A tandemly-oriented late gene cluster within the vaccinia virus genome

Scott L. Weinrich and Dennis E. Hruby

Center for Gene Research, Department of Microbiology, Oregon State University, Corvallis,
OR 97331-3804, USA

Received 1 November 1985; Revised and Accepted 3 March 1986

ABSTRACT

The nucleotide sequence of a 5.1 kilobase-pair fragment from the central portion of the vaccinia virus genome has been determined. Within this region, five complete and two incomplete open reading frames (orfs) are tightly-clustered, tandemly-oriented, and read in the leftward direction. Late mRNA start sites for the five complete orfs and one incomplete orf were determined by S1 nuclease mapping. The two leftmost complete orfs correlated with late polypeptides of 65,000 and 32,000 molecular weight previously mapped to this region. When compared with each other and with sequences present in protein data banks, the five complete orfs showed no significant homology matches amongst themselves or any previously reported sequence. The six putative promoters were aligned with three previously sequenced late gene promoters. While all of the nine are A-T rich, the only apparent consensus sequence is TAA immediately preceding the initiator ATG. Identification of this tandemly-oriented late gene cluster suggests local organization of the viral genome.

INTRODUCTION

Vaccinia virus (VV), the prototype of the orthopoxviruses, is characterized by a complex morphology; a large, cross-linked, double-stranded DNA genome; and a cytoplasmic site of replication (1). The technical advantages offered by cytoplasmic replication have facilitated investigations of several aspects of the VV life cycle, some of which may parallel related processes in higher cells. One of these aspects that has received particular attention is the regulation of VV gene expression. In addition to the basic biological relevance of elucidating genetic regulatory mechanisms, the continued development of VV as a eukaryotic expression vector and recombinant vaccine relies on the availability of such information (2-5).

With the capacity to encode approximately 200 average-sized polypeptides, the 185-kilobase-pair (kbp) genome of VV is expressed under tight temporal control in infected cells (6). Within minutes after infection of a susceptible host cell, early genes are transcribed by a viral RNA polymerase contained in the virion (7). About an hour later, with the onset of viral DNA

replication, late genes are expressed, and the expression of most early viral genes is repressed. The molecular mechanisms underlying modulation of the early and late modes of viral gene expression remain unknown. An initial step towards identifying control elements is to locate, compare, and dissect sequences which may be responsible for promoting and attenuating early and late VV gene transcription.

Preliminary transcriptional and translational maps of the VV genome suggested an absence of global organization (8-10). As representatives of the early or late classes, individual genes, particularly those strongly expressed, have been targeted for closer analysis. An increasing number of early and late genes have been mapped, and the nucleotide sequences of several of these have been reported (11-18). Sequences upstream of both early (19,20) and late genes (15,16) have been shown to contain cis-acting regulatory signals. These promoters of VV genes show little or no consensus to other eukaryotic or prokaryotic promoters, and are apparently unique.

We previously reported the genomic location of a 65,000 molecular-weight polypeptide-encoding late gene (L65), as determined by hybrid-selected translation and S1 nuclease mapping (21). The L65 polypeptide was shown to be a major viral product late in infection that does not appear to be associated with the mature virion. Recently, rifampicin resistance was mapped to the L65 gene (22). In the presence of rifampicin, VV replication is arrested late in morphogenesis, and precursors of virion structural proteins are not processed to their mature forms (23). This suggests that the L65 polypeptide may be involved in virion assembly. We now report the nucleotide sequence of 5.1 kbp of DNA which contains and flanks the L65 gene locus. The open reading frames (orfs) found correlate closely with transcriptional and translational mapping results. Identification of a tight cluster of late genes all read in the same leftward direction suggests local organization of the VV genome.

MATERIALS AND METHODS

A recombinant plasmid containing the HindIII D fragment of VV DNA, strain WR, was originally obtained from B. Moss. Purified VV DNA was digested with SalI and ligated to SalI-digested pUC13 to obtain a clone of the SalI J fragment, which spans the HindIII D-A junction. Subfragments of HindIII D or SalI J were cloned into pUC13. Details of these plasmid constructions have been described (21). Selected VV sequences were recloned from pUC13 derivatives into M13 mp18 and M13 mp19 phage. DNA sequencing was by the dideoxynucleotide chain termination method (24), or by the chemical sequencing

method (25). Enzymes used for construction of recombinant plasmids or phage, and all materials used in sequencing reactions were obtained from Bethesda Research Laboratories, New England Nuclear Corp., and Aldrich Chemical Company, Inc. Isolation of VV late RNA, and S1 nuclease reactions were performed exactly as described in detail (21). Sequences were analyzed (26) and compared for homology with the National Biomedical Research Foundation (NBRF) Protein Data Bank using Microgenie software obtained from Beckman Instruments, Inc. on an IBM personal computer.

RESULTS

Nucleotide Sequence

HindIII and SalI restriction enzyme digests of the VV genome result in manageable numbers of conveniently-sized DNA fragments. The complete HindIII map and the location of a specific SalI fragment of the VV genome are shown in Fig. 1. The 10 kbp SalI J fragment spans the HindIII site between the 16 kbp HindIII D fragment and the 50 kbp HindIII A fragment. With respect to the HindIII D-A junction, an SstI site 1647 nucleotides (nt) to the right and an EcoRI site 3468 nt to the left define the limits of the 5115 nt segment that was sequenced.

After extensive restriction mapping analysis, selected fragments were cloned into M13 mp18 and/or M13 mp19 for dideoxynucleotide chain termination sequencing (Fig. 3). The sequence that was determined is shown in Fig. 2. For most regions (85%), the sequence of both strands was determined. Several areas were confirmed by chemical sequencing (e.g., Fig. 4).

