CHAPTER III

DNA Sequence Assembly and Analysis

3.1 INTRODUCTION

Large-scale sequencing projects, for example those of the already-sequenced poxvirus genomes, make use of automated sequencing machines directly connected to a computer. The amount of sequence that can be read from one sequencing reaction is limited, and therefore it is first necessary to cleave large sequences, using a frequent-cutter restriction endonuclease, such as *Tsp509I* (Afonso *et al.*, 1999, 2000), or by a physical method such as sonication (Cameron *et al.*, 1999). DNA fragments in the regions of between 1.5 and 3 kbp are then isolated and "shotgun"-cloned into bacterial vectors. Once these random, overlapping fragments are sequenced, it is possible to reconstruct the original sequence by searching for regions in common between the gel readings, using specialised software. Missing regions are sequenced by primer-walking, using oligonucleotides designed from partially assembled DNA sequences. Gap4, which is a commonly-used sequencing project management program in the Staden package, helps in the management of sequencing projects not only by assembling gel readings, but also by searching and removing vector sequences, repeat sequences and poor quality sequence regions which can cause problems when assembling the fragments (Staden *et al.*, 1998).

The cloning approach for the LSDV genome was more systematic than the abovementioned "shotgun" approach, which was not possible due to limited infrastructure. The major difference between this genome project and the abovementioned, is that none of the LSDV clones overlap, making a contig assembly program, such as Gap4, virtually useless. The linear order of the *PstI* clones relative to each other, were known from the restriction map compiled by Perlman (unpublished data presented in his MSc dissertation 1993, UCT), but the orientations of the *PstI* fragments were not (the directionally-cloned *PstI*-E-fragment being an exception). The subsequent sub-cloning strategies furthermore produced non-overlapping subclones, and the sequences had to be fitted together in the correct orientations and orders, in a manner that could not rely on overlapping regions. Additionally, very small fragments were overlooked on the agarose gels, and consequently were not cloned.

LSDV clones were sequenced as they were constructed. Analysis identified partial LSDV ORFs which were blasted against the public databases, and shown to be particular regions of homologous full-length poxvirus proteins. GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Benson *et al.*, 2000). There were approximately 11 101 000 000 bases in 10 106 000 sequence records as of December 2000, and the figures are expected to double in size approximately every 14 months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which is comprised of the DNA DataBank of Japan (DDJB), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These organisations exchange data on a daily basis (http://www.ncbi.nlm.nih.gov/Genbank). Once a full-length poxvirus ORF was retrieved from GenBank, it was possible to fit the LSDV LSDV nucleotide sequences of the same protein together in the correct orders and orientations. It is known from the literature that the ORFs near the left-hand terminus of the poxvirus genome tend to read towards the left terminus, while those at the right-hand terminus read towards the right. This phenomenon gave clues as to which direction the LSDV ORFs should be orientated. By fitting the fragmented LSDV ORFs from the clones onto a full-length homologous protein from GenBank, it was furthermore possible to identify small missing DNA sequences spanning the *AseI* juctions, which were later determined by primer walking sequencing. Frameshifts in protein reading frames occur with the insertion or emission of a nucleotide, or where a nucleotide is incorrectly called by the sequencing software, and had to be corrected.

For the purposes of this study, a clone was regarded to present an ORF if the translated peptide initiated with a methionine codon, and was at least 50 aa in length. The LSDV ORF nomenclature includes the *PstI*-fragment containing initiating methionine codon, the linear position of the ORF in the *PstI*-fragment (from left to right), and the direction of the open reading frame (L or R). For example, ORF E3L initiates in the *PstI*-E fragment, and is the third ORF, reading towards the left. The degree of amino acid identity to the top hit was also taken into consideration, with amino acid identities exceeding 20-25% being considered probable evidence of homology, a common approach in other publications of poxvirus sequences.

3.2 MATERIALS AND METHODS

3.2.1 Hardware

Output from ABI377 sequencing managed on Power Macintosh 7600/132 with Mac OS version 7.5.5. Electropherograms (EPGs) were processed by ABI Prism Sequence Navigator Version 1.0.1 (Department of Microbiology, UP).

3.2.2. DNA sequence editing

Forward and reverse sequences were edited in the following manner:

1) Both forward and reverse sequences were imported into the layout, and EPGs opened.
2) base-calling was matched with EPG peaks
3) vector sequence of the multiple cloning site was cut off at the appropriate site: the recombined *AseI/NdeI* sites have the sequence CATAAT in a forward sequence and a TATGGATT sequence in the reverse sequence
4) the reverse sequence was reverse-complemented , and both sequences selected
   a) the comparative alignment option was used off the drop-down menu or
   b) a "shadow" was created , and the reverse sequence slid along the forward sequence until the alignment was focused
5) the reverse sequence was copied onto the forward sequence at the appropriate junction, and the full-length edited sequence was exported to a file and diskette in Staden format.

3.2.3. DNA sequence analysis - GCG

(Wisconsin Package version 10.2, Genetics Computer Group (GCG), Madison, Wisc.), installed on Irix (UNIX).

Staden-format sequences were transferred from diskette, to the UNIX machine by File Transfer Protocol (FTP). GCG converts a specified file from Staden format, in which it was saved in Apple, to GCG format, prompted by the line-command >**fromstaden**.

Map (>**map**) displayed both strands of DNA sequence with restriction sites shown above the sequence, and all six possible amino acid translations shown below.

Open reading frames were visually identified (beginning with a methionine codon, and being at least 50 amino acids in length), and the specific nucleotide sequence, which had to be specified, was translated to amino acids (>**translate**). The output file was written as a *.pep* file. It was worthwhile to review the *.pep* file using the command >**more filename.pep**, to ensure that the correct reading frame was translated (one nucleotide position too many or too few would cause a frameshift) before the peptide sequence was blasted.

The Blastp algorithm (>**blastp**) was used to search the *.pep* amino acid sequence against the Swissprot protein database. The output file can be quite large, and has four parts: an administrative header (filename, copyright information, database size and references), followed by a list composed of hits with a) database accession number, b) brief description, c) score and d) the E- value.

The E-value gives some information on the probability of finding such a hit in the searched database. For example, and E-value of 8.4e-08 on the last line of the output implies that the probability of finding a match as good as this, by chance, in the current database is 8.4 e^{-08}. Put in another way, the BLASTP E-value is a probability value calculated for each sequence, and gives an estimate of the probability that the match happened by chance and is not significant (the lower the number, the better the match) (Lloyd *et al.*, 1999). This value should normally fall below 0.01 or exceed 0.1 (except in the case of very small proteins).

The third part to the blastp output file is a number of alignments of the query sequence with significant hits, and the fourth and final part gives more administrative and statistical information.

The accession numbers of top hits were used to **>fetch** the GenBank entry from the databases. The full-length, top-hit protein retrieved in this way was used for fitting LSDV protein fragments on to. Missing regions at fragment junctions could be identified in this way, and the GenBank protein was also used for fixing frameshifts. To fix frameshifts, conserved motifs from the top hit protein were used to track the true reading frame in three frames (depending on the orientation of the ORF in the LSDV genome). The linear sequence of genes from published poxvirus genomes also gave important clues as to the orders and orientations of the peptide sequences.

To assemble the fragments whose orders and orientations are now known from the peptide sequence, GCG files in which frameshifts had been edited (by referring back to the original electropherogram of the region) were opened in Sequence Navigator once again. Adjacent fragments were joined by cutting and pasting. Back in UNIX, additional sequences discovered by primer walking were edited in (**>nedit**), re-checked for frameshifts, and full-length ORFs blasted.

3.2.4 SMART (V3.1) analysis of ORFs (Schultz *et al.*, 1998; 2000)

Detecting non-enzymatic regulatory domains is essential to predict a novel protein's cellular role, binding partners and subcellular localisation. SMART (a Simple Modular Architecture Research Tool), accessible at http://coot.embl-heidelberg.de/SMART, is able to detect more than 400 domain families found in signalling, extra-cellular and chromatin-associated proteins. These domains are extensively annotated with respect to phyletic distributions, functional class, tertiary structures and functionally important residues. Information on more than 400 domain types in more than 54 000 different proteins is stored in SMART using a relational database management system. For each domain hit, boundaries, a score and an E-value are recorded. SMART is furthermore able to search for all proteins that have an identical domain architecture as the query (i.e. having all the domains of the query protein in the same colinear order), or an identical domain composition (Schultz *et al.*, 2000).

