The regional integration of retroviral sequences into the mosaic genomes of mammals

Alla V. Rynditch a,b, Serguei Zoubak a,b, Ludmilla Tsyba a,b, Nathaly Tryapitsina-Guley a,b, Giorgio Bernardi a,*

a Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France
b Institute of Molecular Biology and Genetics, Ukrainian Academy of Sciences, 150 Zabolotnogo str., 252627 Kiev, Ukraine

Received 27 July 1998; accepted 18 August 1998; Received by J. Svaboda

Abstract

We have reviewed here three sets of data concerning the integration of retroviral sequences in the mammalian genome: (i) our experimental localization of a number of proviruses integrated in isochores characterized by different GC levels; (ii) results from other laboratories on the localization of retroviral sequences in open chromatin regions and/or next to CpG islands; and (iii) our compositional analysis of genes located in the neighborhood of integrated retroviral sequences. The three sets of data have provided a very consistent picture in that a compartmentalized, isopycnic integration of expressed proviruses appears to be the rule (‘isopycnic’ refers to the compositional match between viral and host sequences around the integration site).

The results reviewed here suggest that: (i) integration of proviral sequences is targeted initially towards ‘open chromatin regions’; while these exist in both GC-rich and GC-poor isochores, the ‘open chromatin regions’ of GC-rich isochores are the main targets for integration of retroviral sequences because of their much greater abundance; (ii) isopycnicity is associated with stability of integration; indeed, even non-expressed integrated retroviral sequences tend to show an isopycnic localization in the genome; (iii) transcription of integrated viral sequences (like transcription of host genes) appears to be associated, as a rule, with an isopycnic localization, as indicated by transcribed sequences that show an isopycnic integration and act in trans; (iv) selection plays a role in the choice of specific sites within an isopycnic region; in exceptional cases [such as mouse mammary tumor virus (MMTV) activating GC-rich oncogenes], selection may override isopycnicity. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Isochores; Retroviruses; Transcription

1. Introduction

The stability of transformation caused by Rous sarcoma virus (RSV), observed by Temin (1960, 1962, 1964), and the persistence in rat tumor XC cells of non-infectious but rescuable RSV sequences that were transmitted to the progeny (Svoboda et al., 1963), led to the prediction of integration of retroviral sequences into the host genome. In turn, the sensitivity of retroviral replication to inhibitors of DNA synthesis led to the provirus hypothesis, which postulated the formation of a DNA copy of retroviral RNA and its stable integration into the cell’s genome. The later discovery of reverse transcription (Baltimore, 1970; Temin and Mizutani, 1970) was followed by the demonstration that the integration of viral genomes as proviruses into the genomes of host cells is a critical step in the life-cycle of retroviruses (Temin, 1976; Varmus, 1984). Indeed, the replication of retroviruses is dependent upon integration (see Goff, 1992; Brown, 1997, for reviews), and several pathogenic effects of retroviruses are associated with their integration.

In fact, integration may inactivate host cell genes by disruption, or activate them by the action of viral promoters and enhancers (reviewed in Kung et al., 1991; Tsichlis and Lazo, 1991; Athas et al., 1994; Jankers and
Berns, 1996) and may be associated with the modification of cellular genes and with the formation of new transforming viruses by acquisition of cellular gene sequences (reviewed in Bishop, 1980; Neil, 1983; Hughes, 1983). These events may lead to cell transformation and cancer development. It is obvious, therefore, that retroviral integration is a very important phenomenon. Retroviral integration is, in fact, just a special case of integration of foreign DNA into the genome of host cells, and the understanding of this case has a crucial importance in gene therapy, especially when retroviruses are used as vectors (reviewed in Bushman, 1995; Hodgson, 1996; Anderson, 1998).

Integration of retroviral genomes is a site-specific process as far as the viral sequences are concerned, since long terminal repeat (LTR) sequences are involved (reviewed in Varumus and Brown, 1989; Grandgenett and Mumm, 1990). As for the host genome, local features of integration loci (macronuclear versus non-macronuclear DNA, bent versus unbent DNA, etc.) have been investigated, but it is not clear to what extent local effects can account for the pattern of integration over the whole genome (Brown, 1997). On the other hand, it has been concluded that 'the distribution of integration sites across the whole genome remains to be adequately characterized' (Brown, 1997), thus implicitly questioning: (i) 'the anecdotal data' suggesting that some genes are used more (Frankel et al., 1985; King et al., 1985) or less (Hubbard et al., 1994) frequently as integration targets than expected from a uniform distribution; (ii) 'the analysis of a small sample' of sequenced HIV-1 target sites revealing a high frequency of highly repetitive elements in close proximity (Stevens and Griffith, 1994, 1996); and (iii) the reports of high frequencies of RSV integration sites near DNase-hypersensitive sites, transcriptionally active regions and CpG islands (Vijaya et al., 1986; Rohdewohld et al., 1987; Mooslehen et al., 1990; Scherdin et al., 1990). However, as far as integration of retroviral sequences into the host genomes is concerned, a substantial body of evidence exists that comes from a completely different approach developed in our laboratory. This approach has led to a novel, very consistent picture of retroviral integration that will be the main topic of the present review.

