Anatomy of Highly Expressing Chromosomal Sites Targeted by Retroviral Vectors[†]

Christian Mielke, Karin Maass, Meike Tümmler, and Jürgen Bode*

GBF, Gesellschaft für Biotechnologische Forschung mbH, Genregulation und Differenzierung/Genetik von Eukaryonten, D-38124 Braunschweig, Mascheroder Weg 1

Received October 5, 1995; Revised Manuscript Received December 6, 1995[⊗]

ABSTRACT: The eukaryotic genome contains chromosomal loci with a high transcription-promoting potential. For their identification in cultured cells, transfer of a reporter gene has to be performed by a technique that grants the integration of individual copies. We have applied retroviral vectors in conjunction with inverse polymerase chain reaction techniques to reconstruct a number of these sites for a further characterization. Remarkably, all examples conform to the same design in that the process of retroviral infection selected a scaffold- or matrix-attached region (S/MAR) that was flanked by DNA with high bending potential. The S/MARs are of an unusual type in that they show a high incidence of certain dinucleotide repeats and the potential to act as topological sinks. The anatomy of retroviral integration sites reveals principles that can be exploited for the development of predictable transgenic systems on the basis of expression and targeting vectors.

Gene transfer into mammalian cells is an important tool for production of recombinant proteins for functional studies and therapeutic applications. While episomal systems allow for large expression levels owing to high copy numbers, these systems frequently suffer from a lack of long term stability (Klehr & Bode, 1988). As an alternative, the synthesis of recombinant proteins can be directed by sequences that have integrated and will hence propagate as a parts of the host's genome. The relative stability of integrated sequences is advantageous or even required by safety regulations, but there is increasing evidence that it depends on the presence of few or even a single copy of the transgene. Standard gene transfer techniques favor the integration of multiple copies which in many cases proved to be intrinsically unstable. This instability has been ascribed to the characteristic head-to-tail mode of integration which promotes the loss of coding sequences by homologous recombination (Weidle et al., 1988), especially in cases where these are transcribed (McBurney et al., 1994). In addition, there are epigenetic defense mechanisms directed against multiple copy integration events. This phenomenon has been termed "cosuppression" for plants (Allen et al., 1993), where it is not uncommon that the level of expression is inversely related to copy numbers. Similarly, a case has been desribed for mammalian cells where each 10-fold increase in gene copy number was accompanied by a 10-fold decrease in per-copy expression (Kalos et al., 1994). These observations are in line with findings that multiple copies become inactivated by methylation (Mehtali et al., 1990) and subsequent mutagenesis (Kricker et al., 1992) or silenced by heterochromatin formation (Dorer & Henikoff, 1994). Since vectors based upon the principles of retroviral infection are free from most of these drawbacks and since they are applicable for a wide variety of cell types, they have recently become the preferred tool with which to approach problems in molecular biology and, in particular, gene therapy.

A unique feature of retroviruses is their existence as inherited elements in the germ line of many vertebrates where they behave as reasonably stable Mendelian genes. It has been estimated that as much as 5−10% of the mammalian genome consists of elements introduced by mechanisms involving reverse transcription. Although the deleterious effects of retroviral integration have been the focus of many investigations, the amplification of a selected provirus indicates a beneficial function for the host species, as the expression of *env* proteins can induce interference toward infection by viruses that use the same receptor or tolerance to antigens related to the encoded ones. The extended coevolution of complex organisms with proviruses therefore suggests a highly optimized relationship owing to the kind of integration sites and expression patterns (Taruscio & Manuelidis, 1991).

From the viewpoint of a retroelement, it would clearly be disadvantageous to integrate into a chromosomal context that prevents transcription. Retrotransposons such as the Ty elements of yeast have the capacity to insert specifically into active chromatin regions, i.e. into positions within 1−4 base pairs (bp) of the initiation site for tRNA gene transcription (Craigie, 1992). This mode of target site selection seems to involve direct interaction of the integration machinery with the Pol III transcription apparatus and may help to minimize deleterious effects on the host caused by integration into essential coding or regulatory regions. Murine retroviruses have extensive structural and functional similarities to retrotransposons (Kulkosky et al., 1992), and the still prevailing view is that transcriptionally active regions and regions associated with DNase I hypersensitive sites are likewise preferred by their integration machinery. While these findings have in some cases been ascribed to selective pressure which may have conferred a growth advantage to some cells, such an influence was excluded in other examples (Scherdin et al., 1990).

[†] This work was supported in part by a grant from Deutsche Forschungsgemeinschaft (Bo 419/5-1) and EG funds PL 933069.

* Address correspondence to Dr. Jürgen Bode, GBF, Gesellschaft für Biotechnologische Forschung mbH, Genregulation und Differenzierung/Genetik von Eukaryonten, Mascheroder Weg, D-38124 Braunschweig. Telephone: (0531)6181251. Fax: (0531)6181262. E-mail: jbo@gbf-braunschweig.de.

[⊗] Abstract published in *Advance ACS Abstracts,* February 1, 1996.

0006-2960/96/0435-2239$12.00/0

Until recently, most methods for isolating proviral flanking sequences involved plasmid rescue, and by such an approach, a small number of highly preferred integration targets has been identified (Shih et al., 1988). With the application of PCR[1] techniques designed to study integration without prior selection by molecular cloning, this picture has started to change because these very targets could not be recovered. Possibly, the originally defined loci shared some features that allowed them to be selectively packaged by bacteriophage λ (Withers-Ward et al., 1994). The emerging view is that virtually all regions of the genome are initially accessible to the retroviral integration machinery although there exist pronounced localized preferences within the inspected areas (Engelman, 1994). Such a global accessibility could be explained by the perfect correlation between cellular division and integration which means that the target is a chromatin that has not yet attained its compact structure in the subsequent G1 phase. Alternatively, the IN complex could be specifically directed toward replication sites (Hajihosseini et al., 1993) which might be preferred due to their association with the nuclear matrix (Berezney, 1991).

Another way to reconcile these apparently conflicting results is based on a closer inspection of the integration process. Retroviruses are a family of single stranded RNA (ssRNA) viruses that replicate through a DNA intermediate. The viral DNA is initially found in the cytoplasm of the host cell, assembled as a 160 S integrase (IN) complex which contains all enzymatic functions for the integration reaction except for the activities required for the terminal stages; before nuclear transport, IN specifically removes a dinucleotide from the 3′-ends of both DNA strands, and this is a prerequisite for integration (Panganiban & Temin, 1983). A staggered cut is introduced into the target, and the 3′-hydroxyl termini of the processed retroviral DNA are joined to the resulting 5′-protruding ends. The process is completed by degradation of the unpaired viral nucleotides and gap repair to fill the single strand connections, creating a duplication of the terminal bases of cellular DNA; these functions are provided by the repair system of the host cell, residing at the nuclear matrix [reviewed by Boulikas (1995)]. Whereas an immediate PCR amplification detects all insertion events directed by IN, a functional provirus will only arise after these concluding steps.

How does IN recognize its host target site? While purified IN serves the basic functions, the complete nucleoprotein complex is thought to be required for the recognition of specific chromatin structures and the high fidelity of integration in vivo. Although it is possible that some provirus finds itself in a location that discourages a high expression, favorable integration events can be isolated by application of a selection system. In this paper, we describe a combined infection/selection protocol and the subsequent characterization of highly expressing genomic loci. All recovered sites showed common molecular features which are thought to enhance their recombinogenic and transcriptional potential.

MATERIALS AND METHODS

Plasmids

Retroviral vectors, the inserts of which are shown in Figure 1A, are derivatives of pM5neo (Laker et al., 1987). This vector comprises retroviral sequences necessary for an efficient transcription and packaging derived from the myeloproliferative sarkoma virus (MPSV). For pM5capa, the PAC (puromycine *N*-acetyltransferase) gene is driven by the 5′-LTR, whereas the CAT cassette is under the control of the SV40 promoter/enhancer. For the construction of pM5sepa, the coding sequences were excised and the gap was filled with the aid of linkers by SEAP (secretory alkaline phosphatase), IRES (internal ribosomal enty site), and PAC sequences to yield pM5sepa (Mielke, 1993).

Cell Culture and Retrovirus Infection

Cell Culture, Transfection, and Electroporation. NIH 3T3 murine embryo fibroblasts (ATCC CRL 1658) and ψ-2 packaging cells (Mann et al., 1983) were cultured in 1,2-dimethoxyethane (DME) medium containing 10% fetal calf serum, 20 mM glutamine, 60 μg/mL penicillin, and 100 μg/mL streptomycin and passaged at the time of confluence. Transfection and electroporation routines were as described before (Mielke et al., 1990). Infectious retrovirus particles were generated after transfection of pM5capa or pM5sepa under stable expression conditions.

Infection. The viruses produced by the line ψ-2 (Mann et al., 1983) were used to infect murine 3T3 target cells, applying conditions that rendered 1 in 1000 cells resistant. Such a low titer reduces the risk of multiple infections in a given cell. In detail, virus-producing cells were seeded such that they barely achieved confluence 2 days later. After day one, the medium was replaced by a small amount (10 mL per 75 cm^2 flask) of fresh medium which was withdrawn after another 24 h to be used as infectious medium following passage through a 0.45 μm filter.

3T3 cells were seeded 1 day prior to infection (5 × 10^4 cells per 25 cm^2 flask) and were then treated with dilutions of the infection medium containing 8 μg/mL polybrene. The selection was initiated 24−48 h postinfection; the infected cells were passaged onto large plates (140 cm^2) in a selection medium containing 2.5 μg/mL puromycin which was changed every third day before clones appeared after 8−21 days. When clones reached a diameter of 1 mm, they were immobilized in a mixture of 4 mL of agarose (1.8% in water) and 12 mL of cold medium. After incubation overnight, clones were collected with a syringe and transferred onto a 48-well plate. Twenty-four individual clones were selected for further characterization.

Reporter Assays

SEAP was quantified as described by Berger et al. (1988) or using an overlay assay according to Kirchhoff et al. (1995). Use of PAC as a reporter gene followed the procedure of Mielke et al. (1995).

Copy Number Determination

Copy numbers were quantified by a refined Southern analysis using two probes, one directed toward an intrinsic gene element (murine IFN-β promoter) as the internal control

[1] Abbreviations: FLP, site-specific recombinase from 2 μm plasmid of *Saccharomyces cerevisiae*; FRT, FLP recognition target; IN, retroviral integrase; IPCR, inverse polymerase chain reaction; IRES, internal ribosome enty site; LTR, long terminal repeat; MAR, matrix-attached region; ORI, origin of replication; SAR, scaffold-attached region; S/MAR, consensus term covering SARs and MARs; PAC, puromycine *N*-acetyltransferase; SEAP, secretory alkaline phosphatase; SINE, short interspersed repetitive element.

