

Site-specific inversion sequence of the herpes simplex virus genome: Domain and structural features

(nucleotide sequence/recombination specific sequence/inversion assay/direct and inverted repeats)

EDWARD S. MOCARSKI AND BERNARD ROIZMAN

The Marjorie B. Kovler Viral Oncology Laboratories, The University of Chicago, 910 East 58th Street, Chicago, Illinois 60637

Contributed by Bernard Roizman, July 27, 1981

ABSTRACT The genome of herpes simplex virus-1 consists of two covalently linked components, L and S, that invert relative to each other. The L and S components consist of unique DNA sequences bracketed by inverted repeats. The inverted repeats of the L component are designated *ab* and *b'a'* and those of the S component are designated *a'c'* and *ca*. The number of *a* sequences at the termini and at the L-S component junction varies from one to several copies. Insertion into the middle of the L component of a DNA fragment consisting of 156 base pairs (bp) of the *b* sequence, an entire *a* sequence of 501 bp, and 618 bp of the *c* sequence created a new site through which additional inversions in the genome occurred. Comparison of the nucleotide sequences of DNA fragments containing one and two *a* sequences defined the domain of the *a* sequence. The single *a* sequence consists of two 20-bp direct repeats (designated as DR1) bracketing a region that contains 19 tandem direct repeats of a 12-bp sequence (DR2) adjacent to three direct repeats of a 37-bp sequence (DR4), in addition to short stretches of unique sequences. The fragment with two tandem *a* sequences contained three copies of DR1—i.e., the intervening DR1 was shared by the two *a* sequences. Furthermore, one *a* sequence contained 22 copies of DR2 and two copies of DR4 whereas the second *a* sequence contained 19 copies of DR2 and two copies of DR4. These observations suggest that (i) amplification of the number of terminal and internal *a* sequences is the consequence of intramolecular or intermolecular recombination through DR1, (ii) the number of copies of DR2 and DR4 within the *a* sequence is not fixed and may vary as a consequence of unequal crossing over or slippage, and (iii) inversion results from intramolecular recombination between terminal and inverted *a* sequences.

The genome of herpes simplex virus (HSV)-1, a linear double-stranded DNA molecule with a M_r of 96×10^6 (1, 2) consists of two covalently linked components, L and S (Fig. 1). Each component consists of predominantly unique sequences (U_L or U_S) bracketed by large inverted repeats (4). The L component is bracketed by the sequence *ab* and its inversion *b'a'*, and the S component is bracketed by the sequences *a'c'* and *ca* (2). The *a* sequence is in the same orientation at both ends of the genome. It was operationally defined as the sequence that permits circularization of the genome on exposure of complementary cohesive termini by digestion of intact DNA with a processive exonuclease (5). The number of *a* sequences at the termini and at the junction between the L and S components may vary from one to several tandem copies (3, 6). A remarkable characteristic of the HSV-1 genome is that the L and S components invert relative to each other. Thus, the DNA extracted from virions consists of four equimolar populations of molecules differing only with respect to the relative orientations of the two components (7, 8).

The key questions concerning the inversion of the L and S components relative to each other are the identification of the requirements of the replication cycle satisfied by the inversions, the biochemical mechanism by which inversion is effected, and the nucleotide sequence through which inversion occurs. A model proposing a specific requirement satisfied by a function that obligatorily generates inversions has been published (9, 10). It was also reported elsewhere that insertion of a 4-kilobase pair DNA fragment spanning the L-S junction into the middle of the L component as a second junction resulted in new genome inversions through the inserted junction (11). Invariably, the DNA sequences that inverted were bracketed by inverted repeats, *ab-b'a'* or *a'c'-ca*. Inversions did not occur when DNA fragments from U_L or U_S regions were inserted in either direct or inverted orientation relative to their authentic counterparts. The results showed that HSV DNA contains site-specific inversion sequences and that only DNA flanked by inverted repeats of these sequences can invert.

In this paper, we report the nucleotide sequence of a 1275-base pair (bp) DNA fragment shown to contain the site-specific inversion sequence. By comparing the nucleotide sequences of junction fragments containing one and two *a* sequences, we determined the domain of the *a* sequence.