Indeed, the development of methods for the compositional fractionation of DNA at high resolution (Corneo et al., 1968) led not only to the discovery that mammalian (and avian) genomes are mosaics of isochore families, namely of long (> 300 kb), compositionally homogeneous DNA segments, belonging in a small number of families characterized by different base composition (Filipski et al., 1973; Thiery et al., 1976; Macaya et al., 1976; see also the following section), but also to the possibility of localizing any nucleotide sequence (obviously including proviral sequences) for which an appropriate probe is available in compositional DNA fractions. When first applied (Kettmann et al., 1979), this 'compositional approach' showed that retroviral integration is not random, as shown by the demonstration of a regional specificity of bovine leukemia virus (BLV) integration. More precisely, the same approach showed that viral integration is compartmentalized, because it occurs only in some regions of the host genome, and is isopycnic, because it takes place in host genome sequences which show a compositional match with the viral sequences. This demonstration was subsequently extended to other retroviral systems, and was confirmed by other approaches that will be described later.

2. The sequence organization of the host genomes

As already mentioned, the genomes of warm-blooded vertebrates are mosaics of isochore families (Fig. 1), long DNA segments, originally estimated to be larger than 300 kb (Macaya et al., 1976) and now known to range from 200 to over 6000 kb (De Sanzo et al., 1996, 1997). Isochores are fairly homogeneous in base composition (above a 3 kb level) and belong to a few families characterized by different GC levels which cover an extended spectrum of base composition (reviewed in Bernardi, 1995). The large DNA fragments of standard preparations (approx. 100 kb in size) derive from random physical and enzymatic breakage occurring during extraction and purification and reflect the base composition of the isochores from which they originate. DNA fragments can be fractionated by preparative equilibrium centrifugation in CsCl density gradients in the presence of sequence-specific DNA ligands [Ag(I)- and BAMD were used; BAMD is bis(acetatomercuри-methyl)dioxane], and the GC level of the fractions can be assessed by analytical centrifugation in CsCl.

The compositional distribution of large DNA fragments from the human genome, which is a good representative of the majority of mammalian genomes (Sabour et al., 1993), is characterized by the presence of two families of fragments derived from the corresponding isochores. In the human genome (Fig. 1), GC-poor isochore families L1 and L2 (collectively indicated as L) represent about 65% of the genome, GC-rich H1 and H2 families correspond to about 24% and 7.5% of the genome, respectively, and the very GC-rich H3 family forms almost 5% of the genome, the remaining DNA consisting of satellite and ribosomal DNAs (Zoubak et al., 1996).

The compositional distributions of isochore families (Fig. 1), of coding sequences (and of their third codon positions; see Fig. 3) and of introns represent 'genome phenotypes' (Bernardi and Bernardi, 1986), which are different not only between cold-blooded and warm-blooded vertebrates (the former showing a narrower compositional
in such fractions and exploring chromosomal regions up to twice the size of the DNA fragments making up the DNA samples. Indeed, since the population of DNA molecules is the result of random degradation, the sequences that are probed may be located at different positions on the DNA molecules. As the DNA fragments under consideration are approx. 100 kb in size, hybridization explores regions up to 200 kb around the probed sequence.

Recently, a rapid procedure for the compositional fractionation of the genome of warm-blooded vertebrates has been developed which uses preparative CsCl gradients with an extremely shallow slope, as obtained with vertical rotors (De Sario et al., 1995). In this approach, cells are lysed directly in the CsCl solution, and the sequence of interest is detected by hybridization with an appropriate probe. The modal buoyant density of the DNA molecules hybridizing the probe, as determined by comparison with buoyant density DNA markers, is used to estimate the GC level of the sequence under consideration. This approach lends itself to the analysis of minute amounts of DNA from infected cells and has been applied recently to investigating viral integration in the genomes of individual patients.

Gene localization revealed that GC-rich and GC-poor coding sequences are located in GC-rich and GC-poor isochores, respectively, and that a linear compositional correlation exists (Fig. 2) between GC (and GC•) levels of coding sequences and GC levels of the isochores carrying them (Bernardi et al., 1985; Assani et al., 1991; Clay et al., 1996). This correlation allowed an analysis of the distribution of genes (Fig. 3), which was shown spectrum extending much less towards high GC values), but also between mammals and birds (the latter reaching higher GC levels compared to the former), and even among mammals, in which the GC-richest isochores of some sub-orders (like Myomorpha) do not attain the same high GC levels as those of other mammals (reviewed in Bernardi, 1995).

Hybridization of appropriate probes on compositional fractions of genomic DNA allows localizing sequences...
Transcription is clustered and very active in GC-rich regions in which GC-rich genes are very close to each other, are associated with CpG islands and comprise most or all of housekeeping genes (which are constitutively expressed), whereas it is much more scattered in GC-poor regions where GC-poor genes are dispersed over wide expanses of intergenic sequences, are mostly associated with TATA box promoters and are largely expressed in a tissue-specific manner. Finally, recombination is very active in the former regions and almost absent in the latter ones. Indeed, the most transcriptionally active compartments are also the most recombinogenic, since they are located in H3+ and H3* chromosomal bands, where most recombination events are known to occur.

