REVIEW ARTICLE

Structure, Replication, and Recombination of Retrovirus Genomes: Some Unifying Hypotheses

By JOHN M. COFFIN

Department of Molecular Biology and Microbiology and Cancer Research Center,
Tufts University School of Medicine, 136 Harrison Avenue,
Boston, Massachusetts 02111, U.S.A.

INTRODUCTION

The study of the molecular biology of retroviruses has revealed a number of features not found in other groups of viruses. In recent years, a large number of excellent reviews of the field have appeared (Table 1). In writing this article, it is not my intention to cover ground already well treated in review form. Instead, I will consider only a limited number of interrelated subjects dealing with the organization of information in tumour virus genomes and its transfer from parent to progeny. In general, I will ignore primary data and present only the inferred structures and mechanisms and point out experimental approaches that might fill in gaps in our knowledge.

Retroviruses form a large and diverse group, with isolates from many different vertebrates. In most cases, our knowledge of the molecular biology of the viruses is derived from one model system, the avian sarcoma–leukosis viruses (ASV). This bias is simply because these are the longest known of the retroviruses and, in general, the easiest to deal with in the laboratory. Except in a few cases where murine leukaemia viruses (MuLV) give cleaner results, all of the examples I will use come from the ASV system. I believe that the general picture that emerges will extend to all retroviruses, although details will differ in different viruses.

By way of background, the following is a summary of some principal features of the molecular biology of retroviruses.

Genes. The retrovirus genome has a coding capacity for about $3 \times 10^6$ daltons of protein (Beemon et al. 1974; Billeter et al. 1974). All non-defective viruses have at least three genes (in this case, a gene is defined as the region translated into a primary protein product): gag, which codes for a precursor which is cleaved to yield four internal structural proteins of the virion; pol, which codes for the virion RNA-directed DNA polymerase (reverse transcriptase); and env, which codes for the virion envelope glycoproteins. These three genes may code for all the information necessary for virus replication. A fourth region, c (for 'common'), is found in a number of viruses (Wang et al. 1975, 1976). Although there is very tentative evidence that it may have its own mRNA (Krzyzek et al. 1978) and gene product (Purchio et al. 1977), there is no genetic evidence for such an extra gene. A number of retrovirus strains contain an additional gene which appears to have been added to the virus genome by recombination with host cell information (Stehelin et al. 1976). In general, this added information appears to confer on the virus the ability to induce neoplastic disease with a very short latent period, although genetic proof of this point exists only for the src gene of ASV (Toyoshima & Vogt, 1969; Martin, 1970). In non-defective avian sarcoma viruses, the transforming gene is present in addition to gag, pol, env, and c. In other cases, it replaces a
virus gene – env – in replication defective ASV; possibly parts of gag-pol in transforming avian leukosis virus (such as MC29; P. Mellon, P. Duesberg & P. K. Vogt, personal communication) and murine sarcoma virus (Hu et al. 1977).

Replication. The general outline of retrovirus replication has been known for a long time (Temin & Baltimore, 1972). Soon after entering a cell, the single-stranded virion RNA is copied into double-stranded DNA by the polymerase carried in the virion. This DNA provirus is transported to the nucleus and integrated into the cell genome. The integrated provirus serves as a template for host RNA polymerase which synthesizes progeny virion and messenger RNA. These RNA species are processed into forms resembling cell mRNA and transported to the cytoplasm where virus protein synthesis and assembly of virions occur.

A prominent feature of this overall mechanism is that all steps in virus nucleic acid replication seem to be carried out either by cellular systems or by virus coded enzyme(s) in the virion itself and not by virus coded enzymes synthesized after infection. It should also be pointed out that, in general, retrovirus replication does not lead to death of the host cell. Instead, the cell usually survives the infection, continues to divide and becomes permanently virus producing. When a steady state is reached (one to two days after infection) approx. 1% of cell RNA and protein synthesis, and $10^{-3}$ to $10^{-4}$% of DNA synthesis, is of viral species.

Diversity. Retroviruses appear to have a genetic plasticity unparalleled in other RNA viruses. It is possible that no two isolates of exogenous virus from the same species are identical to one another. Different laboratory strains of ASV, for example, differ in a number of respects from each other, including point mutations, and additions and deletions of information. While the point mutation rate for ASV may not be significantly different from other RNA viruses, the adaptability of virus populations to new selective conditions seems to be quite high (J. Coffin & C. Barker, unpublished data). This phenomenon is likely to be due in part to the extraordinarily high frequency of recombination between retrovirus genomes, even as compared to DNA viruses. Additionally, deletion of non-essential information (i.e. src) is an extremely frequent event (Vogt, 1971a; Martin & Duesberg, 1972) and it is quite difficult to maintain a population of ASV free of such deletion mutants. Frequent deletions of essential information are also found (Shields et al. 1978). Addition of cellular to virus information does not appear to be as frequent, but is at least historically detectable, as discussed above. It is likely that recombination of host genes and virus information happens quite frequently, but lack of a suitable selective system prevents its detection.
In the sections that follow I will attempt to relate these various features of retrovirus biology to one another. The underlying theme of the article will be the interrelationship of structural features of the virus genome and the mechanism of the enzyme systems responsible for its replication. In this way it is possible to develop an overall picture of retrovirus biology which relates many diverse features in a reasonably simple way.

**Structure and expression of the viral genome**

The genetic structure of ASV has been worked out primarily by oligonucleotide mapping of deletion mutants and recombinants between different strains of viruses (Joho et al. 1975, 1976; Wang et al. 1975; Coffin & Billeter, 1976). The best available evidence yields the map shown in Fig. 1(A). This map is drawn to approximate scale. Although it is not yet known exactly where the boundaries between the different coding regions lie, they are certainly in the order 5'-gag-pol-env-src-c-3'. The order in which the four gag and two env polypeptides are encoded can be deduced from pactamycin mapping experiments which locate them on their precursor polypeptides (Vogt et al. 1975; Diggelman & Klemenz, 1978; Shealey & Rueckert, 1978). The coding portion of the genome must also contain control and recognition sequences for mRNA processing, translation initiation and so forth, but these have not yet been characterized and probably will not be until major portions of nucleotide sequence have been determined.

The genome resembles typical eukaryotic mRNAs in its pattern of post-transcriptional modifications. The 3' end is polyadenylated with approx. 200 A residues (Lai & Duesberg, 1972), the 5' end is capped with an m7G4ppp3Gm group (Furuichi et al. 1975; Keith & Fraenkel-Conrat, 1975) and the RNA is methylated internally with approx. 10 m6A residues (Beemon & Keith, 1976). The function of these modifications is not clearly understood.

In the virion, the native state of the genome is apparently a dimer of two such molecules, linked by hydrogen bonds near their 5' ends (Kung et al. 1976). Such a structure has not been visualized for ASV, presumably because of the weakness of the dimer linkage and has been inferred by analogy to other retroviruses. To obtain such a linkage between identical molecules, there must either be a nucleic acid 'linker' base paired to both genomes, or a palindromic sequence in each molecule at the linkage point. A linkage structure for ASV has been proposed on the basis of the nucleotide sequence of the 5' end of the genome (Haseltine et al. 1977). This structure uses both a palindromic sequence and the tRNA primer as a linker. Such a structure is too near the 5' end to account for the structure seen with other viruses and remains to be tested.