MATERIALS AND METHODS

Virus, Infection, and Transfection. The procedures for infection, plaque purification and titration of HSV-1(F⁺) in Vero cells have been described elsewhere (11). The transfection of rabbit skin cells with HSV DNA and subsequent selection by plaque purification of thymidine kinase-deficient virus with AraT (Raylo Chemical, Edmonton, Alberta, Canada) are also described elsewhere (11).

Experimental Design. The experimental design (11) had two steps. First, a DNA fragment spanning the L-S junction was inserted into the coding sequence of the thymidine kinase contained in the *Bam*HI Q fragment of HSV-1(F). Then, the cloned chimeric fragment was recombined into the HSV-1 genome by cotransfection of cells with intact HSV-1(F⁺) DNA. Thymidine kinase-deficient recombinant viruses carrying the insertion were selected for ability to grow in the presence of AraT (11).

Cloning the 1275-bp *Hae* II Junction Fragment Containing a Single *a* Sequence. The 1275-bp *Hae* II fragment (*Hae* II-a) inserted into the *Sac* I site of *Bam*HI Q (clone pRB103; see Fig. 2) was derived from the *Bam*HI SP₁ fragment of HSV-1(F) DNA (clone pRB115) (12). *Hae* II-a was subcloned in the plasmid vector pACYC177 (clone pRB601). To make the insertion (clone pRB359), 0.2 μ g of *Sac* I-digested pRB103 plasmid carrying the *Bam*HI Q fragment was mixed with 1 μ g of *Hae* II-digested pRB601 ethanol-precipitated DNA suspended in 30 μ l of 20

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: HSV, herpes simplex virus; bp, base pairs(s); DR1...DR6, direct repeats 1-6; IR1 and IR2, inverted repeats 1 and 2, respectively.

