[\delta\left(r - (Y_{j+1} - Y_j)\right)\right] = \int_0^L dz \int_0^z dy\, \delta(r - (z - y)) g(y, z)$. This may be evaluated in closed form, and is given by

$$E\left[\delta\left(r - (Y_{j+1} - Y_j)\right)\right] = \left(n/L\right)\left(1 - \frac{r}{L}\right)^{n-1} . \qquad (4)$$

When this result is used in Eq. (3) we obtain one of our theoretical results, namely that the probability density of distances between adjacent elements is $f(r) = (n/L)(1 - \frac{r}{L})^{n-1}$.

Next, let us consider the expected number of IRAP bands observed. With $\Theta(r)$ denoting a Heaviside step function ($\Theta(r) = 1$ for $r \geq 0$ and vanishes otherwise), an IRAP band is observed when adjacent retrotransposons have 5'LTRs facing and their tips are less than D base pairs apart. The expected number of bands is given by

$$E[B] = \sum_{j=1}^{n-1} E\left[\left(\Theta\left(D - (Y_{j+1} - Y_j)\right)\right)/4\right] \qquad (5)$$

where the factor 1/4 follows from the assumption that all 4 four possible orientations of adjacent elements are equally likely, so, on average, only 1/4 of all adjacent retrotransposons have 5'LTRs facing.

Equation (5) can also be written as $E[B] = \frac{1}{4}\sum_{j=1}^{n-1}\int_0^D E\left[\delta\left(r - (Y_{j+1} - Y_j)\right)\right]dr$ and using Eq. (4) yields $E[B] = \frac{n-1}{4}\left[1 - \left(1 - \frac{D}{L}\right)^n\right]$ which is our second theoretical result.

Fig 1

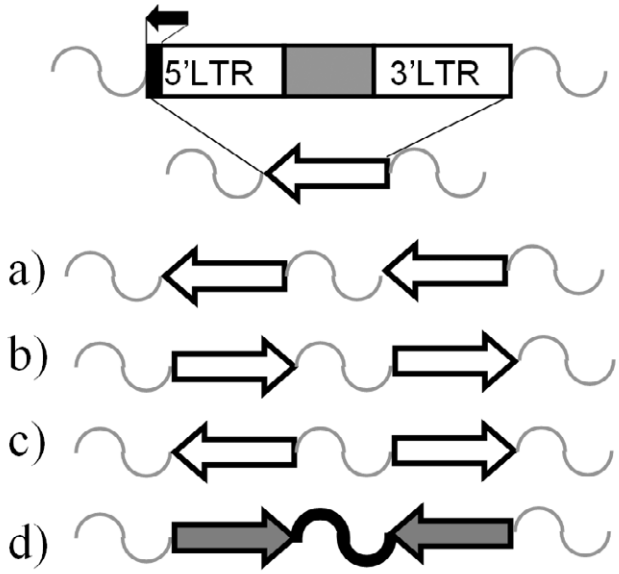


Fig 2

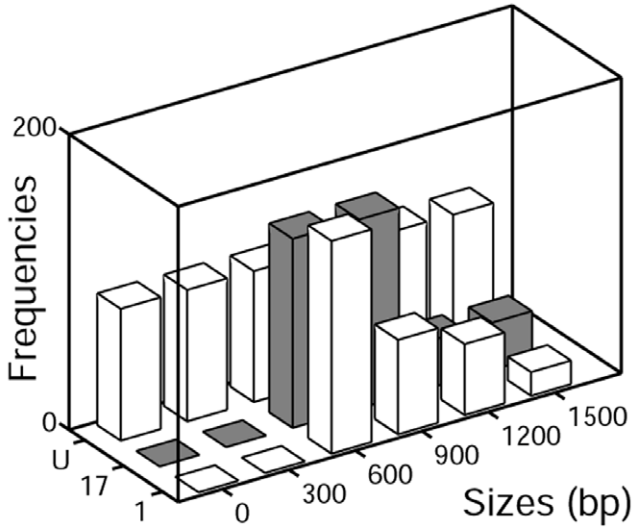
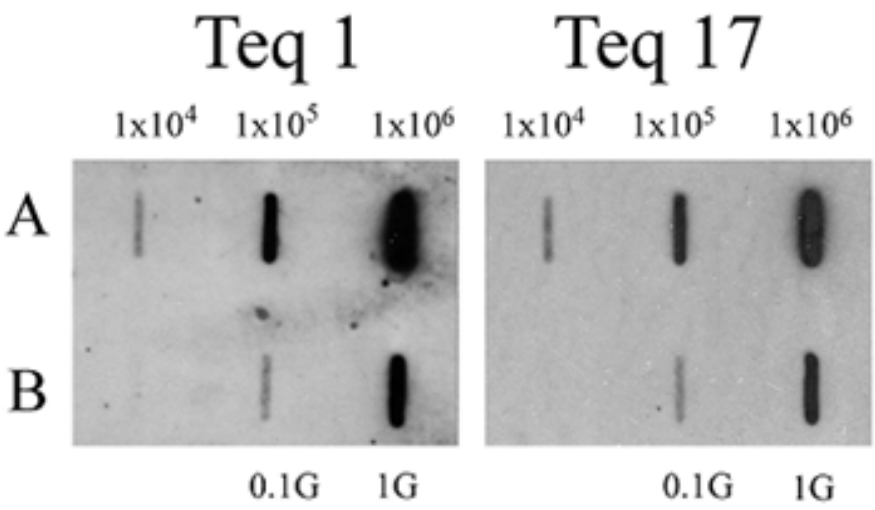


Fig 3



Table

	L	D	n	E[B]	B_{obs}
Teq 1	8.4×10^9	2×10^3	$0.5 \times 10^5 - 2 \times 10^5$	$297 - 1190$	272
Teq 17	8.4×10^9	2×10^3	$0.5 \times 10^5 - 2 \times 10^5$	$297 - 1190$	320