The EMBO Journal vol.11 no.1 pp.291-303, 1992

Retroviral integration into minichromosomes *in vitro*

Peter M.Pryciak[1], Anita Sil[1] and Harold E.Varmus[1,2]

Departments of [1]Biochemistry & Biophysics and [2]Microbiology & Immunology, University of California San Francisco, San Francisco, CA 94143-0502, USA

Communicated by P.Berg

We describe here the use of chromatin as a target for retroviral integration *in vitro*. Extracts of cells newly infected with murine leukemia virus (MLV) provided the source of integration activity, and yeast TRP1ARS1 and SV40 minichromosomes served as simple models for chromatin. Both minichromosomes were used as targets for integration, with efficiencies comparable with that of naked DNA. In addition, under some reaction conditions the minichromosomes behaved as if they were used preferentially over naked DNAs in the same reaction. Mapping of integration sites by cloning and sequencing recombinants revealed that the integration machinery does not display a preference for nucleosome-free, nuclease-sensitive regions. The distributions of integration sites in TRP1ARS1 minichromosomes and a naked DNA counterpart were grossly similar, but in a detailed analysis the distribution in minichromosomes was found to be significantly more ordered: the sites displayed a periodic spacing of ~10 bp, many sites sustained multiple insertions and there was sequence bias at the target sites. These results are in accord with a model in which the integration machinery has preferential access to the exposed face of the nucleosomal DNA helix. The population of potential sites in chromatin therefore becomes more limited, in a manner dictated by the rotational orientation of the DNA sequence around the nucleosome core, and those sites are used more frequently than in naked DNA.

Key words: chromatin/minichromosomes/MLV/nucleosome position/retroviral integration

Introduction

Not all DNA sequences in the cellular genome are in the same state or environment. For example, it is well established that DNA in eukaryotic cells is organized into nucleosomes and various higher degrees of chromatin packaging (for review, see Pederson *et al.*, 1986a). Such differential packaging of chromatin has often been postulated to serve both organizational and regulatory roles. Evidence suggests that different chromatin structures correlate with different physiological states, such as transcriptionally active versus inactive DNA (for review, see Gross and Garrard, 1988). What is the direct effect of chromatin on proteins that need to have access to cellular DNA? Particularly in the case of transcriptional initiation, it has been suggested that nucleosomes can participate in inhibitory regulation, by sterically restricting access of control sequences to transcription factors (for review, see Grunstein, 1990). Whether all DNA based activities are necessarily inhibited by chromatin is less clear; processive phenomena like DNA replication and transcriptional elongation appear not to be inhibited by nucleosomes (Lorch *et al.*, 1987; Bonne-Andrea *et al.*, 1990), although the function of a replication origin can be inhibited by nucleosomes *in vivo* (Simpson, 1990).

Retroviruses integrate a DNA copy of their genome into cellular DNA as an essential part of their life cycle (for review, see Varmus and Brown, 1989). Thus, retroviral integration can potentially provide a paradigm for understanding how various cellular functions are affected by the packaging of DNA into chromatin. But the role of chromatin structure in the choice of integration site has seldom been studied. While it is clear that integration can occur at many positions in the cellular genome and that little sequence preference is displayed (Varmus and Brown, 1989; Sandmeyer *et al.*, 1990), some studies suggest that the choice of integration site is not completely random. It has been observed that integration sites tend to map in or near transcriptionally active regions and nuclease-sensitive regions of chromatin (Vijaya *et al.*, 1986; Rohdewohld, 1987; Scherdin *et al.*, 1990; Mooslehner *et al.*, 1990). In addition, the frequency at which gene expression can be interrupted by retroviral integration can differ markedly from expectation based on random insertion (King *et al.*, 1985). In perhaps the most compelling example, Rous sarcoma virus (RSV) DNA has been observed to integrate at an unusually high frequency into certain preferred regions of the chicken genome, and within those regions insertions tend to occur into the exact same site (Shih *et al.*, 1988). Other studies have argued that most of the genome is available for integration (Reddy *et al.*, 1991).

Few attempts have been made to study these issues *in vitro*. *In vitro* integration assays generally use simple, naked DNA targets (Brown *et al.*, 1987; Craigie *et al.*, 1990; Katz *et al.*, 1990) and are therefore poorly suited to address the effects of complex physiological changes such as chromatin packaging, replication, or transcription on integration. We have begun to address such issues by modifying our previous *in vitro* integration assay to include chromatin targets. As simple chromatin models to serve as targets for integration *in vitro* we chose minichromosomes based upon a yeast plasmid and a mammalian viral genome. Because previous *in vivo* studies suggested that integration sites correlate with nuclease-hypersensitive regions, we used minichromosomes with both nucleosomal and nuclease-sensitive, nucleosome-free regions; this allowed us to design the *in vitro* experiments to test the effect of structural accessibility on choice of integration sites. Using such a strategy, we have been able to show that chromatin can serve as an integration target *in vitro*, that integration is not grossly inhibited by the presence of nucleosomes on the target DNA, and that

the availability of target sites is restricted, favoring some sites over others, when DNA is wrapped around nucleosomal cores.

Results

Characterization of minichromosome targets

We chose two types of well-characterized, relatively small minichromosomes (MCs) as simple models to serve as targets for retroviral integration into chromatin: the TRP1ARS1 plasmid (TA) from yeast, and the SV40 genome (SV) from acutely infected monkey cells (Figure 1). When isolated from yeast cells, the TA MC contains seven nucleosomes precisely arranged on structurally and functionally distinct regions of the 1.5 kb DNA molecule (Thoma *et al.*, 1984; Pederson *et al.*, 1986b; Thoma and Simpson, 1985). Two nucleosomal regions, three and four nucleosomes in size, correspond to untranscribed and transcribed regions, respectively, and they are separated by two nucleosome-free regions which map with the transcriptional control region and the origin of replication. The more heterogeneous SV MCs contain 20–27 nucleosomes distributed on 5.2 kb DNA molecules (Shelton *et al.*, 1980; Sogo *et al.*, 1986; Ambrose *et al.*, 1990). A nuclease-sensitive, nucleosome-free region is often present in the region encompassing the origin of replication and transcriptional control region for the two divergent transcription units (Varshavsky *et al.*, 1978; Saragosti *et al.*, 1980; Ambrose *et al.*, 1986).

These MCs were purified to varying extents, as described in Materials and methods. Because of the nature of the integration assays (described below), the MCs do not have to be highly purified. However, the MC preparation should be free of significant amounts of other DNA, such as genomic or mitochondrial DNA, since these will compete with MCs as integration targets; the MCs should be sufficiently concentrated to serve as targets for the integration machinery; and the vast bulk of the MCs should be present as chromatin, rather than naked DNA from disassembled MCs, to ensure that any integration events will occur into MCs *per se*.

Sedimentation in sucrose gradients demonstrated that the bulk of the DNA in the preparations was present as MCs, clearly sedimenting more rapidly than naked DNAs in the same gradients (Figure 2A and B). Digestion of TA MCs with micrococcal nuclease revealed the expected nucleosomal ladder pattern, with protected regions of ~150–200 bp (Figure 2C), whereas naked DNA was digested into randomly sized fragments and rendered undetectable at relatively low concentrations of nuclease. Hybridization of the digestion products of TA MCs with a probe corresponding to only a small region of the DNA molecule (*Eco*RI–*Xba*I, shown; others, not shown) revealed an unequal pattern consistent with the arrangement of nucleosomes determined by the more extensive analysis of Thoma *et al.* (1984; see Figure 1). Digestion of SV MCs with DNase I showed a peak of cleavage at the expected hypersensitive region located near the origin (Figure 2D), generating fragments of 2.5–2.7 kb after digestion with *Bam*HI, in a manner consistent with many earlier studies (Varshavsky *et al.*, 1978; Saragosti *et al.*, 1980; Ambrose *et al.*, 1986).

Minichromosomes as integration targets

In the experiments described here, the source of viral integration activity was extracts of cells newly infected with MLV; these extracts contain a large viral nucleoprotein complex ready to integrate linear viral DNA into an exogenous target (Brown *et al.*, 1987; Bowerman *et al.*, 1989). The product of the integration reaction can be detected by restriction enzyme digestion and Southern blot analysis (Figure 3A). Enzymes that cleave the viral DNA but not the target DNAs (e.g. *Bst*EII) generate molecules in which full-length target DNA is attached to the ends of viral DNA. Since the size of the crucial digestion product depends upon the size of the target, targets of different sizes can be assayed in the same reaction to assess relative use.

An example of this assay, using MCs and naked, heterologous DNAs as targets present separately or together in the same reaction, demonstrates integration into the TA or SV MCs (Figure 3B): the large fragments of different sizes (5.6, 7.0, 9.3 or 9.5 kb), indicative of integration, reflect the presence of differently sized targets. However, this experiment does not rigorously address how well the MCs were used compared with the naked DNA targets, or whether the MCs *per se*, rather than naked DNA from disassembled MCs, were serving as the integration target. To address these issues, we first examined the effect of target concentration upon integration efficiency (Figure 3C). Integration was sensitive to the amount of target, whether the target was naked ϕX174 DNA or SV40 MCs. More importantly, a comparison of the two targets shows that the amount of integration product is similar at similar concentrations of target. Thus, we cannot ascribe the MC integration products to use of a small fraction of naked DNA in the MC preparation.

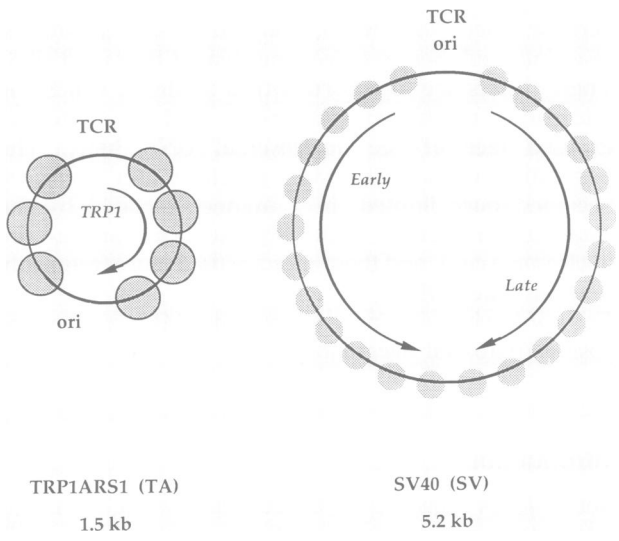

Fig. 1. Schematic representation of minichromosome targets. Nucleosomal and nucleosome-free regions, corresponding to transcriptional control regions (TCR) and origins of replication (ori), are indicated along with transcription units (curved arrows). The 1453 bp TRP1ARS1 (TA) minichromosome (left) is drawn with well defined, solid bordered nucleosomes to indicate their precise positioning into two regions separated by nuclease-sensitive, nucleosome-free regions (Thoma *et al.*, 1984; Pederson *et al.*, 1986). The 5243 bp SV40 (SV) minichromosome (right) is drawn with non-bordered nucleosomes to indicate their imprecise positioning (Ambrose *et al.*, 1990), in an array that is often punctuated by a nuclease-sensitive, nucleosome-free region near ori (Varshavsky *et al.*, 1978; Saragosti *et al.*, 1980; see discussion in Ambrose *et al.*, 1986).