3.2.5 The identification of promoters

Several promoters could be identified visually, but to take a more systematic and objective approach to locate these regions, which can deviate somewhat from the published consensus for vaccinia, blast alignments were implemented. William Pearson's *lalign* program (http://.ch.embnet.org/cgi-cin/LALIGN) allows one to find multiple matching sub-segments in two sequences. The program implements the algorithm of Huang and Miller (1991). To create a consensus sequence for an early, intermediate and late promoter, the vaccinia consensus and its characterized promoters were considered, along with obvious promoter sequences identified visually by inspection of the regions upstream of LSDV ORFs. Assembled LSDV *PstI* clone DNA sequences were searched with consensus sequences for early, intermediate and late promoters. To accommodate ORFs on the non-coding DNA strand, reading towards the left, the promoter consensus sequences were reverse-complemented. Positive ORFs found by lalign were further verified by visual inspection, and those that had substitutions at the most conserved residues were excluded.

3.3 RESULTS AND DISCUSSION

3.3.1 Nucleotide sequences of LSDV *PstI* fragments

The nucleotide sequences of the *PstI* fragments were submitted to GenBank as two annotated sequences: *PstI*-E (7024 bp), *PstI*-M (2225 bp) and *PstI*-K (3281 bp) were combined under the accession number **AF336128**, whereas *PstI*-F (8553 bp) and *PstI*-G (7803 bp) were combined under the accession number **AF336131.**

Table 5: Identified LSDV ORFs and blastp matches of known specific genes with known or unknown function. Please see table legend (pg. 85) for details of the abbreviations and column headings.

LSDV ORF	Size		Top Hit					Putative Structure/ function	Promo ter type
	Position	(aa)	E value	% Identity	Positives	Length (aa)	ORF description		
EIL	E, 1-718	238	2e-34	82/224 (36%)	122/124 (53%)	274	SPV hypothetical protein H6 P32226 VV interferon gamma receptor	Interferon-binding protein	I/L
E2L	E, 755-1445	230	4e-05	32/121 (26%)	64/121 (52%)	170	VV O93116 VAR N2 protein	Alpha-amanitin sensitive protein	E
E3L	E, 1498-1984	162	3e-26	49/150 (32%)	82/150 (54%)	155	SPV hypothetical protein C7 P32225 Murine herpesvirus BHV4-IEI homolog		I
E4L	E, 2030-3173	380	e-170	306/380 (80%)	308/380 (80%)	381	CAP Q2/3L partial ORF Q86917 *Macaca mulatta* chemokine receptor	G-protein coupled chemokine receptor homolog	E
E5L	E, 3283-3919	212	e-111	192/196 (97%)	194/196 (98%)	202	CAP Q2/2L full ORF Q86916 Orf virus ankyrin-like repeat protein	Actin - binding	E
E6L	E, 3966-4893	309	0.0	304/309 (98%)	306/309 (98%)	317	CAP Q2/1L partial ORF Q86915 VV interleukin-1 binding protein precursor	Interleukin-1 receptor	E

E7L	E, 5053-5320	88	1e-21	42/85 (49%)	68/85 (79%)	86	SPV C8 hypothetical protein P32224 VV hypothetical 11.1 kD protein		L
E8L	E, 5309-5792	161	2e-27	56/129 (43%)	89/129 (68%)	134	SPV hypothetical protein C9 P32223 *Homo sapiens* interleukin-18 binding protein A precursor	Interleukin-18 binding protein	L
E9L	E, 5811-6045	78	0.004	21/53 (39%)	30/53 (55%)	80	SFV growth factor H2 P08441 *Xenopus laevis* neureglinα-1	Growth factor	(?)
E10L	E, 6093-6621	176	0.004	37/139 (26%)	62/139 (43%)	167	SPV hypothetical protein C10 P32222 MsEPV ORF 71	Putative early transcription factor-small subunit	L
M1L	E, 6665-7024 M, 1-85	147	5e-46	(73%)	(83%)	142	SPV dUTPase P32208 VAR dUTPase	dUTPase, Nucleotide metabolism	I
M2L	M, 129-1611	494	5e-97	187/466 (40%)	187/466 (40%)	500	SPV hypothetical protein C13 P32206 VV 55 K kelch-like protein	Actin-binding	I
K1L	M, 1713-2225 K, 1-487	321	3e-69	242/321 (75%)	283/321 (88%)	319	SPV ribonucleoside-diphosphate reductase-small chain P20493 VV F4L	Ribonucleoside reductase-nucleotide metabolism	E
K2L	K, 529-787	86	8e-06	20/47 (43%)	33/47 (69%)	86	SPV hypothetical protein C15 P32220		L
K3L	K, 830-1166	112	9.4	15/39 (38%)	22/39 (55%)	435	*Nicotiniana tabacum* A membrane-associated salt inducible protein Q40452		E
K4L	K, 1398-1614	72	0.011	18/40 (45%)	26/40 (65%)	65	VV protein F8 P29887 VAR protein F8		E
K5L	K, 1691-2429	218	5e-72	129/216 (59%)	163/216 (74%)	215	SPV hypothetical protein C9 P32207 VV proteinF9		L
A1L partial	K, 2319-3281	318	e-137	232/314 (73%)	271/314 (85%)	440	SPV protein kinase C20 P32216 VV protein kinase 2	Protein kinase	?
AXR partial	F, 1-386	128	4e-48	91/128 (71%)	104/128 (81%)	338	MYX poly(a) polymerase regulatory subunit (VP39) P18628 VV poly(a) regulatory subunit	Poly(a) polymerase regulatory subunit	?
FIR	F, 303-858	185	7e-78	135/185 (72%)	163/185 (87%)	185	MYX DNA-directed RNA polymerase 21 kD polypeptide	DNA-directed RNA	E

							P18620 VAR DNA-directed RNA pol 22 kD polypeptide	polymerase subunit	
F2L	F, 855-1272	139	1e-39	72/118 (61%)	90/118 (76%)	133	VV J5L protein P33055 VAR protein J5		L
F3R	F, 1484-5113	1210	0.0	920/1215 (75%)	1052/1215 (85%)	1286	VV DNA-directed RNA polymerase 147 kD polypeptide P20504 VAR Rpol47	DNA-directed RNA polymerase subunit	L
F4L	F, 5097-5712	205	8e-76	131/171 (76%)	134/171 (78%)	172	MYX tyrosine phosphatase (I1L) Q85297 RFV tyrosine phosphatase	Tyrosine phosphatase	L
F5R	F, 5728-6299	190	2e-74	123/188 (65%)	153/188 (80%)	189	VV late protein H2 VV putative 21.5 k protein		L
F6L	F, 6300-7269	323	1e-45	99/326 (30%)	167/326 (50%)	325	VV protein H3L Q85385 VAR immunodominant envelope protein	Envelope protein	I/E
G1L	F, 7300-8553 G, 1-1135	799	0.0	566/799 (70%)	656/799 (81%)	797	VV H4L protein MYX m72L	RNA polymerase-associated protein	L
G2R	G, 1277-1955	226	0.001	42/73 (43%)	29/43 (66%)	220	VAR (=VV protein H5R) Q85386 orf virus envelope antigen homolog	Envelope protein	I/L
G3R	G, 1998-2949	317	e-117	210/317 (66%)	247/317 (77%)	314	SFV DNA topoisomerase I P16472 VV late protein H6	DNA topoisomerase I	L
G4R	G, 2971-3337	122	9e-18	44/106 (41%)	71/106 (66%)	129	VAR ORF 85R Q89221 VAR late protein H7		L
G5R	G, 3388-5650	754	0.0	466/719 (64%)	587/719 (80%)	836	SFV mRNA capping enzyme, large subunit P25950 VV mRNA capping enzyme	mRNA capping enzyme	E
G6L	G, 5684-6149	155	5e-22	57/154 (37%)	94/154 (64%)	143	SFV protein D4 P25951 VAR protein D2		?
G7R	G, 6468-6774	102	2.1	20/58 (34%)	32/58 (58%)	569	*Plasmodium falciparum* throphozoite cysteine proteinase precursor P25805 SFV protein D5 (length = 241 aa)		L
G8R	G, 6852-7505	218	5e-99	160/218 (73%)	187/213 (85%)	218	SFV uracil-DNA glycosylase	uracil-DNA glycosylase	E

							VV uracil-DNA glycosylase		
G9R	G, 7551-7803 (partial)	83	1e-16	37/57 (64%)	45/57 (78%)	791	FPV 92.6 kD protein VV 90.4 kD protein		E

Legend to Table 5: Summary of the blast output files for each of the LSDV ORFs identified. The lengths and positions of the respective ORFs in their respective *PstI* clones are indicated, with the ORF size in amino acids (aa). The Top Hit columns give information of the homologous ORFs in the GenBank database. Two references are given. The first reference, with a GenBank accession number, refers to the "top hit"- the protein with the highest homology at the top of the output file. The information in the other top hit columns relates specifically to the top hit, i.e. the E-value, % identity, positives, and length in amino acids. The second entry was the second hit in most, but not all cases, included to give extra information about the other types of proteins that demonstrated significant homology. Some ORFs are exclusive to poxviruses, particularly those of the terminal regions, while others, like the DNA topoisomerase I gene has counterparts with high homology in eukaryotic organisms like mice and humans. The "% identity" refers to the number of matches in the particular alignment that match precisely, at that point. Insertions or deletions of amino acids are therefore penalised in this value. "Positives" refers to the amounts of matching amino acids in the linear peptide sequence. The promoter column again refers to the LSDV ORF, where "I" represents Intermediate, "E" represents Early, and "L" represents a Late promoter. These results are discussed in 3.3.3, LDSV ORFS.