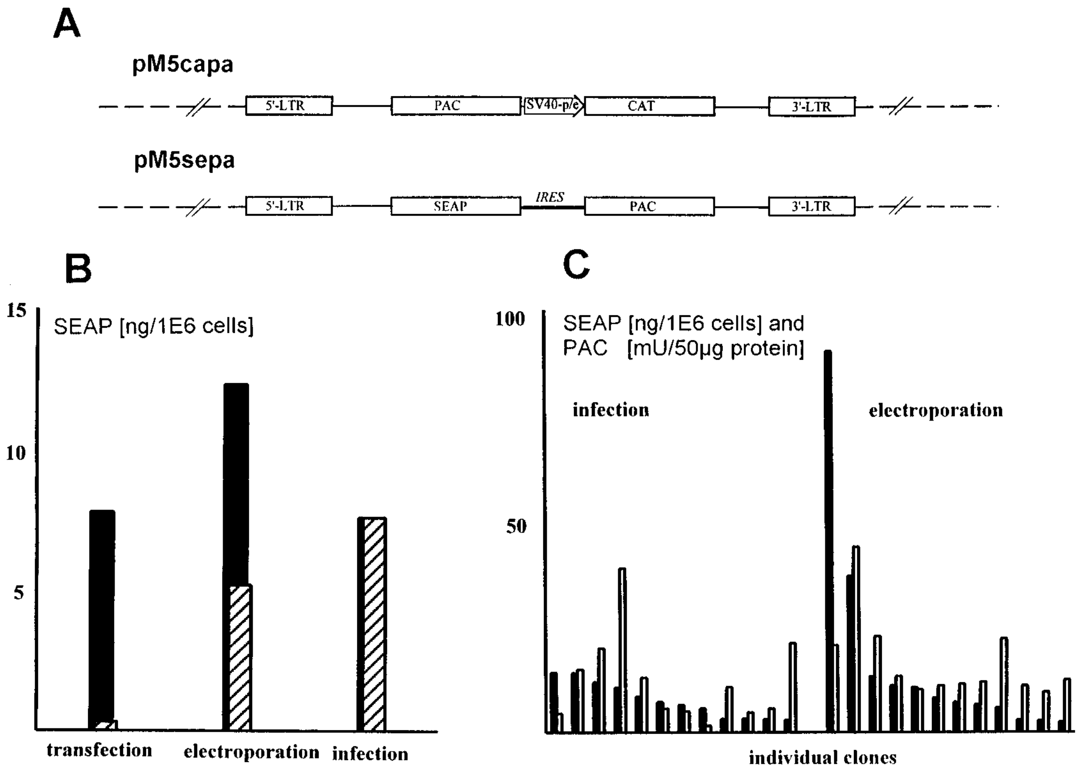


FIGURE 1: Expression levels obtained for retroviral vectors by three gene transfer techniques under selective pressure (2.5 μg of puromycin/mL of DME). (A) Vectors used. For pM5capa, the PAC gene is driven by the 5′-LTR, whereas the CAT cassette is under the control of an SV40 promoter/enhancer. Transcriptional activity of two closely linked promoters is required whereby the second promoter may be shut off by transcriptional interference. The pM5capa derivative, pM5sepa, was obtained by excising the coding sequences and filling the gap by SEAP, IRES, and PAC sequences to yield pM5sepa. Here, both ORFs are part of one bicistronic transcript; translational initiation of the downstream ORF is mediated by an internal ribosomal entrance site. (B) Expression levels obtained from pM5sepa after transfer by transfection, electroporation, and infection. Solid bars represent the total expression level of the entire pool of PAC-resistant clones. Hatched bars represent the corresponding levels referenced to the average copy level. (C) Expression for SEAP (solid bars) and PAC (light bars) determined for a selection of individual clones. In case of electroporation, the left-hand clones contain 13 and 20 copies, respectively. All other clones represent single copy integration events.

and the other specific for a part of the transferred gene. The sequences of both probes were cloned into a single vector. Restriction of this vector with the enzymes used for genomic DNAs and its simultaneous application for Southern analyses grants the presence of both fragments at a stoichiometric ratio of both probes on each filter. This provides an excellent control that permits the exact correction of different signal-to-noise ratios that would otherwise occur. Quantification of autoradiographs was achieved with a Molecular Dynamics phosphoimaging system.

Purification of Genomic DNA

Cells from a 75 cm^2 culture flask were washed with PBS, harvested in TEN buffer, and pelleted for 5 min at 1000 rpm in a table top centrifuge. The pellets were suspended in 1 mL of TEN and provided with 9 mL of extraction buffer [0.5% sodium dodecyl sulfate (SDS), 100 mM ethylenediaminetetraacetic acid (EDTA), 10 mM Tris/HCl (pH 8.0), and 20 μg/mL RNAse A]. After 5 h of incubation at 37 °C, 50 μL of proteinase K solution (20 mg/mL in water) was added and incubation was continued overnight. Following a standard phenolization step, solutions were extensively dialyzed against five changes of TE.

Inverse PCR Techniques

Primers 1−8 (Figure 2), used for amplifying genomic sequences adjacent to the 3′- or 5′-end of the retroviral insert, had the following composition (in parentheses, positions in pM5capa; lowercase letters, tails; underlined, sequences providing additional *Pst*I or *Hin*dIII sites, respectively, cf. Figure 2): 1, gggcggctgcagtaGCTTGCCAAACCTACAGG (25−42); 2, GCAAAATGGCGTTACTTAAGC (45−65); 3, GTTCCCGCCTCCGTCTG (952−968); 4, GCTTTCGGTTTGGGACCG (976−993); 5, GCCTGGACCACCTGATATCC (3192−3211); 6, GGTGATATTGTTGAGTC (3221−3237); 7, GTTCCTTGGGAGGGTCTCC (3838−3856); 8, gggcggaagctttcTGAGTGATTGACTACCCG (3860−3876).

Genomic DNA (500 μg) was incubated with 20 units of the appropriate restriction enzyme (*Pst*I, *Hin*dIII, or *Eco*RI, respectively) in 1x PCR buffer for 4 h in a total volume of 100 μL. After inactivation (20 min at 68 °C), the mixture was supplied with 900 μL of 1x PCR buffer plus 10 units of ligase without further additions of ATP [see Weiss et al. (1968)]. Following an overnight incubation at room temperature, PCR was initiated by transferring 5 μL (the equivalent of 2.5 ng of DNA) of the ligation mix without further purification to the following premixed components: 4.5 μL of 10x reaction buffer [100 mM Tris/HCl (pH 8.3), 500 mM KCl, 15 mM $MgCl_2$, and 0.01% w/v gelatine], 4 μL of a dNTP mixture (containing 2.5 mM of each nucleotide), and 30−100 pmol of each of the primers. The mixture was filled up to 45 μL with water. After an initial heating step (10 min at 94 °C) and cooling to 72 °C, Taq polymerase was added. Amplification occurred during 30

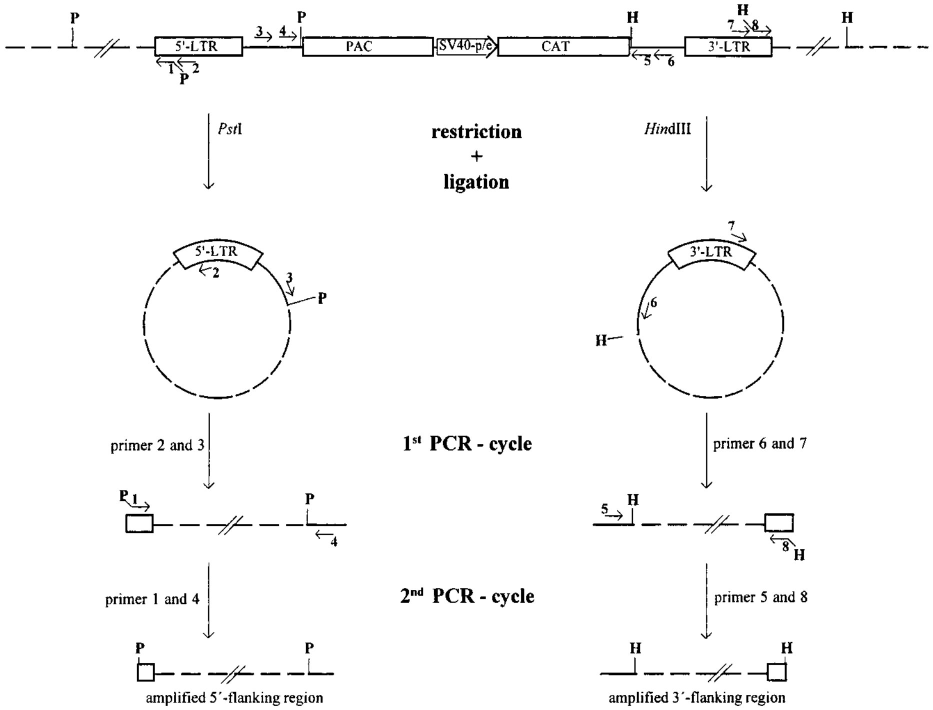


FIGURE 2: Experimental strategy for the amplification of genomic regions flanking integrated proviruses. NIH3T3 cells were infected with pM5capa-based infectious particles expressing the selection marker PAC after integration. The structure of the proviral state is shown on top. Flanking genomic regions are depicted by dashed lines, while the known retroviral sequences embracing the PAC and CAT expression cassettes are represented by solid lines and open bars. Digestion of the genomic DNA of PAC-resistant cell clones with the appropriate restriction enzymes (5′:*Pst*I and 3′:*Hin*dIII) leads to obligatory cuts in the proviral DNA and complementary cuts at various distances in the flanking host DNA. The subsequent ligation yields circular intermediates, the known (proviral) sequences of which serve as primer binding sites for two consecutive PCR reactions (nested PCR). During the first reactions (primers 2,3 or 6,7, respectively), the regions of interest are "preamplified", whereas the subsequent reactions (using primers 1,4 or 5,8, respectively) lead to the logarithmic amplification of the correct products. Since the latter contain the restriction site used for the initial digestion and an internal one provided by primers 1 and 8, respectively, these sites can be used for the convenient cloning of the PCR products in a cloning vector.

cycles of the following kind: 94 °C (1 min)/40−60 °C (1 min)/72 °C (3 min).

The products obtained from nested PCRs according to Figure 2 were restricted by *Pst*I (5′-flanking sequences) or *Hin*dIII (3′-flanking sequences) and cloned into the respective sites of cloning vector pTZ18R (Pharmacia). For further analyses, fragments were usually excised by *Eco*RI-*Hin*dIII cuts; the localization of curvature could also be performed using the two *Pvu*II vector sites instead (cf. Figure 3).