When translated in all three reading frames in the rightward direction (from EcoRI to SstI), there were no significant orfs. Translation in the

Figure 1. Restriction maps of vaccinia DNA. HindIII fragments are designated A to O according to size. The SalI J fragment is expanded, a partial restriction map is shown, and the limits of the 5115 nucleotide fragment that was sequenced are indicated.

Sst I

GAGGCTCTCAAAACCCAGTACAAGAAATATCAATGGATTGACAGGATGGTTTTTATGAATGGTTGAATGGCCAGCGTTTCAGTATTTGCTGCCTCAAAACA
 AlaLeuLysAsnGlnTyrLysLysLeuSerMetAspSerAspAspGlyPheTyrGluTrpLeuAsnGlyAspGlySerValPheAlaAlaSerLysGln
 200
 GCAATGTTGATGAATCAGCTTGCTAACTTGACGACGATCTTCTAACTATGGAAGAAGCCATGCGATGTTTCGAGACATTGTGTATCTTAATTTAT
 nGlnMetLeuMetAsnHisValAlaAlaAsnPheAspAspAspLeuLeuThrMetGluGluAlaMetSerMetHisSerMetHisCysCysIleLeuIleTyr
 300
 GCACAGGATTATGATCAATATATTAGCGCTAGACATATTACAGAACTATTTTAGATTATGATATTTAAATGAGTTGGTAGGAAAAATAACACTGTACT
 AlaGlnAspTyrAspGlnTyrIleSerAlaArgHisIleThrGluLeuPheEndMetSerTrpTyrGluLysTyrAsnIleValLe
 400
 GAAGTGGCTTAAGCGGTGTTCTTTGCGATGTCGGGATAATTTAACTACTATATTGGCGGAAGCGGTAACCATATTAGGCGGACTCTTATTACAGAGCCC
 uLysSerProLysArgCysSerPheAlaCysAlaAspAsnLeuThrThrIleLeuAlaGluAspGlyAsnHisIleArgAlaIleLeuTyrSerGlnPro
 500
 AAAAACTAAAAATATTACAGGATTTCTGGCAAGCTCTAGAAATAAAATGTTTTATATAAAATATTGGACGACGAGATACGTAGAGTGTAACTGAA
 LysLysLeuLysIleLeuGlnAspPheLeuAlaThrSerArgAsnLysMetPheLeuTyrLysIleLeuAspAspGluIleArgArgValLeuThrEnd
 MetAs
 600
 TCTACGATTATGACGGTTGTAGACACAAAGGTATAGTATCTGAACAAGGATATGAATATGTTATTTTGGCAGTCGGTATTTCAAAAAATGACAAAG
 nLeuArgLeuCysSerGlyCysArgHisAsnGlyIleValSerGluGlnGlyTyrGluTyrCysIlePheCysGluSerValPheGlnLysCysThrLys
 700
 GTACAAAAAAGTCAAACTTCATGTGCTTAATAAACTTATTCATTGAGAAATGTATTGGCGGAGATTATGCTCTCATCAATGTTCTGGAGAAATATCT
 ValGlnLysLysSerAsnPheHisIleSerAsnLysLeuIleHisLeuArgAsnValLeuArgArgLeuLeuSerHisGlnCysSerGlyGluIleIleS
 800
 CGAACTCTTGGATATCATGGAATAAAATCAATATCCACGGATGATGATAGTCAAAATTTGTATCTAGTTTCTTAAGCCTAACGAGAGATAAATAA
 erGluLeuLeuAspIleMetGluLysAsnGlnIleSerThrAspAspValAspAlaAsnPheValSerSerPheLeuLysAlaAsnGluArgIleAsnLys
 900
 AAAGGATTATAAGTACTCTTGAATAATCAATCAAGTAAAGATGAGAACTGAATCTGACTACAGAAAAGATTAACTCAAGTAGTAAAAATATTAG
 sLysAspTyrLysLeuValPheGluIleIleAsnGlnValLysAspGluLysLeuAsnLeuSerThrGluLysIleAsnGluValValGluIlePheLys
 1000
 CACTTGGTATCTTTTGGCAAGAAACACTCCTCTAAGACAATTAATTATTCTTTTGGATAAAATATTGGATATCACTCTGTAACATAAAATC
 HisLeuValPhePheCysGlnGluAsnThrProSerLysThrIleAsnTyrSerPhePheLeuAspLysIlePheAspIleThrSerValThrLysAsnL
 1100
 TAAACCTCAAACTGTTAAAAATATACGAAAAATATAGTAACCAATTAGTATGGGAAACCTTTTAGCACATATGAGACTCAAAAAACCTGTAACCTAT
 eLysProGlnThrValLysAsnTyrThrLysAsnAsnSerAsnGlnLeuValTrpGluAsnPheLeuAlaHisMetArgSerLysLysArgValThrMe
 1200
 GGTAGAGGATTATGGACACGACTATGTTTTGTAGATGACAGGTTTTCTACTTGCTCATTAGAAGTATAAAAAATAGTTCGGTATTAATGGCTAAGC
 tValGluAspTyrGlyHisGluTyrValPheValAspGluArgPheSerThrCysSerLeuGluValEndMetAlaLysA
 1300
 GAGTAAGCGCTTCAGATGCTGGTTATTTGACGACCTAAAGCAGTTTTTAAGCCCGCTAAAGAAGAACCTCGCTGTATACCAAGATTATATAAATC
 rgValSerLeuProAspValValIleSerAlaProLysAlaValPheLysProAlaLysGluGluAlaLeuAlaCysIleLeuProLysTyrTyrLysSe
 1400