Since each molecule in the dimer contains the full genetic complement of the virus, the virus is at least chemically diploid. This conclusion was first reached by dividing the weight of high mol. wt. RNA present per virion by its sequence complexity as determined from oligonucleotide fingerprinting (Beemon et al. 1974; Billeter et al. 1974). As I will discuss later, it remains to be established that the virus is genetically diploid – that is, that both subunits make available all their genetic information to the progeny. It should be noted here that, although there are on average two genome subunits per virion, there may exist in a virus population substantial proportions of virions with three or more subunits as well as virions with no genome RNA. This possibility is supported by the observation that, under certain experimental conditions, more or less normal 'empty' virions can be prepared (Levin et al. 1974; M. Linial, personal communication), suggesting that there is no stringent requirement of precisely one dimer for virion assembly.

In addition to its genome, the virion contains numerous small RNA molecules derived from the host cell (Sawyer & Dahlberg, 1973). The vast majority of these are probably
Table 2. Terminally redundant sequences in ASV*

<table>
<thead>
<tr>
<th></th>
<th>5' terminus</th>
<th>3' terminus</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>m' G p p Gm</td>
<td>G C C A U U U U A C C A C A C A A A200</td>
</tr>
<tr>
<td>(2)</td>
<td>m' G p p Gm</td>
<td>G C C A U U U U A C C A C A A A A200</td>
</tr>
<tr>
<td>(3)</td>
<td>m' G p p Gm</td>
<td>G C C A U U U U A C C A C A A A200</td>
</tr>
<tr>
<td>(4)</td>
<td>m' G p p Gm</td>
<td>G C C A U U U U A C C A C A A A200</td>
</tr>
</tbody>
</table>

* Sequences 1 and 2 have been determined exactly. Sequences 3 and 4 are tentative and derived by fitting partial products to give minimum differences from the first two. Sequence 1 is found in Pr-A, B, and C, B77 and RAV-o (Haseltine et al. 1977; Shine et al. 1977; Coffin et al. 1978a); sequence 2 in AMV (Stoll et al. 1977); sequence 3 in RAV-6 (Joho et al. 1978); and sequence 4 in RAV-1, RAV-2, and SR-D (Wang et al. 1977).

† Indicates the minimum points of sequence divergence. Nucleotides whose order is uncertain are in parentheses and sequences not analysed but inferred from the other end are in square brackets.

fortuitous contaminants with no role in the life of the virus, but one of them, a molecule of tRNA\textsuperscript{trp}, serves the important function of primer for synthesis of proviral DNA (Dahlberg et al. 1974). The observation by Taylor & Illmensee (1975) that the tRNA\textsuperscript{trp} is hydrogen bonded very near the 5' end of the genome clarified the problem (which in fact occurs no matter where the primer is located) that polymerization of DNA by the virion polymerase proceeds toward the 5' end of the template and therefore must reach the end of the molecule before copying the entire genome. This finding led to a flurry of model building (Junghaus et al. 1975; Coffin, 1976; Collett & Faras, 1976; Haseltine & Baltimore, 1976; Stoll et al. 1977). To solve this problem these models postulated a terminally redundant sequence in retroviral genomes. Therefore, by copying the 5' terminal region of the virus genome, the polymerase creates a new primer which is complementary to nucleotide sequences near the 3' end and can form a new template-primer pair allowing re-initiation and elongation of the nascent chain from the 3' end. This prediction was quickly verified by nucleotide sequence determination of the terminal regions of ASV and MuLV (Coffin & Haseltine, 1977a; Haseltine et al. 1977; Schwartz et al. 1977; Stoll et al. 1977; Wang et al. 1977).

The terminally redundant sequences of a number of ASV strains are shown in Table 2. Although some of these sequences are tentative and based only on the composition of RNase A and U\textsubscript{g} digestion products of the terminal oligonucleotides, it is clear that there is similarity between different strains of both exogenous and endogenous viruses. There are also, however, several points at which sequence variability occurs with no detectable difference in function. It may be that this sequence serves no other function than to match its opposite number and that therefore almost any pair of matching sequences will work. Consistent with this idea, we have found variants of Pr-RSV-B, which arise during passaging of the virus, to display greater variability in terminally redundant than in coding sequences (S. Voynow, C. Barker & J. Coffin, unpublished data). In Moloney murine leukaemia virus (Mo-MuLV) the redundant sequence is much longer – approx. 60 nucleotides as compared to about 21 for ASV – and has no obvious sequence or structural relationship to any of the ASV sequences (Coffin et al. 1978b).

As Table 2 shows, each strain of virus has exactly the same sequence at the 5' and 3' termini of its genome although there may be some heterogeneity in the length of the
redundant sequence in a virus population (Stoll et al. 1977). This identity (or ‘concordance’) of terminal sequences is also found in recombinants between strains with differing termini (Wang et al. 1977; Joho et al. 1978) and in spontaneous mutants arising in cloned virus populations (S. Voynow, C. Barker & J. Coffin, unpublished data). As we shall see, these results are almost certainly a direct consequence of the mechanism of replication and do not provide any information on either the mechanism of recombination or any selective disadvantage of molecules with slightly mismatching terminal sequences.

Fig. 1 shows, on a greatly expanded scale, the terminal regions of the genome of probable importance for replication. These include, from left to right, the 5′ copy of the redundant sequence (Rs, 16 to 21 nucleotides); a sequence unique to the 5′ end (Us, 80 nucleotides); the primer binding site (PB, 16 nucleotides) complementary to the 3′ terminal 16 nucleotides of tRNA\(^{\text{trp}}\) (Cordell et al. 1976; Eiden et al. 1976); the unique sequence near the 3′ end (Us, approx. 200 nucleotides); and the 3′ copy of the redundant sequence (Rs, 16 to 21 nucleotides). No other sequences are known to be required for replication and it would be interesting to see if any RNA flanked by these nucleotide sequences could be replicated as a ‘virus’ when complemented with virus gene products.

As mentioned above, the Us sequence, at least in ASV, may contain the dimer linkage structure (Haseltine et al. 1977). This sequence is almost exactly the same length in ASV and MuLV (Haseltine & Coffin, 1978) and may also serve as a ‘spacer’ to alleviate steric problems that could arise when the nascent DNA chain is transferred from one end to the other during DNA synthesis. Since the Us sequence is spliced intact onto mRNA molecules (Mellon & Duesberg, 1977; Weiss et al. 1977; M. Champion & J. Coffin, unpublished data),
it may have a role in ribosome binding, although adjacent sequences do not agree with the amino acid sequence of the N-terminal peptide of p19 [which is usually blocked (Herman et al. 1975) and therefore must be the first peptide synthesized]. Protein synthesis must therefore be initiated to the right of this sequence (R. Eisenmann, personal communication).