DNA Sequencing

Sequencing reactions were performed according to the Sequenase protocol provided by the manufacturer (USB) using fluorescently labeled primers according to Ansorge et al. (1986, 1987) and analyzed on an ALF sequencing apparatus (Pharmacia) using M13 universal or T7 promoter-derived primers.

DNA-Bending Analyses

The procedure was essentially as described by von Kries et al. (1990). Accordingly, DNA samples of 0.5 μg were electrophoresed in parallel on a 1.2% agarose gel (11 × 14 cm) in TPE [36 mM Tris (pH 8.0), 30 mM NaH_2PO_4, and 1 mM EDTA] for 4 h and on prerun 6% acrylamide−methylenebisacrylamide (40/1) gels at 4 °C in TBE [45 mM Tris/borate and 1.25 mM EDTA (pH 8.6) and 7 V/cm for 14 h]. Mobilities were referenced to a DNA ladder (multimers of a 123-fragment BRL).

SAR Activities in Vitro

The assay followed the procedure detailed by Kay and Bode (1995). PCR fragments, cloned into pTZ18R (Pharmacia), were excised by *Hin*dIII plus *Eco*RI or via other internal sites where applicable. They were labeled using the fill-in reaction mediated by T4 polymerase (Kohwi-Shigematsu & Kohwi, 1992). To this end, a 1 min reaction with T4 polymerase was allowed to proceed at 37 °C in 20 μL of restriction buffer A (Boehringer) to extend 5′-protruding ends. This was followed by the fill-in reaction in the presence of 50 μCi of [α-^{35}S]dATP and 100 μM each of the other deoxynucleotides. Finally, the reaction was chased with 100 μM cold dATP for another 15 min. Reassociation reactions were performed using nuclear scaffolds from two 150 cm^2 plates of mouse-L or 3T3 cells for a series of seven assays (7 × 10^6 nuclear equiv each). Each restriction fragment (10 000 cpm, equivalent to 0.3 ng) and competitor (75 μg) (*Escherichia coli* genomic) DNA were present per 150 μL sample. The results proved to be independent of the cell type (mouse-L or 3T3 cells) from which the scaffolds were derived. For evaluation, the distributions of the fragments of interest were referenced to the SAR standard which was 99% bound (see Figure 4). The competition experiments summarized in Figure 7 were done correspondingly, but in the presence of different topological forms of the SAR vector pCl. In this case, the nonbinding control (i.e. the 98% unbound vector fragment) had to serve as the reference. The competitors were (i) pCL cut by *Eco*RI plus *Hin*dIII to generate the unlabeled equivalent to fragment

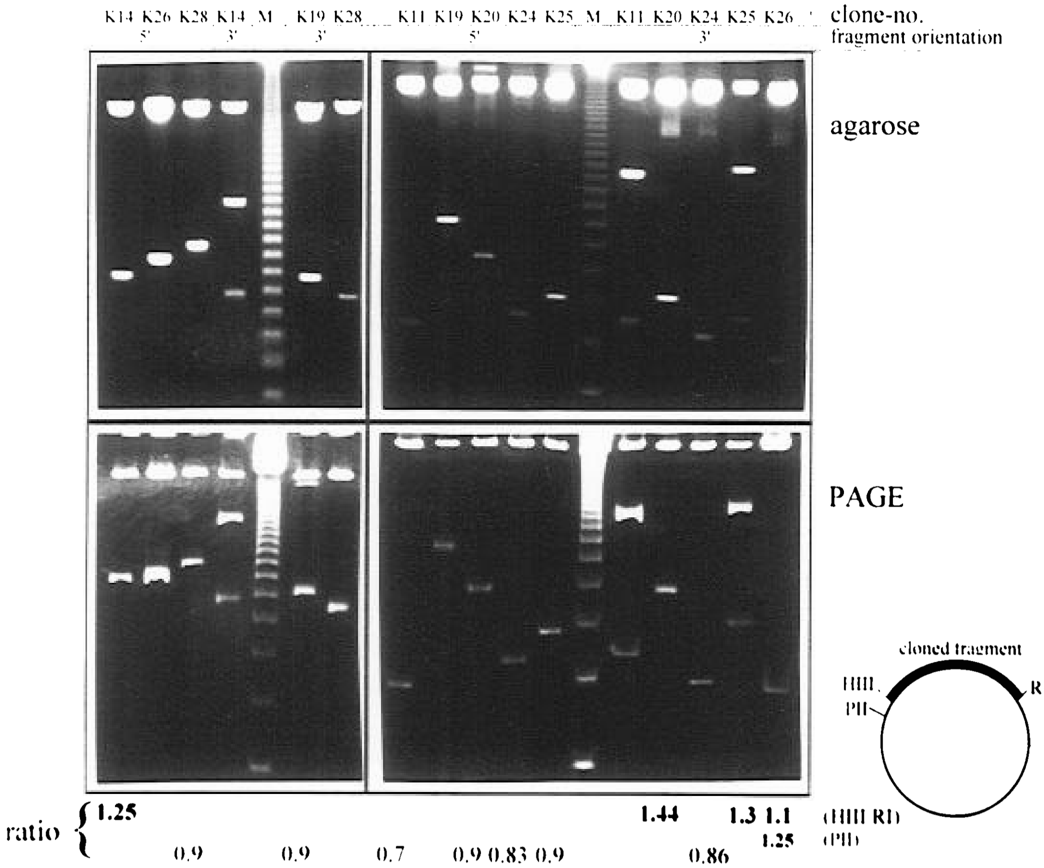


FIGURE 3: Bending analyses of the genomic flanks. Bending was derived from mobility of flanking DNAs on 6% polyacrylamide (PAGE) gel relative to a 1.2% agarose gel. The electrophoretic behavior is referenced to that of an internal control from a DNA ladder (M, multimers of a 123 fragment). DNA fragments containing curved DNA exhibit a slow mobility in polyacrylamide, i.e. a ratio of mobilities [otherwise termed "k-factor", see von Kries et al. (1990)] > 1. In contrast, fragments with an antibent domain show an increased mobility on PAGE, leading to a ratio <1. An alteration in mobility of 10% is considered indicative of bending (Milot et al., 1994). Significant ratios are indicated below the lanes. Excision of the "cloned fragment" K26-3′ by *Pvu*II (PII) rather than the usual *Hin*dIII/*Eco*RI (HIII/RI) cuts increases its mobility ratio from 1.1 to 1.25, demonstrating that the center of curvature resides in the terminal region of the cloned fragment (cf. representation of the vector next to the PAGE gel). Fragments K14-3′, K11-3′, and K25-3′ have been subdivided by an additional restriction cut for a closer analysis. Analysis of the original, reconstructed integration sites according to the scheme in Figure 6 yielded the following ratios (k-factors): K11, 0.8; K14, 0.8; K20, 1; K24, 0.9; K25, 1; and K26, 1. These data are summarized in Figure 5.

SAR, (ii) the same form after heat denaturation and chilling, and (iii) the untreated, supercoiled pCl vector. The competitor concentrations were adjusted such that supersaturatingamounts (1000-fold molar excess relative to the labeled SAR) were present.

RESULTS

Gene Expression Relative to the Technique Used for Gene Transfer

The traditional methods used to introduce DNA fragments into cells for the purpose of their stable integration involve transfection or electroporation procedures. While in the first case the DNA is transferred as a precipitate, permitting the concomitant use of physically unlinked plasmids for expression and selection, the second technique employs transfer from a solution and results in the uptake of fewer DNA molecules and relatively low copy numbers, in particular when linearized plasmids are introduced (Mielke et al., 1990). Both procedures involve nonphysiological processes, raising doubts about the physiological state of the integrated copies. An alternative concept applies vectors with retroviral functions, in particular two long terminal repeats (LTRs), a packaging signal, and sequences for reverse transcription to create infectious particles with the aid of a helper cell line. Such vectors can then be introduced by infection whereupon they will integrate, catalyzed by virally encoded activities.

Our present work is based on two retroviral expression vectors based on the LTRs and packaging signals of the murine myeloproliferative sarkom virus (MPSV). The retroviral coding sequences have been replaced by a selection marker (PAC) and a reporter gene (CAT or SEAP, respectively, see Figure 1). For a direct comparison, pM5sepa was transferred by the three techniques mentioned above. After integration, PAC-resistant cells could be selected. Figure 1B reflects the raw data along with expression levels after copy number correction, demonstrating that at a given selection pressure (2.5 μg/mL puromycin) electroporation and infection lead to a comparable expression per copy. Although transfection yields similar absolute levels, it is clearly inferior if data are corrected for the number of gene copies. These findings strongly suggest the requirement of a threshold PAC expression level to permit the formation of resistant cell clones. We have to conclude that our aim, i.e. the characterization and systematic use of transcription-promoting chromosomal integration sites, depends on a technique favoring or even enforcing low-copy integration events such that expression levels become a reflection of the properties associated with the respective sites.