3.3.2 SMART analysis

Each peptide sequence, in FASTA format (Appendix III), was cut and pasted into the SMART program for the detection of protein motifs and domains. The results were as follows:

Identification of a RING finger motif. A RING finger motif was detected in ORF E3L beginning at aa 11 and ending at aa 54. An E-value of 1.81e+01 was assigned by SMART.

Table 6: E5L - Identification of Ankyrin repeats

MOTIF	BEGIN	END	E-value
ANK	21	50	2.79e+01
ANK	51	82	4.13e-02
ANK	87	120	9,92e+02
ANK	124	155	3.60e-02
ANK	159	189	1.70e+02

Five ANK repeats were detected in ORF E5L. The ANK domain in the ORF E5L query of 212 amino acids starts at 87, and ends at position 120. The ANK repeat sequence in ORF E5L was:
NNLSALAHYLSFNKNVEPEIVKILIIDSGSSVTE

Table 7: E6L- Identification of immunoglobulin-like domains

MOTIF	BEGIN	END	E-VALUE
IGc2	4	83	4.79e+00
IG	6	94	1.87e+00
IG-like	71	166	1.37e+01
IGc2	102	171	8.42e-02
IG	96	201	3.38e-01

Due to overlapping domains in ORF E6L, there were three representations of the proteins.

IGc2: Immunoglobulin C-2 type
IG: Immunoglobulin
IG-like: Ig domains that cannot be classified into one of IGv1, IGc1, IGc2 or IG.

Table 8: F4R- Identification of tyrosine phosphatase motifs

MOTIF	BEGIN	END	E-VALUE
PTPc	2	169	24e+01
DSPc	26	169	1.05e-47
PTPc	69	170	6.45e-07

Three tyrosine phosphatase motifs were identified in ORF F4R.

PTPc: Protein tyrosine phosphatase, catalytic domain
DSPc: Dual specificity phosphatase, catalytic domain

Identification of an Epidermal Growth Factor (EGF)-like domain. An EGF-like domain was identified in ORF E9L, beginning at aa 20 and ending at aa 65. An E-value of 1.30E+01 was assigned by SMART.

Table 9: Identification of TRANSMEMBRANE regions

ORF	BEGIN	END
E31	79	99
	123	143
E4L	17	37
	89	109
	127	147
	165	185
	207	227
	256	276
	294	314
	334	354
E10L	48	68
	113	133
F2L	119	139
F5R	24	49
F6L	286	306
K1L (probable)	74	94
K2L	16	36
K3L	47	67
K5L	180	200

Table 10: Identification of SIGNAL PEPTIDE sequences (collective)

ORF	BEGIN	END
E1L	1	17
E2L	1	22
K2L	1	44

3.3.3 LSDV ORFs (Please refer to Table 5 and 3.3.2 SMART Analysis)

Proteins involved in transcription and mRNA biosynthesis. Five LSDV ORFS that could be involved in viral transcription and mRNA synthesis have been identified. All occur in the central regions as would be expected for genes essential for the viability of the virus. ORF G1L encodes the RNA-polymerase-associated transcription specificity factor (RAP94), spanning the *PstI* junction between *PstI* clones -F and -G. The RAP94 of variola virus has been shown to associate with the RNA polymerase, and is required for the transcription of the early genes, where it possibly mediates the binding of the core RNA polymerase to viral early transcription factors (VETF) (Massung et al., 1993). The poxvirus DNA–directed RNA polymerase is known to consist of at least eight subunits (Jackson and Bults, 1990). In this study, two of these have been identified for LSDV, namely the RNA polymerase subunits encoded by ORFs F1R and F3R, the homologs of the MYX 21 kD polypeptide and VV 147 kD polypeptide, respectively. Although ORF F3R shows high homology to its VV counterpart (85% positives), the LSDV ORF appears to be 76 amino acids shorter at the C-terminus. A regulatory subunit of the RNA polymerase, encoded by partial ORF AXR, is expected to extend for another estimated 210 amino acids (including the amino terminal) in the uncloned LSDV *PstI*-A fragment. ORF G5R encodes the large subunit of the mRNA-capping enzyme. The mRNA capping enzyme is a heterodimer of a large and a small subunit, and catalyses the first two reactions in the mRNA cap formation pathway (Upton *et al.,* 1991). The LSDV ORF is 82 aa shorter than its SFV homolog, with approximately 10% of the C-terminus truncated. ORF E10L encodes the small subunit of a putative early transcription factor. It has a relatively low percent identity (26%) to the SPV homolog, and appears to be some 24 amino acids shorter. Two transmembrane regions were predicted by SMART; the first from aa residues 48 to 68, the second occuring from residues 113 to133.

Other poxvirus enzymes also involved in transcription and mRNA biogenesis, but not identified in these LSDV clones include additional RNA polymerase subunits, mRNA transcription, initiation elongation and termination factors, enzymes that direct posttranscriptional processing of viral mRNA, other transcription factors and additional poly(A)polymerase subunits. All of the transcription and mRNA biogenesis protein homologs identified so far in LSDV have been found in the central region represented by *PstI* clones -F and -G, and additional members of the group are therefore expected in the flanking *PstI* fragments of the central conserved region: -A, -D, -L, -I, -H and -C (Figure 5).

Proteins involved in nucleotide metabolism. Enzymes known to be involved in poxvirus nucleotide metabolism include the thymidine kinase, thymidylate kinase, dUTP pyrophosphatase (dUTPase) and ribonucleotide reductase. ORF M1L encodes a viral dUTPase with high (83% positives) homology to the SPV counterpart. This enzyme produces dUMP, the immediate precursor of thymidine nucleotides which decreases the intracellular concentration of dUTP, so that uracil cannot be incorporated into DNA (Massung *et al.*, 1993). The ribonucleoside reductase is a heterodimer of a large and a small chain. Poxvirus ORFs LSDV K1L, SPV and VV F4L all belong to the ribonucleoside diphosphate reductase small chain family (Goebel *et al.*, 1990). These enzymes catalyze the first reaction in the viral DNA replication pathway, providing precursors necessary for DNA synthesis. The catalytic activity involves the conversion of 2'deoxyribonucleoside diphosphate, oxidised thioredoxin and water to ribonucleoside dishosphate and reduced thioredoxin, with a cofactor containing two iron ions (Goebel *et al.*, 1990). ORF K1L contains a probable transmembrane region, predicted by SMART, from aa residues 74 to 94.

Proteins involved in DNA replication and repair. Uracil DNA glycosylase is an important enzyme for DNA replication and repair, because it excises uracil residues from the DNA arising as a result of misincorporation of dUMP residues or due to deamination of cytosine (Upton *et al.*, 1993). ORF G8R encodes such a putative enzyme of 212 amino acids, corresponding precisely in size to the SFV homolog with 75% identity. DNA topoisomerase is able to relax both positively and negatively supercoiled DNA in the absence of an energy cofactor, and the poxvirus topoisomerase homologs, such as ORF G3R and its SFV and VV homologs, which it closely resembles, has the properties of a cellular type I enzyme, except that it is resistant to camptothecin (Bauer *et al.*, 1977; Fogelsong and Bauer, 1984; Shaffer and Traktman, 1987; Shuman *et al.*, 1988). The failure of attempts to construct a mutant vaccinia virus with deletions within the topoisomerase gene suggests it plays an essential role in the life cycle of poxviruses (Shuman *et al.*, 1989). Other replication and repair enzymes of LSDV expected to be encoded in adjacent *PstI* clones include the DNA ligase, uracil DNA glycosylase, DNA polymerase, and replication-essential protein kinase.