The U₃ sequence has recently attracted attention because of its presence near both ends of the integrated provirus (Hughes et al. 1978; J. Taylor, personal communication). As discussed in a later section, it is likely that this sequence contains some signals for initiation and/or termination of RNA synthesis. At least portions of the U₃ sequence are identical in all strains of exogenous ASV examined to date. In particular, all such viruses contain a specific oligonucleotide (oligonucleotide C; Wang et al. 1975, 1976) which is immediately adjacent to the R₃ sequence (D. Schwartz & W. Gilbert, personal communication) and contains the AAUAAA sequence found near the poly(A) of many eukaryotic mRNAs (Proudfoot & Brownlee, 1976). [This sequence appears within the R₃ sequence of Mo-MuLV RNA (Coffin et al. 1978b).] It is interesting that this region of the genome of RAV-0 and related endogenous avian oncoviruses does not contain oligonucleotide C and these viruses have a U₃ region which contains no detectable sequence relationship to that of exogenous viruses (Coffin et al. 1978a). This different sequence apparently confers a selective disadvantage to RAV-0 relative to exogenous viruses since all recombinants between the two (about 30 examined to date) invariably have this part of the genome from the exogenous virus (P. Tsichlis & J. Coffin, unpublished data). It is tempting to speculate that the difference in the U₃ sequence is related to the lower rate of virus production observed in cells infected with or spontaneously producing RAV-0 as compared with exogenous virus (Linial & Neiman, 1976; Robinson, 1976).

**Virus mRNAs**

Several species of virus RNA, tentatively identified as mRNAs, can be found in retrovirus infected cells. These are shown in Fig. 1(B-D). The mRNAs for env and probably src are portions of the total genome and appear to consist of the translated portion and all information to the 3' end of the genome (Hayward, 1977; Weiss et al. 1977). Both of these species have a segment which is derived from the 5' end of the genome which includes U₅ and probably the PB segment (Weiss et al. 1977; Mellon & Duesberg, 1977).

The mRNAs for the pr76 gag precursor and the pr180 pol precursor are inseparable from each other and from genome RNA in sucrose gradients (Opperman et al. 1977) and are therefore similar or identical in size to the total genome. In infected cells and in vitro the two protein precursors are made in a ratio of approx. 10:1 and the pr180 seems to contain pr76 as its N-terminal portion. These results suggest that there is a weak terminator at the junction between the gag and pol regions, which, in normal cells is suppressed some of the time to allow read-through of pol. Such a mechanism is known to operate in the RNA bacteriophage Qβ (Weiner & Weber, 1973) and serves to regulate the relative amounts of coat protein molecules. The same mechanism may also be used by Mo-MuLV, since addition of amber suppressor tRNA to an in vitro translation system increases the proportion of pr180 to pr76 (Philipson et al. 1978). This result suggests the presence of an amber terminator at the MuLV gag-pol boundary. In the case of ASV, suppressor tRNAs do not increase translation of pr180 although addition of amber suppressor tRNA leads to synthesis of a molecule slightly lower than pr76, suggesting that pr76 terminates naturally at a UAG codon (Bishop et al. 1978). This result must mean that there are other ‘anti-termination factors’ in eukaryotic cells, or that the mRNAs for gag and pol are, in fact, different species.
The latter possibility is much more likely and two alternatives for their structure are shown in Fig. 1 (D1 and D2). In each case the two molecules differ by a splice near the gag-pol boundary. In the first case (D1) the mRNA for pr180 more closely resembles the genome RNA and is devoid of a terminator between gag and pol. A splice at the boundary would create a termination codon to make the mRNA for pr76. In the second case, splicing removes a terminator in the genome at the end of gag or shifts the reading frame to remove terminator(s). Note that MuLV and ASV may be the same in this respect, differing only in that in ASV, more than one termination codon is removed and in MuLV splicing to create the pol mRNA does not induce a frame shift and removes only an amber terminator. It should be noted that neither gag or pol mRNA may be identical to the genome RNA. One reason that the mRNAs might be distinct from genome RNA would be to surmount the problem that seems to occur soon after infection if the genome RNA is incorporated into polysomes. Since RNA is translated 5' to 3' and copied 3' to 5', it is difficult to see how the genome could be copied into DNA with the ribosomes and polymerase moving in opposite directions. This apparent problem could be eliminated if an additional splicing event created the gag and pol initiation (or ribosome binding) sequence, or, alternatively, removed a blocking sequence. Other solutions involving virion proteins are also possible.

The considerations discussed above raise the possibility that the high mol. wt. RNA isolated from virions may be a rather complex mixture of different species, differing only slightly in size, including the authentic genome and spliced mRNAs. It is not known whether the virus has some mechanism for excluding mRNA species from virions. The presence in the 70S complex of occasional molecules of env (Stacey & Hanafusa, 1978) and probably src (Purchio et al. 1978) mRNA has been established. Since these molecules do not appear as peaks in gels or gradients of virion RNA and do not contribute detectable amounts of capped oligonucleotide to fingerprints (Coffin & Billeter, 1976) they must be in substantially lower concentration relative to high mol. wt. RNA than they are in infected cells (Hayward, 1977; Weiss et al. 1977). The mRNAs in the virion are detectable by in vitro translation experiments and at least the env mRNA can apparently be stably expressed following infection of cells with virions containing this species (D. Stacey, personal communication). This result is probably a reflection of the fact that splicing of the 5' terminal sequence of the genome on to smaller RNAs could create a 'mini genome' containing all the structures necessary to be replicated by virus systems. Heterodimers containing such minigenomes may be important intermediates in formation of recombinants between virus genomes and partial virus information expressed in certain cell types (Coffin et al. 1978a).

Most or all synthesis and processing of virus RNA takes place in the nucleus of infected cells (Parsons et al. 1974; Rymo et al. 1974), presumably using the integrated DNA as template. The major enzymatic activities are almost certainly from the host cell – for example, the sensitivity of synthesis to <i>a</i>-amanitin implicates RNA polymerase II (Jacquet et al. 1974; Rymo et al. 1974). It has been suggested that ASV RNA is synthesized as a precursor slightly larger than genome size (Bishop et al. 1976), but this question needs to be re-examined in light of the recent progress in understanding cell mRNA processing.

The presence of both spliced (mRNA) and apparently unspliced (genome) virus RNA species as finished products raises the problem of how the splicing is regulated. It seems likely that the genome is the precursor to spliced mRNAs, since micro-injection of 35S ASV virion RNA into chicken cell nuclei yields active env mRNA (Stacey & Hanafusa, 1978). In other systems where splicing has been observed, such as globin (Tilghman et al. 1978), the precursor seems to be quite short lived. If this is a general phenomenon, the problem arises of preserving a portion of the tumour virus genome molecules against
splicing. It has been suggested that one of the gag proteins (p19) regulates the extent of splicing by binding to possible splice points in the virus RNA (Leis et al. 1978). This model would require that a portion of intracellular p19 be found in the nucleus of infected cells; however, such a finding has not been reported.