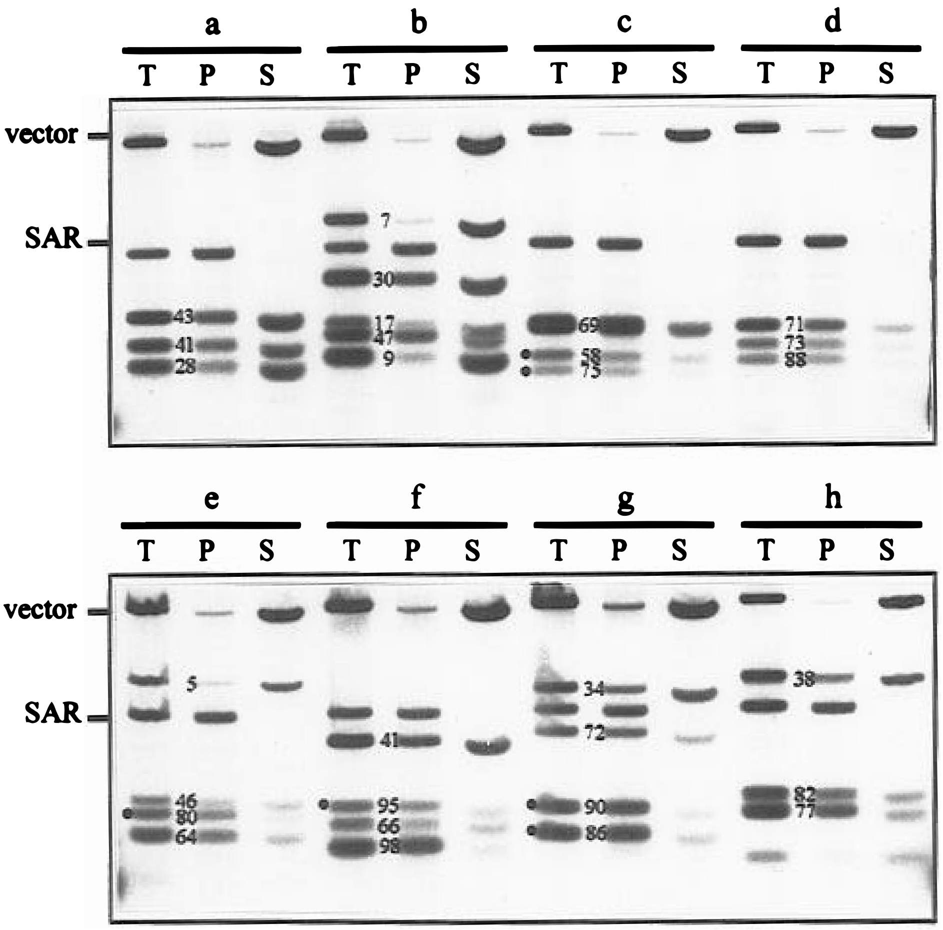


FIGURE 4: Genomic sequences at the sites of retroviral integration belong to a special class of SARs. Mixtures (a−h) of end-labeled fragments (restricted according to Figure 5) were subjected to a standard binding assay according to Kay and Bode (1995). Each assay comprises three lanes: T, total input mixture; P, pellet fraction reflecting fragments binding to the scaffold; and S, supernatant fraction for unbound species. All input mixtures contain a positive control [the 830 fragment IV from Mielke et al. (1990), denoted SAR, 99% of which are bound] and a negative control (vector sequence, 98% unbound). Numbers next to the P-lane bands mark the percentage of retained fragments, derived by referencing to the SAR control. The mixtures a−h, composed for an optimal separation on 2% agarose, are as follows (cf. Figure 6): a, SAR, K11-5′, K14-5′d, K14-5′p; b, K14-3′d (*Nhe*I/*Eco*RI), SAR, K14-3′d (*Eco*RI/*Hin*dIII), K19-5′p, K14-3′p, K19-5′d; c, SAR, K19-3′d/p (same size), K11-c, K14-c; d, SAR, K20-3′, K24-5′, K24-3′; e, K25-3′d, SAR, K25-3′p, K20-c, K25-5′p; f, SAR, K11-3′d, K24-c, K11-3′p, K26-3′; g, K28-5′, SAR, K28-3′, K25-c, K26-c; h, K11-3′d, SAR, K20-5′d, K11-3′p, K20-5′p. Fragments are assigned according to decreasing size and to their proximal (p), distal (d), or central (c) position relative to the site of integration (cf. Figure 6). The six fragments (c) marked with a solid dot correspond to the reconstructed integration sites as shown in Figure 6 and analyzed in the competition assay in Figure 7.

For a more detailed analysis, a dozen selected clones, each of an infection and an electroporation, were individually tested for both copy number and expression. Expression data were based on the quantification of the SEAP marker gene as well as the PAC marker/selector gene for which we have developed a sensitive assay (Mielke et al., 1995). Both genes are linked by an internal ribosomal entry site (IRES) to enable a coupled, cap-independent translation (Mountford et al., 1995) of the second (PAC) gene which resides in the 3′-part of the bicistronic message. Except from the two left-most clones in the electroporation series (Figure 1C) which harbored 13 or 20 copies, respectively, all other examples represent single copy integration events. Regarding their expression, these clones are comparable whether they are derived from electroporation or infection. Therefore, both techniques disfavor multiple integrations to an extent that most (electroporation) or all (infection) clones growing at a certain selection pressure reflect integration into sites with a high transcriptional potential.

Highly Expressing Sites Tagged by Retroviruses

Early observations by Hwang and Gilboa (1984) have suggested that retroviral infection leads to transcription levels 10−50-fold higher than those achieved by transfection, and this was ascribed in part to a preferential methylation of the 5′-LTR in the latter case. In our hands, electroporation yields results comparable to those of infection, and therefore, this technique appears to be free from these shortcomings in cases where selection pressure provides for the survival of cells with a minimum expression level. However, only the retrovirus-mediated gene transfer grants the integration of intact single copies which are immediately flanked by cellular sequences, while for conventional techniques, the integrates are subject to degrading cellular activities which cause the loss of variable portions of vector sequences. These considerations and the fact that only the infection process provides a number of controls required for an unambiguous assignment of corresponding 5′- and 3′- flanking sequences (see below) led to the strategy shown in Figure 2 for the recovery of integration sites. This procedure is a variant of the inverse PCR (IPCR) method originally described by Triglia et al. (1988) which enables the direct amplification of host flanking sequences.

The sequence of pM5sepa is known; therefore, two sets of primers could be designed, enabling the amplification of sequences adjacent to the 5′-LTR (primers 1−4) or the 3′-LTR (primers 5−8), respectively by IPCR. In two parallel experiments, cellular DNA was cleaved by either *Pst*I (P) or *Hin*dIII (H) to produce known cuts within the provirus and cuts of an unknown location within the adjacent host sequences. A ligation step followed to produce circles including the respective LTR and a stretch of host sequence. The first sets of primers (2,3 or 6,7) were then used for a preamplification and the second sets (1,4 or 5,8, respectively) for the final amplification, mostly of the cellular sequence. Built-in restriction sites (P for primer 1 and H for primer 8) enabled the cloning of the respective fragments for sequencing and further characterization. Application of two sets of primers in a "nested PCR" approach was dictated by the abundance of endogenous retroviral sequences in the murine genome which gives rise to the amplification of many nonspecific sequences during the preamplification step.

Assignment of Corresponding Flanks

For 24 clones, this procedure led to the amplification of at least one flanking sequence, yielding a total of 34 sequences in the range of 127 bp to 4.5 kb. In a few cases, the amplification produced more than one fragment which by Southern analysis could either be ascribed to a deletion concomitant with the integration process or to a double infection. With one exception (3′-flank of K28, see below), these clones were omitted from further analyses (data not shown). The remainder was subjected to a number of critical tests which will be summarized in Figure 5. For eight individual clones, it could be established that a pair of corresponding 5′- and 3′-flanks was isolated. (i) Southern blots were used to show that the PCR amplification products reflect the length of the authentic fragment in genomic DNA. (ii) Prior to integration, the target DNA is cleaved in a staggered fashion with a length (MPSV, four nucleotides) that is characteristic for each virus. Repair of the resulting gapped intermediate by filling in and removing nonhomologous nucleotides from the 5′-end of the viral DNA by cellular repair enzymes results in a duplication of host target DNA at the site of integration. PCR fragments were sequenced from both sides. Because of the duplication of a 4 bp piece of host DNA, these sequences had to be present on both members of a putative couple of fragments. Two other sequence features were also established, i.e. a short stretch of LTR sequences beyond primers 1 and 8 demonstrating that the primer had hybridized with the correct target sequence and a loss of the LTR's two terminal A residues typical of an authentic insertion due to the viral integrase. (iii) Ultimately, sequence information from the putative flanks of the mentioned eight clones was used to construct new sets of PCR primers and to amplify the original target sites from the DNA of noninfected cells as 200−300 bp pieces of DNA (see the center bars in Figure 5 and summary in Figure 6 below). For two cases (K19 and K28), a successful reconstruction was prevented by the special sequence features discussed below.

Transcriptional Properties

Previous work has indicated that the majority of Moloney murine leukemia virus integration events in mouse fibroblasts occurred in regions of transcriptional activity and/or in the vicinity of CpG islands (Scherdin et al., 1990). These islands are characterized by an abundance of CpG equaling the reverse sequence GpC due to the fact that in the 5′- or 3′-regions of certain (housekeeping) genes there is a protection from methylation and subsequent mutagenesis. This situation is clearly found in one example, i.e. the sequence left to the integration event marked K20c (Figure 6).

Regarding active transcription, we have labeled by ^{32}P the poly(A)-enriched nuclear RNA fractions from uninfected NIH3T3 cells by random priming and reverse transcription in vitro and used them as a probe for transcriptional activity within the isolated fragments [see also Scherdin et al. (1990)]. With an α-actin sequence as an internal control, we derived strongly hybridizing signals for 3 out of 28 fragments, among these K19-3′. Quite remarkably, these are the same fragments which in a preliminary screen qualified to contain repetitive DNA, later identified as members of the B1 group (Figure 5). The abundant occurrence of various deleted versions of this sequence throughout the murine genome accounts for our inability to amplify the authentic integration site by PCR (see above). Being a homologue to the human Alu sequence, B1 is the prominent short interspersed repetitive element (SINE) of the mouse which has been shuttled by reverse transcription of a polymerase III transcript and subsequent retrotransposition. As B1 signals have been generated from polyadenylated messages, their occurrence in the present test could be a simple reflection of SINES present in certain introns of unspliced hnRNA (Ullu et al., 1982; Ullu & Tschudi, 1984). In summary, there is hardly any evidence for an active transcription in the immediate vicinity of integration, and for B1, it is a rather indirect one.

Structural Properties

Integration event K19 (Figure 5) is remarkable in that it occurred within a complete B1 element. B1, in contrast to the abundant and highly repetitive centromeric DNA, appears to be a preferred target which may be due either to its forming a particular structure or simply to the fact that it is a marker for an open site that still favors subsequent integration events adjacent to or within the SINE (see below).