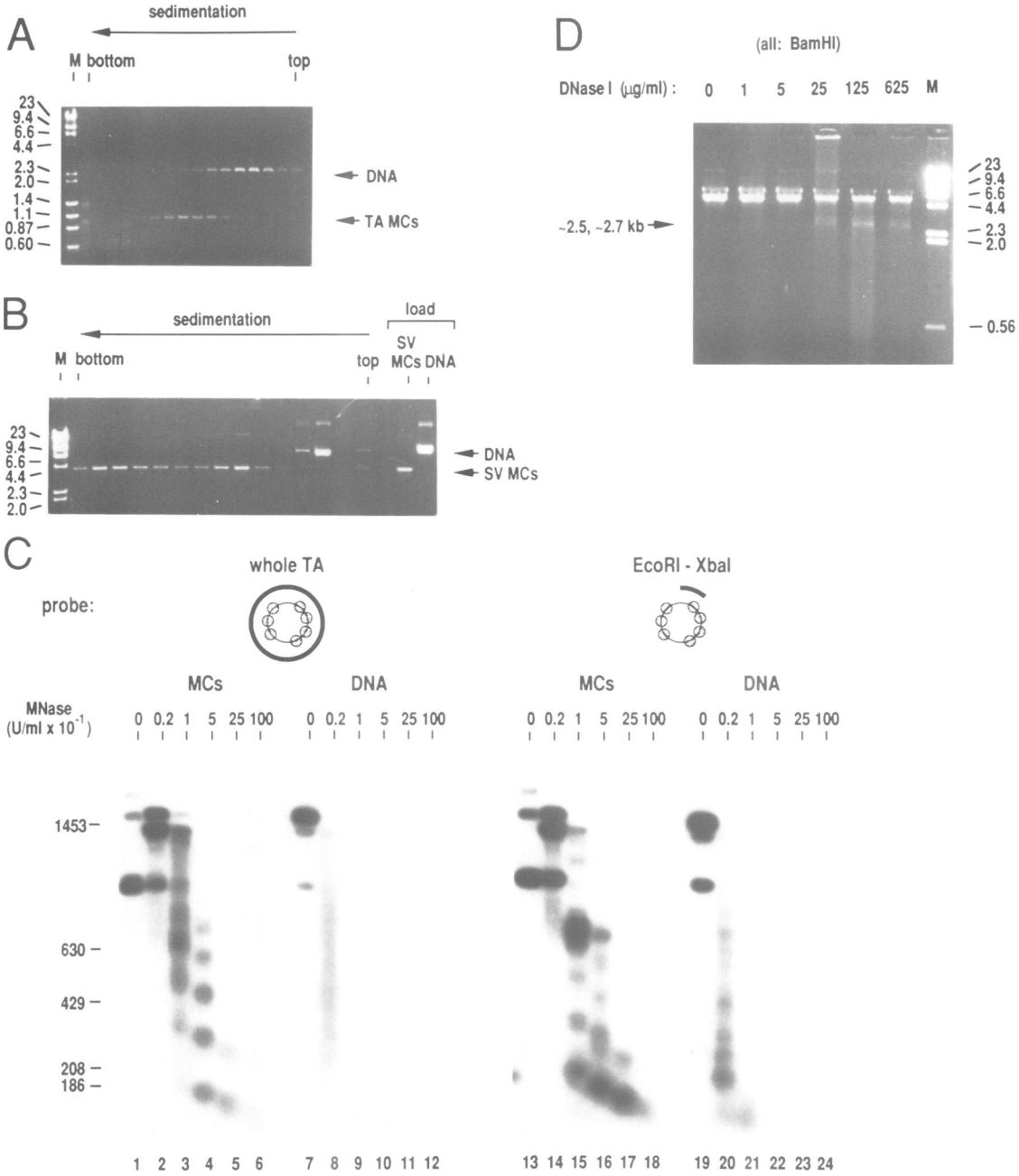

Fig. 2. Characterization of minichromosome targets. All analyses were performed with the same MC preparations used for the subsequent integration assays. **A.** Sedimentation of TA MCs mixed with a naked DNA plasmid (pTA-R; 4.2 kb) in a 10−30% sucrose gradient, for 3 h at 40 000 r.p.m. (Beckman SW50.1 rotor). After sedimentation, 16 equal volume fractions were collected from the bottom of the gradient and DNA prepared from these fractions was then separated on an 0.8% agarose gel and visualized by ethidium bromide staining. Lane M contains molecular weight standards of λ/*Hin*dIII and φX174/*Hae*III digests. **B.** Sedimentation of SV MCs mixed with a naked DNA plasmid (pBS-SVR; 8.2 kb) in a 5−30% sucrose gradient for 2 h at 40 000 r.p.m. (Beckman SW41 rotor). After sedimentation, 15 equal volume fractions were collected from the bottom of the gradient and DNA prepared from these fractions was then separated on an 0.8% agarose gel and visualized by ethidium bromide staining. Also included in the gel are samples of the SV MCs and plasmid DNA before loading on the gradient (load), and molecular weight standards of a λ/*Hin*dIII digest (M). The SV MCs show two peaks, probably corresponding to a 75S MC peak and a 180S 'previrion' peak (Boyce *et al.*, 1982). **C.** Mapping chromatin structure of TA MCs. Micrococcal nuclease (MNase) digestion of TA MCs and DNA (from deproteinized MCs) with the indicated MNase concentration was for 5 min at 37°C in the presence of 5 mM $CaCl_2$. Digestion products were separated on parallel 0.8% gels which were then blotted and hybridized with probes corresponding to the whole TA molecule (left, lanes 1−12) or to a small region between the *Eco*RI and *Xba*I sites (right, lanes 13−24), as schematically depicted above. Molecular sizes are indicated on the left, as determined by the products of restriction enzyme digestion of TA DNA run in the same gel (not shown). The greater protection from nuclease and the 150−200 bp ladder of protected fragments observed in the MC lanes are indicative of nucleosomal DNA. The strong enrichment for the tetranucleosome sized product when probing with the *Eco*RI−*Xba*I fragment is indicative of preferential cleavage at the two nucleosome-free regions, releasing products of three and four nucleosomes in size, of which only the latter will hybridize with the probe used. This experiment, and analysis with other probes and cleavage of MNase digestion products with restriction enzymes (not shown), is consistent with the nucleosome structure shown in Figure 1, as determined by Thoma *et al.* (1984). **D.** Mapping nuclease-sensitive region of SV MCs. DNase I digestion of SV MCs was performed with the indicated concentrations for 5 min at 37°C in the presence of 5 mM $MgCl_2$. DNase I digestion products were digested with *Bam*HI and then separated on a 0.8% agarose gel, along with λ/*Hin*dIII size markers (M), and visualized by ethidium bromide staining. *Bam*HI cleaves SV40 DNA (5.2 kb) at position 2533, and hence the appearance of DNase I-dependent products in the 2.5−2.7 kb size range (arrow) is indicative of preferential cleavage of the SV MCs near position 0, corresponding to the ori region indicated as often being nuclease-sensitive and nucleosome-free in Figure 1.

We also used direct competition assays to compare the behavior of MCs and naked DNA as integration targets (Figure 4). In these assays as many as three different targets of three different sizes were present in the same reaction, and their relative use was again measured by the different sizes of their products. Increasing amounts of a naked DNA (#2) were added as competitor to reactions containing another naked DNA (#1) and TA MCs; the relative use of #1 DNA and TA MCs matched very closely throughout the range of competition (Figure 4A). This provides additional evidence that the primary relevant target in the MC sample was the MCs *per se* since, in the presence of competitor, the level of integration into a small fraction of contaminating naked TA DNA from disassembled MCs would be expected to reflect its relative proportion of the total available DNA target.

A surprising but useful difference between MC and naked DNA targets was revealed when the integration reactions were carried out in the presence of polyethylene glycol (PEG; Figure 4B). Under these conditions, excess naked DNA (#2) did not interfere with the use of TA MCs as targets, even though it competed effectively with another naked DNA (#1) in the same reaction. Thus, in the presence of PEG, MCs behaved differently from naked DNA, again arguing that the vast majority of integration events into the MC target must have been into the MCs *per se*. Analogous experiments revealed similar differences in behavior between SV MCs and naked DNA in the presence of PEG (Figure 4C). The differences were dependent upon maintenance of the MC target as chromatin, since deproteinized SV MCs (SV DNA) behaved like other naked DNAs in response to increasing amounts of competitor DNA (Figure 4D). While we have made use of these differences, we have not elucidated their underlying mechanisms; nor have we attempted to assign values for absolute or relative target efficiencies, since they depend upon reaction