Proviral DNA synthesis

Analysis of intermediate structures in proviral DNA synthesis is proceeding rapidly in many laboratories, but a complete picture of the mechanism of virus DNA synthesis is not yet available. One major problem is that retroviruses use a system different from all other known groups of viruses to ensure that all portions of the genome are copied into double stranded DNA. This system combines direct terminal repeats in the genome and the ability of the polymerase to move, along with the nascent chain, from one template to another. As we shall see, all the consequences of this surprising mechanism are probably not yet appreciated.
Fig. 2 shows the genome of ASV (A; with the terminal regions expanded for clarity) and the forms of DNA observed in vitro and in vivo which are currently considered to be intermediates. All the forms are shown aligned with the portion of the genome from which they are believed to be copied and molecules of the same sense as the virus gene (+ strand) are labelled with capital letters; complementary molecules (− strand) with small letters. The 3′ end of each strand (and therefore the direction of its synthesis) is shown by an arrowhead.

The first two DNA forms shown (B and C) have been detected only in in vitro endogenous or reconstructed polymerase reactions. When the total product of such a reaction is analysed by polyacrylamide gel electrophoresis, various homogeneous DNA species can be isolated, each of which is an elongation product of smaller species (Haseltine et al. 1976). The most prominent product (B), variously called ‘strong stop DNA’, ‘short stop DNA’, ‘iDNA’, and ‘cDNA3’, is a copy of the U5 and R5 region of the genome terminating opposite the Gm (Shine et al. 1977; Haseltine et al. 1977; Coffin & Haseltine, 1977a). In the case of Pr-C ASV it is 101 nucleotides long, in Mo-MuLV its length is 135 to 145 nucleotides. Its prominence among the reaction products is presumably because initiation and elongation of DNA chains are rapid events relative to the transfer from end to end. In the case of MuLV it has been shown that DNA molecules longer than strong stop DNA (C) contain a single copy of the redundant sequence, flanked by the U5 sequence at their 5′ end and sequences copied from the 3′ end of the genome at their 3′ end (Haseltine & Coffin, 1978). The presence of only one copy of the repeated sequence shows that this sequence must be used as a bridge to create a template-primer pair near the 3′ end of the genome, as predicted by the models. In the case of ASV, the same type of experiment has not yielded the same result. Instead, DNA species only slightly longer than strong stop DNA have been proven to contain sequences derived from many parts of the genome, particularly a region near the middle (Cashion et al. 1976; J. Coffin & W. Haseltine, unpublished data). Although this result was first attributed to additional initiation sites in the virus genome, it is more likely to be a consequence of ‘improper’ elongation, since we have found that DNA in each single band from polyacrylamide gels hybridized to both 5′ terminal and internal sequences (J. Coffin & W. Haseltine, unpublished data). It is possible that the difference between MuLV and ASV in in vitro DNA species is a consequence of difference in length of the terminally redundant sequences. The stability of the hybrid of strong stop DNA and the 60-nucleotide MuLV R5 sequence is much higher than that of the corresponding 21-nucleotide ASV hybrid (Coffin & Haseltine, 1977a; Coffin et al. 1978b). It is probable, under in vitro reaction conditions, that the specificity of hybrid formation between strong stop DNA and the ASV R5 sequence may be low relative to the affinity of the polymerase for any sequence with perhaps a few nucleotides of homology to the 3′ end of the strong stop DNA. This difference probably also has some historical interest, since it has traditionally been much more difficult to prepare cDNA probes complementary to the whole genome of MuLV as compared to ASV using only the endogenous virion polymerase and primer. The ASV polymerase does not seem to have such a problem in vivo, since the majority of the DNA found (at sufficiently late times after infection) is in full-genome length forms (Varmus et al. 1976) although some smaller species can be detected which might be analogous to the ‘improper’ elongation products seen in vitro (Guntaka et al. 1976).

In the case of Mo-MuLV, an infectious product can be synthesized by virions in vitro (Rothenberg et al. 1977), but this species is only a small minority of the total reaction product. This species is not made if actinomycin D is present in the reaction, suggesting that a portion of the infectious molecule might be synthesized using a DNA template. The
longest DNA molecule synthesized in such reactions contains a complete copy of the genome including the terminal regions (Rothenberg et al. 1978), and it has been suggested that it is somewhat longer than full length (Shank et al. 1978) and corresponds to the minus strand of the form shown in Fig. 2(E).

The remaining forms of DNA shown in Fig. 2 have been detected and characterized in cells shortly after infection, usually by hybridization of labelled virus RNA or cDNA with cell DNA fractions often using the ‘blotting’ procedure of Southern (1975). Unfortunately, only relatively long-lived forms can be observed in this way and many key intermediates have not been found. Two of the DNAs are found in the cytoplasm 0 to 24 h after infection (Varmus et al. 1978). The first of these (Fig. 2D) is a partial (−) strand of heterogeneous length whose 5′ end is the strong stop sequence. A short (250 to 300 nucleotides) piece of (+) strand DNA is found before full length (−) strands appear (Varmus et al. 1978). This piece apparently is synthesized from a unique starting point and contains both U5 and U3 sequences. It must therefore be synthesized using the 5′ terminal region of the (−) strand DNA as a template before completion of (−) strand synthesis. At a somewhat later time (3 to 4 h in ASV) double stranded DNA with a sedimentation coefficient similar to that expected for a copy of the virus genome appears (E). This species consists of an intact (−) strand hydrogen bonded to short fragments of (+) strand DNA (Varmus et al. 1976). Very recent evidence (Shank et al. 1978; J. Taylor, personal communication) suggests that this DNA has a surprising structure. As expected, its right end is a copy of the U5-R5-U3 combination as found in structure (C). Its left end contains an additional copy of the U3 sequence adjacent to R5-U5. This juxtaposition of sequences can be created only by a second 5′ to 3′ transfer of polymerase plus growing chain.

The presence of the remaining forms (Fig. 2F) exclusively in the nucleus of infected cells suggests that some cell enzymes which are exclusively nuclear must be involved in their formation. Such cell enzymes must include a DNA ligase, since the (+) strand is now in a continuous chain (Varmus et al. 1976). The nuclear forms are covalently closed circles of two different sizes (not including defective forms). The larger circle is similar to a closed form of the linear molecule and has the sequence -C-U3-R5-U5-R5-U5-PB-. The smaller circle contains only a single copy of the U5-R5-U3 repeat. One or both of these forms is presumed to be the form which is subsequently integrated into the host DNA, although there is as yet no direct evidence on this point.