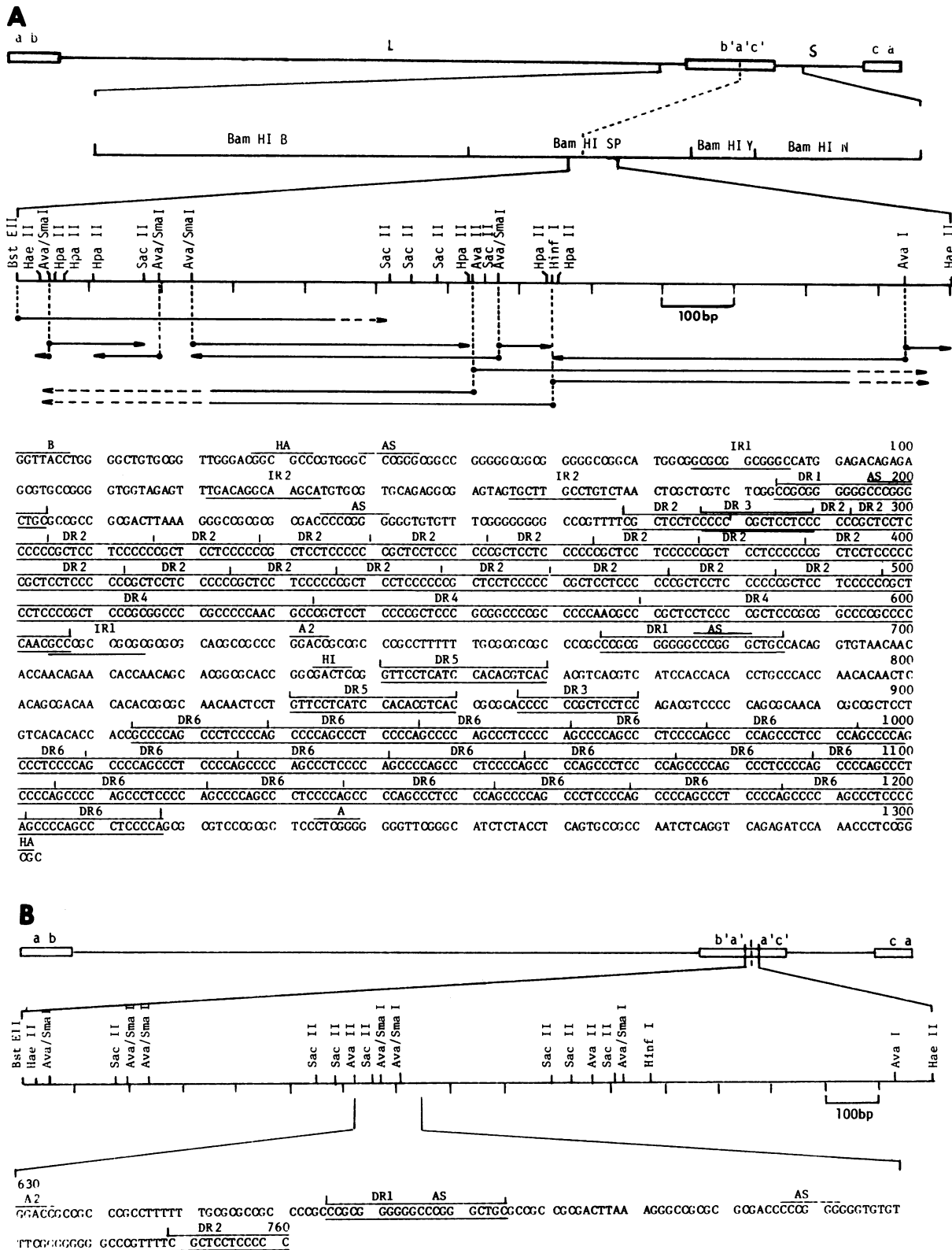


FIG. 1. Nucleotide sequence analysis of the *Hae* II junction-spanning fragment. (A) Nucleotide sequence of the *Hae* II-a fragment. The top line represents the HSV-1 genome, showing the inverted repeats as open boxes. Sequence *ab* and its inverted repeat each contain 6% of total DNA whereas *ca* and its inverted repeat each contain 4.3% of total DNA. The second line is a *Bam*HI map of the DNA region encompassing the L-S junction (3). The third line is a restriction enzyme map for the DNA region between the *Bst*EII site in the *b* sequence and the *Hae* II site in the *c* sequence. The lines below the map indicate the sites of initiation and the directions of nucleotide sequencing. The sequence shown includes the 28-bp *Bst*EII-*Hae* II fragment in the *b* sequence adjacent to the *Hae* II-a fragment. (B) Nucleotide sequence across the *a*-*a* junction in the 1686-bp *Hae* II-*aa* fragment. Below the restriction map of the region from the *Bst*EII site is the nucleotide sequence from the *Ava* II site at 630 bp through the junction of the two tandem *a* sequences. All 10-bp or larger perfect inverted repeats (IRs) are underlined and direct repeats (DRs) are both under- and overlined. B, *Bst*EII; HA, *Hae* II; A, *Ava* I; S, *Sma* I; A₂, *Ava* II; HI, *Hinf*I.

mM Tris·HCl at pH 7.6/10 mM MgCl₂/1 mM dithiothreitol/33 μM each dATP, dCTP, dGTP, and TTP containing 2 units of T4 polymerase (P-L Biochemicals) and incubated at 37°C for 30 min as described (13). Frozen C600SF8 were transformed (14) with ligated plasmids and screened as described (12). All restriction enzymes were obtained from New England BioLabs.

Cloning the 1686-bp *Hae* II Junction Fragment Containing Two *a* Sequences. The 1686-bp *Hae* II fragment (*Hae* II-aa) subcloned in pACYC177 (clone pRB602) was derived from the pRB104, a pBR322-derived clone of the *Bam*HI SP₂ fragment of HSV-1(F) DNA (12).