3. The localization of retroviral integrated sequences in the host genome

The localization of integrated proviruses in the isochore regions of the host genome can be approached in exactly the same way as the localization of genes described in the preceding section. In fact, hybridization experiments localize provirus-carrying fragments in compositional DNA fractions, thus defining the GC levels of those fragments. Restriction enzymes that do not cut proviral sequences are typically used, so that the number of different-size bands reflects the number of proviral copies. Hybridization patterns are generally characterized by different distributions of proviruses in the compositional fractions, different spreads of same-size bands and different intensities of the bands. Usually, gene concentration parallels GC levels, being very low in GC-poor isochores and increasingly higher in GC-rich isochores. Investigations on the chromosomal localization of human isochores have shown that the GC-richest (and gene-richest) isochores of the H3 family are concentrated in two sub-sets of R(verse) bands, the 25 H3+ bands and the 31 H3* bands (also called T and T∞ bands, respectively), the former comprising larger amounts of H3 isochores compared to the latter; another much larger sub-set of 140 R bands, the H3− bands, do not contain detectable amounts of H3 isochores (Saccone et al., 1992, 1993, 1996; this analysis was done at a resolution of 400 bands).

Isochores showing different gene concentration correspond to different chromatin structures, which are ‘open’ in the GC-rich, gene-rich isochores, and ‘closed’ in the GC-poor, gene-poor isochores. The open chromatin structures are characterized by accessibility to DNases (Kerem et al., 1984), as well as by a larger spacing of nucleosomes, absence of histone H1 and acetylation of histones H3 and H4 (Tazi and Bird, 1990). Replication is late in the cell cycle in GC-poor isochores and early in GC-rich isochores, the replication timing order being H3+, H3* and H3− bands (Federico et al., 1998).
mouse mammary tumor virus (MMTV)), that also do not contain oncogenes. The second, major, class includes all oncoviruses containing oncogenes (like RSV) and some oncoviruses that do not contain oncogenes [like BLV and human T-cell leukemia virus (HTLV-I)], except for those of B type and for some of the D type. The striking bimodal compositional distribution of retroviral genomes is accompanied by a remarkable compositional homogeneity of genes (gag, pol and env genes, as well as oncogenes and genes for regulatory proteins) within each retroviral genome. Even LTRs are GC-rich or GC-poor according to the GC level of the viral sequences.

The compositional localization of integrated sequences from four retroviruses was investigated for three GC-rich viruses, BLV (Kettmann et al., 1979, 1980), HTLV-I (Zoubak et al., 1994), and RSV (Rynditch et al., 1991); and for one GC-poor virus, MMTV (Salinas et al., 1987). Some common features of proviral integration have been observed, such as compartmentalization, isopycnicity, and correlation between transcription of the proviral sequences and its localization in the host genome. These features will be illustrated by describing the results obtained with different viral systems in this and in the following section.

3.1. BLV

A compartmentalized integration of BLV proviral sequences (Fig. 6) was demonstrated in the genomes of infected cells obtained from three animals, one with persistent lymphocytosis and two in the tumor stage. Proviral copies were found at a large number of genomic sites in the leukocytes of the animals with persistent lymphocytosis, and at one or very few sites in the genomes of leukocytes or tumor cells of animals in the tumor stage of the disease. In all six samples investigated (from leukocytes, lymph-node tumors and spleen), BLV sequences were integrated only in DNA fragments present in GC-rich compositional fractions (49% GC), which represent much less than 10% of the bovine genome, if one neglects the presence of several satellite DNAs (that make up 25% of the bovine genome) and of rDNA in the same buoyant density range (Filipski et al., 1973; Macaya et al., 1978; Meunier-Rotival et al., 1979). As BLV has a GC-rich genome (54% GC), such integration into the GC-rich compartments of the host genome was defined as 'isopycnic', meaning that integration is found in isochores approximately matching the viral genome in base composition.

3.2. RSV

In this case, retroviral integration was investigated in the genomes of six well-characterized clones of hamster
in DNA fractions having GC levels equal to 50–53% (Ryndich et al., 1991). The direct sequencing of extended DNA stretches flanking the provirus in one such cell line, H-19, confirmed its integration in GC-rich DNA regions (Machon et al., 1996). In the case of cell lines H9 and H42 comprising both complete and defective proviral copies, complete copies were present in fractions centered at about 50% GC, whereas defective copies were found in these fractions as well as in GC-poorer ones.

### 3.3. HTLV-I

The localization of this GC-rich provirus (54% GC) was investigated (Fig. 7) in five immortalized cell lines, containing a total of 22 integrated complete or defective proviruses, and in seven T-cell clones obtained from patients with TSP/HAM (tropical spastic paraparesis/HTLV-I-associated myelopathy), in which case each clone comprised one to three integrated proviral sequences for a total of 18 proviruses (Zoubak et al.,...
In all cases, HTLV-I sequences were found in the GC-rich 60% of the human genome. More precisely, 40 proviruses were found in the 39–54% GC range of the human genome, whereas no HTLV-I sequences were found in isochores lower than 39% GC, which represent 40% of the human genome.