Possible mechanisms of proviral DNA synthesis

The mechanisms which relate one proviral DNA structure with another are not yet fully understood. The principal problems are: the mechanism of the transfers of the growing chain from end to end; the nature of the primers for (+) strand synthesis; and the mechanism of circle formation. From the in vitro results discussed above, the first end-to-end transfer of the nascent DNA chain must utilize base pairing between sequences at the 3′ end of the DNA and the 5′ end of the genome, i.e. r5-R5. It seems most probable that the DNA becomes available for base pairing as a result of hydrolysis of the copied RNA by the polymerase-associated RNase H (Mölling et al. 1971) and recent in vitro evidence (Collett et al. 1978) supports this view. For this purpose, the RNase H has two useful properties: it degrades only as an exonuclease (Leis et al. 1973) and therefore will not attack the RNA prior to completion of the DNA strand; and it hydrolyses only the RNA strand of DNA–RNA hybrids (Mölling et al. 1971) and therefore should not separate the tRNA<sup>trp</sup> primer from the PB region of the genome. This specificity would ensure that end-to-end transfer of the growing DNA chain is a unimolecular reaction. Loss of the U5-R5 sequence by RNase H
Fig. 3. Possible intermediates in synthesis of complete linear proviruses. (a) According to a suggestion by J. Taylor et al. (personal communication). The strong stop sequence is copied from one genome (italic letters) and transferred to a second to synthesize a complete (−) strand (D4) which is then transferred to the 3' end of (probably) the first genome (D3) to add the U3 sequence (D2), yielding the complete structure shown in D4. An experimental prediction of this model is the distribution of sequences from each parent in the progeny. (b) According to a suggestion by Shank et al. (1978), before completion of the (−) strand, a short (+) strand is synthesized near its 5' end (D1). The sticky end (B) created by copying the primer allows a template-primer pair to be formed (D2) and elongation of the (−) strand on the short (+) strand template leads to the structure shown (D3).

hydrolysis would appear to create a major problem for synthesis of the structure shown in Fig. 2(E), since the information which must be copied again has been lost. Two mechanisms have been proposed to eliminate this problem (Fig. 3). In the first (J. Taylor, personal communication), shown in Fig. 3(a), DNA synthesis is initiated on only one of the two genomes in the complex and is subsequently transferred to the 3' end of the other which retains its 5' terminal region. D1 shows the structure resulting from elongation to the end of the genome. In this molecule the left-hand r5-u5 sequence and all the coding regions are derived from one parent genome and the right hand r5-u5 sequence from the other (shown by italics). The U3 sequence would then be added to the DNA by using the identical mechanism of transfer as before (D2) and terminating specifically at the end of the U3 sequence (D3). This process yields the structure shown in D4. The second model, Fig. 3(b), proposed by Shank et al. (1978), need use only one genome as template. Initiation of synthesis, transfer of the growing chain and elongation to the (now truncated) 5' end lead to the (−) strand shown in D4. If at the same time the short (+) strand initiated near the 5' end is elongated to copy the tRNAtrp primer, a new redundancy (the PB sequence) is created. Removal of the primer by RNase H creates another sticky end and allows the (+) strand fragment to be used as template (D2) to arrive at the structure shown in D3.
At the moment there is too little information to decide between these models. The first (Fig. 3a) has two unattractive features: the sacrificing of a complete genome for the sake of 100 nucleotides of sequence and the problem of why the polymerase should terminate at a specific point the second time it is copied, but not the first. In the second model (Fig. 3b), the specific initiation site for (+) strand synthesis is somewhat difficult to explain, but such a piece of DNA has been found. The sensitivity to actinomycin D of in vitro synthesis of infectious MuLV DNA, alluded to above, lends some support to model (b) since DNA-directed synthesis would be required for a complete (−) strand. A possible experimental test is suggested by the drawings in Fig. 3. In the case of model (a), the R3 sequence of progeny virus will always be derived from one parent and the R4 from another. Thus, examination of progeny of cells infected with heterozygotes between viruses with distinguishable redundant sequences should allow us to distinguish between the two models.

The most reasonable guess for the mechanism of initiation of (+) strand synthesis is that the genome RNA, after being copied into an RNA–RNA hybrid is nicked by a cellular (endo) RNase H (or is broken before DNA synthesis, see later; Keller & Crouch, 1972). The virion enzyme could then use the 3′ ends of the fragments to initiate synthesis and simultaneously remove the remaining genome RNA and elongate the new (+) chains in a sort of ‘nick translation’ synthesis (thus providing the only example I know of in nature of wholesale destructive replication of genetic information). A less likely possibility for initiation of (+) strand synthesis would be association of the (−) DNA with small cell RNAs after removal of the genome. These could act as primers in a manner analogous to the tRNA\textsuperscript{trP}. A particular problem is posed by the small piece of (+) strand DNA found at the left hand end of the replicating molecule (Varmus et al. 1978) (Fig. 2D), since it would seem that a very specific initiation site would have to be used. Either a site-specific cleavage, or attachment of some very specific primer must be required for this step.

The last detectable intermediate prior to integration is believed to be one of the two closed circular forms (Guntaka et al. 1975; Fig. 2F). The smaller circle could be created by recombination of the identical sequences at the ends of form E as:

\[
\begin{align*}
- - - R_5 U_5 \\
\times \\
U_3 R_5 U_5 - - - \rightarrow - - - R_5 U_5 - - - + U_3 R_5 U_5 \text{ (lost)}
\end{align*}
\]

For the larger circle there are at least three possible mechanisms: (1) direct ligation of the ends of the linear molecule; (2) illegitimate recombination of non identical sequences near each end of the molecule; or (3) recombination through a short sequence repeated at each end (S. Hughes, personal communication). The latter possibility is illustrated by sequence ‘X’ in Fig. 2. If this short sequence is present in the genome near the 5′ end of the U3 sequence and repeated at the 3′ end of the U5 sequence, then the linear form (Fig. 2E) would have it near each end. Recombination to close the circle would then proceed:

\[
\begin{align*}
- - - R_5 U_5 X \\
\times \\
X U_3 R_5 U_5 X - - - \rightarrow - - - X U_3 R_5 U_5 X U_3 R_5 U_5 X - - - + X \text{ (lost)}
\end{align*}
\]

Note that the X sequence is still present more than enough times for synthesis of progeny containing this sequence twice. The complete sequence of the 3′ terminal region of ASV should soon be available for comparison with the U5 sequence to validate this hypothesis. In principle it could be tested by hybridization; however, the presence of sequences in the genome other than near the 5′ terminus hybridizable to strong stop DNA is in dispute.
Using a method which would have detected homologous regions of 30 nucleotides or more, we could find no such sequence (Coffin & Haseltine, 1977b). Other hybridization protocols, which should not have seen the R5-R5 hybridization, have sometimes detected sequences complementary to strong stop DNA near the 3' end of the genome (Friedrich et al. 1977; W. S. Hayward, personal communication) and sometimes not (Weiss et al. 1977; Krzyzek et al. 1978).

From the previous discussion, it should be clear that the pathway from parental virus single stranded RNA to closed circular DNA is still a fruitful area for speculation. While I have made no effort to hide my bias in this matter, it is unlikely that any of the mechanisms proposed will be completely correct. Nevertheless, such model building has in the past proved useful for designing experiments and as a pastime for tumour virologists. The reader is invited to join in by following a few simple rules:

1. Use virion systems (polymerase, RNase H, tRNAtrp) as much as possible, bearing in mind the in vitro biochemistry of the enzymes.
2. Use cell systems (additional enzymes and primers) as little as possible.
3. Rely on base pairing rather than protein molecules to join and hold the nucleic acids together.
4. Favour intramolecular events over intermolecular events.
5. Do not be afraid to invoke directly repeated sequences in the genome when it seems necessary or useful.
6. Be careful to ensure that all parts of the genome are passed on to the progeny.

Integration

Recently much has been learned about the structure of the integrated DNA (Fig. 2 G) of ASV.