Proteins involved in protein modification. Protein tyrosine phosphatases are recognized as important regulatory enzymes, which together with protein tyrosine kinases regulate protein phosphorylation within a cell. Although structurally diverse, both tyrosine phosphatases and kinases can be classified into two major groups: transmembrane or receptor-like and cytosolic or non-receptor-like. Both classes of tyrosine phosphatases contain one or two homologous catalytic domains where an active site cysteinyl residue is located within a conserved sequence motif (Guan and Dixon, 1991). Mossman *et al.* (1995) demonstrated the importance of poxviral phosphatases by showing that MYX I1L is essential to virus viability in tissue culture. Poxviral tyrosine phosphatases constitute proteins that are highly conserved, both within and between different genera (Mossman *et al.*, 1995). The open reading frames which encode for all the known poxviral phosphatases are located within the central portion of the virus genome, adjacent to another essential virus gene, RPO147. ORF F4L encodes a tyrosine phosphatase, flanked by an LSDV RPO147 (F3R), with 76% identity to the MYX tyrosine phosphatase (I1L). SMART analysis additionally identified three tyrosine phosphatase motifs: two PTPc (protein tyrosine phosphatase, catalytic domain) domains from aa residues 2 to 169 and 69 to 170, and a DSPc (dual specificity phosphatase, catalytic domain) motif from aa residues 26 to 169. At a length of 205 aa, the LSDV ORF is slightly longer at the C-terminus than the MYX homolog of 172 aa. Homologous phosphatase genes have also been reported in the central regions of eight other orthopoxviruses to date (Mossman *et al.*, 1995). Vaccinia H1L phosphatase has been reported to be packaged into virions and required for early virus gene expression *in vivo* (Mossman *et al*., 1995; unpublished data), further supporting the essential role of this enzyme. Partial ORF A1L encodes a protein kinase. A predicted 122 amino acids of the N-terminus is expected to complete the ORF in *PstI* fragment A. At 73% identity, the homology to the SPV ORF is relatively high. It is most likely that poxvirus-encoded phosphatases and kinases could influence a variety of cellular functions, ranging from the replication of the viral genome to the disruption of cellular signaling pathways critical for the host immune response.

Structural proteins. Poxviral structural proteins include core proteins, membrane-associated proteins and proteins of the envelope complex. SMART analysis identified several proteins containing transmembrane regions, suggesting membrane association (Table 9). These include ORFs E3L (two regions), E4L (eight regions), E10L (two regions), F2L, F5R, F6L, K2L, K3L, and K5L each with one transmembrane region, and K1L with a probable transmembrane region from amino acid residues 74 to 94. F6L is a putative LSDV envelope protein with 30% identity to VV ORF H3L and the VAR immunodominant envelope protein. No function is assigned to ORF E3L, a homolog of SPV ORF C7. ORF E4L is a G-protein-coupled chemokine receptor homolog and is discussed further on. ORF E10L is a putative early transcription factor subunit. F2L's VV J5L homolog has no assigned function, neither has a function been assigned to the F5R VV putative late protein H2 homolog. The LSDV ORF K1L is a small subunit of the ribonucleoside reductase. ORF K2L, the LSDV homolog of SPV protein C15 has no assigned function. Furthermore, SPV C15 was the only hit, suggesting that this protein is unique to SPV and LSDV. K3L is a salt-inducible protein with no homology to any other poxvirus proteins identified to date, and is also discussed later on. ORF K5L has no assigned function. ORF F6L is a homolog of a known envelope protein, VV H3L.

Excluding ORFs E4L, K1L and F6L, no definite functions for the ORFs containing transmembrane regions are known. These proteins presumably associate with the cell membrane, and might be components of the complex poxvirus capsid structure. Some of them could possibly encode antigenic determinants on the virion surface, which could be involved in host cell receptor recognition and binding. The reasons why an early transcription factor subunit, which theoretically binds to a DNA molecule, would have a transmembrane region are unknown; it is possible that the transmembrane signal plays some role in the packaging of the early transcription factor during viral assembly. ORF G2R is an envelope protein, a homolog of VV H5R, additionally showing similarity to the orf virus envelope antigen. No transmembrane regions were detected in the 226 aa peptide sequence, suggesting that ORF G2R is one of the outer capsid proteins. Outer capsid proteins assemble onto another structural proteins and can be presented on the surface of the virus, possibly playing a role as antigenic determinants.

Host-related functions:

Proteins with immune evasion functions. Growth factors, soluble chemokines, chemokine receptors, interleukins, serpins and epidermal growth factor (EGF) are all implicated in the interference of host immune mechanisms. All of the identified LSDV protein homologs suspected of playing a role in host-related functions and the evasion of the immune response occur in the left-hand terminal regions represented by the the *PstI* clones -E, -M and -K, although the possible role of the tyrosine phosphatase and protein kinase, which occur in *PstI* clones -F and -G respectively, have previously been mentioned. ORF E9L encodes homolog of a SFV growth factor H2. It furthermore has an EGF-like domain from aa residues 20 to 65, identified by SMART. The EGF receptor has been implicated as the portal of entry for vaccinia viruses (Eppstein *et al.*, 1965; Marsh and Eppstein, 1987). ORF E1L encodes an interferon gamma receptor homolog. The first 17 amino acids were identified as a signal peptide sequence, which will determine its destination in the host cell. Two proteins with probable interleukin-binding functions were identified: ORF E8L is an IL-18 binding homolog, and ORF E6L shows homology to the VV interleukin-I binding protein precursor. SMART furthermore identified five immunoglobulin-like domains, from three representations of the proteins due to overlapping domains (refer to Table 7). Two immunoglobulin C-2 type domains were detected, from aa residues 4-83 and 102-171. "C-2" refers to a particular type of domain, found specifically in the constant region of an immunoglobulin heavy chain (Roitt, 1997). Two immunoglobulin domains were identified, from residues 6 to 94 and 96-201, and one immunoglobulin-like domain, which cannot be classified into one of IGvI, ICcI, IGc2 or IG is found in residues 71-166. The poxviral interferon-γ-binding proteins contain relatively low (~20%) homology to the ligand binding domain of the human and murine IFN-γ receptors, although similar lengths are observed within the chordopoxvirus homologs (Mossman et al., 1995). ORF E1L contains nine cysteinyl residues (Appendix IV), eight of which are known to be important for the ligand binding domain. The MYX homolog contains all eight, whereas VV and VAC counterparts have only six, with the first two cysteinyl residues replaced by tyrosine in each case (Mossman et al., 1995).

ORF E4L shows 80% identity to ORF Q2/3L, identified as a G-protein coupled chemokine receptor homolog in sheeppox isolate KS-1 *HindIII* fragment (Cao *et al.*, 1995). The authors were able to demonstrate, through Southern blot analysis, that all three species of the capripox genus (SP, GP and LSDV) contain copies of this ORF. ORF Q2/3L showed 38% sequence identity to SPV ORF K2R (Massung *et al.*, 1993), and 27% identity with HMCV ORF US28 (Human cytomegalovirus chemokine receptor homolog) (Chee *et al.*, 1990). The authors report that a highly conserved motif, DRYLAIVHA, at the end of the third transmembrane domain is present in ORFs Q2/3L and SPV-K2R. Other hits of the same protein family ORFs CKR-1, NY-3R, IL-8ra, and 8rb contain the motif, but it is absent from ORF C5a, fmlp binding proteins (which are also chemokine receptor homologs) and ORF HCMV-US28. SMART identified eight transmembrane regions in ORF E4L (Table 9). A similar motif, DRYLAVVHP, was found from aa residues 186-194, which corresponds to the end of the fourth transmembrane region in LSDV. Certain features of ORF Q2/3L (and LSDV ORF E4L) and the related proteins make them distinct from classical seven transmembrane-spanning receptors: the carboxyl terminus is relatively short and lacks the cysteinyl residue involved in membrane anchorage (O'Dowd *et al.*, 1989) and the segments between the transmembrane domains are quite short (Cao *et al.*, 1995). By comparison of the ORF E4L aa sequence (Appendix III) to ORF Q2/3L aa sequence (Cao *et al.*, 1995), the conservation of several potential N-linked glycosylation sites located mainly in the N-terminus is evident. An additional site in the transmembrane domain is unlikely to be glycosylated, owing to its predicted position within the membrane (Cao *et al.*, 1995).