A striking example for a structural feature attracting the integrase complex is K28, where the provirus has integrated right at the border of a $(GAA)_{43}$ repeat, closely followed by a $(GA)_{20}$ repeat on the opposite strand. This motif is found exclusively at the 3′- end of the provirus, and although sequences of the general $(GA)_n$ type are also present in the 5′-portion, this observation supports the relevance of a junction structure as an integration target. The facts that (i) two different fragments (300 and 800 bp, respectively) were PCR-amplified from the 3-flank, (ii) these fragments share the same 4 bp repeat (GAAG), and (iii) both fragments could readily be detected at equal intensity on a Southern blot demonstrate that the retroviral integration process has triggered a partial deletion of the $(GAA)_n$ stretch. It is noted that $(GAA)_n$ is a prominent member of simple trinucleotide

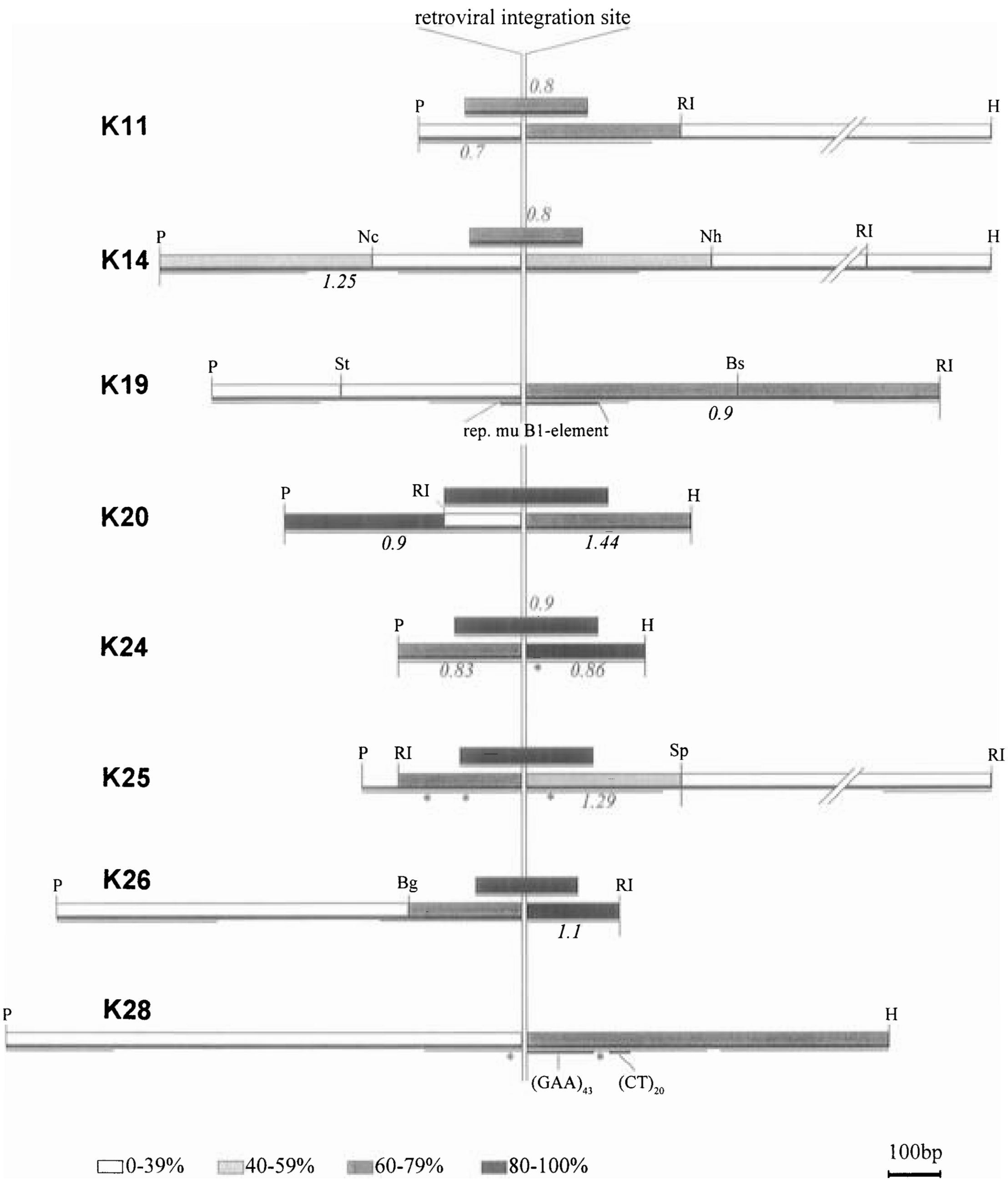
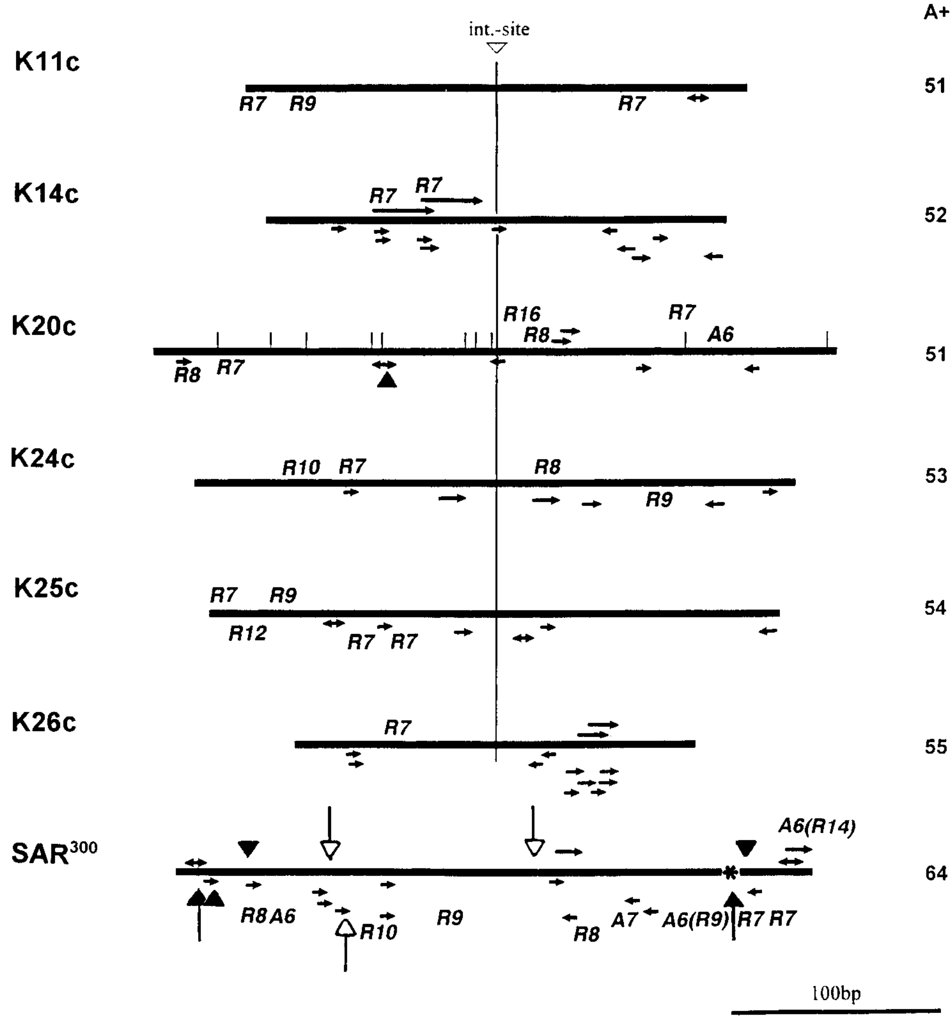


FIGURE 5: Scaffold attachment and bending. Sequenced portions are underlined. Sets of flanking genomic sequences and reconstructed integration target sites are characterized according to their relative mobilities (k-factors) which are >1.1 or <0.9 in the case of significant curvature or antibending features, respectively. Curvature predicted from computing (Boulikas, 1994) has been marked by an asterisk. SAR activities in vitro have been referenced with respect to a SAR standard (99% binding); their relative affinity is indicated by the degree of shading. Sequenced portions of the fragments are indicated by the faint lines below the respective fragment bars. They comprise the reconstructed integration sites shown in Figure 6.

repeats that have gained considerable attention since their intrinsic instability may be the cause of specific genetic defects (Siedlaczck et al., 1993).

Sequences Flanking an Integration Site Show the Phenomenon of Bending. Bending, the tendency for successive base pairs to be nonparallel in an additive manner, is the basis for either of two macroscopic features (Goodsell & Dickerson, 1994). Curvature is the result of a phased array of bends causing the helix axis to follow a nonlinear pathway. In contrast, a rodlike structure follows from a succession of bends canceling each other, i.e. from a nonphased array of bending sites avoiding the 10−11 bp periodicity of the DNA double helix. Whenever DNA is bent, both grooves on the side of the curve narrow due to the associated compression, whereas those on the outside become correspondingly wider. The fact that a widened major groove, whether it is stabilized by a nucleosome or not, serves as a preferred target for retroviral integrases in vitro, (Mueller & Varmus, 1994) justified a screening of retroviral integration sites for their bending potential. This is conveniently done by a comparison of electrophoretic mobilities on agarose and polyacrylamide gels (Figure 3) and expressed as “k-factor” which is identical to the ratio of mobilities (von Kries et al., 1990). Hence, our analyses reveal a significant proportion of fragments with a relatively low mobility on polyacrylamide (k-factor > 1) which is ascribed to curvature and another set of fragments with the reverse phenomenon (fast mobility, k-factor < 1) which is considered typical of stiff rods. The



	A+T (%)	Normalized Doublet Usage (%)
K11c	51	CA/TG (130) AG/CT (120)
K14c	52	AG/CT (144) CA/TG (126)
K20c	51	AA/TT (127) AG/CT (126)
K24c	53	CA/TG (147)
K25c	54	CC/GG (133)
K26c	55	CA/TG (129) AG/CT (129) AA/TT (128) CC/GG (122)
SAR^{300}	64	CC/GG (155) AA/TT (122)

FIGURE 6: Sequence and structural features implicated in SAR function. Reconstructed integration sites (Figures 5 and 6) were analyzed for sequence features implicated in SAR functions. For comparison, a 320 bp section from the 830 bp SAR standard (Figure 4) is represented accordingly. The SAR activities of both the 830 and 320 bp fragments have been determined before [fragments IV and VI, respectively in Mielke et al. (1990)]. The symbols used previously (solid arrowhead, ATATTT; open vertical arrow, *Topo*II consensus; solid vertical arrow, same but including the ATATTT motif; and asterix, base-unpairing region) were maintained. In addition, direct and inverted repeats >8 bp were marked by solid horizontal arrows and immediate palindromes by double arrows. A_n tracts with $n > 6$ and R_n, i.e. polypurine tracts with $n > 7$, are also indicated. CpG residues in fragment K20 are indicated by vertical lines. The CpG/GpC ratio for the fragment's left half is 8/9, characteristic for a CpG island, whereas for the right half, it is 2/6. All fragments show an overrepresentation of certain doublets of bases as shown in the table to the right.

data in Figure 3 and their summary in Figure 5 emphasize that all flanking sequences are associated with either of these phenomena. In case of K26-3′, a bending center appears to be located adjacent to the integration site since its low mobility is increased as more DNA is added to the 5′-end of the cloned fragment (Figure 3), whereby the curvature is shifted to a more central position. In general, bending is only occasionally found for the reconstructed fragments which contain the integration target at a central position. Therefore, the in vivo target does not have to be a bending center itself, but it may require chromatin structures shaped by nearby stretches of bent DNA. We note that these conclusions are in complete agreement with recent findings by Milot et al. (1994).