3.4 MMTV

GC-poor MMTV proviral sequences (43.3% GC) were localized in the DNAs from the livers of five inbred strains of mice and from GR cells derived from primary implants of mammary tumors (Salinas et al., 1986). 12 sets of endogenous MMTV sequences (corresponding to 27 proviruses) were characterized by restriction patterns and chromosomal localizations; exogenous sequences were also localized and produced strong hybridization signals (the total number was 15), or weak hybridization signals corresponding to proviral integrations in a minor cell population (the total number was six). While endogenous sequences were present in the GC-poor 43% of the host genome, exogenous sequences showed a broader distribution, being present in the GC-poor 60% of the genome (Fig. 6), therefore within the L isochores (see Fig. 1). Interestingly, endogenous sequences were centered at 40% GC, whereas exogenous sequences showed a broader distribution of sequences centered at 43% GC.

In conclusion, four retroviruses, three of C-type (BLV, RSV, HTLV-I) and one of D-type (MMTV), were localized in four mammalian species. The data show that integrated retroviral sequences are not spread at random in the host genomes, but are found only in some host genome compartments matching the viral sequences in composition. This conclusion is stressed by a presentation of the results in terms of proviral density in host DNA (Fig. 8). Incidentally, in none of the cases investigated were the proviral sequences found in satellite or ribosomal DNA.

Similar results were obtained for the integration of GC-rich sequences human hepatitis B virus (HBV), a DNA virus (Zerial et al., 1986). Eight out of the nine viral sequences integrated in the hepatoma Alexander cell line were localized in DNA segments having a GC level of 51% (Fig. 7). Likewise, integration of a GC-rich recombinant SV40-adenovirus 5 was observed in the H3 isochore family (Romani et al., 1993).

4. Correlation between the isochore localization of integrated retroviral sequences and their transcription

The results on the compositional distribution of HBV sequences (Zerial et al., 1986) provided the first information about a possible correlation between the isopycnic localization of integrated viral sequences in compositional compartments of the host genome and their transcription. Indeed, the eight HBV sequences integrated in the GC-richest regions of the genome of the Alexander cell line were expressed, whereas the single one located in a GC-poorer region of the genome was not (see Fig. 7).

Likewise, all RSV transcribed sequences were localized in the GC-richest isochores (Fig. 6), whereas the proviral sequences whose expression was not detected were distributed in isochores with lower GC content (Ryndich et al., 1991).

A more detailed analysis was possible on integrated HTLV-I sequences, in which case 40 sequences from 12
hypersensitive sites correlate with gene expression and insertion of MuLV leading to clonal expansion has been base pairs of a DNaseI hypersensitive chromatin region nisms of activation of the same oncogene (Ben-David

1-collagen gene has been localized near a DNaseI hyper- site are rare (reviewed in Rosenberg and Jolicoeur, 1994), although they may also depend upon the observation has been confirmed by the analysis of et al., 1994). Interestingly, the other four unselected compositional analysis of genes located in the vicinity 

and Wyke, 1991). In bursal lymphomas, avian leukemia Indeed, the data on insertional mutagenesis which impli-

which information about the GC level of several genes 

5. Other lines of evidence for the regional integration of retroviral sequences

5.1. An analysis of integration sites in 'open' chromatin and/or near CpG islands

Several observations suggest that retroviral integra-

tion is correlated with the level of transcription, recom-

bination, and with the degree of DNA accessibility in chromatin (Vijaya et al., 1986; Rohdewohld et al., 1987; Scher din et al., 1990; Mooslechner et al., 1990; Fincham and Wyke, 1991). In bursal lymphomas, avian leukemia virus (ALV) integrants were found near each of five major DNasel hypersensitive sites which are immedi-

ately upstream of the coding sequences for c-myc (Robinson and Gagnon, 1986; Schubach and Groudine, 1984). The integration of MoMuLV provirus in the α-

1-collagen gene has been localized near a DNasel hyper-

sensitive site (Brendl et al., 1984). Likewise, MoMuLV integrations in seven regions of chromosomal DNA were mapped near DNase hypersensitive sites (Vijaya et al., 1986). Three of these regions, containing c-erbB, c-myc and dbv-1, were targets for multiple tumor-induc-

ing integrations. Interestingly, the other four unselected integrations also occurred near DNasel hypersensitive sites, suggesting that retroviral sequences preferentially integrate near hypersensitive sites in all cases. This observation has been confirmed by the analysis of chromatin structure of MoMuLV proviral integration sites in early embryonic cells and in differentiated fibroblasts, where integration occurs within a few hundred base pairs of a DNasel hypersensitive chromatin region (Rohdewohld et al., 1987). It is well known that DNasel hypersensitive sites correlate with gene expression and are located preferentially in regions with an open chro-

matin structure (Edmondson and Roth, 1996).