Proteins with other host range functions. ORF E5L shows very high (97% identity) homology to capripox (SP) ORF Q2/2L, but also shows homology to an orf virus ankyrin-like repeat protein. SMART identified five repeats of the motif NNLSALAHYLSFNKNVEPEIVKILIIDSGSSVTE in the 212 aa LSDV protein, namely from residues 21 to 50, 51 to 82, 87 to 120, 124 to 155 and 159 to 189. The presence of these five ankyrin repeat motifs suggests that ORF E5L binds actin elements of the cytoskeleton, thereby playing a role in tissue tropism. ORF M2L shows 40% identity to a SPV C13 ORF, and also has homology with other known kelch-like proteins, such as the VV 55K protein. Kelch repeats are also implicated in actin-binding.

The RING finger is an evolutionarily conserved structure identified in more than 200 proteins to date (Schultz *et al.*, 2000), in which two loops of amino acids are pulled together at their base by eight cysteinyl or histidinyl residues that bind two zinc ions. In order to keep a cell healthy, some proteins have to be destroyed, and these are tagged for elimination by the attachment a small peptide, ubiquitin. The tagging occurs in steps, with one enzyme binding to the protein marked for destruction and another ferrying the ubiquitin label to the target. Recently, RING fingers have been identified as having ubiquitin-protein ligase activity, identifying them as the molecular mediators that couple ubiquitin and E2 ubiquitin-congugating enzymes so that the tagging can take place (Barinaga, 1999). SMART identified a RING finger motif in residues 11 to 54 of ORF E3L. The ability of a virus-encoded protein to mediate the targeting of host proteins, particularly antiviral or immune-signalling proteins for elimination would have obvious benefits for evasion of the host immune response.

Proteins with cellular functions. A fowlpox virus ORF, FPV114, has been shown to share a 180 aa conserved domain with proteins found in plants, yeast, roundworms and bacteria. ORF FPV114 is most closely related to the yeast Hal3 and SIS2 genes and a putative Hal3 homolog from the plant *Arabidopsis thaliana* (Afonso *et al.*, 2000). These proteins function as inhibitory subunits of cellular protein phosphatases, and they promote salt tolerance and affect growth (De Nadal *et al.*, 1998). LSDV ORF K3L, a 112 aa putative protein, has no homologs in poxviruses but includes a region with low (38%)

identity to a 39 amino acid domain of a *Nicotiniana tabacum*, (common tobacco) membrane-associated salt inducible protein. A possible role for ORF K3L in the host could therefore be in the regulation of the host cellular functions. SMART identifed a transmembrane region from residues 47 to 67, suggesting that ORF K3L too is membrane-associated.

Other proteins of unknown function. In the left terminal region, ORF E2L encodes a 212 aa protein with low percentage identity (26%) to a VV alpha-amanitin sensitive protein and ORF E7L encodes a homolog of hypothetical protein C8 of SPV and a VV 11.1kD protein. In the central region, ORF G6L encodes a homolog of similar length to SFV protein D4, whereas ORF G7R encodes a potential protein of unknown function with highest homology to a 58 aa region of *Plasmodium falciparum* throphozoite proteinase precursor, but also shares homology with SFV protein D5. Partial ORF G9R's 92.6kD protein homolog in FPV has no assigned function.

3.3.4 Nucleotide composition (GCG)

>**Composition** was used to calculate the nucleotide contents of the assembled *PstI* clone DNA sequences

Table 11: Nucleotide Composition

Fragment (bp)	#A (%)	#C(%)	#G(%)	#T(%)
E (7, 024)	2,465 (35.1)	1,033 (14.7)	714 (10.2)	2,812 (40.0)
M (2, 225)	774 (34.8)	373 (16.8)	233 (10.5)	845 (40.0)
K (3, 281)	1,126 (34.3)	467 (14.2)	353 (10.8)	1,335 (40.7)
F (8, 553)	3,278 (38.3)	1,133 (13.2)	1,219 (14.3)	2,932 (34.3)
G (7, 803)	3,187 (40.8)	902 (11.6)	1,066 (13.7)	2,648 (33.9)
TOTAL (28, 886)	36.7	14.1	11.9	37.8

74.5% A+T

A total of 28 886 bp of LSDV genomic DNA sequence has been determined. This is the largest amount of sequence data available for the SA vaccine strain (Neethling), or a capripoxvirus, to date. The 74.5 % A+T calculated (GCG) corresponds closely to the value of 72.4 % determined by restriction endonuclease analysis of Gershon and Black (1989).

3.3.5 Comparison of poxvirus genome organizations

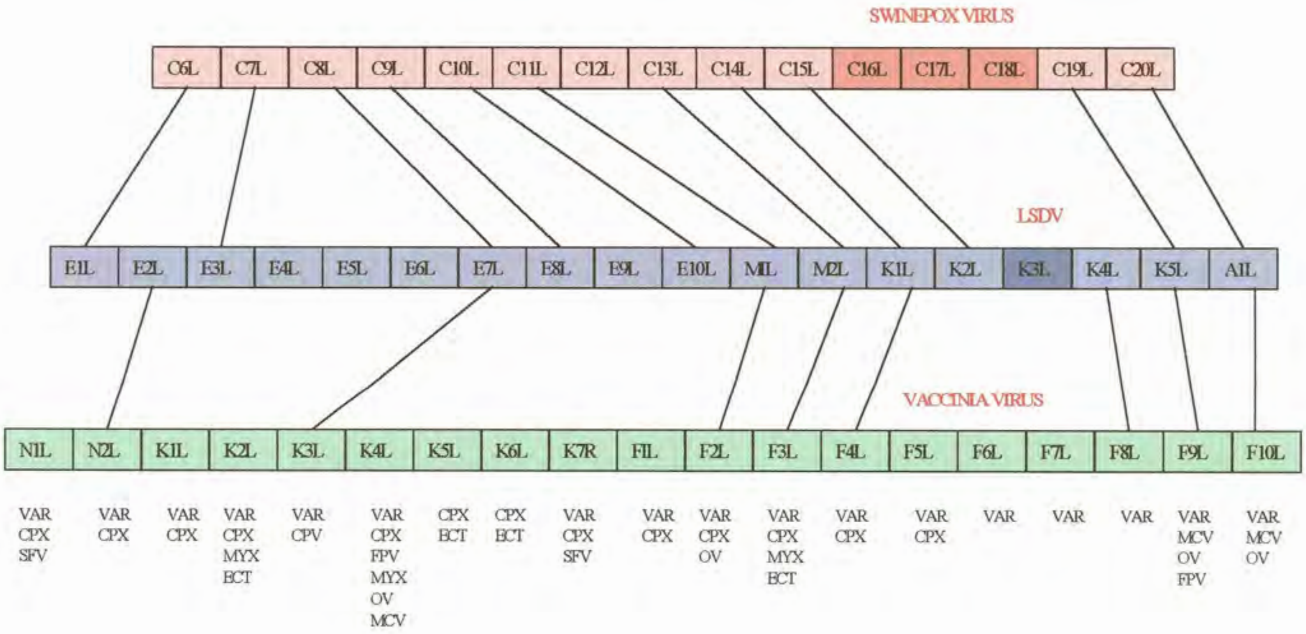
SWINEPOX VIRUS
C6L C7L C8L C9L C10L C11L C12L C13L C14L C15L C16L C17L C18L C19L C20L
LSDV
E1L E2L E3L E4L E5L E6L E7L E8L E9L E10L M1L M2L K1L K2L K3L K4L K5L A1L
VACCINIA VIRUS
N1L N2L K1L K2L K3L K4L K5L K6L K7R F1L F2L F3L F4L F5L F6L F7L F8L F9L F10L
VAR CPX SFV
VAR CPX
VAR CPX
VAR CPX MYX ECT
VAR CPV
VAR CPX FPV MYX OV MCV
CPX ECT
CPX ECT
VAR CPX SFV
VAR CPX
VAR CPX OV
VAR CPX MYX ECT
VAR CPX
VAR CPX
VAR
VAR
VAR
VAR MCV OV FPV
VAR MCV OV

Figure 14: Linear comparison between the left-hand terminal regions of SPV, LSDV and VV genomes. Homologs are indicated by a line; additional poxvirus homologs are indicated in text below the VV genome. The shaded boxes indicate novel genes. Boxes that are neither shaded nor linked by a line have homologs in poxviruses other than SPV, LSDV and VV.