 TATGGCAGATGTGTCTATTAAAGCAAACTAGCTAAATGATAAGTGTGGTTTTGTAATCAAGATTGGTTTTTAAACCTATTAGTATTGAGACATTCAAG
 rMetAlaAspValSerIleLysThrAsnSerValIleAspLysCysTrpPheCysAsnGlnAspLeuValPheLysProIleSerIleGluThrPheLys
 1500
 GGTGGTCAAGTTGGGTATTTCTGTTCAAAATATGATAGGCAATCGTTGGCGTCTATGGTTAAGTCTCAGCTGCTCTAGAGAAGAACCAAAATTTCTT
 GlyGlyLeuValGlyTyrPheCysSerLysIleCysArgAspSerLeuAlaSerMetValLysSerHisValAlaLeuArgGluGluProLysIleSerL
 1600
 GTTGGCTTTAGTATCTATGAAGTAAAGGAAAGGTCAATAATACAATAAACTACTAAGAGATAAAGACGGCGTTTACGGAAGCTGTACTTTAAGGA
 euLeuProLeuValPheTyrGluAspLysGluLysValIleAsnThrIleAsnLeuLeuArgAspLysAspGlyValTyrGlySerCysTyrPheLysGln
 1700
 AAATGACAAATATAGATATTTCTCTACGGAGTTTATGTAAGCTTTTCCATTTTAAATAGAAATGAATAACTACTATTAACTTTCTGATGGGGT
 uAsnSerGlnIleIleAspIleSerLeuLeuEndMetAsnAsnThrIleIleAsnSerLeuLeuGlyG
 1800
 GGGATGACTCTATAAAGCGTCTAATGTCTTCCGACGTAGTCAAACTTCAACTTTATATATGCCGCAATATATTTCTCTATCGGAGTTATGACAA
 lyAspAspSerIleLysArgSerAsnValPheAlaValAspSerGlnIleProThrLeuTyrMetProGlnTyrIleSerLeuSerGlyValMetThrAs
 1900
 CGATGGTCAGACAATCAGGCTATCGCTAGCTTCGAAATTAGCGATCAGTATATTACTGGCTTAATCATTGGTTCTGAGTTTGGAACTCCAGAAATT
 nAspGlyProAspAsnGlnAlaIleAlaSerPheGluIleArgAspGlnTyrIleThrAlaLeuAsnHisLeuValLeuSerLeuGluLeuProGluVal
 2000
 AAAGGTATGGGAAGATTGGTTACGTACCATATGTTGGATATAAATCTATTAATCAGTATCTATCTCTCGTGAACCGGTGTTATTGGGAAATTCAGG
 LysGlyMetGlyArgPheGlyTyrValProTyrValGlyTyrLysCysIleAsnHisValSerIleSerSerCysAsnGlyValIleTrpGluIleGluG
 2100
 GCGAAGAATTATATAAATTTGATCAATAACAAATGCTTTGAAACACTCTGGATATTCTAGTGAACCTTAATGATATTTCTATTGGCCTAACTCGCTAA
 lGlyGluLeuTyrAsnAsnCysIleAsnAsnThrIleAlaLeuLysHisSerGlyTyrSerSerGluLeuAsnAspIleSerIleGlyLeuThrProAs
 2200
 TGACACTATTAAAGAACCATCTACAGTATACGTTTATATATAAACTCCGTTTGTATGTCGGAAGATACATTGACGAGCTTAAACTATCCGATTCAAAAT
 nAspThrIleLysGluProSerThrValTyrValTyrIleLysThrProPheAspValGluAspThrPheSerSerLeuLysLeuSerAspSerLysIle
 2300
 ACCGTAAACCGTAACTTCAATCCAGTATCCGATTCGTTATTCGTGACTCTCGTTGACCTTTGAAACCGTTGCAACAAAGATTGTTTATGTTCTCGAAT
 ThrValThrValThrPheAsnProValSerAspIleValIleArgAspSerSerPheAspPheGluLysLeuGluPheAsnLysGluPheValTyrValProGluL
 2400
 TGAGCTTATGGATATATGCTTAAGAATGTACAAATTAACCATCATTTATAGAGAAACCTACGAGAGTAATAGGTCAATAAACCAACCAACCGGGAC
 euSerPheIleGlyTyrMetValLysAsnValGlnIleLysProSerPheIleGluLysProArgArgValIleGlyGlnIleAsnGlnProThrAlaTh
 2500
 TGTAAGTGAAGTTCATCGCGCAACATCGCTCTCTGTTTAACTAAACCTTATTATGGAAATACGGATAATAAATTTATTTCGTATCCAGGGTACTCAACA
 rValThrGluValHisAlaAlaThrSerLeuSerValTyrThrLysProTyrTyrGlyAsnThrAspAsnLysPheIleSerTyrProGlyTyrSerGln
 2600
 GATGAAAAAGATTATATAGATGCATATGTGAGTACATGTTGGATGATCTAGTTATGTTAGGAGTGGTCCACCGGATGGTTATCCGAGCTCTGCCGAGA
 AspGluLysAspTyrIleAspAlaTyrValSerArgLeuLeuAspAspLeuValIleValSerAspGlyProProThrGlyTyrProGluSerAlaLysI
 2700
 TCGTAGAGGTTCCAGAAGATGCTATCGTTTCTATCAAGATGCTGATGCTATGTAATAAATGATAATGCTCGTGAATATAGCTGTTTATCTTGATAC
 leValGluValProGluAspGlyIleValSerIleGlnAspAlaAspValTyrValLysIleAspAsnValProAspAsnMetSerValTyrLeuHisTh
 2800
 TAATCTGCTAATGTTTGAACACGAAAAATCTTTTATATATAACATTCTAAAAAGTTTCCGCCATTACTGGAACATATAGTATGCCACTAAGAGA
 rAsnLeuLeuMetPheGlyTyrArgLysAsnSerPheIleTyrAsnIleSerLysLysPheSerAlaIleThrGlyThrTyrSerAspAlaThrLysArg