These results prompted the analysis of transcription of cellular sequences flanking the integrated proviruses. In three out of five randomly chosen mouse strains harboring one copy of MoMuLV in their germ lines, the provirus was integrated in the vicinity of DNA regions transcribed in an embryonal stem cell line and in an embryonal carcinoma cell line (Mooslechner et al., 1990). Among nine sequences randomly chosen for MoMuLV integration in NIH3T3 mouse fibroblasts, at least six were in transcriptionally active regions and/or contained CpG-rich islands (Scher din et al., 1990). Finally, the proviral integration sites of transcribed and non-transcribed RSV in rat DNA were examined (Fincham and Wyke, 1991). In this case, the transcribed sequences showed a tendency to integrate close to the 3’ ends of CpG islands, whereas non-transcribed sequences did not.

5.2. An analysis of integration sites near sequenced genes

An analysis of integration sites of proviruses can also be done on the basis of the composition of host cell genes located in the neighborhood of integrated viral sequences, since host cell genes are ‘isochore markers’, in that their composition is correlated with that of the isochores in which they are embedded (Bernardi et al., 1985; Clay et al., 1996; see Fig. 2). This compositional analysis (S. Zoubak, A. Rynditch, G. Bernardi, unpublished results) was done only for viral sequences for which information about the GC level of several genes in the neighborhood of integration sites was available. Indeed, the data on insertional mutagenesis which impli-

cate viral integration in the vicinity of genes important for cell growth and/or differentiation should be consid-

ered only after elucidating the possible rearrangements taking place after integration, even if deletions and rearrangements of cellular sequences near the integration site are rare (reviewed in Rosenberg and Joliceour, 1997).

ALV or murine leukemia viruses (MuLV), which induce hematopoietic tumors by complex, indirect mech-

anisms generally involving insertional activation of cellular protooncogenes, are so far the best sources for compositional analysis of genes located in the vicinity of integration sites. In the case of MuLV, different isolates induce tumors with a large variety of phenotypes which are often dependent upon viral enhancers (Ott et al., 1994), although they may also depend upon the coding sequences of the particular virus used for induc-

tion (Mukhopadhyaya and Wolff, 1992; Nazarov and Wolff, 1995), or may be correlated with different mecha-

nisms of activation of the same oncogene (Ben-David et al., 1991, 1992; Bergeron et al., 1993). An oncogenic insertion of MuLV leading to clonal expansion has been
observed in different chromosomes, but often common integration sites were mapped close to each other on the same chromosome (Lazo et al., 1990; Bartholomew and Ihle, 1991; Lammi et al., 1992; Jiang et al., 1994) and were sometimes associated with the activation of the same oncogene, like c-myb in many cases of promonocyte leukemia (Shen-Ong and Wolff, 1987; Wolff et al., 1991). c-myb in T-cell or B-cell lymphomas (Tschilits and Lazo, 1991), or Evi-1 in myeloid neoplasms (Bordereaux et al., 1987; Mushinski et al., 1988; Morishita et al., 1988).

Among some 50 cases of insertional mutagenesis by MuLV, 28 could be analyzed compositionally taking into account the information about the localization of the proviruses and of neighboring genes. The compositional analysis of host genes located close to MuLV proviruses show that they are distributed in a GC range centered around 55% GC (Fig. 9), a value close to the GC content of MuLV (53%). This means that in the case of insertional mutagenesis, it is possible to observe an isopycnic localization of integrated proviral sequences. The relatively wide compositional range of integration sites exhibited by a minority of proviruses may be due to the fact that the 'reporter' genes may be located in a contiguous isochore characterized by a different GC level. For example, an interesting exceptional integration of MuLV enhances the expression of c-Ki-ras which is 37% GC. The latter is located in a genomic region where another gene, 

\[ \text{vmy} \]  

(57% GC), is a target of MuLV insertional activation (Tremblay et al., 1992; Hanna et al., 1993). As 

\[ \text{vin} \]  

1 is known to be present between 

\[ \text{tsf} \]  

(55.8% GC) and c-Ki-ras (37% GC); Hanna et al., 1993), c-Ki-ras might be located just at the chromosomal border where integration of MuLV occurred and undergo promoter activation as an onco-gene (Hanna et al., 1993).

Similar observations basically hold for ALV (53% GC). Different ALVs, such as Rous-associated virus-1 (RAV-1), chicken syncytial virus (CSN) and reticuloendotheliosis virus (REV), as well as ring-necked pheasant virus (RPV), induce B-cell lymphomas by activating c-myb (63% GC) (Hayward et al., 1981; Payne et al., 1982; Pachl et al., 1983; Noort-Daoloi et al., 1981; Boerkoel and Kung, 1992; Westaway et al., 1984; Simon et al., 1984; Isfort et al., 1987). Among ALV viruses, the same viral sequences can induce different neoplasias through activation of different oncogenes, such as RAV-1 inducing nephroblastoma, B-cell lymphoma, or erythroblastosis by activating c-fos (70.6% GC), c-myb (51% GC), c-erbB (Fung et al., 1983; Nielsen et al., 1985; Hayward et al., 1981; Payne et al., 1982; Pizer and Humphries, 1989; Kanter et al., 1988; Collart et al., 1990; Kabrun et al., 1990). Other ALV representatives integrate also into c-Heras (53% GC; Westaway et al., 1986) and c-ret (46.7% GC; Kabrun et al., 1990).