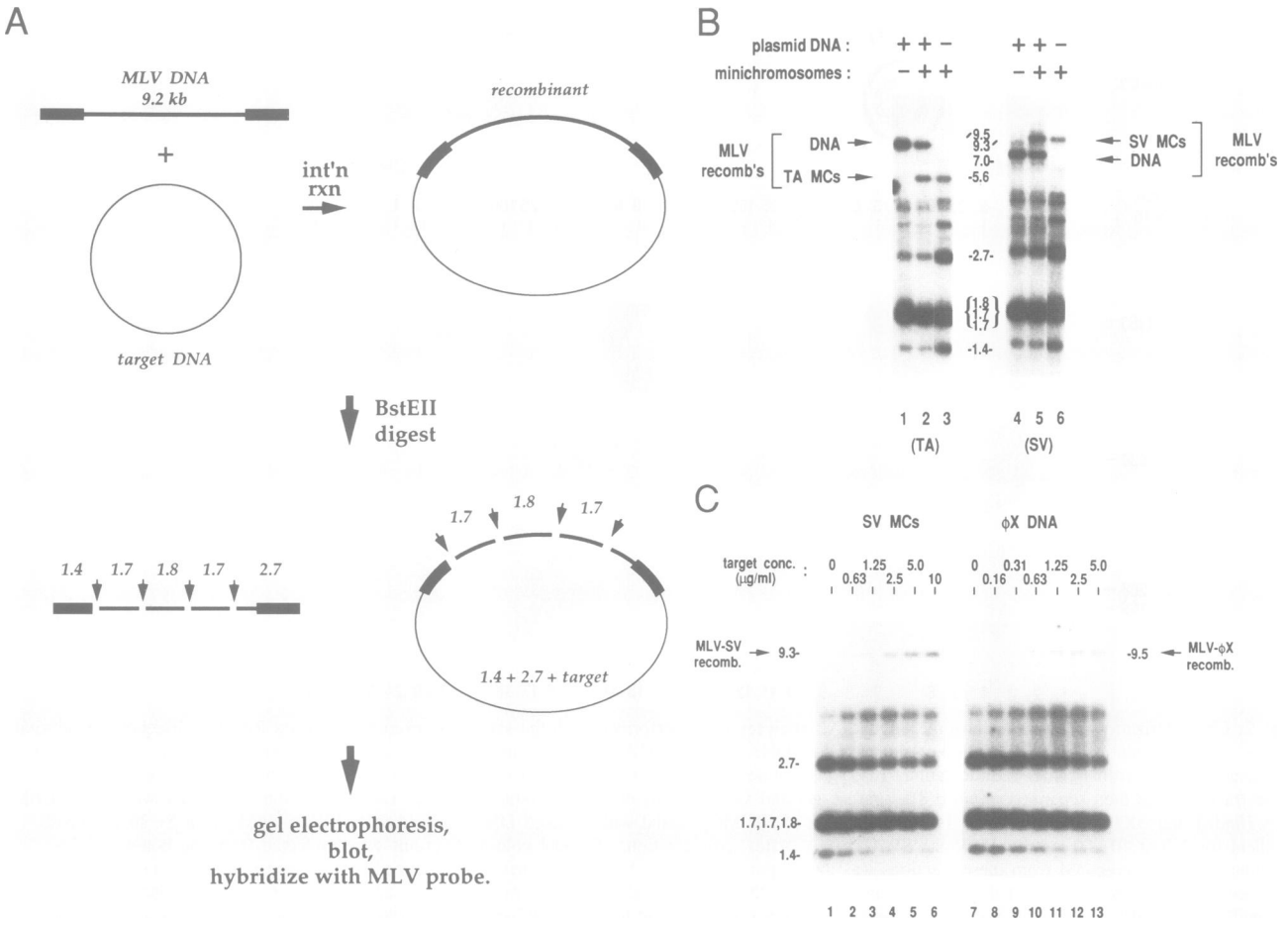

Fig. 3. Integration assays. **A.** Schematic depiction of the integration reaction and assay. The integration reaction gives a product with the MLV DNA (MoMLV-SupF, 9.2 kb) inserted into the target DNA (whose size is variable). Rarely does all the MLV DNA participate, so a subsequent *Bst*EII digest of the integration reaction products yields fragments originating from both unintegrated and integrated DNAs. However, a recombinant-specific fragment is generated from attachment of the ends of MLV DNA (1.4 and 2.7 kb in size) to the target DNA [and will be of a size equal to the sum of 1.4 + 2.7 + target size (in kb)]. **B.** Example of integration into MCs. Reactions contained either 1 μg/ml TA MCs, 10 μg/ml φX174 DNA or both as targets (left), or 10 μg/ml SV MCs, 10 μg/ml pGEM-2 DNA or both as targets (right). All reactions were carried out in parallel. Recombinants resulting from integration into TA MCs (1.5 kb), φX174 DNA (5.4 kb), SV MCs (5.2 kb) and pGEM-2 DNA (2.9 kb) are visualized by the appearance of 5.6, 9.5, 9.3 and 7.0 kb bands, respectively. Hybridizing fragments of sizes between those from recombinants and those from unintegrated DNA are usually observed, resulting (data not shown) from '1-LTR' and '2-LTR' circles and intramolecular 'autointegrants' (Shoemaker *et al.*, 1980, 1981; Varmus and Brown, 1989) that are not important for the experiments presented in this study. For a recent discussion of autointegrant formation *in vitro*, see Lee and Coffin (1990). **C.** Response of integration efficiency to target concentration. Reactions contained the indicated concentration of either SV MCs (left) or φX174 DNA (right), increasing in two-fold steps and recombinants (arrows) are visualized as described for (B). All reactions were carried out in parallel. The reactions in (B) and (C) were carried out in the presence of PEG (see Figure 4).

conditions. Also notable in the presence of PEG was the distinct boost in use of low amounts of MC targets upon addition of naked DNA (Figure 4B, lanes 2 and 3; Figure 4C, lanes 2 and 3); this is in accord with a consistently observed increase in all integration activity upon addition of DNA to these MLV integration extracts (P.M.P. and H.E.V., unpublished observations; Patrick Brown, personal communication), and is not pursued further here.

Because the addition of PEG strongly affected the relative use of MCs and DNA as targets, we tested other additives, including some non-ionic polymers that might be expected to change the apparent concentration of solutes by 'volume exclusion' (Tanford, 1961) and the polyamine spermidine (Figure 5A). Parallel reactions were performed using either SV MCs alone or SV MCs and naked plasmid DNA as targets. Of several reaction conditions examined, addition of spermidine was the most effective in stimulating use of both MCs and DNA (Figure 5A). Moreover, in the presence

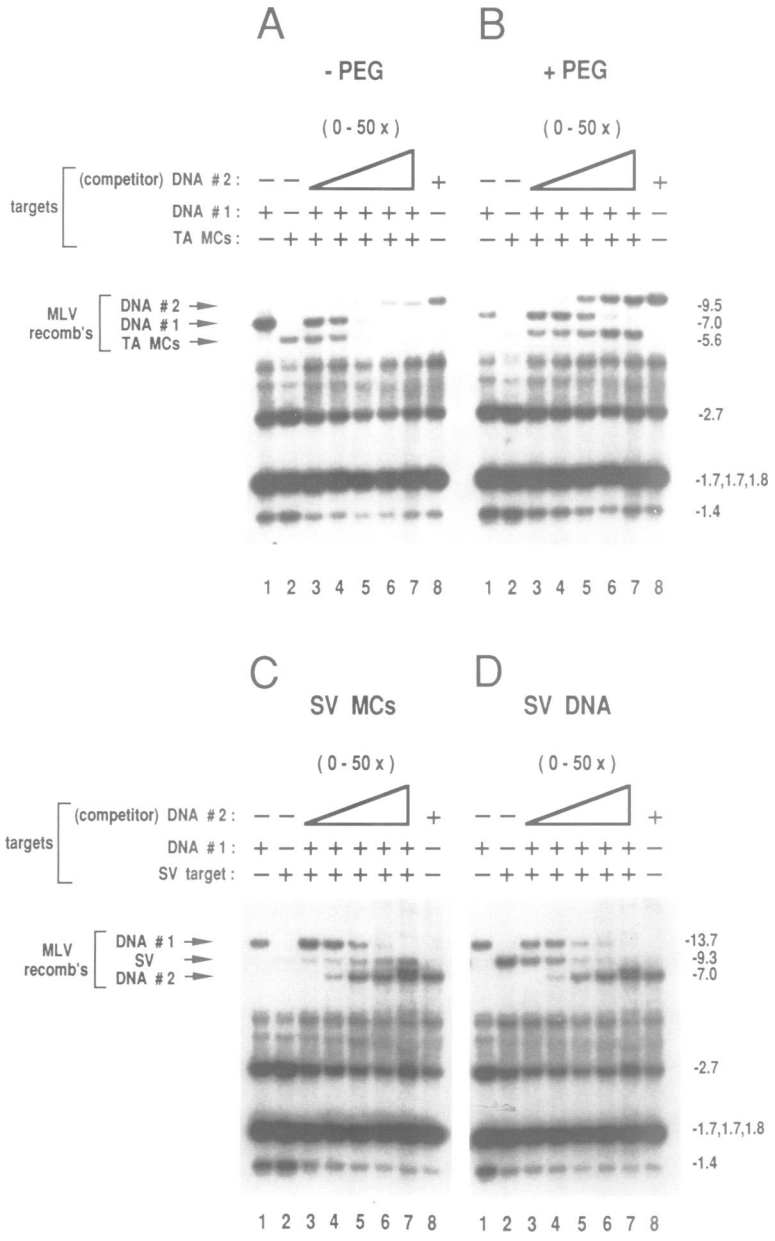

Fig. 4. Comparison of MCs with DNA as integration targets in response to excess competitor DNA. Integration reactions contained one, two, or three potential targets in the same mixture. In lanes 3–7, an MC target and a naked DNA (#1) target were present in constant amounts, and a varying amount of a third target (competitor; naked DNA #2) was present. **A** and **B.** TA MCs were absent (−) or present (+) at 0.8 μg/ml; pGEM-2 DNA (DNA #1) was absent (−) or present (+) at 1 μg/ml; φX174 DNA (DNA #2) was absent (−), present at 10 μg/ml (+) or present at 0, 0.4, 2, 10 or 50 μg/ml in lanes 3, 4, 5, 6 and 7, respectively, representing an increase from 0 to 50-fold excess over each of the other two targets. Recombinants (arrows) into TA MCs (1.5 kb), DNA #1 (pGEM-2; 2.9 kb), and DNA #2 (φX174 DNA; 5.4 kb) were identified by the appearance of 5.6, 7.0 and 9.5 kb bands, respectively. Reactions were carried out in the absence (A) or presence (B) of PEG. **C** and **D.** SV target was absent (−) or present (+) at 1 μg/ml; pSV-RI DNA (DNA #1) was absent (−) or present (+) at 1 μg/ml; pGEM-2 DNA (DNA #2) was absent (−), present at 10 μg/ml (+) or present at 0, 0.4, 2, 10 or 50 μg/ml in lanes 3, 4, 5, 6 and 7, respectively, representing an increase from 0 to 50-fold excess over each of the other two targets. Recombinants (arrows) into DNA #2 (pGEM-2; 2.9 kb), SV (5.2 kb) or DNA #1 (pSV-RI; 9.6 kb) were identified by the appearance of 7.0, 9.3 and 13.7 kb bands, respectively. Reactions were carried out in the presence of PEG, and the SV target used was either MCs (C) or DNA from deproteinized MCs (D). All reactions in A–D were performed in parallel.

of spermidine, as with PEG, DNA competed poorly with MCs as integration targets (Figure 5B). As a result, we added spermidine to all subsequent reactions.