2900

ACAATCTTTGCTCACATATCACATAGTATCAACATCGATACATCTATTCTGTAACTCTTGGACTAGTCAACGTCTATAACGGAGATAATA
 ThrIlePheAlaHisIleSerHisSerIleAsnIleIleAspThrSerIleProValSerLeuTrpThrSerGlnArgAsnValTyrAsnGlyAspAsnA
 3000
 GATCAGCGGAATCAAGGCGAAGGATTGTTCATTAACGATCCCTTCATCAAGGGAATAGATTTTAAAGAATAAGACCGATATTATTTCTAGACTAGAACT
 rgSerAlaGluSerLysAlaLysAspLeuPheIleAsnAspProPheIleLysGlyIleAspPheLysAsnLysThrAspIleIleSerArgLeuGluVa
 3100
 TAGATTGGAAATGATGTTCTATATTACAGAAACGGACCATCTCGAGAAATTTATAATGAACACTGACAAAAAGGAATAATGGAAAGAAGCCCTAACT
 lArgPheGlyAsnAspValLeuTyrSerGluAsnGlyProIleSerArgIleTyrAsnGluLeuLeuThrLysSerAsnAsnGlyThrArgThrLeuThr
 3200
 TTTAACTTTACACGAAGATATTCTTTAGGCGGCAACTATTACGGCTAATGTATCTAGGGGGAAGATAAATACTCTGTCGAGTAGTTTATTCGACCA
 PheAsnPheThrProLysIlePhePheArgProThrThrIleThrAlaAsnValSerArgGlyLysAspLysLeuSerValArgValValTyrSerThrM
 3300
 TGGATGTCACCATCGAATCTATTATGTACAAAAACAATGGTACTGTATGTATGACCTGTATAAGGTACTTACGATCAAGGGGTAACTATTACGAA
 etAspValAsnHisProIleTyrTyrValGlnLysGlnLeuValValValCysAsnAspLeuTyrLysValSerTyrAspGlnGlyValSerIleThrLys
 3400
 GATTATGGGAGATAATACTAATAATAATGAAACAACTATAGACTCTTAAATGGATGAAATGTAAAAATATCGGGAGGGAAGCCATGCTCTCTT
 sIleMetGlyAspAsnAsnEndEndEndEndMetAspGluIleValLysAsnIleArgGluGlyThrHisValLeuLeu
 3500
 CCATTTTGAACACTGCGGAGAATTAATCTGCTCTAGTAAAGCCCATACCTACTGCTGGAATACGGAGCTAATTAATCTTCTGAGATTCTAGAG
 ProPheThrGluLeuProGluLeuAsnLeuSerGlyLysSerProLeuProSerLeuGluTyrGlyAlaAsnTyrPheLeuGlnIleSerArgV
 3600
 TTAATGATCTAAATAGAATCGGACCGACATGTTAAACATTTTACACATGATATCATGTTACGAGAAAGGATCTAGATAAAGCTATGAAATTTTAA
 alAsnAspLeuAsnArgMetProThrAspMetLeuLysLeuPheThrHisAspIleMetLeuProGluSerAspLeuAspLysValTyrGluIleLeuLys
 3700
 GATTAAATACGCTAAAGTATTATGGGAGGACTACTAAAGCGGACCGGCTAGTTCGGACCTCAGCGGACCGCAATAAATCTGTTCAACCTGAACGAGATGCT
 sIleAsnSerLysIleLysLeuProThrLeuArgSerThrLysAlaAspAlaValValAlaAspLeuSerAlaArgAsnLysLeuPheLysGluArgAspAla
 3800
 ATTAATCTAATAATCATCTCACTGAAACAACTATACATTACGGATTATAATATGTTAACCTTCGACGTGTTTCGACCATTTATGATTTTGTAAAGC
 IleLysSerAsnAsnHisLeuThrGluAsnAsnLeuTyrIleSerAspTyrAsnMetLeuThrPheAspValPheArgProLeuPheAspPheValAsnG
 3900
 AAAAAATGTTATTAACACTTCAACTTATTTCGGTAGAGGTGTAATCGATACATAGAGAATATATGTAAGTCTCTTTAAAAATGTTAGACTGCTAA
 lulysTyrCysIleIleLysLeuProThrLeuArgGlyValIleAspThrMetArgIleTyrCysSerLeuPheLysAsnValLeuLeuLeuLys
 4000
 ATCGCTAACCGATAGCTGGTTAAAGATACGCCCATATTGCTGGCTAGTGATGTTGTAAAAAAATTTGGATTATTTATGTCATGTTAAAGTCGCT
 sCysValSerAspSerTrpLeuLysAspSerAlaIleMetValAlaSerAspValCysLysLysAsnLeuAspLeuPheMetSerHisValLysSerVal
 4100
 ACTAAGTCTCTCTCTGGAAGGATGTGAACAGTGTCAATTAGTATTTTAAACAATCGAGTGATACGGAATTCATTAAATAGTCTTAGAGTTTTCGA
 ThrLysSerSerSerTrpLysAspValAsnSerValGlnPheSerIleLeuAsnAsnProValAspThrLysSerAlaLeuArgPheSerA
 4200
 ATAGACTATAGCAAGCTCTCTATACGTTCACTGCTGCTTTATTTCTAGTATGACTCTGATTCAAAAAGTATCGAAAAACAAACATCAGAGAAGACTAGT
 anArgValTyrGluAlaLeuTyrTyrValHisSerLeuLeuTyrSerSerMetThrSerAspSerLysSerIleGluAsnLysHisGlnArgLeuVa
 4300
 TAAACTACTGCTGTGATTTTTAAACATAGTTATTACTTATCACTACTAAATGAGTAAATCACAGCGGGCCCTATATCGATTATGCAATCGGAGAACTAC
 lLysLeuLeuLeuEndMetSerLysSerHisAlaAlaTyrIleAspTyrAlaLeuArgArgThrTh
 4400
 TAAATGCTCTGTTGAAATGATGGGCTGGAGCTAGTACCGCTCAAGGATTATCAACATTTGTAGCAAGGTTTCTTAGGATTAGACAGTATGATTCT
 rAsnMetProValGluMetMetGlySerAspValValArgLeuLysAspTyrGlnHisPheValAlaArgValPheLeuGlyLeuAspSerMetHisSer
 4500
 GTTTTATGTTCCGATGAAACGGGTGTCGGTAAAAAATGACTACTGTATATATTCTCAACATCTTAAAGATATTATACGAATTCGGCTATTATCTTAT
 lLeuLeuPheHisGluThrGlyValGlyLysThrMetThrValTyrIleLeuLysHisLeuLysAspIleTyrThrAsnTrpAlaIleLeuLys
 4600
 TGGTAAAAAGGCTCTGATAGAGATCTGGATGAGACACTATACCTAGATACCGCTCCAGAGATAACGAAGGATGCTATTTTATTAATACGATGATCA
 euValLysLysAlaLeuIleGluAspProTrpMetAsnThrIleLeuArgTyrAlaProGluIleThrLysAspCysIlePheIleAsnTyrAspAspG
 4700
 AAATTTTAGAAAATAATTTTCTAATATCAAACTATTAAATTCAGAGTAGAATATGCGGTCAATTATGATGAATGTCATAAATCTTCTAAATCA
 nAsnPheArgAsnLysPhePheThrAsnIleLysThrIleAsnSerLysSerArgIleCysValIleIleAspGluCysHisAsnPheIleSerLysSer
 4800
 TTAATCAAGAAGATGCTAGATCGCTGCTAGTATATAATTTTATCTAAGACGATGCGAATTAACCAAGATGATTGTTGTTATGCTGG
 LeuIleLysGluAspGlyLysIleArgProThrArgSerValTyrAsnPheLeuSerLysThrIleAlaLeuLysAsnHisLysMetIleCysLeuSerA
 4900
 CTACACCTATGCTCAATAGTGTGCAAGAAATTCACGATGTTGCTTAACCTACTACGACAGGATCGCTACAAACACCAATGGTATTGAGAATAAAGCTCT
 laThrProIleValAsnSerValGlnGluPheThrMetLeuValAsnLeuLeuArgProGlySerLeuGlnHisGlnSerLeuPheGluAsnLysArgLe
 5000
 AGTTGATGAAAAAGAAATAGTCTCCAACTAGGAGGCCCTATGTTGCTAGATAGTTAATAACGAATTTTCTATTTTGTATGATGACGAGGCTGTCATCA
 uValAspGluLysGluLeuValSerLysLeuGlyGlyLeuCysSerTyrIleValAsnAsnGluPheSerIlePheAspAspValGluGlySerAlaSer
 5100
 TTCCTAAGAAAAACAGTATTAAATCGGATACGTTAATGTCGAAAAAGCAAGAAATTTATCAAAAGGCTAAATCGCTGAAATAAAAACAGGTATAT
 PheAlaLysLysThrValLeuMetArgTyrValAsnMetSerLysLysGlnGluGluIleTyrGlnLysAlaLysLeuAlaGluIleLysThrGlyIleS