As far as MMTV integration is concerned, genes critical in regulating the development of mouse embryo have been shown to be activated by MMTV in multiple independent mouse tumors: Int-1/Wnt-1 (65.5% GC; Nusse and Varmus, 1982; Nusse et al., 1990; Van Ooyen and Nusse, 1984; Clauss et al., 1993); Int-2/Fgf3 (63% GC; Peters et al., 1983, 1989b; Dickson et al., 1984, 1990; Clauss et al., 1993; Morris and Dutra, 1997); Int-3/Nothd (65% GC; Gallaham and Callahan, 1987; Gallaham et al., 1987; Robbins et al., 1992; Callahan, 1998); Int-4/Wnt-3 (60.9% GC; Roelink et al., 1990); Fgf4/Hstl (64.5% GC; Peters et al., 1989a; Shackleford et al., 1993); Fgf-8 (65.3% GC; MacArthur et al., 1995); Wnt-10b (Lee et al., 1995); Int-5/Int5 (42% GC; Gray et al., 1986; Durgam and Tekmal, 1994) and Int-6 (41% GC; Marchetti et al., 1995). Most insertional activations occur upstream or downstream of these genes.
genes, sometimes at distances larger than 20 kb (Peters et al., 1989a; Morris et al., 1990), and lead to the deregulation of their transcription. Insertional activations are also observed within the 3' untranslated region (Dickson et al., 1990), within introns, and even within coding regions (Marchetti et al., 1995; Miyazaki et al., 1995; Robbins et al., 1992; Deilla et al., 1997). Since, except for two of them, the listed genes are GC-rich, they should be located in GC-rich isochores.

The isopycnic location of transcribed integrated sequences of human hepatitis B virus found in the Alexander cell line (Zerial et al., 1986) is confirmed by the fact that, in other cases, the integration sites of HBV have been observed near or inside GC-rich genes, like the retinoic acid receptor gene (Dejean et al., 1986; de Thé et al., 1987), the cyclin A gene (Wang et al., 1990), the erbB gene (Zhang et al., 1992) and the mevalonate kinase gene (Graef et al., 1994). The insertional activation of the GC-rich myc family genes by another representative of GC-rich hepadnaviridae, woodchuck hepatitis virus (WHV), also suggests an isopycnic integration, both the proviral sequence and the myc genes being GC-rich (Fourel et al., 1990, 1994).

6. Conclusions

An ample body of data (see Bernardi, 1995, for a review) has conclusively demonstrated that the genomes of mammals (and, more generally, of warm-blooded vertebrates) are not only structural mosaics, made up of compositional compartments, the isochores, but also are functional mosaics. Indeed, it should be recalled (i) that gene concentration is strikingly non-uniform, being very low in the GC-poor isochores and increasingly higher in increasingly GC-rich isochores; (ii) that the gene-rich, GC-rich isochores, which represent a small minority of the genome, are more often and more widely transcribed (because of the high concentration of housekeeping genes in those regions) compared to the gene-poor, GC-poor isochores, which represent the vast majority of the genome and mainly contain tissue-specific genes, and (iii) that the genes of GC-rich isochores are very frequently associated with CpG islands (Aïssani and Bernardi, 1991a,b) which contain Sp1 binding sites, preventing methylation and gene silencing (Svoboda, 1998); more generally, the gene-rich, GC-rich isochores have a lower relative methylation than the gene-poor, GC-poor isochores (Jabbari and Bernardi, 1998). Schematically, one could say that the former can be visualized as a dense array of chromatin boxes that tend to be permanently open and the second as a sparse array of chromatin boxes that are open only occasionally (Fig 10). It should be stressed that open chromatin regions are certainly much more frequent in the scarce GC-rich, gene-rich isochores than in the abundant GC-poor, gene-poor isochores. Indeed, the latter, which represent 63% of the genome, comprise only about 22% of human genes (Zoubak et al., 1990) and, as already mentioned, contain more tissue-specific and developmentally regulated genes than the former. Since only a very small percentage of GC-poor genes are active at any given time and in any given time, open chromatin regions of GC-poor isochores correspond to a much smaller part of the genome than those of GC-rich isochores. Given this organization of the mammalian genome, the question concerning the distribution of integrated viral sequences in the mosaic genome of the host is a most relevant one.

The data reviewed here belong in three major groups: (i) our experimental localization of a number of retroviral sequences integrated in isochores belonging to different families; (ii) results from other laboratories concerning the localization of retroviral sequences in open chromatin regions and/or next to CpG islands; and (iii) our compositional analysis of genes located in the neighborhood of integrated retroviral sequences. The three sets of data have provided a very consistent picture in that a compartmentalized, isopycnic integration of proviruses in the mammalian genome appears to be the rule, a more scattered distribution being the exception. In particular, the data on the preferential integration of GC-rich viruses in the regions characterized by abundant CpG islands and/or by ‘open’ chromatin structure confirm the integration pattern found for GC-rich retroviruses based on experiments involving the compositional approach.

A key question then concerns the reasons for the compartmentalized, isopycnic integration of retroviral sequences. An obvious possibility is that this situation is the result of selection for certain integration sites, the activation of an oncogene providing a replicative advantage to the infected cell.