Distribution of integration sites in MCs and DNA

To examine the distribution of integration sites within MCs, we cloned a large number of independent recombinants and sequenced the viral DNA−target DNA junctions, as described in Materials and methods. Most clones were sequenced at only one of the two junctions. Seven randomly picked MLV-TA MC recombinant clones were sequenced at both junctions, and all seven showed the expected loss of 2 bp from both ends of viral DNA and duplication of 4 bp of target DNA (Varmus and Brown, 1989), indicating that

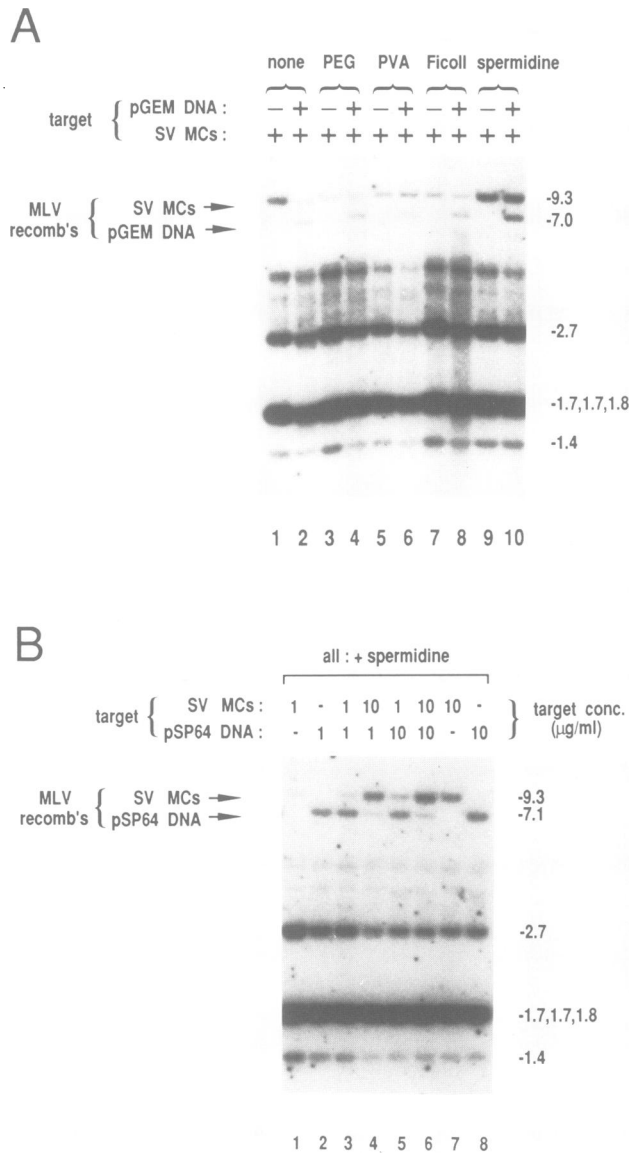

Fig. 5. Effect of reaction conditions on use of MC and DNA targets. **A.** All reactions contained 2 μg/ml SV MCs and either 0 (−) or 10 mg/ml (+) pGEM-2 DNA. Also, reactions were supplemented with additives as indicated (at concentrations given in Materials and methods). Recombinants (arrows) were identified as described in Figures 3 and 4. **B.** All reactions were supplemented with spermidine and were performed in the absence or presence of the indicated concentrations of SV MCs or pSP64 DNA. Recombinants (arrows) were identified as described in Figures 3 and 4 (pSP64 is 3.0 kb).

the insertions into MCs represent the products of legitimate MLV integration events (data not shown). The positions of 89 independent insertions into TA MCs, compared with 77 into TA DNA, reveal no strong preference for the nucleosome-free, nuclease-sensitive regions of the TA MCs (Figure 6A). This observation was supported by mapping 30 insertions into SV MCs (Figure 6B). In general, the distributions of insertions into TA MCs and DNA were grossly similar, with no major clustering or excluded regions. However, there was slightly more clustering of integration sites in MCs than in DNA, and there seemed to be a mild preference for, rather than a bias against, the nucleosomal regions in the MCs, but we have not attempted to analyze these features more rigorously here.

Because we mapped insertions by cloning and sequencing, allowing precise identification of insertion sites, we were able to recognize several non-random aspects of the distribution of integration events into MCs. All pairs of insertions that mapped within 25 bp of each other were identified, and the separation distance for each pair was measured (Figure 7A). Insertions into the TA MCs tended to be spaced in a regular fashion, mapping either very close (within 2 bp), or ~10 or 20 bp from each other, following a 10 bp periodic distribution. Such periodicity was not observed when naked TA DNA was used as the target. Because of its larger size and fewer insertions mapped, there were too few proximal pairs of insertions into SV MCs to analyze by these methods. A χ^2 goodness-of-fit test applied to these spacing data (see Materials and methods) suggested that the spacings of insertions into the TA MC were inconsistent with a random distribution ($P<0.001$), unlike the spacings of insertions into TA DNA ($P>0.10$). By summing the values in the histograms, the degree of bias towards this periodicity in TA MCs was estimated to be nearly 3-fold (Figure 7A). The insertion site period of ~10 bp correlates with the period of a DNA double helix, suggesting that the integration machinery displays a preference for one face of the DNA helix over the other in MCs. A model that can explain these observations (Figure 7B) is considered at length in the Discussion.

Two additional non-random features were observed in the distribution of integration sites in the TA MCs. First we compiled the coincident insertions, or independent integration events that map to the exact same position (Table I). In general, there were more frequent coincident insertions in MCs than in DNA: 34 of 89 in MCs versus eight of 77 in DNA. The fact that many coincident pairs were of different orientation and the careful steps taken in the cloning (see Materials and methods) argue for the independent origin of the coincident insertions. A statistical analysis (see Materials and methods) suggests that the probability that the number of coincident events observed with TA MCs would result from a random (Poisson) distribution is <0.001, in contrast to when TA DNA was the target ($P>0.50$).

Second, we looked for sequence bias among the sites used for insertions into TA MCs and DNA (Figure 8). The sequences of the target sites were aligned according to the position at which the four-base staggered break was made during the integration reaction. We then looked for bias among the aligned bases; specifically, we looked for frequent presence at any position of particular bases, A/T- or G/C-richness, or A/T-rich di- and trinucleotides that are known to show the strongest preferred positions in nucleosomal

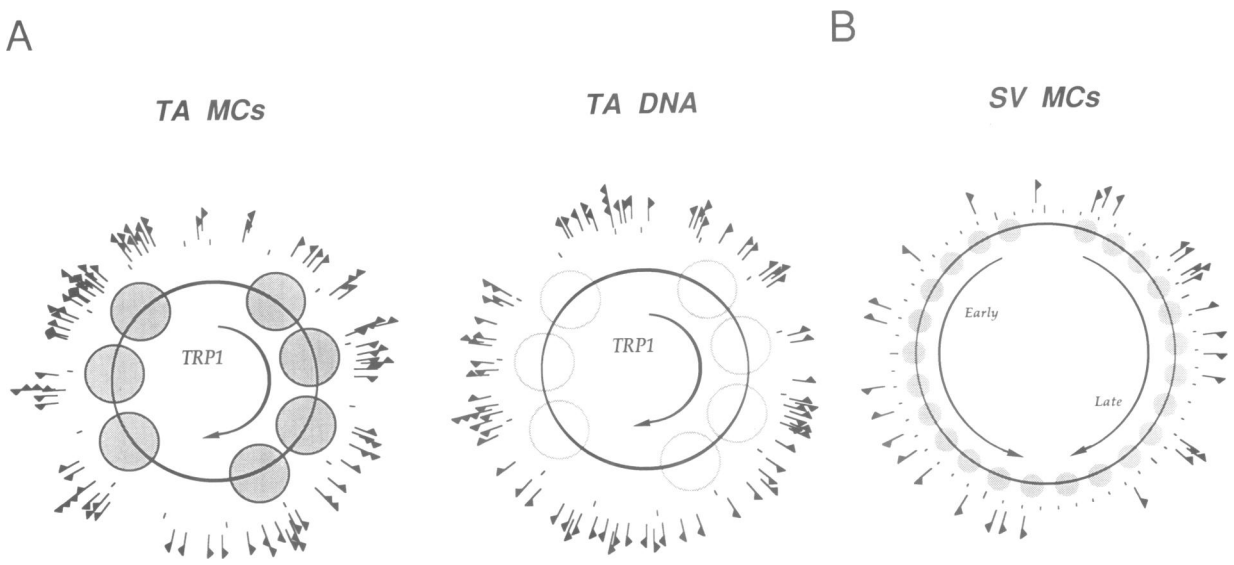

Fig. 6. Distributions of integration sites. Schematic representations of targets are similar to those in Figure 1. Flags designate mapped positions of independent insertion events (determined by cloning and sequencing as described in Materials and methods) and the orientation of each insertion. **A.** Distribution of 89 insertions into TA MCs (left) and 77 into TA DNA (right). Hash marks designate the progression along the TA sequence in 100 bp intervals, with position 0/1453 at top. In the TA DNA diagram, empty dashed-border circles indicate the positions of nucleosomes before deproteinization. **B.** Distribution of 30 insertions into SV MCs. Hash marks designate the progression along the SV sequence in 100 and 1000 bp intervals, with position 0/5243 at top.