Reconstructed Integration Targets Behave as SARs. It is possible that specific interactions between host DNA binding proteins and IN direct the integration machinery to certain regions of the host genome as suggested for the yeast retrotransposon Ty3 (Craigie, 1992). The retroviral integrase complex contains all enzymatic functions for the integration reaction except the activities required for the terminal steps which are thought to be a component of the nuclear matrix [reviewed by Boulikas (1995)]. Although, at first glance, the host sequences next to the provirus do not remind us of classical scaffold/matrix-attached regions [S/MARs, reviewed by Bode et al. (1995, 1996)], S/MARs are sometimes found to colocalize with bent DNA (Anderson, 1986; Homberger, 1989; von Kries et al., 1990). This led us to inspect the SAR activity for the available fragments in vitro. The results in Figure 4 were derived by the standard scaffold reassociation approach detailed in Kay and Bode (1995). Accordingly, they include a binding standard (marked SAR, 99% bound) and a nonbinding control (vector, 98% unbound). Out of a selection of separable restriction fragments (T), those fragments with an affinity to the scaffold become concentrated in the pellet (P trace), whereas the weakly or nonbinding complement accumulates in the supernatant (S). Figure 5 demonstrates that a SAR character is found for all retroviral chromatin targets that could be reconstructed (marked with a dot next to the T lanes in Figure 4), and this activity may or may not extend into the more remote 5′- or 3′-regions.

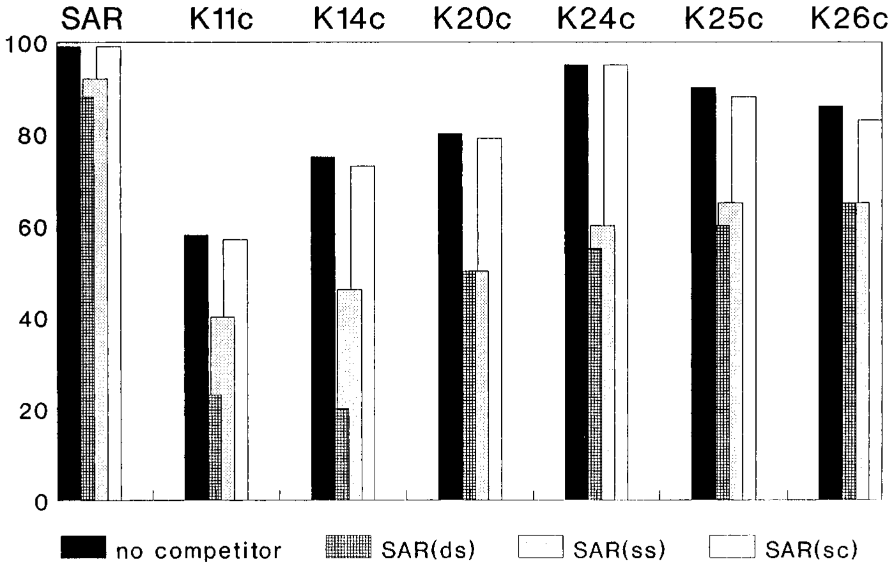


FIGURE 7: Integration fragments and a standard SAR fragment share scaffold binding sites. Binding assays were performed as for Figure 4 but in the presence of different topological forms of the SAR vector (pCl). Competitors were applied at a concentration corresponding to an amount of double-stranded SAR fragment (excised from pCL by *Hin*dIII plus *Eco*RI) which is sufficient to displace some of its radioactive counterpart (12% in the supernatant). The same concentrations were also maintained for the supercoiled and single stranded forms of pCl.

The now classical SAR elements that have been detected at the putative borders of numerous chromatin domains (Laemmli et al., 1992), adjacent to enhancer elements (Cockerill & Garrard, 1986; Cockerill et al., 1987), and at the site of a T-DNA integration event (Dietz et al., 1994) share a number of sequence motifs which are in part a direct consequence of their AT richness (>65%). Among these are oligo(dA) tracts of various extension (A_n in Figure 6) and sequences which conform to the topoisomerase II consensus of *Drosophila* (open arrows). This consensus is compatible with a central motif ATATTT (solid arrowhead) which in some cases forms the nucleation center of a base-unpairing region [BUR, see the region marked with an asterisk in Figure 6 and Bode et al. (1992)]. While all these features are repeated several times on a classical SAR (Cockerill & Garrard, 1986), they are virtually absent at the sites of retroviral integration (Figure 6). Having a balanced content of A+T versus G+C, this new group of SARs is nevertheless rich in certain dinucleotide repeats (CA/TG and AG/CT, cf. Figure 6) which have been correlated with a high incidence of cruciform structures (Boulikas & Kong, 1993a,b; Bode et al., 1996). In addition, the present selection of elements comprises numerous direct repeats with extensions of 8−29 bp and many polypurine tracts [R_n, see Kohwi and Kohwi-Shigematsu (1993)] which together provide a high potential of secondary structure formation.

We have found previously that the nuclear scaffold offers separate sets of binding sites for single stranded and supercoiled DNA regardless of their sequence content. Among these, there is a partial overlap of the sites for single stranded and for SAR DNA, but neither of these species is competed off by supercoils (Kay & Bode, 1994; Bode et al., 1996). In Figure 7, we have explored the relation of the integration sequences to these species. To this end, we have performed reassociations as for Figure 4, but in the presence of an authentic SAR in its (standard) double-stranded form, as part of a supercoiled plasmid or as single strands. For example, the left-most group of bars is from a self-competition of the labeled SAR which is normally 99% bound (black bar). Presence of the unlabeled counterpart (cross-hatched, dark) yields 12% displacement, at the given input concentration. If corresponding concentrations of single strands or supercoils are applied, the displacement levels are 9 and 0%, respectively, indicating the overlap with the ssDNA but not the supercoil sites. The remaining experiments involve the labeled integration targets in a corresponding setup. While, in general, all these species follow the above trends, the competition is most pronounced for the fragment with the lowest affinity, whereas the competition of the high-affinity elements becomes progressively harder. These data are compatible with a simple model in which the SAR character is, at least in part, based on the propensity of exposing single strands, be it by unwinding or by the formation of cruciform, slippage, or triple helical structures which may all follow the buildup of negative superhelical tension.

DISCUSSION

We have previously studied the interplay of selection pressure and the generation of highly expressing clones by transfection techniques (Wirth et al., 1988). As in the present case (Figure 1B), selection favored the recovery of clones carrying high copy numbers, although some marked exceptions were also noted which suggested integration into transcriptionally active chromatin. Electroporation is an alternative technique used for gene transfer resulting in largely reduced copy numbers (Mielke et al., 1990). After selection, a high proportion of clones contains single copies, permitting convenient access to events that have favored integration into highly expressing sites. This is demonstrated in Figure 1B,C where we have compared the effect of gene transfer techniques upon expression. All experiments were based on a retroviral vector, pM5sepa, which contains the elements required for its packaging into retroviral particles and subsequent transfer by the infection route, i.e. the long terminal repeats providing promoter and polyadenylation functions, and the ψ packaging sequence. At a certain multiplicity of infection (MOI), this transfer route strongly favors single copy integration of the sequences flanked by LTRs, and this was particularly true under the conditions applied here which led to only 1000 resistant clones among the 10^6 cells used for infection. In part, this can be ascribed to an interference phenomenon by which infected cells become protected from superinfection by viruses using the same receptor (Coffin, 1991).

It has been reported repeatedly that expression of artificially introduced proviral DNA is much lower than that of the same DNA introduced by infection. In a careful study, Hwang and Gilboa (1984) compared G418-resistant clones arising from infection and selected clones harboring a single copy of the construct after transfection. Their conclusions can only partially be verified in our system. While the overall expression is mostly independent of the technique used for transferring PM5sepa, the per-copy level for transfection is clearly inferior to electroporation and infection which are rather comparable by both criteria (Figure 1B,C). These findings tend to emphasize the importance of selection pressure which is satisfied either by multiple copies or by individual ones in a favorable genomic environment. The only technique that virtually guarantees the incorporation of single copies and for which, in addition, a long term stability of clones has been established is the retroviral infection route (D. Schübeler, unpublished). While this indicates that there is no gradual inactivation due to methylation, the eventual loss of the single copy would be selected against. Equally important, only this technique provides (i) the presence of a transgene devoid of plasmid DNA but nevertheless with defined ends, (ii) a number of built-in controls to establish an authentic integration event, and finally (iii) possibilities for an efficient tagging procedure by which these sites can be marked for repeated use (see below).

A striking demonstration for the applicability of a combined infection/selection protocol has been provided by the insertion of three recombinant retroviruses into the same region of the mouse F2 locus which elevates the level of LTR-driven *neo*^{r} gene expression owing to the transcriptional potential of cellular 5′-flanking sequences (Petersen et al., 1991). Analysis of the cellular RNAs encoded by F2 indicated that the F2 promoter is exceptionally strong. Although no function has yet been assigned to these transcripts, F2 homologous sequences are found in the genomes of human, goat, horse, rabbit, and chinese hamster.

What Do in Vitro Studies Tell about Retroviral Integration Sites? Since integration sites conform to no obvious consensus, the choice of where integration into the host cell genome occurs may be dictated by the secondary structure of DNA, host proteins, or by both. While in vivo, the retroviral integrase (IN) performs the central cutting and joining steps as part of a 160 S nucleoprotein complex, purified IN protein, and model substrates can be used in vitro to study aspects of the integration process except from the terminal resolution steps which require host functions. The intermediate created by retroviral functions alone can be assayed by PCR techniques to measure the distribution of sites that are preferred for initiation of an integration event. These studies have emphasized that the choice of retroviral integration sites is strongly influenced by chromatin structure as it occurs more efficiently into nucleosomal rather than naked DNA. Preferred nucleosomal sites are regions where the major groove is on the exposed face of the histone core, but interestingly, a similar effect is also seen if DNA is intrinsically bent in the absence of proteins (Mueller & Varmus, 1994).

In vivo, by necessity, most proviral experiments assay cell systems where viral integration is reasonably rapid and often limited to a single or few integration sites. PCR techniques monitoring the state immediately after IN action do not check for complete or functional integration since signals will arise also in cases where the terminal steps of trimming and gap repair are not complete. Rather unexpectedly, studies of this type showed that all regions of the genome examined are accessible to retroviral integration (Withers-Ward et al., 1994). There were, nevertheless, localized preferences within these 500 bp regions, some sites being used up to 280 times more often than expected for random integration (Withers-Ward et al., 1994). These hot spots may arise from the same criteria that guide the IN protein to preferred targets in vitro. In conclusion, this type of in vivo study emphasizes the importance of DNA structure for integration but beyond that tells little about the transcriptional properties of sites that have been screened for their integrity by selection.