3.3.6 Relationship of LSDV to other poxviruses. Figure 14 illustrates the linear order of the genes in the left near-terminal regions of SPV (top), LSDV (middle) and VV, the protypal poxvirus (below). The vaccinia virus (Copenhagen strain) genome was determined to be 192 kb in length (Goebel *et al.*, 1990), whereas the swinepox virus genome is ~175 kb in length (Massung *et al.*, 1993) and LSDV is estimated to be 152 kb in length (unpublished data). SPV was chosen for comparison because its ORFs came up as the top hit most frequently in the blast searches of the LSDV ORFs against the SwissProt protein database, and show the highest homology to the LSDV ORFS identified, suggesting that the capripoxviruses are more closely related to the suipoxvirus

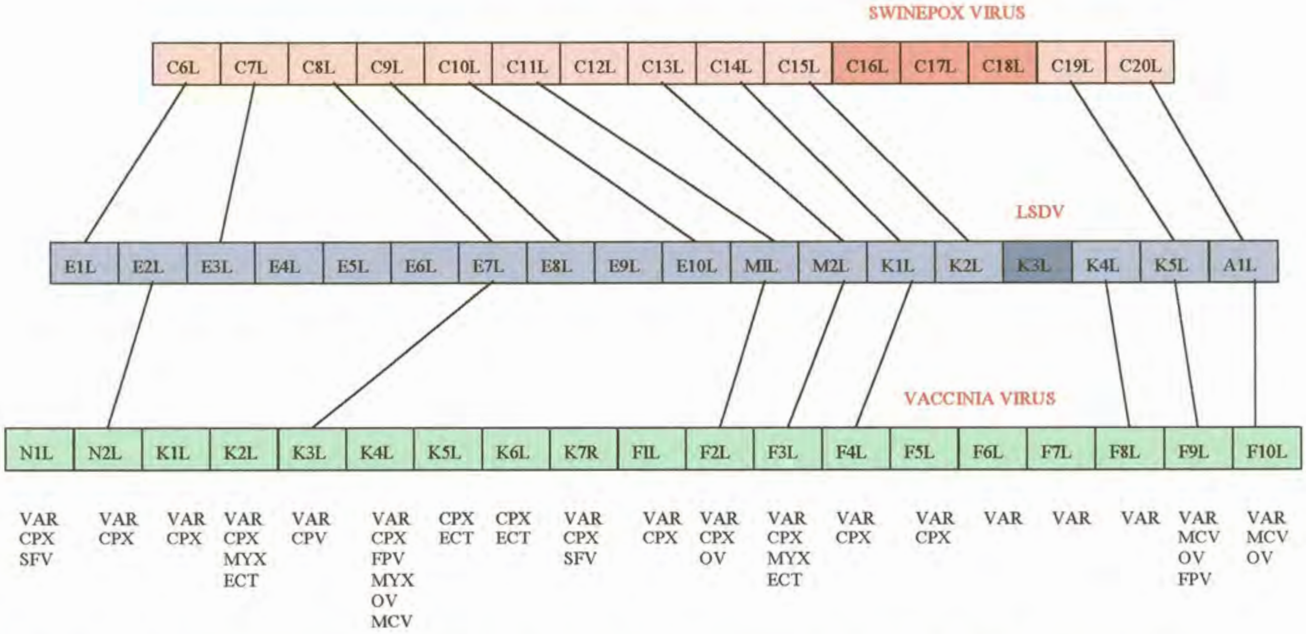

Figure 14: Linear comparison between the left-hand terminal regions of SPV, LSDV and VV genomes. Homologs are indicated by a line; additional poxvirus homologs are indicated in text below the VV genome. The shaded boxes indicate novel genes. Boxes that are neither shaded nor linked by a line have homologs in poxviruses other than SPV, LSDV and VV.

SPV than to any of the other ChPV genera, despite differences in genome sizes, although shope fibroma virus (a leporipoxvirus), at 159.8 kb (Willer *et al.*, 1999) is closer to LSDV in size.

Central region genome organisation of LSDV and the other poxviruses is indeed similar when the linear order of genes (refer to Table 5) is compared to similar tables in publications. This colinearity was also noticed by Gershon *et al.* (1989) who showed that a 100-115 kb piece of the genome is co-linear in organisation according to cross-hybridisation studies of capripoxviruses and vaccinia virus. No cross-hybridisation was detected between VV DNA and the respective left- and right-hand terminal 8 and 25 kb of capripoxvirus DNA, or between capripoxvirus DNA and the respective left- and right-hand terminal 38 and 35 kb of VV, but Figure 14 illustrates that the two regions do indeed have several ORFs in common.

As indicated in Fig. 14, homologous ORFs have been linked by a line. Starting from the left, SPV ORF C6L has a homolog in LSDV ORF E1L, although this protein is not found in VV. LSDV has a homolog of VV ORF N2 (ORF E2L) which is not found in SPV, but also occurs in VAR and CPX. ORF E3L again has a homolog in SPV but not VV. ORFs E4L, E5L and E6L are not the equivalent of VV K1L and K2L, although these ORFs have homologs in poxviruses other than VV and SPV (data not shown). This region of ~2000 bp seems to be absent from SPV, and the three missing ORFs, a G-protein chemokine receptor homolog, an actin-binding protein and an interleukin-I receptor homolog respectively, are all implicated in host-range functions in LSDV. These are presumably not required for the infection of leporid host cells, unless the genes have translocated and are present elsewhere in the SFV genome. The block of ORFs from LSDV ORFs E7L to K2L corresponds closely to SPV ORFs C8L to C15L, although LSDV ORF E9L is absent from SPV but present in other poxviruses (data not shown). ORF E9L is a growth factor implicated in the LSD virus/host interaction. LSDV ORFs M1L, M2L and K1L, a dUTPase, actin-binding protein and ribonucleoside reductase enzyme are conserved in most poxviruses. ORFs M1L and K1L are involved in nucleotide metabolism and therefore essential to poxviruses. SPV ORF C12L is similarly absent from LSDV, but is homologous to VV ORF K7R (no line indicated in Fig. 14). It is interesting to note that VV ORF K4L which seems to be widely conserved among genera, encodes a 37k major EEV antigen IMCBH sensitive protein (palmityl protein) (data not shown) which is absent from the suipoxviruses and capripoxviruses. Furthermore, three genes of SPV: C16L, C17L and C18L, are not present in this region of LSDV, and do not correspond to a similar location in VV (containing VV ORFs F5L, F6L and F7L). These last three genes seem to be unique to orthopoxviruses (VV, VAC, CPX) as no homologs for members of other genera have been identified as yet. LSDV ORF K3L has no homology to any other poxvirus protein, and has therefore been highlighted. As a potential gene unique to LSDV, it possibly contributes to the bovine-specific nature of LSDV infection. The ORF resembles salt-inducible proteins, with roles in cellular functions. LSDV ORFs K4L, K5L and A1L again have homologs in VV ORFs F8L, F9L and F10L. ORFs K5L (unknown function) and A1L (encoding an essential protein kinase) are seemingly conserved among poxvirus genera.

3.3.7 Discussion of the regulatory elements

```
                    critical              spacer     initiation
                     region               region       region

               -28               -13 -12          -2 -1      +6

E2L            ATAAAAATGAAAAAAAA  TATTACACTTT  TTAAAAA  ...19  ATG
E5L            ACAAAAATGAAATATAA  CTTTTTATTAA  TAATAAG  ...26  ATG
K1L            AAAATAATGAAATATAA  TTTTTATGGTA  TAAAATA  ...7   ATG
K4L            TTAAAAATGAAACATAA  ATCATAGCTAA  TAACATA  ...23  ATG
G9R            GCTTTAGTGAAATTTTA  ACTAACTTGTG  TATTAAA  ...1   ATG +L
F1R            TAAAATATGAAAAAAAG  ATGTTTTATTT  TAATAAA  ...193 ATG
G8Ra           GGTGTAATGAAATTATT  CAAAAAAATGA  TAATATC  ...50  ATG

G8Rb           TAGAAAATGAAAAGGTA  AAAAAAAAATA  CAATGAT  ...4   ATG
F6L            GGATAGATGAAAACATA  ACTAAAATTAG  AGAGCTA  ...51  ATG

E4L            AAAAAAATAAAATAAAA  GTGTATGATTA  AGTAAAG  ...16  ATG
K2L            TAAAAAATAAAAAAAAG  TATAAATTTTA  CAATAGT  ...8   ATG
K3L            AAAAAAATAAACTAAGT  TATACGTTATT  AGATAGT  ...9   ATG
G5R            AACAAAATAAAAAAATA  ATTAACGCAAG  TAAAAAA  ...32  ATG

E6L            GATAAGATGGAAAAAGT  AACGACATTTA  TTTTCTA  ...68  ATG


VV early promoter
consensus      AAAAAAATGAAAAAAAA
VV strong
promoter       AAAAAAATGAAAAACTA


LSDV promoter
consensus      AAAAAAATGAAAAAAAA  AATAAAATTTA  TAATAAA
```

Figure 15: Alignment of LSDV early promoters with the VV consensus (Davison and Moss, 1989). The deduced LSDV consensus is indicated at the bottom.