CATGATTAGAATTC
 erSerPheArgIle
 EcoRI

Figure 2. Nucleotide sequence of the central 5.1 kbp of the Sali J fragment. Contiguous sequence of the leftward-reading strand is shown, beginning at the leftmost SstI site of HindIII A, and extending to an EcoRI site within the HindIII D fragment. Open reading frames (orfs) are numbered 0 through 5. Late RNA start sites are marked with arrowheads. Putative control signal TAA's are underlined.

Figure 3. Diagrammatic representation of open reading frames.

A) Cleavage sites of rarely-cutting restriction nucleases within the sequenced fragment. b) Organization of open reading frames designated 0 through 5 and corresponding molecular weights of the translated sequences. C) Sequencing strategy. Arrows represent the direction and length of sequence determined from M13 clones.

leftward direction revealed five complete orfs and two incomplete orfs (Fig. 3). For convenience, the five complete orfs were designated orf 0 through orf 4 in order of appearance from SstI to EcoRI. The leftmost, incomplete orf was designated orf 5. Approximate molecular weights of the translated sequences appear above each orf. The organization is tandem, with spacer sequences of 14, -4, 20, 23, 30 and 34 nt between adjacent orfs.

Transcript Mapping

In order to map the 5' ends of transcripts originating from this region of the genome, restriction fragments spanning the first ATG of each open reading frame and 5' end-labeled at the downstream restriction site were gel isolated. The size, termini, and 5' end labeled for these probes are indicated in Fig. 4. Each probe was denatured and annealed to 20 µg of total RNA isolated from uninfected cells or from infected cells at late times post infection, and then treated with S1 nuclease (21). No S1 nuclease-protected fragments were detected when RNA from uninfected cells was hybridized to the probes (data not shown). S1 nuclease-protected fragments resulting from hybridization with late RNA are shown in lanes S1, for each of 6 open reading frames, alongside Maxam-Gilbert sequencing reaction A + G (lanes AG) performed on the same labeled DNA fragment.