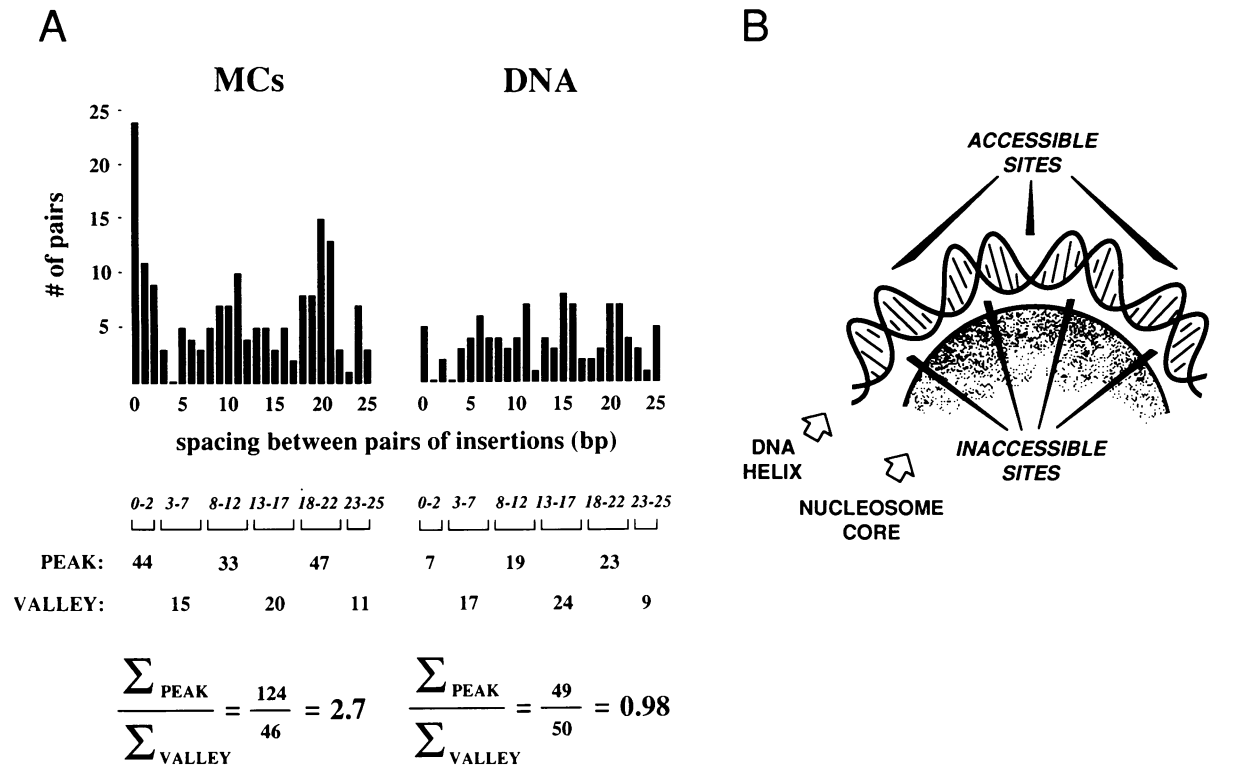

Fig. 7. A. Periodic distribution of integration sites in TA MCs. The distances between all pairs of independent insertions that map within 25 bp of each other were measured, and the histograms (above) show the number of pairs with each spacing value. In TA MCs (left), an ~10 bp periodic distribution of spacings is observed that is not observed in TA DNA (right). The MC case is inconsistent with a random distribution ($P<0.001$), in contrast to the DNA case ($P>0.10$; see text and Materials and methods for calculations). In addition, the numbers were combined in groups indicated by the brackets, corresponding to the 'peak' and 'valley' regions of the MC distribution, and the ratio of these sums was determined (below). In the DNA case, the ratio of these sums was close to 1 (0.98), as expected for a random distribution. In the MC case, however, the ratio of sums was 2.7, providing an estimate of the degree of bias (~3-fold) for the periodic distribution. **B.** Model for restriction of integration sites in nucleosomal DNA. The diagram depicts a segment of a DNA helix wrapped around a nucleosome core. The outside, or exposed, face of the helix is suggested to be preferentially accessible to the integration machinery. The implications of the model and its utility in explaining the experimental results are discussed in the text.

Table I. Higher frequency of coincident insertions in MCs than in DNA

	TA MCs		TA DNA	
	No. of inserts	No. of sites (orientations)	No. of inserts	No. of sites (orientations)
Sites not used	0	1364	0	1381
Sites used once	55	55	69	69
Sites used twice	24	12 (7+,−; 2+,+; 3−,−)	2	1 (−,−)
Sites used three times	6	2 (1+,−,−; 1+,+,−)	6	2 (2−,−,−)
Sites used four times	4	1 (+,−,−,−)	0	0
Total:	89	1453	77	1453
Frequency of coincident inserts	34/89=38%		8/77=10%	
Probability from random[a]		<0.001		>0.50

[a]Calculated as a χ^2 goodness-of-fit to a random (Poisson) distribution (see Materials and methods).

DNA (Satchwell *et al.*, 1986). As expected, insertion sites in naked DNA showed little preference at most positions, although a T was strikingly frequent at the second base from the site of cleavage on each strand. In contrast, insertion sites in TA MCs revealed additional base bias at several positions. Specifically, A/T-rich regions were favored in small clusters, following a roughly symmetrical organization about the target site: at the immediate center of the integration site and 10−12 bases to either side of the center. Such a bias pattern was independently revealed by noting positions where A/T-rich mono-, di- or trinucleotides were frequently present.

Thus, three distinct non-random features arose when comparing the integration sites in MCs with those in DNA: a periodic spacing of integration sites, an increased frequency of coincident insertions and an increased sequence bias at the target sites.

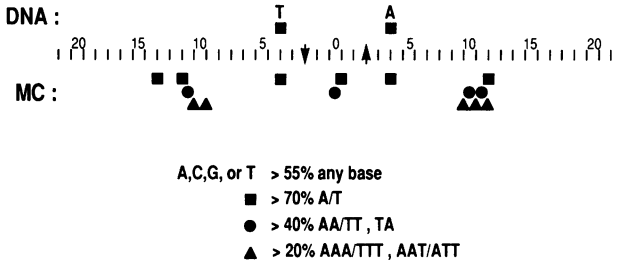

Fig. 8. Insertions into MCs show increased sequence bias. Insertion sites in TA MCs (86 total) and TA DNA (70 total) were separately aligned according to the position of the 4 bp staggered cut made in the target DNA (arrows), such that the orientation of MLV DNA was always in the same direction. At each position within the 4 base stagger, and 20 bases to each side, the number of occurrences of each base was counted (presented only for the top strand for simplicity). Also counted was the number of occurrences of A or T (squares), of the dinucleotides AA/TT or TA (circles), and of the trinucleotides AAA/TTT or AAT/ATT (triangles), and the frequencies of each occurrence calculated by dividing by the total number of sequences analyzed. In order to summarize the measured frequency distributions consisely, the positions at which these frequencies exceeded selected threshold values are shown. The probabilities of exceeding these thresholds at any one position are $P<0.001$ for the '>55% any one base' example and $P<0.05$ for the other examples (see Materials and methods for statistical calculations). Position 0 represents the exact center of the integration site, and positions 5, 10, 15 and 20 bonds from the center in each direction are shown.

Discussion

Minichromosomes are used as targets for integration in vitro

We have developed a system for studying the integration of retroviral DNA into chromatin *in vitro*. Using viral nucleoprotein integration complexes from infected cells and two kinds of minichromosomes (MCs), we have found that integration occurs in both nucleosomal and nucleosome-free regions of chromatin. Moreover, the distribution of insertion sites in nucleosomal regions suggests that the integration machinery shows a preference for the exposed face of the DNA helix.

Several observations argue that the MCs *per se*, rather than naked DNA from disassembled MCs, were used as integration targets: (i) integration efficiency responded to similar concentrations of MC and DNA targets (Figure 3C); (ii) MC and DNA targets were sensitive to similar ratios of competitor target DNA (Figure 4A); (iii) MCs differed from DNA as targets under certain reaction conditions (Figures 4 and 5); and (iv) the distribution of insertion sites into TA MCs was less random than the distribution of those into TA DNA (Figures 7 and 8; Table I). The sequences of the virus−target junctions in the integration products confirm that the insertions into MCs resulted from legitimate integration events (not shown).

We were initially concerned that it might be difficult to observe integration events into MCs because much of the DNA would be unavailable when incorporated into nucleosomes. Instead, the MCs were used as integration targets with efficiencies similar to those of naked DNA targets. There were, however, complex effects of reaction conditions on the utilization of MCs and DNA as integration targets (Figures 4 and 5) that, although experimentally useful, are difficult to understand. It is possible that higher apparent concentrations of all components in the reaction mixture (e.g. in the presence of PEG or spermidine) favors integration into MCs. Alternatively, some partitioning (e.g. Yamamoto *et al.*, 1970; Hoopes and McClure, 1981) of integration machinery from target into different phases might occur, such that a greater proportion of MCs than DNA partitions with the integration machinery. Finally, it is possible that distinct classes of integration complexes exist, each specific for either MCs or DNA, although we disfavor this idea because of the observation that MCs can compete with DNA (e.g. see Figure 3B, and Figure 5B, lanes 3−8).

Retroviral integration: a function that does not prefer nucleosome-free DNA

The distributions of integration sites into MCs and DNA were determined by sequencing individual clones of many integration products. The results show there is no strong preference for integration into nucleosome-free regions in either SV40 (SV) or TRP1ARS1 (TA) MCs; instead, nucleosomal DNA can be used as frequently as, and perhaps even more frequently than, nucleosome-free DNA in the same molecule. In contrast, nucleases such as micrococcal nuclease, DNase I, and restriction enzymes tend to show a strong preference for nucleosome-free regions, and, therefore, are used to determine nucleosome placement. However, such enzymes may not be able to interact productively with nucleosomal DNA. Retroviral integration, while a viral activity, may provide a useful paradigm for classes of cellular functions, such as general recombination or replication, that are active on (and may even prefer) nucleosomal DNA. Because retroviral integration can be studied both *in vivo* and *in vitro*, with many different DNA regions serving as targets, it is especially well suited to serve as such a paradigm.

A model for integration into nucleosomal DNA: the nucleosome influences choice of target sites

Although integration is not grossly inhibited by the presence of nucleosomes on the target DNA, a detailed analysis of the insertion sites suggests that the population of potential sites becomes more limited in MCs. Three distinct consequences of site limitations were noticed in our comparisons of integration sites in TA MCs and TA DNA: in MCs there was a periodic spacing of insertion sites at ~10 bp intervals, an increase in the frequency of coincident insertions and an increase in the apparent sequence bias. All three of these independent observations can be explained by the same model (Figure 7B). In this model, the orientation of the DNA sequences about a nucleosome core limits potential integration sites. We suggest that the face of the DNA helix against the nucleosome core is poorly accessible to the integration machinery, and the face away from the nucleosome core remains accessible. Recognizable features of the DNA helix, such as the major groove, would then be available only according to the period of the helix itself—roughly 10 to 10.5 bp—and would give rise to the observed periodic spacing of insertion sites (see Figure 7A). Alternatively, rather than inhibition of the inside face, enhancement of the outside face of the nucleosomal DNA helix, e.g. through bend-induced perturbations of DNA structure, could also result in its preferential use. One prediction from our model might be that the periodic spacing would not be observed in the nucleosome-free regions of the TA MCs. Unfortunately our collection contains too few proximal insertions in these regions (only four pairs in total) to address this point.