Early promoters. Nine examples of early promoters which closely resemble the VV consensus have been identified and aligned (Fig.15). These are situated upstream of ORFs E2L, E5L, K1L, K4L, G9R, F1R, G8Ra, G8Rb and F6L. The ORF E2L promoter region is the best example of this, with the closest homology to the VV consensus sequence, having only one nucleotide difference at -27 in the critical region. With the exception of upstream regions of ORFs G8Rb and F6L, all of these contain a T at –1 and an A at +4. There is no evidence to suggest that these specific nucleotides are critical for early promoter function, but they do provide a useful guideline in the blast identification of LSDV early promoters. The initiation region for ORF F1R's promoter is unusually large, and it is possible that the true site for initiation occurs further to the initiation site in other A/T- rich tracts closer to the ATG start site. However, none of these A/T-rich tracts were found to resemble the early promoter consensus as closely as the presented sequence. The promoter type for the corresponding homolog in SFV happens to be unknown (Willer *et al.*, 1999). The region immediately upstream of the ORF G9R initiation codon, presented in the early promoter alignment as an early promoter also has the characteristics of a late promoter (Fig. 16). This would suggest that ORF G9R can be expressed both at early and late times, and that the product, of unknown function, may be required in large amounts by the virus.

Cao *et al.* (1995) reported that although the sheeppox Q2/3L ORF has an upstream A+T –rich region. The Q2/3L ORF shares an 80% nucleotide sequence homology with LSDV ORF E4L. There is no obvious sequence upstream of ORF E4L matching the conserved motifs found in early or late poxvirus promoters. Although the proposed promoter for ORF E4L has a CAAAAT motif adjacent to the RNA start (Fig. 16), the upstream region is relatively G+C-rich, whereas this region is characteristically A+T-rich in late and intermediate promoters. The early promoter sequence in the alignment was therefore considered to be the best candidate for a regulatory element. Although the critical region contains a TAAAA motif instead of a TGAAA motif for promoter regions of ORFs E4L, K2L and K3L and G5R, the VV 42 kD protein (Venkatesan *et al.*, 1982) promoter, which has been characterized as a definite early promoter (sequence not shown), shows a similar motif.

The region immediately upstream of ORF E6L lacks any motif to suggest a late or intermediate promoter. However, the regions further upstream are A+T-rich, resembling the early promoter motif. A TGGAA sequence here, is reminiscent of the TGAAA early promoter motif. Although the region upstream of ORF G5R has a TGTAA motif instead of a TGAAA motif, the VV RNA polymerase gene (Broyles and Moss, 1986) shares the same pattern and has been characterized as a definite early promoter.

```
            -40                 -19                   +1  +4    +9

F3R           GTTAATATATTGTATTTGTATTAATAAAAAACCTAGTAA TAAAT GGCAG
F4L           ATTAACATTTAATGTTGTTTTATCCATTTGTAGCCATTA TAAAT GGATA
K5L           AAAAAATTATTTCAAAGAAAAACATTTTCTCAACGTGGA TAAAT GGAAA
G4R           TAGTAAATAGAGACACTATATAAAAAAAACCTTAAACGA TAAAT GGATG
G1L           ATAAATCTATTATAAATTTTGCCTTTTGTATATAATATT TAAAT GGAAA

F5R           GTTTTCGTAAAGACTTTTTTTATCCATTTATAATGGCTA CAAAT GGATA
K2L           AAAAAATAAAAAAAAGTATAAATTTTACAATAGTGGTAG CAAAT GCATC

F2L           CTGATTTTTTCTTTTTTTCTTTTTTTTTCTTTTATTATA AAAAT GTTAG
G3R           AAACATCAAAATTATGATTTTTATTTTTATTAGCGAATA AAAAT GAGAG
G9R           GTTTAAGCTTTAGTGAAATTTTAACTAACTTGTGTATTA AAAAT GGCAG
E10L          GTTTTTTATAATAAAGGAAAATATAAGTTACTTTTTGTA AAAAT GGATA

E7L           AAAAAACTTTAATATTGAACGTTTCAGAAAGTGAGAGCA AAAAT TCATC
E4L           AAATAAAAGTGTATGATTAAGTAAAGCACAACTCCAACA AAAAT GAATT
E8L           TTTTAAAAAAAAAATATAAAATATAGTTCTAAAAAATAA AAAAT GATCT

G7R           GATCTAATAATAATCAAGATATTCTTATTTTGAAAAAAA AATAT GGAAG


VV CONSENSUS  GGATTTAATAAAAAATATTTTAAAAAAAATTTTCAAATA TAAAT GGATA
              ATC A   C  T            T T T    A T              G
              TA      A


LSDV CONSENSUSATTAAAATATAATATATTTTTTTTTTTTATTATAGATTA AAAAT GGAAG
               AA         A    A    A         G A A
```

Figure 16: Alignment of LSDV late promoters with the VV consensus sequence (Davison and Moss, 1989). The deduced LSDV consensus is indicated for comparison.

Late promoters. By visual examination and blast searches of the regions immediately upstream of the RNA start sites for all the LSDV ORFs identified, upstream regions of ORFs F3R, F4R, K5L, G4R and G1L were found to closely match the VV late promoter consensus. All four contain the typical TAAATG site within which transcription initiates. The four nucleotides at +5 to +9 of F4L, the putative tyrosine phosphatase gene, are an exact match with the VV consensus.

ORFs F5R and K2L promoter sequences initiate with CAAATG instead of TAAATG, although this specific mutation have been shown in vaccinia virus to have a minimal detrimental effect on transcription (Davison and Moss, 1989).

The region upstream of ORF F6L (Fig. 17) closely resembles the late/intermediate motif. Position –19 is however occupied by a C, a mutation which has been shown to be detrimental to late promoter function. Although it lacks the upstream AAANAA motif characteristic of intermediate promoters, the region upstream of ORF F6L has been aligned with the intermediate promoters. An early promoter further upstream was identified for ORF F6L and is presented in Figure 15. The VV P35 immuno-dominant envelope protein, to which this LSDV ORF F6L shows 30% identity, has however been described as a late protein (Goebel *et al.*, 1990).

The regulatory region upstream of ORF E7L lacks motifs resembling the early promoter consensus, but initiates with CAAAAATG. There are no motifs to suggest intermediate promoter status, and the region upstream of the SPV homolog, ORF C8L, has additionally been described as a late promoter Massung *et al.*, 1993). Similarly, the region upstream of SPV ORF C9L, the homolog of LSDV E8L, initiates with an AAAATG motif and was also classified as a late promoter.

ORF G7R has no obvious upstream early promoter motif, and the A+T-rich initiation region ATAATG has previously been cited as a poxvirus late motif for SPV ORFs C1L, C8L, C12L, C16L and K4R (Massung *et al.*, 1993). ORF E10L too contains this upstream ATAATG motif, and an upstream region is additionally presented in the intermediate promoter alignment (Fig. 17). The AAANAA motif characteristic, of intermediate promoters, lies upstream at the unusually high position of –20.

```
F6L             ATTTTAATAACTAAATAGTGCTTTAGATAACTAATTATC AAAAT GGCAG
MIL             AAAAAACAAAAAAAAAAATATCCTAATTTTAAAAAATAAC AAAAT GGATC (I/L)
E1L             TTTTTTGACGAAATAAGTCATTTTAGAGAAAAGAATTTG AAAAT GGGCG (I/L)
E3L             AAATAAAAAAATAATAACAACAAATAATGATATTTTATA AAAAT GGAAG (I)
M2L             ATAAAGAATATGTAGAAGAACTAAAGTACATGGTATTAA AAAAT GATTA (I/L)
E10L            AGAATAGAATCATTAAATAATAATGAGTTGGCAATTTTA ATAAT GATAC (I/L)
G2R             ATTATAAAAATAGCTTCATATAAATACAAACAATTATTA AAAAT GTCGT (I/L)

VV I3                        GGTTAAACAAAAACATTTTTATTCT CAAAT GAG
VV I8                          GGTAACTCAAAAACATCATATAT AGAAT GGTAA

                ATAATAAAAAAAAAAAATAACTAAAGATAATAATTTTA AAAAT GGTAG
                                     A   A    A                A
                                              C                C
```

Figure 17: Alignment of LSDV intermediate promoters with two *characterize*d VV intermediate promoters, I8 and I3 (Hirschman *et al*., 1990) and the LSDV consensus sequence.