Nucleotide Sequence Determination. Plasmid clones pRB115 or pRB104 were digested with *Bst*EII and then labeled by using T4 polynucleotide kinase as described (15) except that the reaction volume was 50 μl and included 1 mCi (1 Ci = 3.7 × 10¹⁰ becquerels) of crude aqueous [γ -³²P]ATP (ICN, Irvine, CA) added directly to the reaction mixture. After incubation at 37°C for 30 min, the DNA was ethanol precipitated, digested to completion with *Rsa* I, and electrophoretically separated on a 3.5% polyacrylamide gel, and its sequence was determined (16). Additional sequence determination was done on *Hae* II-a and *Hae* II-aa fragments after they were end labeled at *Ava* I, *Ava* II, or *Hinf*I sites and either digested with a second restriction enzyme or strand separated (16) as depicted in Fig. 1.

Hybridization to Immobilized Restriction Enzyme Fragments. The methods for transfer of electrophoretically separated DNA fragments to nitrocellulose and hybridization ³²P-labeled plasmid probes are reported elsewhere (12).

RESULTS

Inversion Through the *Hae* II-a Fragment. The 1275-bp *Hae* II fragment was inserted into the unique L sequences as illustrated in Fig. 2. On the basis of previous studies (11), if the inserted *Hae* II-a fragment were to cause additional inversions, it would be expected that the viral progeny DNA would consist of 12 isomers instead of the 4 present in wild-type virus. Of the 12, 6 are depicted in the form of *Bam*HI cleavage maps in Fig. 2. The remaining six can be produced by inverting the S component. Because the *Hae* II fragment was inserted into *Bam*HI Q, the *Bam*HI digests of R359 DNA would be expected to contain a new fragment, designated as *Bam*HI IIΦ, in place of *Bam*HI Q. If all predicted isomers were present, two new junction fragments and two new terminal fragments would be formed. The new junction fragments would be *Bam*HI ΦP and *Bam*HI SII, derived from inversion of the L_B segment of the L component. The predicted new terminal fragments would be *Bam*HI Φ and *Bam*HI II. The terminal *Bam*HI Φ could arise in two steps—i.e., inversion of the entire L component followed