In order to investigate the effect of selection, two sets of experiments were carried out studying RSV integration under non-selective conditions. As described by Brown (1997), one set of experiments suggested that certain rare sites were used as integration targets at a frequency one-million-fold greater than the average site in the whole genome (Shih et al., 1988). Yet, a second, more recent, set of experiments, using a virtually identical experimental system, led to the conclusion that there was little variation across the genome in the frequency with which individual intervals were used as integration targets (Withers-Ward et al., 1994). Moreover, the previously reported putative ‘hot spots’ for RSV integration appeared to be no ‘hotter’ than average sites in the genome. In other words, no conclusion, one way or another, was reached by the experiments just mentioned.

We will now discuss the possible explanations for the
compartmentalized, isopycnic integration in the light of the results presented in this review.

If one considers the best known and most widespread case of GC-rich proviruses, like MuLV or ALV, acting in cis and the fact that oncogenes are mostly GC-rich (S. Zoubak, A. Rynditch, G. Bernardi, unpublished results), one might think that the compartmentalized, isopycnic integration is just the result of selection for integration sites conferring a replicative advantage to the cell through the activation of a nearby oncogene. According to this interpretation, the GC-richness of both the viral sequences and of the oncogenes would be purely coincidental. Some exceptions to the ‘isopycnicity rule’ seem to lend further support to an explanation based on ‘selection’. Indeed, in rare cases, the provirus is not in an isopycnic chromosomal environment, but is still next to an oncogene which is activated by it (an example being the activation of GC-rich genes by MMTV), suggesting that selection for a replicative advantage can override isopycnic integration. It should be noted, however, that the examples of non-isopycnic integration practically only concern a GC-poor provirus, MMTV, being integrated next to GC-rich genes. Indeed, the opposite case, namely that of GC-rich proviruses activating GC-poor oncogenes, has been found only once (in the doubtful case of MuLV activating c-ki-ras), in spite of the fact that the number of GC-rich proviruses explored was very much larger than that of GC-poor proviruses.

The point just made indicates that selection can only be, at best, a partial explanation for the compartmentalized, isopycnic integration, and that other explanations should be found. This conclusion is supported by three independent lines of evidence which lead to a different, broader picture of retroviral integration.

(1) As far as oncogene activation is concerned, the target of proviral integration, namely the regions flanking oncogenes, is so small (see the scheme of Fig. 10) that it is impossible that selection for those integration sites operates on an initially completely random integration pattern. In contrast, if genome regions accessible to viral integration are limited to those undergoing active transcription, this obviously greatly restricts the potential integration sites and strongly reduces the load imposed on selection. Moreover, one would also expect that, since open chromatin regions are much more frequent in GC-rich isochores, targeting would be primarily directed to them.

(2) A compartmentalized, isopycnic integration is also found for proviruses acting in trans (such as BLV and HTLV-I). This indicates that selection for a growth advantage may be associated with the expression of genes located far away from the integration site. Since this gene expression is dependent upon the transcription of a provirus and the production of trans-activating factors, the isopycnic location of the provirus indicates a dependence of its own transcription on a matching compositional environment. One could also say (Peckham et al., 1989) that preference for integration in transcriptionally active genome regions may reflect selection for proviral expression and for the pre-existing differences in transcription properties of the regions in which integration takes place. This situation is very understandable because it mimics that of host
genes, which are compositionally correlated with their genomic environment, this correlation being the result of an adaptive process that has taken place over evolutionary times.

(3) Non-transcribed and recently acquired viral sequences showed both an isopycnic integration and a broader distribution in the host genome. While the broader distribution indicates that, initially, a large number of potential integration sites may be explored, still a preference for an isopycnic integration is evident under non-selective conditions. This strongly suggests that stability of the integrated sequence in a compositionally matching environment plays an important role in isopycnic integration.

In conclusion, the results reviewed here suggest that integration of proviral sequences is initially targeted towards ‘open chromatin regions’ because of their much greater abundance, the ‘open chromatin regions’ of GC-rich isochores are the main targets for integration of retroviral sequences; as expected, no observation of integration of proviruses in heterochromatic regions, in which repeated sequences are characterized by a closed chromatin structure, was reported so far; (ii) that isopycnicity is associated with stability of integration; indeed, ‘open chromatin regions’ exist in both GC-rich and GC-poor isochores, yet even non-expressed integrated retroviral sequences tend to show an isopycnic localization in the genome; along the same line, the majority of interspersed repeated sequences, which are not expressed, show their highest densities in isopycnic isochores (Meunier-Rotival et al., 1982; Soriano et al., 1983; Zerial et al., 1986; Jabbari and Bernardi, 1998); (iii) that transcription of integrated viral sequences appears to be associated, as a rule, with an isopycnic localization, as very clearly indicated by transcribed integrated sequences that act in trans; (iv) that selection plays a role in the choice of specific sites within an isopycnic region; (v) that in exceptional cases (such as MMTV activating GC-rich oncogenes) selection may override isopycnicity.