After correcting for 1 to 1.5 bases to account for the elimination of the modified terminal nucleotide during chemical sequencing, the 5' ends of major late transcripts were mapped upstream of putative initiator ATGs as follows:

Figure 4. 5' S1 nuclease mapping of late mRNA start sites. DNA probes (solid line) with indicated size (nucleotides), cleavage termini, and 5' end labeled (star), and late transcripts (waving line) with indicated size of protected fragments appear below each panel. Lanes SI are S1 nuclease-resistant hybrids from the indicated probes and late RNA. Lanes AG are Maxam-Gilbert A + G reactions of the same probes. Late RNA starts are marked with arrowheads within the sequence context of the initiator ATG. Panels 0 to 5 represent orfs 0 to 5.

orf 0, 1-3 nt; orf 1, 35-37 nt; orf 2, 2-4 nt; orf 3, 9-11 nt; orf 4, 2-4 nt; orf 5, 2-4 and 30-31 nt. These positions are indicated in Figure 2. We conclude that genes represented by orfs 0 to 5 are transcribed late in infection.

Open Reading Frames

As an indication of their protein-encoding potential, codon utilization in the orfs was analyzed. Orf 3 was used as a basis for comparison since the corresponding polypeptide has been identified in infected cells (21). Orfs 1, 2, 4 and the incomplete orf 5 exhibited codon preference which closely matched that of orf 3, while orf 0 did not.

The closely-spaced organization suggested the possibility of gene duplication. Homology and alignment analyses between the five orfs and the incomplete orf 5 showed no significant relationships. Additionally, the NBRF Protein Data Bank was searched and no strong homologies were found. The polypeptides represented by orf 0 through orf 5 appear to be unique.

Analysis of the putative polypeptides was extended in order to identify similarities or differences between the five. When translated from the first ATG, orfs 0 to 5 represent polypeptides of molecular weights 8,956; 26,291; 16,908; 61,697; 33,342; and 33,235 kilodaltons. Orfs 0, 1, 2, and 4 encode polypeptides composed of more basic than acidic residues, and the majority of the amino acids present are predicted to be involved in α -helical secondary structures (27). The polypeptide encoded by orf 3 is dissimilar by containing more acidic than basic residues, and by preferring β -sheet structures. Asparagine residues were scanned to determine the number of potential N-linked glycosylation sites according to the consensus sequence (Asn-X-Ser/Thr). Polypeptides encoded by orfs 1 and 3 are the most likely candidates for N-linked glycosylation.

Upstream Sequences

Assuming that transcriptional control signals for VV lie just 5' to their RNA start sites and that these sequences might share common features, 100 nucleotides immediately upstream of the transcriptional start closest to the first ATG for orfs 0 through 5 were analyzed for A-T richness, for the presence of direct or inverted repeats, and for homology and alignment with the other five.

The entire 5115 nt segment sequenced contains 66% A-T which is in close agreement with 63-64% for total VV DNA (1). The 100 nt strings contain 68 to 74% A-T. In every case, the putative promoters increased 3 to 7% in A-T richness for the first 50 nt upstream of the transcriptional start.

Imperfect direct repeats were found upstream of orfs 0, 4, and 5. Imperfect inverted repeats were found upstream of orfs 0, 2, and 5. When individual repeat elements were compared between the 5' sequences, no obvious correlations with respect to sequence, placement, or spacing were evident.

Homology and alignment analyses between the six putative promoters were performed under high stringency; to be considered a match, at least 10 nt with at least 75% homology with a maximum loopout allowance of 3 nt was required. No homologies found were perfect.

In several comparisons, similar strings of nucleotides were found in similar positions in two or three different promoters. Orf 0, 1, and 2 promoters contain an approximate sequence of ATTTATGCACAG at nucleotide position -80 to -60 when 0 is the transcriptional start site. Part of this element is directly repeated with two mismatches in the orf 0 promoter at -20 to -10. Orf 0 and 2 promoters contain an additional string of homology at -30 to -20 with two mismatches from AGGATTATGATCA. Promoters for orf 1 and 2 show homology in position -100 to -80 having 3 mismatches from GTAACCATATTAG. Orf 1, 3, and 4 promoters show homology at -60 to -40 diverging from ACAATATTATAG.

More frequently found were homologous strings with dissimilar positions. For example, the homology described above between orf 1, 3, and 4 promoters is also found in different positions in orf 2 and 5 promoters. An approximation of AAAGTATCGAAAA is found at different positions in promoters of orf 2, 3, 4, and 5.

DISCUSSION

Correlation of open reading frames with polypeptides and mRNAs

Preliminary mapping data in this region of the VV genome by hybrid-selected translation (9,10,28) and by 5' S1 nuclease protection (18,29) show correlations with the sequence analysis. Results of recent work in our laboratory (21), focused on specific mapping of polypeptides and transcripts within the 5115 nt DNA segment, show strong correlations with the identified orfs. The rightmost sequences of the HindIII D fragment hybrid-selected late mRNAs encoding 65K and 32K polypeptides (21), which are in close agreement with polypeptides represented by orfs 3 and 4, respectively (Fig. 3). Using the leftmost HindIII A sequences to select late mRNAs, we were unable to detect polypeptides corresponding to orfs 0, 1, and 2. However, approximate locations of the 5' ends of late transcripts were recently reported for orf 0 (18), orf 1 (18), and orf 2 (18,21,29), in agreement with the exact locations reported here. Thus, the current data clearly indicates that genes represented by orfs 0 through orf 5 are transcribed late in infection, although polypeptides corresponding to these transcripts have only been identified for orf 3 and orf 4.

Late gene expression

Novel mechanisms of regulation are perhaps responsible for the peculiar features which characterize the postreplicative mode of VV gene expression. While most of the genome is being transcribed late in infection (30,31), late polypeptides are preferentially synthesized (32,33). Late mRNAs are found to include early sequences (34), and are capable of forming intermolecular duplexes with themselves or with early RNAs (35,36). These characteristics are perhaps explained by the recognition that mRNAs encoding single late polypeptides vary several-fold in length due to 3'-terminal heterogeneity (15,37), unlike the early transcripts, which have discrete sizes (13,38,39). Northern analysis of late mRNAs within the 5.1 kbp region detected a complex mixture of complementary species lacking any length specificity (data not shown). It is for these reasons that we made no attempt to map the 3' ends of these transcripts. The complications of mapping heterogeneous transcripts are in part responsible for the lack of precise information on the majority of VV late genes.