Our study demonstrated a relatively frequent integration into repetitive DNA (8/24 integration sites). Out of these, three sites were associated with B1 sequences which represent the prominent SINE element of the mouse. This appears to be a more widespread phenomenon. As an example, Stevens and Griffith (1994) showed that seven out of eight HIV proviruses integrated directly into or within a nucleosome's distance from SINES, corresponding to a 6-fold higher frequency compared to random. Like retroviruses, these SINES, L1 and Alu, include RNA and cDNA intermediates in their replication cycle, and L1Hs even harbor an open-reading frame (ORF) for reverse transcriptase. During the time SINES have coexisted with mammalia, they may have continuously scanned the genome for the most attractive integration sites. For these reasons, SINES may not present favorable targets per se, but both retroelements may sense the same features of chromatin structure. In light of the discussion to follow, it is remarkable that there is a frequent integration of these elements next to S/MARs (Cockerill, 1990) while B1 sequences themselves have little if any affinity for the matrix (see the K19 fragments in Figure 4, traces b and c).

Bending and Scaffold Attachment for the Sites Selected in Vivo. Given the relevance of bending for the IN-mediated integration in vitro, we screened a selection of reconstructed integration targets and flanking regions for this property (Figures 3 and 5). Where investigated, it is seen that all integrations are associated with at least one flank having a mobility which differs by more than 10% between polyacrylamide and agarose gels, whereas this criterion was only fulfilled by three out of the six reconstructed targets. Hence, while integration is certainly associated with bending, the bending center itself cannot represent the preferred target in vivo. This is also reflected by curvature maps of these sequences which have been derived by the procedures of Boulikas (1994) and were kindly constructed by this author. In complete agreement with our observations, Milot et al. (1994) found six out of seven independent murine leukemia virus integration sites to have bent or antibent DNA elements in close proximity. They have estimated the incidence of curved elements in chromosomal DNA to be one per 11 kilobases (kb) and that of antibent elements one per 12 kb which underlines the relevance of these structures.

Bending has eventually been implied in the function of scaffold- or matrix-attached regions (S/MARs), i.e. the elements that subdivide the eukaryotic genome into functional units (domains) by mediating interactions with the nuclear matrix (Anderson, 1986; Homberger, 1989; Boulikas, 1995). In one of these examples, the bending center was associated with a S/MAR but did not coincide with the region of DNA binding most strongly (von Kries et al., 1990). Using an authentic 800 bp SAR element as an internal standard, we examined the affinity of reconstructed targets and the corresponding flanks for scaffolds of two murine cell lines (3T3 and L-cells) and obtained a striking answer: without exception, all reconstructed targets behaved as SARs (Figures 4 and 5). Some flanking sequences did also, but as a rule, the center of affinity coincided with the site of integration. The affinity of the 200−300 bp targets typically ranged between 75 and 95% of the 800 bp standard. In one case (fragment K11-c), it was lower which appears to correlate with the lack of certain sequence features (Figure 6 and below). Regarding the relative strength of association, it should be noted that all fragments are comparable to a 300 bp subfragment [designed VI in Mielke et al. (1990) and depicted in Figure 6, bottom] of the SAR standard.

A Conceivable Setup. Common features of cellular DNA flanking functional, integrated DNA and RNA viruses are nuclease hypersensitivity, transcriptional activity, the potential to form DNA superstructures, and the presence of topoisomerase II sites which are compatible with the presence of both supertwisted and SAR-type DNA (Howard & Griffith, 1993; Laemmli et al., 1992). A closer inspection has disclosed structural and functional features involved in genetic recombination like extended foldback structures (de Ambrosis et al., 1992). The frequent viral integration in the proximity of CpG islands can also been seen in this context as these are clustered in the Giemsa light bands of metaphase chromosomes, i.e. in regions that are a frequent target for chromosomal break points and translocations. Generally, simple tandem repeats undergo mutational events much more frequently than do single copy sequences such that considerable diversity results at these loci. Certain simple trinucleotide repeats have gained considerable attention as they are correlated with specific genetic defects, and among a panel of these repeats, $(GAA)_6$ revealed predominant signals in fungal, plant, and animal species which for the majority of offspring are not inherited according to Mendelian rules (Siedlaczck et al., 1993). A polypurine stretch of the same type has attracted a retrovirus in the case of the K28 integrant (Figure 5). We note that the $(GAA)_{43}$ and $(CT)_{20}$ repeats within its 3′-flank are parts of a SAR with nearly 80% binding potential and that both structural components are recognized by a prototype SAR binding protein, SATB1, that senses regions with an unwinding potential (Dickinson et al., 1992; Nakagomi et al., 1994; T. Kohwi-Shigematsu, private communication). $(CT)_n$ repeats are also the basis of certain DNAse I hypersensitive sites (Lu et al., 1992, 1993).

An inspection of Figure 6 tells us that homooligonucleotide stretches and inverted and direct repeats are the most prominent common properties shared by AT-rich prototype SARs and the SARs recovered here. These motifs promote the formation of triple helical DNA, cruciforms, and slippage structures which have in common the exposure of single stranded regions. These in turn are recognized by nuclear scaffolds (Kay & Bode, 1994, and references therein) and can hence be competed off by denatured ssDNA (Figure 7). Interestingly, the SAR character of the original (unperturbed) integration sites is only fully appreciated after their reconstruction via PCR primers designed according to the flanking sequences. This observation together with the fact that the affinity may or may not extend into the flanks warrants the consideration of SARs as primary targets for functional proviruses with a marked transcriptional potential. Moreover, we believe that both the recombinogenic and the transcriptional potential of these sites may be assisted by DNA bending which is consistently found (Figures 3 and 5). The beneficial effect due to the bent flanks may be several-fold. (i) Bending might create a high-affinity site for Mn^{2+}, the metal ion necessary for the integration reaction. (ii) Weak or poor nucleosome positioning sequences have an advantage as SARs and ORIs. Computer simulations show that even in cases where bending produces curvature there is a strong interference with the occurrence of positioned nucleosomes. This means that DNA−histone interactions are easily interrupted if secondary structure formation or establishment of alternative protein contacts are required (T. Boulikas and J. Bode, in preparation). (iii) Probably most important, bending accumulates backbone strain that can subsequently be retrieved to reduce the energy required for overcoming base−base interactions (Ramstein & Lavery, 1988). In fact, it has been shown that for unpairing to occur it must be supported by bending and/or torsional stress (Khan et al., 1994). Bending-induced unpairing has been implicated in a number of processes like open complex formation in transcription and ORI unwinding in replication (Eckdahl & Anderson, 1990).

How does the process of retroviral integration profit from this setup? After the IN complex has introduced a staggered cut into the target and performed the strand transfer step joining the 3′-ends of viral to 5′-ends, host enzymes are required to complete the process. In particular, they have to perform a resolution step that requires melting between the sites of joining, degradation of the unpaired 5′−ends of viral DNA, and gap repair to fill in the single strand connections between viral and host DNA. All these repair functions are found in close association with the nuclear matrix [Mullenders et al., 1983a,b; McCready & Cook, 1984; reviewed by Boulikas (1995)], and they are greatly aided by the fact that integration has occurred into matrix-associated DNA which, in the examples above, has recombinogenic potential. Finally, the realization of the retroviral information might be supported by an augmenting effect of SARs upon transcription [reviewed by Bode et al. (1995)].

A somehow related setup may have assisted the integration/expression of T-DNA which invaded a plant SAR element following gene transfer by *Agrobacterium tumefaciens* (Dietz et al., 1994), although in that case, a standard SAR element of the AT-rich type was selected. This raised the question of the transcriptional potential of the particular group of SARs in Figure 6 and, of course, of the reasons prohibiting an integration into standard SARs. We have introduced subfragments of the human interferon-β SAR as well as members of the present collection of elements into the luciferase test construct used by Schlake et al. (1994) and derived the following enhancement factors [see Mielke et al. (1990) for designations; numbers in brackets mark SAR length in bp]: I (2200), 14x; IV (800), 10x; VI (300), 5x. The corresponding series for the nonstandard, approximately

300 bp long elements in Figure 6 yielded enhancements of 5.34(±0.36)-fold, and therefore, they behaved exactly as expected for a SAR of this extension (data not shown). This reemphasizes our proposal that SARs act due to their unwinding potential which in turn is only indirectly related to their AT content (Bode et al., 1995, 1996).

Since SARs are found to be beneficial for integration, why can AT-rich elements not serve these functions? Grandgenett et al. (1993) have compared the ability of IN to utilize specific plasmid sequences for binding and subsequent integration. While it was found that IN prefers to bind AT-rich regions in the absence of strand transfer conditions, it did not favor these regions for integration. In solution, IN appears to have an available binding site which differs from the target binding site of the enzyme engaged with recessed LTR termini and primed for the transfer of single strands. In fact, there is evidence that IN harbors two different DNA binding regions, one of which prefers ss- over dsDNA (Kahn et al., 1991; Woerner et al., 1992). Therefore, it is proposed that the integration machinery depends on sequences which are not of the AT-rich type but nevertheless primed for strand separation.

Conclusions and Perspective. A better understanding of the mechanisms involved in the expression of transgenes can be expected to reveal new aspects of nuclear organization and may lead to improved tools for genome manipulation. Retroviruses have long been thought to sense the transcriptional potential of their target, but relatively few studies of retroviral integration sites have provided data of extended sequences surrounding integration sites. We have applied selection pressure to screen for functional integration and used the intrinsic advantages of the system for an extensive characterization of target sequences.

The first generation of vectors constructed for these purposes (pM5capa) led to the recovery of 28 flanks ranging between 0.2 and 4.5 kb by IPCR techniques. Certain restrictions experienced during its use were overcome by construction of a derivative, pM5sepa (Figure 1A). While the stringency of the PAC system already enforces high transcription of the resistance gene, for pM5sepa, the very best producers can be screened for by an agarose overlay assay which provides for their convenient isolation (Kirchhoff et al., 1995). Systematic tests have proven that the long term stability of expression is granted irrespective of expression levels (D. Schübeler, private communication). Coexpression with the selective marker is achieved by an IRES sequence, and another simple assay is available for monitoring PAC expression levels in parallel as a criterion of provirus integrity (Mielke et al., 1995).

pM5capa showed a principal restriction in offering just two sites (*Pst*I and *Hin*dIII) which have to recur at a certain distance in the genomic DNA to enable circularization for the IPCR approach. pM5sepa was therefore provided with multiple cloning sites in the respective positions. Together with advanced PCR techniques, based on a combination of polymerase and proofreading activities, we hope to extend the scope of integrants that can be tested by this concept.

Since the favorable properties of retroviral integration sites are obvious, we have started to tag these sites for reuse. Our tool is the FLP recombinase from yeast (Schlake & Bode, 1994) which recognizes two 48 bp FRT sites in order to excise the intervening sequence as a circle. As this process is reversible, an analogous second FLP-catalyzed event can be used to reinsert another FRT-tagged expression cassette of choice into the remaining single FRT. Our experiences using the corresponding derivatives of pM5sepa will be reported elsewhere (Bode et al., 1996; Schübeler and Bode, in preparation).