Intermediate promoters. (Fig. 17) The region upstream of ORF M1L resembles that of a late promoter, but it could also has characteristics of an intermediate promoter as position –19 is occupied by a C. The AAANAA motif starts at +3. ORF E1L is preceded an upstream AAAGAA motif, and is therefore also considered to be transcribed in the intermediate time interval. The ORF M2L upstream region lacks the AAANAA motif, but position –19 is a C, whereas the ORF E3L upstream region contains both the intermediate upstream motif, and the C at –19. ORF G2R again only has the intermediate promoter motif. The VV I3 and I8 promoter sequences (Hirschman *et al*., 1990) which have been characterized as intermediate promoters in VV, are presented for comparison. In both cases, a C-residue is present at 18 instead of -19. A consensus for an LSDV intermediate promoter is also presented.

3.4. Conclusions

In this chapter, the edited fragment sequences described in Chapter II were assembled into two continuous nucleotide sequences, one from the LHT of the LSDV genome, and the other from the central region, which were translated to amino acids, from which the ORFs were identified. To be able to recreate the full-length sequence, the individual fragments were first translated to amino acid sequences, and blasted against public protein databases. Consequently, it was possible to assemble the fragments into the correct orders and orientations, based on full-length poxvirus homologs from the databases. Gaps revealed in the LSDV nucleotide sequences were filled by primer walking sequencing, also described in Chapter II.

At present, these 12530 bp from the LHT and 16356 bp from the central region represent the largest amount of continuous sequence data available for the LSDV genome (Neethling str.). The 3600 bp *PstI* J-fragment, of the same isolate has also been sequenced, by an Australian group (personal communication). Further analysis of the LSDV ORFs included the identification of specific protein motifs (SMART) which can aid in the assignment of possible protein function in the host cell, and the identification of upstream promoter sequences for the regulation of transcription.

In total, 35 conserved gene homologs were identified in the sequenced regions, including proteins involved in transcription and mRNA biosynthesis, nucleotide metabolism, DNA replication and repair, protein processing and virion structure. Comparison of the LSDV genome with those of other ChPVs (Table 5 and Fig. 14) furthermore revealed extensive genome colinearity in the central regions, and lesser colinearity in the terminal region. Predicted proteins with putative functions involving immune evasion included a G-protein-coupled chemokine receptor, an interleukin-18-binding protein, an interleukin-I receptor homolog, a growth factor and a putative interferon-binding protein. Other potential LSDV host range proteins included homologs of those involved in tissue tropism (e.g. an ankyrin repeat-containing protein, a kelch repeat-containing protein and a ring finger repeat- containing protein). In addition to a novel LSDV protein, several proteins with homologs but unknown function, such as those containing immunoglobulin domains were identified. This diverse complement of genes with likely host range functions in LSDV suggests that it is highly adapted for replication in the bovine host.

CHAPTER IV

Concluding Remarks

There can be no doubt that poxviruses have played an important role in the history of humankind, A.D The smallpox scourge of previous centuries devastated entire civilisations, and caused massive loss of life in others. The directed efforts at control and the eventual eradication of the disease in 1977, was a social and scientific triumph. The means thereof, viz. the deliberate inoculation of healthy individuals with the related but non-pathogenic vaccinia virus, coined the term and founded the principles that are currently universally adopted to prevent diseases of many other pathogens from spreading. Despite the eradication of the human pathogen (smallpox), other poxvirus species continue to be a veterinary threat, with serious impact on animal health and national economies. As a first step to combating serious diseases however, a thorough knowledge of the pathogen is crucial. Furthermore, it is necessary to be able to understand how and why it replicates in specific host cells, and how transmission from these cells occurs.

In this regard, genome projects are becoming increasingly important, and poxvirus genomes are no exception. Computer-assisted genome analysis of the sequence data is providing a basic knowledge of what types of roles the putative pox proteins identified might be playing in the host cell. Furthermore, as increasingly more completed poxvirus genome sequences become available, comparisons can reveal which proteins are specifically responsibly for certain abilities of the particular virus. Despite the great variations in genome sizes between and within poxvirus genera, a central core of between 100 and 115 kb is conserved in gene composition and length (Gershon *et al.*, 1989), encoding the genes known to be essential for poxviral growth and survival, and deletions in these regions have often been shown to be fatal to the viruses. Deletions in the terminal regions however, whose varying lengths are responsible for general variations in genome sizes in different poxvirus species, still permit virus growth in tissue culture (Wittek *et al.*, 1980). The genes of these regions are implicated in the ability of different poxviruses to infect different types of cells and cause a particular severity of infection. Furthermore, each poxvirus genome so far sequenced has been shown to encode genes unique to that particular virus. A knowledge of the locations of these host range specificity and virulence genes is particularly important, as poxviruses are considered to have enormous potential as viral veterinary expression vectors: poxviruses have a large capacity to accommodate foreign genes (antigens), and the limited replication of a poxvirus in a non-host species still seems to lead to a protective immune response in the host (Taylor and Paoletti, 1988). Capripoxviruses, possessing some of the smaller poxviral genomes due to their somewhat shorter termini and have received much attention in recent years, as the sequencing of their genomes is expected to shed some light on the specific factors that influence host-range specificity and virulence of poxviruses.

In this study, the two largest continuous stretches of nucleotide sequence for a capripox virus genome to date (one region from the LHT of the genome and the other from the central region) are presented, along with the identification and computational analysis of the genes encoded in those regions. Specifically, the subject of this study was the Neethling strain of LSDV, an avirulent bovine-specific capripoxvirus used as a vaccine for the past 50 years in South Africa. The general poxvirus genome organisation, initially implied by the results of hybridisation studies undertaken by Gershon and Black in 1989, was verified again. The central region represented by *PstI* clones -F and -G, closely resembles corresponding regions in many other Chordopoxviruses, in gene content, homology and linear organisation. The near-left terminal region however has revealed some unique features. Firstly, the *PstI*-E fragment is expected to be within 6000 bp of the left hairpin loop, yet no inverted repeat sequences have been identified. If LSDV inverted terminal repeats could be short and few in number, and are expected to occur within the yet un-cloned and -sequenced regions. Secondly, one gene unique to LSDV has been identified, so far: ORF K3L has no homolog in any other poxvirus, and only low homology to a salt-inducible protein of the tobacco plant *Nicotiniana tabacum*. The poxvirus to which LSDV seems to share the closest homology with in terms of gene content, organisation and percent identity, is the *suipoxvirus* swinepoxvirus (SPV). This virus is host-restricted to pigs. One would have assumed closer homology to poxviruses such as variola virus and cowpox virus (both orthopoxviruses) which are known to infect the same hosts as capripoxviruses. Two host-restricted viruses (infecting totally different species) therefore appear to be more closely related to each other, than either is to a larger genus with a broader host range. This may suggest a common ancestor in the evolution of

LSDV, sheeppox, goatpox and SPV. If the gene composition is compared, LSDV proteins differ slightly in length and amino acid sequence, with truncations (where applicable) occuring mostly at the C-terminus. Seemingly slight variations such as these might introduce subtle changes in the protein folding, enabling the recognition of specific bovine receptors and not those of sheep and goats, or indeed other animal species, in the recognition and infection of cells. In this way, slight changes over a broad array of virus proteins may have a greater effect on the infectivity of the virus for cattle, than the presence of one or more absolutely unique LSDV proteins that recognise a bovine cell-surface receptor. The presence of such genes cannot be ruled out, however, as only a relatively small percentage of the total LSDV genome has been sequenced and analysed in this study.

In the event of the discovery of such a gene, whose absence or disruption would cause a drastic change in the infectivity of LSDV for cattle or an enhancement of its infectivity for another species or range of species, an excellent target for the engineering of novel vaccine viruses and expression vectors with enhanced efficacy and greater versatility would have been located. Additionally, the identification and characterization of specific LSDV virulence and host range genes would contribute to our overall understanding of pathogen-host interactions- information that is likely to have a broad impact on future studies for controlling bovine infectious diseases in general. These objectives will be attainable once the entire LSDV genome is completely sequenced, a grand-scale collaborated project which is likely to be completed within the next two years.