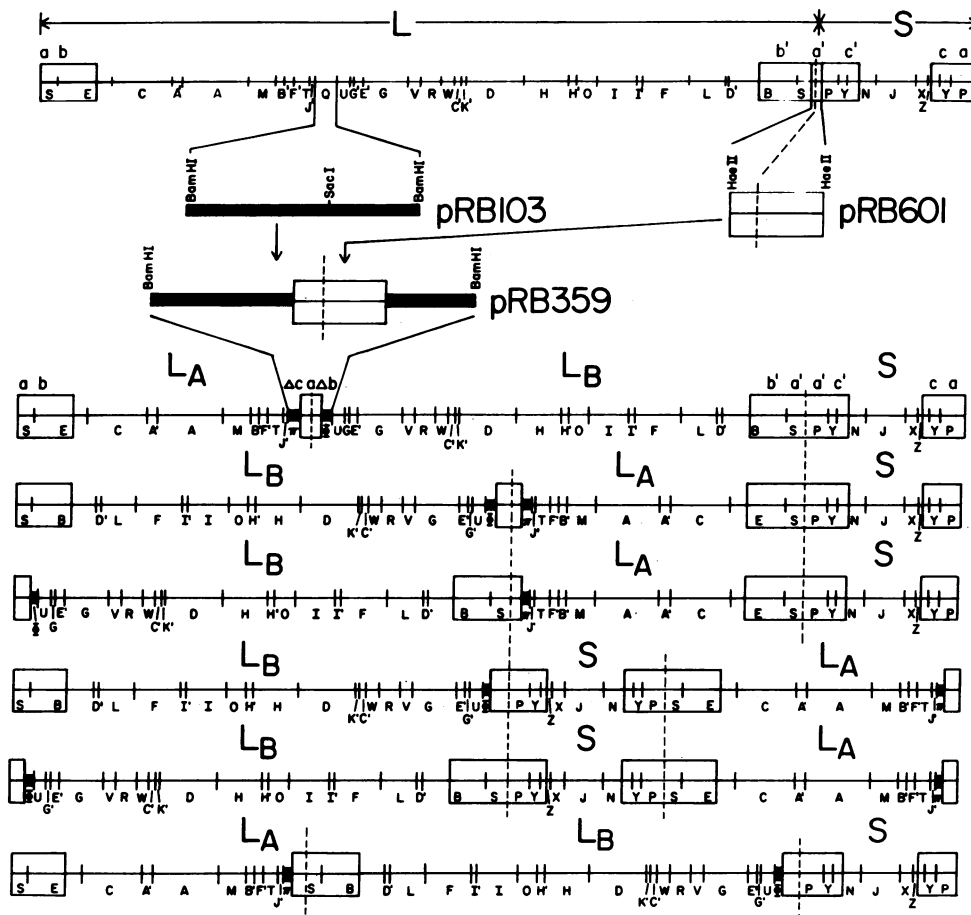


FIG. 2. Construction and DNA sequence arrangements in recombinant R359 composed of the *Hae* II-a fragment inserted in the thymidine kinase gene. (Upper) Construction of the *Hae* II-a insertion. *Bam*HI restriction endonuclease map of the isomer of HSV-1(F) DNA followed by a diagram of the *Bam*HI Q fragment (clone pRB103) showing the position of the *Sac*I site into which the *Hae* II-a junction fragment (clone pRB601) was inserted. The chimeric fragment was recombined into HSV-1(F⁺) DNA. (Lower) New genome inversions in R359. Insertion of a second junction in the L component creates new segments, L_A and L_B. L_B and the combination of L_A and S constitute new components that can invert. For diagrammatic purposes, only *Bam*HI restriction endonuclease maps of the isomers that could arise by inversions of the L_B, L_AS, and L components are shown here. The remaining six isomers can be produced by inverting the S component. The new junction fragment resulting from insertion of the *Hae* II-a fragment in *Bam*HI Q is designated *Bam*HI IIΦ. Any inversion that occurs will give rise to unique terminal *Bam*HI fragments (II or Φ) or junction fragments (IIP and ΦS).

by inversion of the L_B segment. The terminal *Bam*HI II could also arise in two steps—i.e., inversion of the entire L component followed by inversion of the component containing both the L_A segment and S component. The new *Bam*HI fragments can readily be identified by size and sequence because they should contain both the *Bam*HI Q sequences and the inserted *Hae* II fragment sequences. It is further expected that, if the *a* sequence contained in the *Hae* II-a fragment were amplified by tandem reiteration as it normally occurs in wild-type DNA, all of the new junction and terminal fragments containing at least a portion of the *b* sequence would be present in multiple forms differing in a number of *a* sequences and in size by 450-bp increments (3, 6).

To test for these predictions, electrophoretically separated fragments were transferred to nitrocellulose strips and hybridized to 32 P-labeled *Bam*HI Q and *Hae* II-a fragment DNA. As shown in Fig. 3, all predicted bands were present. The *Bam*HI II Φ fragment, resulting from insertion of the *Hae* II-a fragment into the *Bam*HI Q fragment, was present in a family of fragments 4600, 5050, and 5500 bp long. In addition, new *Bam*HI fragments resulting from inversion through the new junction were also present. The *Bam*HI SII junction fragment was represented in a family of fragments 5800, 6250, and 6700 bp long. The *Bam*HI Φ P junction fragment was represented in a family of fragments 4900 and 5350 bp long. The *Bam*HI II terminal fragment was present as a fragment 3200 bp long and the *Bam*HI Φ terminal fragment was present in a family of fragments 1800 and 2250 bp long. The bands within a family differ by \approx 450-bp increments due to amplification of *a* sequences at the *ab* terminus and at *bac* junctions (3, 6, 11).