In the present case of a tandemly-oriented late gene cluster, the 3' terminal heterogeneity exhibited by transcription of the upstream late gene results in readthrough transcription through the next one or several late genes. This was clearly indicated when Rosel and Moss, using a single end labeled probe, detected the 5' ends of late transcripts for orf 2, orf 1, orf 0, P4b, and additional readthrough transcription initiating upstream of the unlabeled end of the probe (18). Thus any probe designed to detect the 5' end of a specific late transcript within this cluster will also be fully protected by readthrough transcription of the upstream late gene(s). This may be responsible for our lack of success in confirming the 5' ends of these transcripts by primer extension analyses; we cannot at this time distinguish between the possibility of three closely-spaced start sites and S1 nuclease nibbling.

Despite these difficulties, the number of precisely mapped late genes has increased (29,37,40), some have been sequenced (17,18), and a few have been subjected to detailed examination (15,16). Weir and Moss obtained evidence that cis-acting regulatory signals reside in 366 nucleotides preceeding a late gene that encodes a 28K polypeptide (15). Bertholet et al. examined the 5' limit of a late promoter and found that 100 nucleotides of 5' flanking sequence from an 11K polypeptide-encoding late gene was sufficient for temporal regulation (16). In both cases, the exact 5' end for the late transcripts was found to be within 5 nucleotides upstream of the first ATG, as

is the case for orf 0, orf 2, orf 4, one of the start sites for orf 5, and two of the start sites for P4b (18). Some early transcripts also have been found to contain short 5' untranslated regions (12,13).

In an attempt to identify recurrent features, we expanded our presumptive promoter analysis to include three late genes previously sequenced: 28K (15), 11K (16), and P4b (18). As for orfs 0 to 5, the first 100 nt upstream of the RNA start closest to the first ATG were analyzed. These three additional promoters vary from 68 to 73% A-T rich, and except for the 28K late gene there is an increase in A-T richness for the first 50 nt upstream of the RNA start, as found for orfs 0 to 5. A high stringency homology search revealed three cases of similar nt strings with similar positions: orf 1 and 11K promoters at -20 to 0; orf 2 and 28K promoters at -20 to 0; orf 4 and P4b promoters at -60 to -40. Again, the more frequent finding was homologous elements in dissimilar positions.

The level of stringency used in our promoter analysis was chosen to identify strong similarities. While relationships were identified, no homologies were found in a similar position for all nine promoters. However, a recurrent feature of this set of late gene promoters is the stop codon TAA immediately preceding the first ATG (Fig. 2). This is found for orf 0, orf 2, orf 4, orf 5, 11K, 28K, and for two additional late genes F2 and F4 for which the sequences, but not the RNA starts, have been reported (17). Besides F2 and F4, this set is additionally related by having RNA starts within 5 nucleotides upstream of the first ATG. Orf 1 and orf 3, which do not contain TAA immediately preceding the first ATG, also diverge by utilizing RNA starts further upstream of the ATG. However, the RNA start for both orf 1 and orf 3 maps to a TAA. In the case of orf 1, the TAA is in phase, and for orf 3, out of phase with the first ATG. Besides the RNA start just preceding the first ATG, orf 5 utilizes a second RNA start upstream which maps to an in-phase TAA. While the extreme A-T richness of these promoter regions supports a coincidental explanation, a more interesting possibility would be a signal function for the TAA.

Concerning the types of homologies identified among this first significant set of late gene promoters examined, we offer the following possible interpretations. (1) There are perhaps 100 VV late genes. The products of these late genes do not appear in a synchronous fashion (32,33). Temporal subsets of the late gene class may contain promoter elements that distinguish them from other postreplicative subsets. The relatively small number of late genes here examined might include representatives from several

subsets. Indeed, a kinetic analysis of transcription of orfs 0 to 5 shows that, while each is transcribed late, there are marked differences in the kinetic patterns of transcription (manuscript in preparation). (ii) Distinct sequences or repeat elements involved in late gene regulation may have flexibility in terms of sequence, placement, or spacing, and this flexibility could be utilized to fine tune a promoter for specific temporal or quantitative modulation of expression. (iii) Interfacing with the flexibility described above could be an inflexible situation where TAA is used to earmark late genes or transcripts for transcriptional or post-transcriptional regulation.

Genomic Organization

Earlier studies of VV transcription indicated a lack of overall temporal organization of the genome (8). Translational mapping surveys suggested a concentration of late genes toward the central portion of the genome (9,10). Mapping and sequencing analyses of large segments of VV DNA have indicated that VV genes can be overlapping, tightly-spaced, and tandemly-oriented (17,29,37,38,41), while evidence for splicing has not been found. The sequence determination of a tandemly-oriented and tightly-spaced late gene cluster adds to the possible organizational motifs of the VV genome. In addition to the late gene cluster we have identified, late genes encoding the structural proteins P4b (18, overlaps the sequence determined here by 442 nt) and P4a (18,29), and perhaps a third late gene (29) have been mapped to the leftmost sequences of the HindIII A fragment, and these also read in the leftward direction. This would indicate a cluster of 9 tandemly-oriented late genes. The genomic region represented by the leftmost HindIII A sequences and the rightmost HindIII D sequences appears to contain preferentially late genes transcribed in the leftward direction, and may illustrate local temporal organization of the VV genome.