1. It is, to within a few base pairs, identical to the double stranded linear form (Fig. 2 E) and therefore contains a complete copy of the genome plus repeats of sequences derived from the ends (Hughes et al. 1978; J. Taylor, personal communication).
2. There is only one or a small number of copies per infected cell (Varmus et al. 1973).
3. Different clones of ASV-infected mammalian and quail cells have the provirus integrated at different sites in the host genome – it has been calculated that there are at least 40 such sites (Hughes et al. 1978).
4. In ASV-infected chicken cells, the integration site is apparently not unique even within a clone of cells (J. Taylor, personal communication; H. Varmus, personal communication).

The simplest mechanism to imagine for integration of the provirus is recombination between the large circular form and cell DNA. In such a case, the integration system must recognize a specific sequence in the DNA at the point of closure of the circle (the U5–U5 junction). Since both the U5 and U3 sequences appear twice in the circle, such specificity would require that features of both sequences be recognized. Additionally, the system may have no specificity for any sequence in the cell DNA of more than a few base pairs in length. It is therefore most unlikely that integration is a result of a general recombination system which recognizes and joins any homologous sequences. This latter conclusion is supported by the failure to find recombinants among the virus progeny from super-infected cells pre-infected with a different virus strain or harbouring a non-expressed endogenous virus genome (Wyke et al. 1974; Weiss et al. 1973).

The system for integration most likely pre-exists in the host cell, although some elements from the virion may be retained in a complex with the DNA. It should be possible to infer
the role of virion proteins in integration by transfection experiments with pure proviral DNA. Although chicken cells transfected with any form of ASV proviral DNA synthesize infectious virus (Fritsch & Temin, 1977), it has been impossible to prove in this system that the transfecting DNA is integrated into the cell genome (G. Cooper, personal communication). This result suggests that a virion protein may be required for some phase of the integration process. However, transfection of ASV or murine sarcoma virus DNA into mammalian cells seems to lead to integration of the transfecting DNA (Smotkin et al., 1976; G. Cooper, personal communication), although the low efficiency of the transfection process and the fact that partial MSV DNA molecules lacking one end are infectious (R. Weinberg, personal communication) leaves open the possibility that the usual mechanism of integration is not involved in transfection.

The failure to find a specific integration site for the ASV genome in clones of infected chicken cells raises the interesting possibility that the integrated genome can be excised and re-integrated at reasonably high frequency. Recombination in the directly repeated sequence (U3-Rs-U5) at each end of the integrated genome would lead to excision of the provirus from the cell DNA. The excised provirus would be the same as the smaller circle in Fig. 2F. It is difficult to see how this form could be re-integrated without loss of one of the repeated sequences.

The absence of specific integration sites and the presence of additional information at each end (U3 at the left and U5 at the right) of the integrated provirus strongly suggests that control sequences for initiation (and possibly termination) of virus RNA synthesis are encoded in the virus genome. If the progeny genome is not cleaved from a longer precursor, these sites are as shown in Fig. 2G (i and t).

As suggested by J. Taylor (personal communication), the promoter for virus RNA synthesis is probably in the U5 portion of the provirus, and the U5 region may contain a termination sequence. These considerations make the U3 sequence of particular interest, since ASV RNA is synthesized at a very high rate – about 1% of total cell RNA or some 10% of total mRNA synthesis (Coffin et al., 1974; Jacquet et al., 1974; Rymo et al., 1974) and reaches a steady state level of some 10000 copies per cell (Coffin & Temin, 1972), as much as the most abundant cell mRNA species in specialized cells. As noted above, the major difference in sequences between endogenous and exogenous ASV strains is in the 3' proximal sequence. This system should therefore provide a very useful model for a genetic test of the ideas outlined above and for study of structure and function of sequences regulating RNA synthesis in eukaryotic cells.

If the hypothesis of initiation and termination controlled by proviral sequences is correct, then an interesting problem arises, since the RNA polymerase must cross the R5-U5 termination sequence twice. The first time the polymerase encounters this, sequence transcription must not be efficiently terminated or there would be no synthesis of the bulk of the genome RNA. An analogous, although not so serious, problem seems to occur at the 3' end in that the U3 sequence might initiate RNA synthesis to the right of the provirus. Similarly, if the final genome size is the result of cleavage at the indicated points, the system responsible would have to recognize the U3-R5-U5 sequence twice and cleave at U3-R5 on the left side of the genome and at R5-U5 on the right side.

A final point in the overall replication scheme concerns the fate of the redundant sequences in the course of replication. From the scheme in Fig. 2 and from in vitro evidence (Haseltine & Coffin, 1978), it should be clear that the R5 sequence possesses no genetic continuity, since it is not copied during replication. Thus, although a provirus with slightly different terminal sequences could be formed as a result of recombination, point mutation
or transfer of the enzyme and DNA from one template genome to another, the termini of any homozygous progeny formed would again be identical after one further round of replication. This phenomenon almost certainly accounts for the 'concordance' of terminally redundant sequence in recombinants (Wang et al. 1977; Joho et al. 1978) and point mutants (S. Voynow, J. Coffin & C. Barker, unpublished data) which have been re-cloned to homogeneity.

Recombination

The excellent review on recombination by Hunter (1978) is still current, so I will refer the reader there for a complete discussion of all the models proposed to date and experimental evidence for and against each. In place of such a complete review I will cover a few major points and propose yet another model, one that seems to mesh well with the mechanisms discussed in the previous section.

Although recombination has been demonstrated between MuLV strains (Elder et al. 1977; Failer & Hopkins, 1978), the phenomenon has been studied extensively only in ASV, due to the large selection of genetic markers available with the avian viruses. A typical experiment (Blair, 1977) is to infect a culture of cells with two genetically different viruses and harvest and examine the progeny with some selective system after a few replication cycles. Although determination of precise recombination frequencies is impossible in such a system, as much as 40% of the progeny of the original cross can be found as recombinants, even when closely spaced markers [for example, two ts mutants in src (Balduzzi et al. 1978)] are used.

Characteristics of recombinant virus genomes

Fig. 4 shows genomes derived from two typical recombination experiments. In both cases, recombinant progeny were selected to received src from one parent and env from another. The figure shows simplified oligonucleotide maps in which regions of the genome derived from the src+ parent are shaded; regions from the env donor are open; and regions of undetermined origin are shown as lines. The hatched and open boxes above the maps show the regions selected from each parent. Examination of the figure reveals several points related to the mechanism of recombination.

(1) As previously demonstrated (Beemon et al. 1974; Joho et al. 1975), the genomes are the products of intermolecular recombination events, since all contain nucleotide sequences derived from each parent and also lack sequences derived from each parent.

(2) Unselected crossovers are very frequent. Of the 21 molecules shown in Fig. 4, only two have no crossover points except between env and src and there is an average of more than two additional detected crossovers per molecule. As a result, there is no detectable physical linkage. This is most noticeable in the genome region immediately to the left of the selected env marker. Within about 10% of the genome (or 1000 nucleotides), the recombinants contain sequences derived about equally from each parent. Thus the probability of recombination between neighbouring nucleotides in such an experiment can be very roughly estimated to be of the order of 10^-8 (perhaps plus or minus an order of magnitude). Even if a high negative interference is considered, this is an extraordinary figure. In T-even bacteriophage, considered to have an extremely high frequency of recombination, the corresponding value is about 2 x 10^-5 (Luria et al. 1978). The result of this high frequency can be seen to randomize the parentage of all unselected regions in the 5' half of the genome.