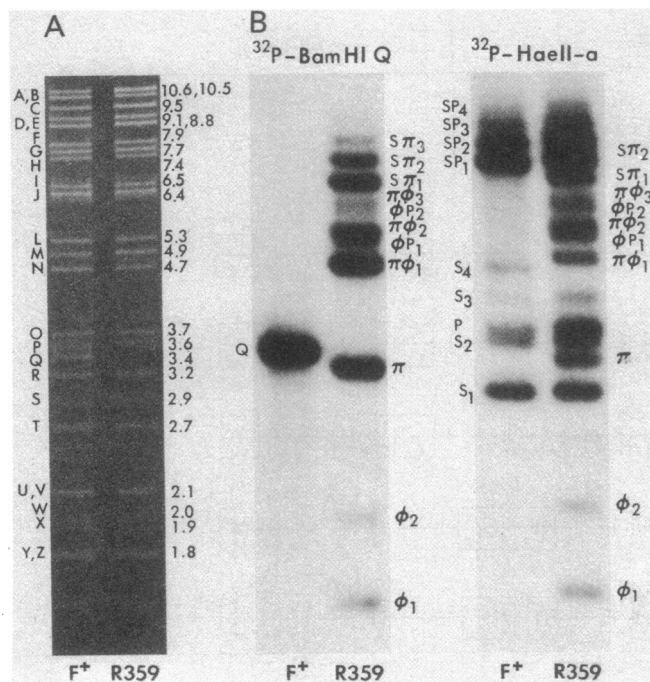


FIG. 3. Analysis of R359 DNA for inversion. (A) Stained *Bam*HI fragments of HSV-1(F⁺) and R359 DNA. (B) Autoradiogram of electrophoretically separated *Bam*HI digests of HSV-1(F⁺) and recombinant R359 DNA immobilized on nitrocellulose and hybridized with 32 P-labeled *Bam*HI Q or *Hae* II-a junction fragment probes. *Bam*HI fragments of HSV-1(F⁺) and R359 DNA were subjected to electrophoresis in parallel and hybridized with either *Bam*HI Q or *Hae* II-a probe as indicated. All standard HSV-1(F⁺) *Bam*HI fragments are designated as described (3) and their sizes are given in kilobase pairs. New fragments arising after insertion of the *Hae* II junction fragment into *Bam*HI Q are designated as described in Fig. 2.

Nucleotide Sequence of the *Hae* II-a Fragment. The strategy used in determining the sequence of the *Hae* II-a fragment spanning the L-S junction is illustrated in Fig. 1A. To orient the fragment, the sequencing was begun from the *Bst*EII cleavage site known from previous studies to be in the *b* sequence to the left of the *Hae* II-a fragment (11, 12).

The 1275-bp *Hae* II-a fragment has a base composition of 76.9% (mol/mol) G + C asymmetrically distributed; thus, the strand shown in Fig. 1 contains 54.6% (mol/mol) cytosine. All of the perfect repeats >10 bp long are indicated in Fig. 1 as DR (direct repeats) or IR (inverted repeats). Most striking are two regions of long tandem arrays of direct repeats. The first includes 19 copies of a 12-bp sequence (DR2) followed immediately by three copies of a 37-bp sequence (DR4), and the second includes 19 copies of a 16-bp sequence (DR6). In addition, there are other perfect direct repeats (DR1, DR3, and DR5), as well as inverted repeats (IR1 and IR2) within this fragment. DR3 is especially noteworthy because it spans adjacent copies of DR2 repeats shown below to be part of the *a* sequence and is also present at positions 856–870 shown below to be part of the *c* sequence. The *Hae* II-a fragment also contains numerous repeats <10 bp long as well as longer repeats with single base-pair mismatches not indicated in Fig. 1.

Determination of the *b*, *a*, and *c* Sequences Within the *Hae* II-a and *Hae* II-aa Junction Fragments. The strategy used for the determination of the domains of the *b*, *a*, and *c* sequences was based on the observation that junction fragments are heterogeneous in that they contain one to several *a* sequences arranged as *bac*, *baac*, *baaac*, etc. The precise nucleotide sequence of the *a* sequence can then be defined by comparing the nucleotide sequence of the 1275-bp *Hae* II-a fragment containing one *a* sequence with that of the 1686-bp *Hae* II-aa fragment containing two *a* sequences. The results (Fig. 1) showed the following.