Mapping the L65 polypeptide to this genomic locus (21) provided the impetus to determining the sequence of the 5115 nt segment. Other results suggesting the importance of this region include conservation of this region in the genome of different orthopoxviruses (42), mapping two late genes encoding major structural polypeptides just upstream of this region (18,29), the observation that several different VV temperature-sensitive mutants are marker rescued by restriction fragments of this region (E. G. Niles, personal communication), and recently, that rifampicin resistance has been mapped to this region (22). Our current line of investigation pursues the dissection of the control sequences responsible for the expression of these late genes and the determination of the function of their encoded polypeptides.

ACKNOWLEDGEMENTS

We wish to thank Jeff Hughes for assistance in M13 cloning and dideoxy-sequencing, and Lisa Wilson, Elcira Villareal, and Chris Franke for their technical support. This work was supported by a grant from the National Institutes of Health (AI-20563) and by a RCDA (D.E.H.), NIH (AI-00666). Oregon State University Agricultural Experiment Station Technical Paper No. 7697.

REFERENCES

1. Moss, B. (1985) in *Virology*, ed. Fields, B. N., Raven Press N.Y., 685-703.
2. Panicali, D., and Paoletti, E. (1982) *Proc. Natl. Acad. Sci. USA* 79, 4927-4831.
3. Mackett, M., Smith, G. L., and Moss, B. (1982) *Proc. Natl. Acad. Sci. USA* 79, 7415-7419.
4. Perkus, M. E., Piccini, A., Lipinskas, B. R., and Paoletti, E. (1985) *Science* 229, 981-984.
5. Franke, C. A., Rice, C. M., Strauss, J. H., and Hruby, D. E. (1985) *Mol. Cell. Biol.* 5, 1918-1924.
6. Wittek, R. (1982) *Experientia* 38, 285-297.
7. Munyon, W., Paoletti, E., and Grace, Jr., J. T. (1967) *Proc. Natl. Acad. Sci. USA* 58, 2280-2287.
8. Cabrera, C. V., Esteban, M., McCarron, M., McAllister, W. T., and Holowczak, J. A. (1978) *Virology* 86, 102-114.
9. Belle Isle, H., Venkatesan, S., and Moss, B. (1981) *Virology* 112, 306-317.
10. Chipchase, M., Schwendimann, F., and Wyler, R. (1980) *Virology* 105, 261-264.
11. Venkatesan, S., Baroudy, B. M., and Moss, B. (1981) *Cell* 25, 805-813.
12. Venkatesan, S., Gershowitz, A., and Moss, B. (1982) *J. Virol.* 44, 637-646.
13. Hruby, D. E., Maki, R. A., Miller, D. B., and Ball, L. A. (1983) *Proc. Natl. Acad. Sci. USA* 80, 3411-3415.
14. Weir, J. P., and Moss, B. (1983) *J. Virol.* 46, 530-537.
15. Weir, J. P., and Moss, B. (1984) *J. Virol.* 51, 662-669.
16. Bertholet, C., Drillien, R., and Wittek, R. (1985) *Proc. Natl. Acad. Sci. USA* 82, 2096-2100.
17. Plucienniczak, A., Schroeder, E., Zettlmeissl, G., and Streeck, R. E. (1985) *Nucleic Acids Res.* 13, 985-998.
18. Rosel, J., and Moss, B. (1985) *J. Virol.* 56, 830-838.
19. Mackett, M., Smith, G. L., and Moss, B. (1984) *J. Virol.* 49, 857-864.
20. Vassef, A., Mars, M., Dru, A., Plucinniczak, A., Streeck, R. E., and Beaud, G. (1985) *J. Virol.* 55, 163-172.
21. Weinrich, S. L., Niles, E. G., and Hruby, D. E. (1985) *J. Virol.* 55, 450-457.
22. Tartaglia, J., and Paoletti, E. (1985) *Virology* 147, 394-404.
23. Moss, B., and Rosenblum, E. N. (1973) *J. Mol. Biol.* 81, 267-269.
24. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.
25. Maxam, A. M., and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* 74, 560-564.
26. Queen, C., and Korn, L. J. (1984) *Nucleic Acids Res.* 12, 581-599.

27. Garnier, J., Osguthorpe, D. J., and Robson, B. (1978) *J. Mol. Biol.* 120, 97-120.
28. Cooper, J. A., and Moss, B. (1979) *Nucleic Acids Res.* 6, 3599-3612.
29. Wittek, R., Richner, B., and Hiller, G. (1984) *Nucleic Acids Res.* 12, 4835-4848.
30. Boone, R. F., and Moss, B. (1978) *J. Virol.* 26, 554-569.
31. Paoletti, E., and Grady, L. J. (1977) *J. Virol.* 23, 608-615.
32. Moss, B., and Salzman, N. P. (1968) *J. Virol.* 2, 1016-1027.
33. Pennington, T. H. (1974) *J. Gen. Virol.* 25, 433-444.
34. Oda, K., and Joklik, W. K. (1967) *J. Mol. Biol.* 27, 395-419.
35. Boone, R. F., Parr, R. P., and Moss, B. (1979) *J. Virol.* 30, 365-374.
36. Colby, C., Jurale, C., and Kates, J. R. (1971) *J. Virol.* 7, 71-76.
37. Mahr, A., and Roberts, B. E. (1984) *J. Virol.* 49, 510-520.
38. Mahr, A., and Roberts, B. E. (1984) *J. Virol.* 49, 497-509.
39. Wittek, R., Cooper, J. A. Barbosa, E., and Moss, B. (1980) *Cell* 21, 487-493.
40. Wittek, R., Hänggi, M., and Hiller, G. (1984) *J. Virol.* 49, 371-378.
41. Cooper, J. A., Wittek, R., and Moss, B. (1981) *J. Virol.* 39, 733-745.
42. Mackett, M., and Archard, L. (1979) *J. Gen. Virol.* 45, 683-701.