(3) There are no obvious 'hot spots' for recombination in which crossovers occur more
frequently. In the left half of the genomes in B, the six detectable crossover positions are used (from left to right) 3, 3, 5, 7, 4 and 8 times.

(4) There are unexpected selected regions. The bars at the top of Fig. 4(a) and (b) show the regions expected to have been selected from each parent on the basis of previously published physical maps (Coffin & Billeter, 1976; Coffin et al., 1978a). The recombinants in Fig. 4(b), however, all have at least one additional selected region from the src parent, shown by the arrow. While a detailed explanation of this phenomenon is beyond the scope of this review, we believe it to be the result of a sequence in the gag region of RAV-o whose gene product is incompatible with the expression of src (P. Tischlis, personal communication). It is very likely that a similar phenomenon is the cause of the frequent appearance of 'abnormal' gag proteins in numerous recombinant viruses (Shaikh et al. 1978a, b).

For these reasons, it is highly improbable that 'standard' genetic or linkage maps can be
Review: Retrovirus genomes

Fig. 5. Recombination by breakage-repair. Two genomes in a dimer are shown. The dimer linkage and other terminal structures are omitted for clarity. In the course of DNA synthesis, the polymerase encounters a break in the template (1); RNase H removes the copied portion (2); and the nascent chain is transferred by base-pairing to the other genome where synthesis continues (3). On encountering another break, the process is repeated, leading to a crossover back to the first genome (4). The final (−) strand is thus the recombinant molecule shown in (5). The primes denote sequences from different genomes. As usual, (+) strand sequences are shown by capital letters, (−) strand sequences by lower case letters.

developed for RNA tumour viruses, except, possibly, within very short distances (for example between ts mutants in src; Balduzzi et al. 1978). While very rough linkage maps have seemed to yield the correct gene order (i.e. gag-pol-env-src; Hunter, 1978) I consider this a fortuitous result of selective effects either of the sort described above, or some more subtle effect, for example a higher specific infectivity of virions of one subgroup than those of another or of wild type than temperature sensitive virus, even at permissive temperature.

A possible mechanism of recombination

The simple model for recombination I wish to propose is outlined in Fig. 5. It rests on the following points:

(1) The virion of a tumour virus protects its genome very poorly from breakage. Years of experience with ASV RNA have shown that, unless extraordinary measures are taken (such as harvesting of virus at very short intervals; Cheung et al. 1972), it is almost impossible to obtain virions with a significant proportion of unbroken RNA (unpublished observations of nearly everyone in the field). Nevertheless, preparations of virus with no detectable intact genomes contain substantial amounts of infectious virus. This is an unusual situation for an RNA virus, since with most other RNA viruses the genomes are relatively easily
isolated as intact, homogeneous, species. This difference could also reflect on the technical skills of RNA tumour virologists, but I do not favour this view, for obvious reasons.

(2) An old observation with ASV is that it is much more resistant to inactivation by u.v. light than other RNA viruses with comparable genome size (Rubin & Temin, 1959; Bister et al. 1977). Furthermore, such inactivation seems to be the result of cross-linking of RNA and virion protein rather than single strand breaks (Owada et al. 1976).

(3) The frequency of recombination is such that it is probable that one or more crossovers occur at each infection cycle.

(4) Recombination between different strains of virus seems to require infection with a heterozygote virus containing a genome of each parent in the same virion (Wyke et al. 1974). This point has been contested recently (Alevy & Vogt, 1978), but I consider the evidence in favour of the heterozygote model to be much stronger than that against it. Note that the requirement for heterozygotes suggests only that recombination is an event preceding integration. At this time in the infection cycle the probability of two genomes (or proviruses) becoming juxtaposed into whatever structure is necessary to mediate recombination is much higher for a heterozygote virion than for two virions entering the cell at different points.

(5) The virion polymerase is capable of being transferred via redundant sequences from one template to another.

(6) Synthesis of (-) strand proviral DNA is very slow compared to other polymerizing systems (Varmus et al. 1978).

Observations (1) and (2) suggest that the virus has a mechanism for surviving breaks in its genome RNA. The other observations suggest what the mechanism might be. A model is shown in Fig. 5. (The terminal sequences are omitted from the model for clarity.) The key assumption is that when a polymerase molecule encounters a break in the template it behaves the same as when it reaches the end of an RNA template. Thus, in Fig. 5 the nick between B and C causes the polymerase to halt (1), and the template to be subsequently degraded by RNase H (2). As at the 5' end of the molecule, this frees the nascent chain for base pairing, but this time with the corresponding sequence on the other molecule in the dimer (3). If, in the course of subsequent synthesis, another break is encountered, the growing chain will be transferred back to the first template (4) and so forth, leading finally to the recombinant shown (5).

The proposed mechanism is a variant both of copy-choice recombination, which has been proposed for numerous systems but never demonstrated, and of nick repair, which occurs in at least some bacteriophage systems (Broker & Doerman, 1975). As in the latter case, the recombination rate should be a function of the number of breaks in the RNA. This is probably the strongest prediction capable of being easily tested. It is expected that this process would be quite slow and if numerous breaks are encountered could lead to a net rate of proviral synthesis much less than that expected from the nucleotide addition rate.

If the breaks which lead to crossing over are random effects of 'aging' of virions, then the expected frequencies of single and multiple crossovers can be calculated in certain cases, since both RNA molecules in a dimer must be of the same age. For example, in virions with an average of one break per genome, or two breaks per dimer, then by Poisson distribution 37% of the molecules will be unbroken and transcribed in one piece. These will not give rise to recombinants. Of the remaining 63%, 37% (or 23% of the total) will transfer to an unbroken molecule and lead to a single crossover. Of the 40% that jump to a broken molecule, slightly more than half (approx. 60%) will transfer to the right of a break in the second genome and lead to a second crossover. This will lead to a final distribution of 37% parental, 39% single recombination and 24% multiple recombinant. Clearly, as the number
of breaks per genome increases, the proportion of single recombinants will decrease very rapidly. The derivation of both genomes from pools with the same breakage history would thus lead to a rather peculiar effect. Recombinants selected for a crossover between markers relatively near one another will come more frequently from more broken and (probably) older molecules in a virus population. Therefore, the closer the selected markers are, the more unselected crossovers should be found in the progeny.