Periodic modulation of accessible sites in nucleosomal DNA has previously been observed with DNase I. This enzyme cuts nucleosomal DNA only at sites where the minor groove of the DNA helix is exposed on the face away from the nucleosome core, producing a 10–10.5 bp periodic pattern (Lutter, 1978; Drew, 1984). In its limitation to one face of the nucleosomal DNA helix, DNase I resembles the retroviral integration machinery, but it differs by strongly preferring nucleosome-free regions. Perlmann and Wrange (1988) and Pina *et al.* (1990) have shown that a transcription factor, the glucocorticoid receptor, can bind its sites in the MMTV LTR even when the DNA is incorporated into a nucleosome. Furthermore, it was suggested that sites in nucleosomal DNA were only bound if the major groove of the helix at that site was facing away from the nucleosome core. This scenario is highly analogous to that which we are proposing for retroviral integration.

The model proposed to explain our findings with retroviral integration (or results of others with glucocorticoid receptor binding) requires that the nucleosomal DNA sequences have a consistent rotational orientation. In other words, the same bases must consistently face either toward or away from the nucleosome core in most or all nucleosomes. Is this a reasonable and usual condition? The answer appears to be yes, at least for the cases which have been directly analyzed. In general there is strong evidence that the primary determinant of the rotational orientation of the nucleosomal DNA helix is the sequence of the DNA segment itself (Linxweiler and Horz, 1985; Drew and Travers, 1985; Satchwell *et al.*, 1986). Particular short DNA sequences display strong preferences for placement either toward or away from the nucleosome core, because of differences in the physical constraints of the compressed (inside) versus the expanded (outside) portions of the wrapped DNA. The incorporation of DNA into the nucleosome appears to occur in a manner which maximizes the placement of as many sequences as possible at their preferred positions (Satchwell *et al.*, 1986). These general principles have even been used to design artificial DNA segments that are incorporated into nucleosomes more favorably than natural sequences (Shrader and Crothers, 1989). In several closely examined cases, such as the MMTV LTR fragment mentioned earlier, DNA sequences are placed into nucleosomes precisely, such that the vast majority of nucleosomes in a population show the same rotational orientation of the nucleosomal DNA (Simpson and Stafford, 1983; Linxweiler and Horz, 1985; Pina *et al.*, 1990). However, the rotational orientations of the seven nucleosomes in the TA MCs used in our studies have not been determined, and the nucleosome positions along the DNA have been determined only at low resolution (±10–20 bp; Thoma *et al.*, 1984). One prediction of our work, therefore, is that the TA MC nucleosomes will show a consistent or major rotational positioning.

Increased frequency of coincident insertions in minichromosome targets

The above sort of specific rotational positioning of nucleosomal DNA sequences, which we hypothesize exists in the TA MCs, could explain the increased frequency of insertions into the exact same site (coincident insertions). A specific or major rotational positioning would make a particular subset of all possible sites, clustered at 10 bp intervals, preferentially available. In the simplest case, fewer actual sites would be available in the MC than in DNA, because many of them would be made unavailable due to their position on the core-proximal face of nucleosomal DNA (see Figure 7B). Thus, with fewer sites and a similar number of events analyzed, more events would occur in the same site in MCs than in DNA, in accord with our observations (Table I). Again, it is also possible that the creation of exceptionally good sites on the outer face of nucleosomal DNA could contribute to the increased frequency of

coincident insertions. We presently lack any compelling evidence to distinguish between these two ideas, although the lack of bias against nucleosomal DNA (and for nucleosome-free DNA) may favor the second explanation.

Some information is available about the frequency of coincident insertions *in vivo*. There are several loci in which numerous proviral insertions have been mapped following selection for insertional activation of proto-oncogenes (Selten *et al.*, 1984; Shih *et al.*, 1984; Raines *et al.*, 1985). In all of these cases, it is clear that many different positions have served as integration sites. However, the locations of some proviruses are indistinguishable at the resolution level of restriction mapping, but the actual viral–host junctions have not yet been sequenced in these cases to determine whether any represent coincident insertions. Shih *et al.* (1988) showed that a disproportionate number of RSV-mediated integration events occur into a subset of chromosomal regions, and that all of the insertions within those regions were at exactly the same base. This represents a degree of bias not easily explained by our model, and may depend on more complicated phenomena. Importantly, when these high frequency sites were used as naked DNA targets in an *in vitro* integration reaction, they were not used preferentially, and insertions occurred at multiple sites (J.Coffin, personal communication). In a large number of insertions of the yeast retrotransposon Ty1 selected on the ability to impair function of the *URA3* and *LYS2* genes, an exceptionally high proportion of sites sustained multiple insertions (Natsoulis *et al.*, 1989). While these observations may suffer from the usual caveats of selection bias, the use of disruption rather than activation insertions would seem the most permissive strategy. In any case, it should now be possible to compare unselected insertions into the same loci *in vitro*, in a manner analogous to the present study, by use of an *in vitro* Ty1 transposition system (Eichinger and Boeke, 1988). Integration of the yeast retrotransposon Ty3 occurs virtually exclusively at the transcriptional start sites of tRNA genes (Chalker and Sandmeyer, 1990). This extreme degree of target site preference is not explainable by target sequence alone, since the tRNA gene must be transcriptionally competent to serve as a high-frequency target (D.Chalker and S.Sandmeyer, personal communication). This may be the clearest example of how the same DNA sequence can be used differently as an integration target depending on its physiological state.

Insertions into minichromosomes show target sequence bias

As expected from previous attempts to define target consensus sequences (Shimotohno and Temin, 1980; Shoemaker *et al.*, 1981), the sites used for insertion into naked DNA bear little resemblance to each other, although there was a striking preference for a T base on each strand at the second position 5′ from the target cleavage site (see Figure 8). More importantly, the insertion sites into MCs showed additional sequence bias. This presents another non-random feature of integration into chromatin that can be explained by the model presented above, without a requirement for specific sequence recognition by the integration machinery. As discussed above, the DNA sequence determines the rotational positioning of the DNA helix around the nucleosome core, i.e. which bases face towards and which away from the core. The most critical features of the DNA sequence are the distribution of A/T- and G/C-rich regions. These will become arranged to maximize the placement of A/T-rich regions at positions where the minor groove of the DNA helix faces inward and G/C-rich regions where it faces outward, with particular di- and trinucleotides showing the strongest periodic modulation of position (Drew and Travers, 1985; Satchwell *et al.*, 1986). The sequence bias of the MC insertion sites is predominantly towards a symmetrical pattern of A/T-rich mono-, di- and trinucleotides. We suggest that this results from integration into the major groove of the DNA helix on the face away from the nucleosome core, such that A/T-rich regions will tend to be present at the center and at ~10 bp intervals to either side of the center of the integration site. Insertion by attack into the major groove is consistent with the fact that the 4–6 bp 5′ staggered cuts of target DNA made during retroviral integration (Varmus and Brown, 1989) can be accomplished by cleavage of phosphodiester bonds that face each other directly across the major groove. Importantly, we argue that the insertion site sequence bias observed for the MC case would not be due to a true bias by the integration machinery for those sequences; instead it would be due to the presentation by the nucleosome of a limited subset of the original potential sites, as determined by their rotational arrangement. We are currently subjecting these ideas to a more thorough test by using as an integration target nucleosomal DNA for which the rotational orientation is known.

The cloning and sequencing method of mapping has allowed us to draw interesting conclusions from a higher density of retroviral integration sites than previously obtained. Nevertheless, the data obtained can be regarded as rather sparse, in the sense that far fewer insertions were mapped than there are potential sites in the targets. We have therefore recently developed a general technique for mapping integration site distributions in large populations of recombinants rather than individual clones. Preliminary results relevant to the findings presented here suggest that the reaction conditions observed to affect the efficiency of integration into MCs and DNA (see Figures 4 and 5) do not affect the distribution of integration sites in these targets or the chromatin structures of the MC targets (P.M.P. and H.E.V., unpublished observations).

Chromatin structure and integration in vivo

Among the motivations for this work were previous suggestions that integration *in vivo* occurs preferentially in nuclease-sensitive and/or transcriptionally active regions. Do our observations conflict with those claims? We think not. In prior efforts to correlate integration sites with nuclease-sensitive sites *in vivo* (Vijaya *et al.*, 1986; Rohdewohld *et al.*, 1987), the criteria for proximity of integration sites to nuclease-sensitive sites allowed distances of up to 500 bp. By this measure, most of the target DNA in our MCs can be considered to be 'near' a nuclease-sensitive region; in the case of the TA MCs, all bases are within 500 bp of a nuclease-sensitive, nucleosome-free region. Most importantly, our results imply that the previous *in vivo* data cannot be explained by preferential use of nucleosome-free DNA over nucleosomal DNA. It seems plausible, however, that some preference could operate by discrimination between uncondensed and more highly condensed levels of chromatin packaging; in our *in vitro* studies we have only

utilized chromatin targets with a relatively low condensation level. Thus, the use of more complex chromatin models as *in vitro* integration targets would seem warranted.