(i) Both *Hae* II-a and *Hae* II-aa contain a single *Hinf*I cleavage site within the domain of the *c* sequence. Whereas the *Hae* II-a fragment contains a single *Ava* II cleavage site between DR4 and DR1, *Hae* II-aa contains two *Ava* II sites—i.e., one within the leftmost *a* sequence and one within the rightmost *a* sequence. The nucleotide sequence from the leftmost *Ava* II site across the *a-a* junction (Fig. 1B) establishes that the *a* sequence begins and ends with DR1. *Hae* II-aa, with two *a* sequences, contains only three copies of DR1, indicating that adjacent *a* sequences share the intervening DR1.

(ii) The left *a* sequence in the *Hae* II-aa fragment contains 22 copies of DR2 followed by two copies of DR4. The right *a* sequence contains 19 copies of DR2 followed by two copies of DR4. As noted above, the *Hae* II-a fragment contains a single *a* sequence with 19 copies of DR2 and three copies of DR4.

DISCUSSION

Elsewhere, we reported that HSV DNA contains inversion-specific sequences (11). In this paper, we report on a 1275-bp *Hae* II-a fragment and a 1686-bp *Hae* II-aa fragment that contain site-specific inversion sequences. The junction-spanning fragment is G + C rich and contains numerous direct and inverted repeats, both dispersed and in tandem arrays. The *Hae* II-a fragment contains 156 bp derived from the *b* sequence, an entire *a* sequence of 501 bp, and 618 bp derived from the *c* sequence. Because the inverting components must be flanked by either *ab* and *b'a'* or *a'c'* and *ca*, the inversion-specific sequence must be in both the *b* and the *c* sequences or only in the *a* sequence. The absence of homologous sequences in the *b* and *c* sequences contained in the *Hae* II-a fragment indicates that either *b* and *c* contain distinct site-specific inversion sequences or, as previously concluded (11), the site-specific inversion sequence is

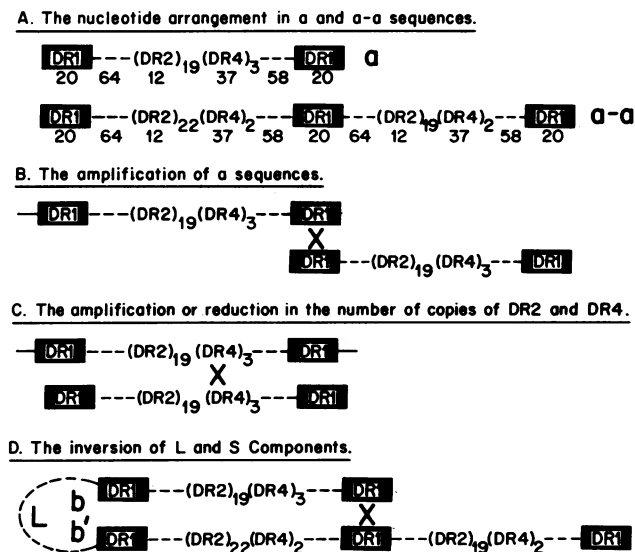


FIG. 4. Diagram of sequence arrangement and recombinational events that could give rise to amplification of the *a* sequence, variable number of copies of DR2 and DR4, and inversion of the L or S component (illustrated for the L component). Subscripts indicate numbers of copies of repeats and numbers below the schematic representation in A indicate numbers of base pairs in each repeat unit and in unique sequences (---).

contained in the *a* sequences only.

The *a* sequence is of particular interest for several reasons. (i) It appears to be involved in circularization of DNA during replication (4). (ii) It varies as a result of amplification or reduction from one to several copies when it is juxtaposed next to the *b* sequence in HSV-1(F) DNA (3, 6). (iii) It probably contains the site for cleavage of head-to-tail concatemers generated during replication of HSV DNA (9, 10). (iv) Finally, it contains the site for inversion (11).