Under the mechanism proposed, a dimeric genome leads to only a single provirus. Thus heterodimers would not be genetically detectable as heterozygotes. Genetic heterozygotes, however, do persist for long periods of time in some mixed virus populations, but not in others (Weiss et al. 1973; Vogt, 1977; P. C. Balduzzi, personal communication; and our unpublished data). As mentioned in the first section of this review, it is possible that these genetic heterozygotes are a consequence of a small proportion of virions with two or more dimers which contain complementing defects not easily repaired by recombination. Particles suggesting multigene virions have been observed in MuLV (Yuen & Wong, 1977). Two types of non-recombining structures could be imagined to lead to persistent complementing heterozygotes. The first would be a deletion in each genome with no common sequence between the deletions. The second could be a type of recombinant suggested by the genomes in Fig. 4. Note that the env region always consists of a large block of information from one parent, with no detectable internal recombination. This result, combined with the observation of substantial sequence homology in this region between different ASV subgroups (Junghans et al. 1975; Tal et al. 1977; Coffin et al. 1978a) suggests that such recombinants should occur but not be viable. In a cross between, for example, ASV strains of env<sup>src−</sup> and env<sup>src+</sup> genotypes, a substantial proportion (probably a majority) of the recombinant progeny would be env<sup>src−</sup> where env<sup>src+</sup> codes for non-functional glycoprotein. Thus the majority of viable progeny detectable by selection of transforming virus of subgroup A could easily come from a minority of multigene virions that are genetic heterozygotes between the two parents or between an env<sup>src−</sup> recombinant and the env<sup>src−</sup> parent.

Note that a corollary of this line of reasoning is that a substantial proportion of cells infected with virions containing heterodimeric genomes coding for two subgroups of env glycoproteins should yield only non-infectious progeny as observed by Vogt (1971b), particularly when the virus is highly at risk for recombination (i.e. the genomes have large numbers of breaks). This effect could lead to an apparent higher recombination frequency between closely spaced point markers (for example, ts mutants in the same gene) than between more distant markers and env.

A similar phenomenon would occur in a cell infected with a heterozygote of a full length genome and a deletion mutant. If a break is encountered by the polymerase in the full-length genome in a region not found in the deleted genome, then synthesis will most likely terminate and the infection will be aborted. This effect will not occur during copying of the deleted genome. Therefore, the majority of the cells infected with such heterozygotes will receive a provirus containing the same deletion, although other parts of the genome may be recombinant. The size of the majority depends on the physical state of the infecting RNA. This would lead to an apparent ‘dominance’ of deletion mutants under appropriate experimental conditions as in the previously rather puzzling observation of Kawai & Hanafusa (1976) who performed a cross of SR–RSV ts68 (env<sup>src+</sup>) and SR–RSV NY8 (env<sup>src+</sup>). They found among the progeny only env<sup>src+</sup> and no env<sup>src−</sup> recombinants. The same sort of dominance could lead to a rapid selection of deletions in mixed virus populations especially those including env, a result we have observed (J. Coffin & C. Barker, in preparation).
An argument that could be made against this proposal is that a structure with numerous breaks would tend to fall apart during replication. This is unlikely for three reasons:

(1) The 70S structure isolated from 'aged' virions retains the identical sedimentation coefficient even though the RNA after denaturation can be shown to consist of pieces only a few hundred nucleotides in length (again, an observation made by almost all workers). Furthermore, such badly degraded structures still have all portions of the genome equally present (Billeter et al. 1974).

(2) During the course of synthesis, the nascent chain could remain attached to the structure via the interaction of the primer with the PB sequence (although, for simplicity, Fig. 2 is not drawn this way).

(3) The strongest dimer linkage point is near the 5' end of the genome (Kung et al. 1976). This would contribute to maintaining the dimer structure until (−) strand synthesis was very nearly complete. Note that the result described in (1) shows that other, weaker linkage points must also exist. Structures suggesting multiply linked subunits can be visualized by electron microscopy (Mangel et al. 1974; Weissman et al. 1974).

Three other possible consequences of this model can be derived by considering a possibility suggested by in vitro experiments with ASV, described in a previous section. If the polymerase plus growing chain are able to transfer to new templates with only limited homology to the nascent chain, then such a mechanism could easily result in a reasonable frequency of deletions or re-duplications. The former would occur by a jump on either template to the left hand side of the region already copied (e.g. from C to E) and the latter by a jump to the right on the other genome (e.g. from C to B'). Deletions of non-essential information (for example, src) have been known for a long time to occur with high frequency (Vogt, 1971a). More recently, it has been recognized that other types of deletions also occur with high frequency (Shields et al. 1978; J. Coffin & C. Barker, unpublished data). Reduplications of sequences have not been reported, although the heteroduplex maps of partial td viruses found by Lai et al. (1977) are as consistent with re-duplication as with the authors' interpretation of insertion of unrelated information.

If the mechanism of proviral synthesis allows such jumps to relatively unrelated sequences, then, as a very rare event, such a jump might take place between a virus genome and unrelated cell mRNA which happened to be nearby. This kind of mechanism could lead from an endogenous virus to a sarcoma virus, since the src gene is apparently closely related to sequences expressed as mRNA in uninfected cells (Spector et al. 1978). As compared to recombination between viruses, such an event would be extremely uncommon, but the selection would be powerful. One virion with an src gene out of perhaps 10^{12} or more in a viraemic animal would manifest itself as a visible tumour for the curious scientist to select for laboratory propagation. Such a crossover would be much more frequent – although still rare – if there were a partial homology between viral and cell sequence. This would account for the apparently frequent reappearance of sarcoma virus after injection of virus with partial td deletions into chickens (Hanafusa et al. 1977). Note that this mechanism would place no constraint on where such an insert occurred in the genome and that the final location would be determined probably by chance small homologies in sequence. Also, the most frequent event would involve replacement of virus with cell information, leading to a defective transforming virus – the usual case.

Although copy-choice recombination is not a new proposal for RNA tumour viruses (Vogt, 1973), I believe the ability to use broken genomes to recombine and even to replicate efficiently is new. This model has some experimental tests to distinguish it from other possibilities. For one, unlike the model proposed by Hunter (1978), it would be possible,
under certain experimental conditions, to repair large deletions, although deletions with no intervening common sequence could not be repaired. Second, the progeny of cells infected with heterozygote viruses should yield either parental or (probably more frequently) recombination types, never heterozygous or partially heterozygous progeny, so long as multigenome heterozygotes are excluded. The third prediction is the relationship between intactness of the virion RNA and recombination frequency mentioned above.

The overall scheme of proviral DNA synthesis that would emerge from this model appears quite complex – the growing DNA chain is transferred not only twice from end to end of the genome, but also repeatedly from one genome to another before a complete (−) strand is synthesized. In spite of this complexity, however, such a mechanism would allow the virus to survive both physical damage to its genome RNA and to undergo rather rapid evolution by recombination. This would be an advantageous situation since it would combine in retroviruses some of the best features of both RNA and DNA viruses. The proposed schemes of replication and recombination also seem to tie together in a fairly neat package many of the apparently diverse features of retroviruses which distinguish them from all other groups of viruses.

I thank numerous workers, cited in the text, for the communication of unpublished material without which this review would be hopelessly out of date. I also want to thank P. Balduzzi, K. Conklin, P. Duesberg, M. Kotler, B. Mermer, S. Voynow, and P. Tschilis for constructive discussion and suggestions to improve the manuscript. Work in my laboratory is supported by Public Health Research Grant CA 17659 from the National Cancer Institute and Grant NPz46B from the American Cancer Society. I am the recipient of a Faculty Research Grant from the American Cancer Society.

Note added in proof. After submission of this article, it came to my attention that a similar mechanism for retrovirus recombination had previously been proposed (Haseltine and Baltimore, 1976).

REFERENCES