Materials and methods

Strains, cells and plasmids

MoMLV-SupF (see Brown *et al.*, 1987) was propagated on SC-1 cells (courtesy of J.Levy, University of California San Francisco) and used to infect NIH-3T3 cells. The triple protease-deficient yeast strain BJ2168 (*MAT***a** *pep4-3, prc1-407, prb1-1122, ura3-52, trp1, leu2, gal2;* Jones, 1991), made cir^0 by selecting for loss of endogenous 2μ plasmid (by J.Thomas and K.Yamamoto, University of California San Francisco), served as the host for the TRP1ARS1 plasmid. This plasmid was introduced by transformation of a self-ligated *Eco*RI fragment cut from the plasmid pTA-R (from J.Thomas), which contains the TRP1ARS1 *Eco*RI fragment cloned into the *Eco*RI site of the pUC18 vector. Following transformation, a clone (BJ2168cir^0:TA-9) that contained monomer-size plasmid (as determined by gel and Southern blot analysis) was picked for further use and was grown in standard synthetic medium without tryptophan (Dean *et al.*, 1989). Wild-type SV40 strain #777 (from E.Shekhtman and N.Cozzarelli, University of California Berkeley) was propogated on CV-1 monkey cells by standard methods (e.g. Oudet *et al.*, 1989). The plasmid pSV-RI (from M.Verderame) contains an *Eco*RI-linearized SV40 genome cloned into the vector pBR322; subsequently, the *Eco*RI SV40 fragment was cut from pSV-RI and cloned into the pBS-KS(+) vector to obtain the pBS-SVR plasmid. The plasmids pGEM-2 and pSP64 (Promega) and ϕX174 (New England Biolabs) were used as integration targets. A plasmid pUC8.2, constructed by cloning the 8.2 kb *Hin*dIII fragment of permuted circular MoMLV DNA from p8.2 (Schwartzberg *et al.*, 1983) into the *Hin*dIII site of pUC18, was used as a probe for Southern blot hybridization. The bacterial strain XAC-1 (F' $lacI^q_{373}$ $lacZ_{ull8am}$ *proB*+/ F$^-$ Δ(lacpro)$_{x111}$ nalA, rif, $argE_{am}$, ara; Normanly *et al.*, 1986) was obtained from J.Miller (University of California Los Angeles).

Integration extract preparation and reactions

MLV-SupF integration extracts were prepared as described (Brown *et al.*, 1987), 12–16 h after infection of NIH-3T3 cells with fresh 12–24 h harvests of virus from chronically infected SC-1 cells and were stored frozen in aliquots at −80°C. Only cytoplasmic extracts were used for the experiments described here. Integration reactions were performed as described (Brown *et al.*, 1987). Briefly, 50 μl of cold extract was combined with target DNA, minichromosome preparation or both (usually at 1–10 μg/ml; for specific concentrations see text) and brought to a final reaction volume of 75 μl with water. Some reactions (see text) also contained 5% polyethylene glycol 8000 (PEG), 3.3% polyvinyl alcohol (PVA), 4.2% Ficoll 400 or 15 mM spermidine as additives; these were always added last. Reactions were then incubated at 37°C for 30–60 min and were stopped; nucleic acids were prepared as described (Brown *et al.*, 1987). To assay the extent of integration reactions (Brown *et al.*, 1989), the purified nucleic acids were digested with *Bst*EII and run on 0.7% agarose gels, which were then blotted onto Hybond-N nylon membranes (Amersham), hybridized (Church and Gilbert, 1984) to an MLV probe [pUC8.2 labelled with ^{32}P by the hexamer-priming method (Feinberg and Vogelstein, 1983)] and analyzed by autoradiography.

Preparation of minichromosomes

TRP1ARS1 minichromosomes (TA MCs) were purified from yeast cells by a method based upon that of Dean *et al.* (1989). Yeast cultures (OD_{600} = 1) were collected, spheroplasts were prepared, collected and lysed, and nuclei were collected according to Dean *et al.* (1989). Nuclear pellets were washed once in 80 mM KCl, 5 mM $MgCl_2$, 10 mM PIPES pH 6.3, 1 mM EGTA, 0.5 mM spermidine, 1% aprotinin and 18% Ficoll 400, and were then resuspended in (1 ml per liter original culture) 200 mM NaCl, 5 mM $MgCl_2$, 10 mM MOPS pH 7.4, 0.5 mM EGTA, 1% aprotinin and 0.1% 2-mercaptoethanol [nuclear elution buffer (NEB)]. MCs were then allowed to elute from these nuclei on ice for 1.5 h, followed by centrifugation in a GSA rotor at 7000 r.p.m. for 5 min at 4°C, collection of the supernatant and then one repeat of resuspension, elution, centrifugation and supernatant collection. The combined supernatants were mixed with an equal volume of 80% Nycodenz (Accurate Chemical) in NEB, and centrifuged for 36 h at 45 000 r.p.m., 4°C, in a Beckman VTi50 rotor. Fractions containing MCs were pooled and re-centrifuged for 18 h at 55 000 r.p.m., 4°C, in a Beckman VTi65 rotor. Peak fractions (250 μl per liter original culture, in ~40% Nycodenz) were pooled, made 8% in sucrose and frozen in aliquots at −80°C. These preparations contained ~8 ng/μl TRP1ARS1 DNA, as estimated by comparison with known standards on agarose–ethidium bromide gels and confirmed by fluorometric measurement.

SV40 minichromosomes (SV MCs) were purified 68 h after infection of CV-1 cells at a multiplicity of infection of 10 by a method based upon Luchnik *et al.* (1982) and Oudet *et al.* (1989). Cells were scraped from culture plates, pelleted and washed once in 10 mM PIPES pH 6.8, 150 mM NaCl and 1 mM EDTA [wash buffer (WB)], then lysed in WB plus 0.05% digitonin. Nuclei were harvested by centrifugation, washed twice in WB plus digitonin, then resuspended in (100 μl/10^7 cells) 10 mM MOPS pH 8.0, 200 mM NaCl and 1 mM EDTA. MCs were allowed to elute for 4 h on ice, then nuclei and debris were pelleted at 10 000 *g* for 15 min. The supernatant was made 8% in sucrose and frozen in aliquots at −80°C. Such (68 h post-infection) preparations contained significantly more DNA than others at shorter times post-infection, although much of the DNA was associated with previrion (180S) structures (Boyce *et al.*, 1982). The concentration of SV40 DNA was determined to be ~100 ng/μl by methods described above for the TA MCs.

Cloning and sequencing of integration products

To prepare integration products for cloning and sequencing, integration reactions were carried out as described above (with spermidine present in the reaction), in two separate experiments: experiment #1 involved separate parallel reactions into SV and TA MCs, and experiment #2 involved separate parallel reactions into TA MCs and TA DNA (which was obtained by deproteinization, phenol–chloroform extraction and ethanol precipitation). Data for TA MCs from the two experiments were combined for all analyses in this study. Integration products were cloned by digestion with *Sac*I (TA) or *Xba*I (SV) and ligation into the corresponding site in λ ZAP DNA (Stratagene), which contains a debilitating amber termination codon (Sam100) that can be suppressed by the suppressor tRNA gene resident in the MLV-SupF genome, allowing growth on a Sup° bacterial host. Ligated DNA was packaged into phage, which were then plated onto XAC-1 cells in the presence of X-gal and IPTG. Phage containing MLV-SupF sequences gave rise to blue plaques, and those blue plaques that contained recombinants (usually 5–10%) were identified by hybridization of plaque lifts (Benton and Davis, 1977) to SV or TA probes (these target DNAs could only be cloned as MLV recombinants, since they do not contain sites for the restriction enzymes used for the cloning). Recombinant plaques were picked from several independent platings of the same packaged phage, and were stored as individual clones. The inserts in these phage were subsequently excised as phagemids (according to the manufacturer's instructions), were grown in cultures, and plasmids were prepared by alkaline lysis miniprep procedures. Lambda phage clones were individually manipulated by this procedure in small groups (no more than 24 clones picked per plate) on several different days (no more than 24 clones per day) in order to ensure the independent origin of each final miniprep DNA sample. Thus, of the 11 cases of same-orientation same-site inserts (see Table I), 10 were from different experiments (#1 or #2 TA MC) or from different plates; only one case was from the same plate of lambda phage plaques.

Double-stranded miniprep DNAs were sequenced using an oligonucleotide primer (SupF-17) complementary to SupF gene sequences in the MLV-SupF LTR sequence and directed toward the viral DNA–target DNA junction. Some reactions gave unreadable sequence and were not resequenced; in total, 30 SV MC recombinants, 89 TA MC recombinants and 77 TA DNA recombinants (196 total) gave readable sequence. The insertion sites were located by matching the sequence immediately past the junction with the known target sequence [using the computer program DNA Strider, (Marck, 1988)]. These were used to generate the insertion site distribution maps. All clones were sequenced at one end of the viral DNA (using the SupF-17 primer). In addition, seven TA MC recombinant clones were picked at random and were sequenced at the other end (using the MoU5L17 primer) to verify that the insertions into MCs were legitimate products of retroviral integration. All others, except eight of 196, displayed a normal junction ending in the viral CA dinucleotide at the one end sequenced; but one of 30 SV MC, two of 89 TA MC and five of 77 TA DNA recombinants had junctions ending in CATT, suggesting that the proximal TT dinucleotide was not removed during the integration process. Seven of these eight clones were then sequenced at the other end, and all showed proper ending of the viral sequence with CA but aberrantly sized duplication of target sequences. These structures are consistent with normal integration, followed by misrepair of the gapped intermediate without removal of the unpaired AA dinucleotide of viral DNA at one end. These clones were not used in the detailed analyses of insertion site periodicity and sequence bias (none were members of coincident insertion sets), because of the ambiguity of the exact site of integration. Also not included in these detailed analyses were three other TA recombinants; two (one MC and one DNA) for which the sequencing gel was slightly obscured in the region

of the target junction so that the insertion site could only be assigned ± 1 bp, and 1 (DNA) that was found to have duplicated 5 bp of target DNA after being randomly picked for sequencing of both junctions. The remaining 86 TA MC and 70 TA DNA recombinants were then analyzed further. The exact insertion sites are not presented here, but can be obtained by written request to the authors.

Statistical methods for analysis of insertion site distributions

The insertion site distributions were analyzed by χ^2 tests of goodness-of-fit to a random distribution (Zar, 1984). Such tests calculated $\chi^2 = \Sigma(f_o - f_e)^2/f_e$, where f_o is the observed frequency and f_e is the expected frequency of an occurrence. The calculated χ^2 was then compared with a table of χ^2 critical values for the appropopriate degrees of freedom, υ. For the analysis of the spacings between insertion sites: f_e is equal to the total number of pairs of insertion sites with spacings of 25 bp or less (170 for TA MCs, 99 for TA DNA) divided by the total number of potential spacing values (=26; thus, f_e is 170/26 for TA MCs and 99/26 for TA DNA); f_o for each spacing value (0 to 26) is equal to the actual observed number of pairs with that spacing value (see Figure7A). Thus, $\chi^2 = 98.3$ for the TA MC distribution and 32.6 for the TA DNA distribution; in both cases $\upsilon = 25$; comparison with a table of critical values reveals that the TA MC distribution is inconsistent with a random distribution ($\chi^2 = 98.3$, $P<0.001$) (i.e. the null hypothesis, that the observed data resulted from a random distribution, should be rejected), but the TA DNA distribution ($\chi^2 = 32.6$; $0.25>P>0.1$) is consistent with a random distribution (i.e. the null hypothesis cannot be rejected).