Two observations concerning the structure of the *a* sequence and adjacent *b* and *c* sequences are of particular significance. First, in the *Hae* II-*a* fragment, the *a* sequence is bracketed by DR1. By comparison with the *Hae* II-*aa* fragment, the number of DR1 repeats equals the number of *a* sequences plus one; thus, the intervening DR1 is shared by adjacent *a* sequences (Fig. 4). This observation suggests that amplification in the number of *a* sequences occurs by recombination through DR1 (Fig. 4B) whereas reduction could occur by recombination between the *a* sequences. In addition, the variability in the number of copies of DR2 and DR4 can be explained by unequal crossover during recombination events involving the *a* sequence (Fig. 4C). Finally, recombination between the terminal and internal *a* sequences or through DR1 would produce inversion of the component bracketed by the *a* sequences (Fig. 4D). Second, there are numerous other direct and inverted repeats throughout the *bac* sequence within *Hae* II-*a*. The *a* sequence contains two tandem arrays, DR2 and DR4, and the *c* sequence contains a single tandem array, DR6. Other direct and inverted repeats are contained entirely within the *a*, *b*, or *c* sequences (IR2, DR5)

or in both *b* and *a* or *a* and *c* (DR3, IR1). The high frequency of recombination that gives rise to inversion could require specific proteins that recognize the *a* sequence and bring together the terminal sequence in apposition to the inverted repeats. A possible function of tandem arrays of repeats is as sites for binding of viral proteins. This hypothesis can be tested by the isolation of specific binding proteins and by determining the biological activity of *a* and *c* sequences in which the repeats have been experimentally modified.

HSV DNA is apparently replicated by a rolling circle mechanism that gives long concatemers (9, 10); these are cleaved to generate unit-length HSV DNA. On the basis of the available sequence, the circularization could occur by annealing of single-stranded exposed complementary sequences of either the entire *a* sequence or of DR1 by itself. In either case, the new junction formed during circularization would not differ in nucleotide sequence from the *bac* and *baac* L-S component junctions sequenced in this study. The putative cleavage site for generation of unit-length HSV DNAs is not apparent from the *bac* and *baac* nucleotide sequences. However, because the HSV-1(F) strain contains only a single *a* sequence at the right terminus of the DNA, the cleavage site must be either within or immediately outside the *a* sequence adjacent to the *c* sequence. In either case, one *a* sequence, either from the L component or from the S component, would have to be repaired to regenerate a complete *a* sequence following replication. A model for such repair has been published elsewhere (9, 10).

We thank Susanna Rudofsky for excellent technical assistance. These studies were aided by grants from the National Cancer Institute (CA 08494 and CA 19264) and the American Cancer Society (ACS MV-20). E. S. M. is a U. S. Public Health Service Postdoctoral Trainee (CA 09241-03).

- Kieff, E. D., Bachenheimer, S. L. & Roizman, B. (1971) *J. Virol.* **8**, 125-132.
- Wadsworth, S., Jacob, R. J. & Roizman, B. (1975) *J. Virol.* **15**, 1487-1497.
- Locker, H. & Frenkel, N. (1979) *J. Virol.* **32**, 429-441.
- Sheldrick, P. & Berthelot, N. (1975) *Cold Spring Harbor Symp. Quant. Biol.* **39**, 667-678.
- Wadsworth, S., Hayward, G. S. & Roizman, B. (1976) *J. Virol.* **17**, 503-512.
- Wagner, M. M. & Summers, W. C. (1978) *J. Virol.* **27**, 374-387.
- Hayward, G. S., Jacob, R. J., Wadsworth, S. C. & Roizman, B. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 4243-4247.
- Delius, H. & Clements, J. B. (1976) *J. Gen. Virol.* **33**, 125-134.
- Jacob, R. J., Morse, L. S. & Roizman, B. (1979) *J. Virol.* **29**, 448-457.
- Roizman, B. (1979) *Cell* **16**, 481-494.
- Mocarski, E. S., Post, L. E. & Roizman, B. (1980) *Cell* **22**, 243-255.
- Post, L. E., Conley, A. J., Mocarski, E. S. & Roizman, B. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 4201-4205.
- Wartell, R. M. & Reznikoff, W. S. (1980) *Gene* **9**, 307-319.
- Morrison, D. A. (1977) *J. Bacteriol.* **132**, 349-355.
- Berkner, K. L. & Folk, W. R. (1977) *J. Biol. Chem.* **252**, 3176-3184.
- Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499-560.