ACKNOWLEDGMENT

The authors express their gratitude to Dirk Schübeler (GBF) and to Michael Breindl (SDSU) for numerous stimulating discussions and a critical review of the manuscript.

REFERENCES

Allen, G. C., Hall, G. E., Jr., Childs, L. C., Weissinger, A. K., Spiker, S., & Thompson, W. F. (1993) *Plant Cell 5*, 603−613.

Anderson, J. N. (1986) *Nucleic Acids Res. 21*, 8513−8533.

Ansorge, W., Sproat, B. S., Stegemann, J., & Schwager, C. H. (1986) *J. Biochem. Biophys. Methods 13*, 315−322.

Ansorge, W., Sproat, B., Stegemann, J., Schwager, C., & Zenke, M. (1987) *Nucleic Acids Res. 15*, 4593−4602.

Berezney, R. (1991) *J. Cell. Biochem. 47*, 109−123.

Berger, J., Hauber, J., Hauber, R., Geiger, R., & Cullen, B. R. (1988) *Gene 66*, 1−10.

Bode, J., Kohwi, Y., Dickinson, L., Joh, R. T., Klehr, D., Mielke, C., & Kohwi-Shigematsu, T. (1992) *Science 255*, 195−197.

Bode, J., Schlake, T., Ríos-Ramírez, M., Mielke, C., Stengert, M., Kay, V., & Klehr-Wirth, D. (1995) in *Structural and Functional Organization of the Nuclear Matrix - International Review of Cytology 162A* (Jeon, K. W., & Berezney, R., Eds.) pp 389−453, Academic Press, Orlando.

Bode, J., Stengert-Iber, M., Schlake, T., Kay, V., & Dietz-Pfeilstetter, A. (1996) in *Critical Reviews of Eukaryotic Gene Expression* (Stein, G. S., Stein, J. L., & Lian, J. B., Eds.) (in press).

Boulikas, T. (1994) *J. Cell. Biochem. 55*, 513−529.

Boulikas, T. (1995) Chromatin domains and prediction of MAR sequences, in *Structural and Functional Organization of the Nuclear Matrix - International Review of Cytology* (Jeon, K. W., & Berezney, R., Eds.) pp 279−388, Academic Press, Orlando.

Boulikas, T., & Kong, C. F. (1993a) *J. Cell. Biochem. 53*, 1−12.

Boulikas, T., & Kong, C. F. (1993b) *Int. J. Oncol. 2*, 325−330.

Cockerill, P. N. (1990) *Nucleic Acids Res. 18*, 2643−2648.

Cockerill, P. N., & Garrard, W. T. (1986) *Cell 44*, 273−282.

Cockerill, P. N., Yuen, M. H., & Garrard, W. T. (1987) *J. Biol. Chem. 262*, 5394−5397.

Coffin, J. M. (1991) Retroviridae and their replication, in *Fundamental Virology* (Field, B. N., & Kniepe, D. M., Eds.) Raven Press Ltd., New York.

Craigie, R. (1992) *Trends Genet. 8*, 187−190.

de Ambrosis, A., Casciano, I., Querzolaq, F., Vidali, G., & Romani, M. (1992) *Cancer Genet. Cytogenet. 60*, 1−7.

Dickinson, L., Joh, T., Kohwi, Y., & Kohwi-Shigematsu, T. (1992) *Cell 70*, 631−645.

Dietz, A., Kay, V., Schlake, T., Landsmann, J., & Bode, J. (1994) *Nucleic Acids Res. 22*, 2744−2751.

Dorer, D. R., & Henikoff, S. (1994) *Cell 77*, 993−1002.

Eckdahl, T. T., & Anderson, J. N. (1990) *Nucleic Acids Res. 18*, 1609−1612.

Engelman, A. (1994) *BioEssays 16*, 797−799.

Goodsell, D. S., & Dickerson, R. E. (1994) *Nucleic Acids Res. 22*, 5497−5503.

Grandgenett, D. P., Inman, R. B., Vora, A. C., & Fitzgerald, M. L. (1993) *J. Virol. 67*, 2628−2636.

Hajihosseini, M., Lavachev, L., & Price, J. (1993) *EMBO J. 12*, 4969−4974.

Homberger, H. P. (1989) *Chromosoma 98*, 99−104.

Howard, M. T., & Griffith, J. D. (1993) *J. Mol. Biol. 232*, 1060−1068.

Hwang, L.-H. S., & Gilboa, E. (1984) *J. Virol. 50*, 417−424.

Kahn, J. D., Yun, E., & Crothers, D. M. (1994) *Nature (London) 368*, 163−166.

Kalos, M., & Fournier, R. E. K. (1995) *Mol. Cell. Biol. 15*, 198−207.

Kay, V., & Bode, J. (1994) *Biochemistry 33*, 367−374.

Kay, V., & Bode, J. (1995) Detection of scaffold-attached regions (SARs) by in vitro techniques; activities of these elements in vivo, in *Methods in Molecular and Cellular Biology: Methods for studying DNA-protein interactions - an overview* (Papavassiliou, A. G., & King, S. L., Eds.) pp 186−194, Wiley-Liss, Inc., New York.

Khan, E., Mack, J. P. G., Katz, R. A., Kulkosky, J., & Skalka, A. M. (1990) *Nucleic Acids Res. 19*, 851−860.

Kirchhoff, S., Koester, M., Wirth, M., Schaper, F., Gossen, M., Bujard, H., & Hauser, H. (1995) *Trends Genet. 11*, 219−220.

Klehr, D., & Bode, J. (1988) *Mol. Genet. (Life Sci. Adv.) 7*, 47−52.

Kohwi, Y., & Kohwi-Shigematsu, T. (1993) *J. Mol. Biol. 231*, 1090−1101.

Kohwi-Shigematsu, T., & Kohwi, Y. (1992) *Methods Enzymol. 212*, 155−180.

Kricker, M. C., Drake, J. W., & Radman, M. (1992) *Proc. Natl. Acad. Sci. U.S.A. 89*, 1075−1079.

Kulkosky, J., Jones, K. S., Katz, R. A., Mack, J. P. G., & Skalka, A. M. (1992) *Mol. Cell. Biol. 12*, 2331−2338.

Laemmli, U. K., Kaes, E., Poljak, L., & Adachi, Y. (1992) *Curr. Opin. Genet. Dev. 2*, 275−285.

Laker, C., Stocking, C., Bergholz, U., Hess, N., DeLamatar, J., & Ostertag, W. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 8452−8458.

Lu, Q., Wallrath, L. L., Allan, B. D., Glaser, R. L., Lis, J. T., & Elgin, S. C. R. (1992) *J. Mol. Biol. 225*, 985−998.

Lu, Q., Wallrath, L. L., Granok, H., & Elgin, S. C. R. (1993) *Mol. Cell. Biol. 13*, 2802−2814.

Mann, R., Mulligan, R. C., & Baltimore, D. (1983) *Cell 33*, 153−159.

McBurney, M. W., Fournier, S., Schmidt-Kastner, P. K., Jardine, K., & Craig, J. (1994) *Somatic Cell Mol. Genet. 20*, 529−540.

McCready, S. J., & Cook, P. R. (1984) *J. Cell. Sci. 70*, 189−196.

Mehtali, M., LeMeur, M., & Lathe, R. (1990) *Gene 91*, 179−184.

Mielke, C. (1993) Ph.D. Thesis, University of Braunschweig, Germany.

Mielke, C., Kohwi, Y., Kohwi-Shigematsu, T., & Bode, J. (1990) *Biochemistry 29*, 7475−7485.

Mielke, C., Tuemmler, M., & Bode, J. (1995) *Trends Genet. 11*, 258−259.

Milot, E., Belmaaza, A., Rassart, E., & Chartrand, P. (1994) *Virology 201*, 408−412.

Mountford, P. S., & Smith, A. G. (1995) *Trends Genet. 11*, 179−184.

Mullenders, L. H. F., van Zeeland, A. A., & Natarajan, A. T. (1983a) *Biochim. Biophys. Acta 740*, 428−435.

Mullenders, L. H. F., van Zeeland, A. A., & Natarjan, A. T. (1983b) *Nucleic Acids Res. 16*, 10607−10623.

Müller, H. P., & Varmus, H. E. (1994) *EMBO J. 13*, 4704−4714.

Nakagomi, K., Kohwi, Y., Dickinson, L. A., & Kohwi-Shigematsu, T. (1994) *Mol. Cell. Biol. 14*, 1852−1860.

Panganiban, A. T., & Temin, H. M. (1983) *Nature (London) 306*, 155−160.

Petersen, R., Sobel, S., Wang, C., Jaenisch, R., & Barklis, E. (1991) *Gene 101*, 177−183.

Ramstein, J., & Lavery, R. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 7231−7235.

Scherdin, U., Rhodes, K., & Breindl, M. (1990) *J. Virol. 64*, 907−912.

Schlake, T., & Bode, J. (1994) *Biochemistry 33*, 12746−12751.

Schlake, T., Klehr-Wirth, D., Yoshida, M., Beppu, T., & Bode, J. (1994) *Biochemistry 33*, 4187−4196.

Shih, C.-C., Stoye, J. P., & Coffin, J. M. (1988) *Cell 53*, 531−537.

Siedlaczck, I., Epplen, C., Riess, O., & Epplen, J. T. (1993) *Electrophoresis 14*, 973−977.

Stevens, S. W., & Griffith, J. D. (1994) *Proc. Natl. Acad. Sci. U.S.A. 91*, 5557−5561.

Taruscio, D., & Manuelidis, L. (1991) *Chromosoma 101*, 141−156.

Triglia, T., Peterson, M. G., & Kemp, D. J. (1988) *Nucleic Acids Res. 16*, 8186.

Ullu, E., & Tschudi, C. (1984) *Nature (London) 312*, 171−172.

Ullu, E., Murphy, S., & Melli, M. (1982) *Cell 29*, 195−202.

von Kries, J. P., Phi-Van, L., Diekmann, S., & Strätling, W. H. (1990) *Nucleic Acids Res. 18*, 3881−3885.

Weidle, U. H., Buckel, P., & Wienberg, J. (1988) *Gene 66*, 193−203.

Weiss, B., Thompson, A., & Richardson, C. C. (1968) *J. Biol. Chem. 243*, 2556−2563.

Wirth, M., Bode, J., Zettlmeissl, G., & Hauser, H. (1988) *Gene 73*, 419−426.

Withers-Ward, E. S., Kitamura, Y., Barnes, J. P., & Coffin, J. M. (1994) *Genes Dev. 8*, 1473−1487.

Woerner, A. M., Klutch, M., Levin, J. G., & Marcus-Sekura, C. J. (1992) *AIDS Res. Hum. Retroviruses 8*, 297−304.

BI952393Y