For the analysis of coincident insertions, f_e was calculated according to a Poisson distribution, such that $f_e(X) = N\mu^X e^{-\mu}/X!$ where X is the number of inserts at the same site (0 to 4), N is the total number of potential sites (1453) and μ is the population mean number of occurrence per site (equal to the number of inserts sequenced divided by N; so $\mu = 89/1453$ for TA MCs and 77/1453 for TA DNA); f_o for each value of X is equal to the number of sites that were observed to receive X inserts at the same site (see Table I); χ^2 was then calculated as above, with pooling of the frequencies in the tails of the distribution such that f_e was never <1.0 (Cochran, 1954; υ was 2 after such pooling); the TA MC distribution ($\chi^2 = 68.3$) was inconsistent with a random (Poisson) distribution ($P<0.001$), while the TA DNA distribution was consistent with a random distribution ($\chi^2 = 0.767$; $0.75>P>0.50$).

For the analysis of sequence bias, f_e was directly measured by counting the frequencies of bases, dinucleotides and trinucleotides in the TRP1ARS1 sequence. This sequence contains 29.8% A, 29.8% T, 20.2% G, 20.2% C, 59.6% A/T, 40.4% G/C, 27.1%AA/TT+TA and 12.3% AAA/TTT+AAT/ATT. χ^2 was calculated for exceeding the selected threshold frequencies that are shown in Figure 8: for >55% any one base, $\chi^2 = 26.2$ (MC) or 21.1 (DNA), giving $P<0.001$; for >70% A/T, $\chi^2 = 3.86$ (MC), giving $P<0.05$; for >40% AA/TT+TA, $\chi^2 = 7.24$ (MC), giving $P<0.01$; and for >20% AAA/TTT+AAT/ATT, $\chi^2 = 4.72$ (MC), giving $P<0.05$.

Acknowledgements

The authors thank Jay Thomas for generous gifts of reagents and advice in minichromosome preparations. We are especially grateful to Pat Brown for advice, discussion and assistance during many stages of this work. We also thank Julien Hoffman, Stanton Glantz, Joan Hilton and Vojtech Licko for advice on statistical analysis, and Hans-Peter Muller for comments on the manuscript. This work was supported by a grant from the NIH to H.E.V., and P.M.P. was supported by a UCSF Chancellor's Fellowship and an NIH Genetics Training Grant. H.E.V. is an American Cancer Society Research Professor.

References

Ambrose,C., Blasquez,V. and Bina,M. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 3287–3291.

Ambrose,C., Rajadhyaksha,A., Lowman,H. and Bina,M. (1989) *J. Mol. Biol.*, **209**, 255–263.

Ambrose,C., Lowman,H., Rajadhyaksha,A., Blasquez,V. and Bina,M. (1990) *J. Mol. Biol.*, **214**, 875–884.

Benton,W.D. and Davis,R.W. (1977) *Science*, **196**, 180–182.

Bonne-Andrea,C., Wong,M.L. and Alberts,B.M. (1990) *Nature*, **343**, 719–726.

Bowerman,B., Brown,P.O., Bishop,J.M. and Varmus,H.E. (1989) *Genes Dev.*, **3**, 469–478.

Boyce,F.M., Sundin,O., Barsoum,J. and Varshavsky,A. (1982) *J. Virol.*, **42**, 292–296.

Brown,P.O., Bowerman,B., Varmus,H.E. and Bishop,J.M. (1987) *Cell*, **49**, 347–356.

Brown,P.O., Bowerman,B., Varmus,H.E. and Bishop,J.M. (1989). *Proc. Natl. Acad. Sci. USA*, **86**, 2525–2529.

Chalker,D.L. and Sandmeyer,S.B. (1990) *Genetics*, **126**, 837–850.

Church,G.M. and Gilbert,W. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 1991–1995.

Cochran,W.G. (1954) *Biometrics*, **10**, 417–451.

Craigie,R., Fujiwara,T. and Bushman,F. (1990) *Cell*, **62**, 829–837.

Dean,A., Pederson,D.S. and Simpson,R.T. (1989) *Methods Enzymol.*, **170**, 26–41.

Drew,H.R. (1984) *J. Mol. Biol.*, **176**, 535–557.

Drew,H.R. and Travers,A.A. (1985) *J. Mol. Biol.*, **186**, 773–790.

Eichinger,D. and Boeke,J.D. (1988) *Cell*, **54**, 955–966.

Feinberg,A.P. and Vogelstein,B. (1983) *Anal. Biochem.*, **132**, 6–13.

Gross,D.S. and Garrard,W.T. (1988) *Annu. Rev. Biochem.*, **57**, 159–197.

Grunstein,M. (1990) *Annu. Rev. Cell Biol.*, **6**, 643–678.

Hoopes,B.C. and McClure,W.R. (1981) *Nucleic Acids Res.*, **9**, 5493–5505.

Jones,E.W. (1991) *Methods Enzymol.*, **194**, 428–453.

Katz,R.A., Merkel,G., Kulkosky,J., Leis,J. and Skalka,A.M. (1990) *Cell*, **63**, 87–95.

King,W., Patel,M.D., Lobel,L.I., Goff,S.P. and Nguyen-Huu,M.C. (1985) *Science*, **228**, 554–558.

Lee,Y.M.H. and Coffin,J.M. (1990) *J. Virol.*, **64**, 5958–5965.

Linxweiler,W. and Horz,W. (1985) *Cell*, **42**, 281–290.

Lorch,Y., LaPointe,J.W. and Kornberg,R.D. (1987) *Cell*, **49**, 203–210.

Luchnik,A.N., Bakayev,V.V., Zbarsky,I.B. and Georgiev,G.P. (1982) *EMBO J.*, **1**, 1353–1358.

Lutter,L.C. (1978) *J. Mol. Biol.*, **124**, 391–420.

Marck,C. (1988) *Nucleic Acids Res.*, **16**, 1829–1836.

Mooslehner,K., Karls,U. and Harbers,K. (1990) *J. Virol.*, **64**, 3056–3058.

Natsoulis,G., Thomas,N., Roghmann,M.-C., Winston,F. and Boeke,J.D. (1989) *Genetics*, **123**, 269–279.

Normanly,J., Masson,J.-M., Kleina,L.G., Abelson,J. and Miller,J.H. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 6548–6552.

Oudet,P., Weiss,E. and Regnier,E. (1989) *Methods Enzymol.*, **170**, 14–25.

Pederson,D.S., Thoma,F. and Simpson,R.T. (1986a) *Annu. Rev. Cell Biol.*, **2**, 117–147.

Pederson,D.S., Venkatesan,M., Thoma,F. and Simpson,R.T. (1986b) *Proc. Natl. Acad. Sci. USA*, **83**, 7206–7210.

Perlmann,T. and Wrange,O. (1988) *EMBO J.*, **7**, 3073–3079.

Pina,B., Bruggemeier,U. and Beato,M. (1990) *Cell*, **60**, 719–731.

Raines,M.A., Lewis,W.G., Crittenden,L.B. and Kung,H.-J. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 2287–2291.

Reddy,S., DeGregori,J.V., von Melchner,H. and Ruley,H.E. (1991) *J. Virol.*, **65**, 1507–1515.

Rohdewohld,H., Weiher,H., Reik,W., Jaenisch,R. and Breindl,M. (1987) *J. Virol.*, **61**, 336–343.

Sandmeyer,S.B., Hansen,L.J. and Chalker,D.L. (1990) *Annu. Rev. Genet.*, **24**, 491–518.

Saragosti,S., Moyne,G. and Yaniv,M. (1980) *Cell*, **20**, 65–73.

Satchwell,S.C., Drew,H.R. and Travers,A.A. (1986) *J. Mol. Biol.*, **191**, 659–675.

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

Schwartzberg,P., Colicelli,J. and Goff,S.P. (1983) *J. Virol.*, **46**, 538–546.

Selten,G., Cuypers,H.T., Zijlstra,M., Melief,C. and Berns,A. (1984) *EMBO J.*, **3**, 3215–3222.

Shelton,E.R., Wassarman,P.M. and DePamphilis,M.L. (1980) *J. Biol. Chem.*, **255**, 771–782.

Shih,C.-K., Linial,M., Goodenow,M.M. and Hayward,W.S. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 4697–4701.

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

Shimotohno,K. and Temin,H.M. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 7357–7361.

Shoemaker,C.S., Goff,S., Gilboa,E., Paskind,M., Mitra,S.W. and Baltimore,D. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 3932–3936.

Shoemaker,C., Hoffmann,J., Goff,S.P. and Baltimore,D. (1981) *J. Virol.*, **40**, 164–172.

Shrader,T.E. and Crothers,D.M. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 7418–7422.

Simpson,R.T. (1990) *Nature*, **343**, 387–389.

Simpson,R.T. and Stafford,D.W. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 51–55.

Sogo,J.M., Stahl,H., Koller,Th. and Knippers,R. (1986) *J. Mol. Biol.*, **189**, 189–204.

Tanford,C. (1961) *Physical Chemistry of Macromolecules*. Wiley, New York.
Thoma,F. and Simpson,R.T. (1985) *Nature*, **315**, 250–252.
Thoma,F., Bergman,L.W. and Simpson,R.T. (1984) *J. Mol. Biol.*, **177**, 715–733.
Varmus,H.E. and Brown,P.O. (1989) In Berg,D.E. and Howe,M.M. (eds), *Mobile DNA*. American Society of Microbiology, Washington, DC pp. 53–108.
Varshavsky,A.J., Sundin,O.H. and Bohn,M.J. (1978) *Nucleic Acids Res.*, **5**, 3469–3478.
Vijaya,S., Steffen,D.L. and Robinson,H.L. (1986) *J. Virol.*, **60**, 683–692.
Yamamoto,K.R., Alberts,B.M., Benzinger,R., Lawhorne,L. and Treiber,G. (1970) *Virology*, **40**,734–744.
Zar,J.H. (1984) *Biostatistical Analysis*. Prentice-Hall, Inc., New Jersey.

Received on August 27, 1991; revised on October 7, 1991