These conclusions have some implications concerning both the compositional evolution of retroviral genome and gene therapy, namely (i) that the bimodal compositional distribution of retroviral sequences may be the result of their compositional coevolution with the sequences of the host genome, from some of which they originally derived; it is conceivable that the compositional transition, which led to the formation of GC-rich isochores of warm-blooded vertebrates (Bernardi, 1995), also led to a similar compositional transition (acting on their integrated forms) of what are now the GC-rich retroviral sequences (Zoubak et al., 1992); and (ii) that the reasons for the poor expression of retroviral constructs used in gene therapy (Anderson, 1998) should be investigated not only at the level of the promoters present in the constructs, but also at the level of the integration sites which are used by the constructs.

References


References

Hayward, W.S., Neel, B.G., Austen, S.M., 1981. AVL-induced lymphoid leukemia induction of a cellular onc gene by promoter inser-


shion and chromosomal rearrangement in a Moloney leukemia virus-


usual retrovirus-induced avian T-lymphoma resembling Marek’s dis-

ease: proviral insertion of 5' of exon one enhance the expression of the intron promoter. Oncogene Res. 2, 81–94.


Kabrun, N., Bumstead, N., Hayman, M.J., Enrietto, P.J., 1990. Char-


Kantor, M.R., Smith, R.E., Hayward, W.S., 1988. Rapid induction of B-


King, W., Pilet, M.D., Lobel, L.I., Goff, S.P., Young-Hsu, M.C., 1985. Insertion mutagenesis of embryonal carcinoma cells by retro-

virus Science 228, 554–556.

Kratzschwil, K., von der Mark, K., Kollare, E.J., Jamshe, R., Moos-


ne 3, 2381–2387.


tion within the Flk-2 locus results in inactivation of the erythroid trans-


Marchetti, A., Bartitta, F., Miyazaki, S., Callahan, D., Smith, C.H., Callahan, R., 1995. Int-6, a highly conserved, widely expressed gene, is mutated by mouse mammary tumor virus in mammary preneopla-


Marcus, S.L., Smith, S.W., Sarkar, N.H., 1981. Quantitative of murine mammary tumor virus-related RNA in mammary tissues of low-


Miyazaki, S., Keszak, C.A., Marchetti, A., Bartitta, F., Callahan, D., Callahan, R., 1995. The chromosomal location of the mouse mam-


Nazarov, V., Wolff, L., 1995. Novel integration sites at the distal 3’ end of the c-myc locus in retrovirus-primed promyelocytic leuke-

mias. J. Virol. 69, 3885–3888.


Nilken, T.W., Maroney, P.A., Goodwin, R.G., Rottmann, F.M., Crit-

insertion result in expression of an amino-truncated EGF receptor.
Cell 41, 719–726.


Olt, D.E., Keller, J., Reu, A., 1994. 10A1 MuLV induces a murine leukemia that expresses hematopoietic stem cell markers by a mecha-
nism that includes L1 integration. Virology 205, 563–568.


Peckham, J., Sobel, S., Comer, J., Jaenisch, R., Barklis, E., 1989. Retrovirus activation in embryonal carcinoma cells by cellular pro-


Peters, G., Brookes, S., Placzek, M., Schieremann, M., Michalides, R., Dickson, C., 1989b. A pituvative int domain for mouse mammary tumor virus on murine chromosome 7 is a 5’ extension of int-2.

Pizer, E., Humphries, E.H., 1989. RAV-1 insertional mutagenesis: dis-

Pizer, E., Humphries, E.H., 1989. RAV-1 insertional mutagenesis: dis-


Peckham, J., Sobel, S., Comer, J., Jaenisch, R., Barklis, E., 1989. Retrovirus activation in embryonal carcinoma cells by cellular pro-


Peters, G., Brookes, S., Placzek, M., Schieremann, M., Michalides, R., Dickson, C., 1989b. A putative int domain for mouse mammary tumor virus on murine chromosome 7 is a 5’ extension of int-2.

Pizer, E., Humphries, E.H., 1989. RAV-1 insertional mutagenesis: dis-

Pizer, E., Humphries, E.H., 1989. RAV-1 insertional mutagenesis: dis-

Pizer, E., Humphries, E.H., 1989. RAV-1 insertional mutagenesis: dis-

Pizer, E., Humphries, E.H., 1989. RAV-1 insertional mutagenesis: dis-

Pizer, E., Humphries, E.H., 1989. RAV-1 insertional mutagenesis: dis-

Pizer, E., Humphries, E.H., 1989. RAV-1 insertional mutagenesis: dis-

Pizer, E., Humphries, E.H., 1989. RAV-1 insertional mutagenesis: dis-

Pizer, E., Humphries, E.H., 1989. RAV-1 insertional mutagenesis: dis-

Pizer, E., Humphries, E.H., 1989. RAV-1 insertional mutagenesis: dis-

Pizer, E., Humphries, E.H., 1989. RAV-1 insertional mutagenesis: dis-

Pizer, E., Humphries, E.H., 1989. RAV-1 insertional mutagenesis: dis-

Pizer, E., Humphries, E.H., 1989. RAV-1 insertional mutagenesis: dis-
of the absence of infectious Rous sarcoma virus in rat tumor XC, whose structurally intact cells produce Rous sarcoma when transferred to chicks. Folia Biol. Prague 9, 77–81.


