Sequence and comparative analysis of the genome of HSV-1 strain McKrae

G. Watson[1,3], W. Xu[4,5], A. Reed[1,2], B. Babra[1,3], T. Putman[1,3], E. Wick[2], S. Wechsler[6,7,8], G. F. Rohrmann[2], L. Jin[1,2]

Department of Biomedical Sciences, College of Veterinary Medicine[1], Department of Microbiology[2], Molecular and Cellular Biology[3], Oregon State University, Corvallis, OR 97331. Supercomputing Institute for Advanced Computational Research[4], University of Minnesota, Minneapolis, MN 55455. Department of Veterinary and Biomedical Sciences[5], 1971 Commonwealth Avenue, Saint Paul, MN 55108. Gavin Herbert Eye Institute[6], Department of Microbiology and Molecular Genetics[7], Center for Virus Research[8], University of California Irvine, Irvine, CA 92697.

Running title: HSV-1 genome sequencing and comparison

Keywords: herpes virus, high throughput DNA sequencing, DNA sequence comparison

Word count: 5117

Figures: 5

Tables: 3

Supplemental Table: 1

Supplemental Figures: 2

Accession No: JQ730035

*Address Correspondence to:

Ling Jin

Department of Biomedical Sciences

College of Veterinary Medicine,

Oregon State University,

Corvallis, OR 97331.

Email: ling.jin@oregonstate.edu

Phone: 541-737-9893

Fax: 541-737-2730

Abstract:

Ocular infection by HSV-1 strain McKrae is neurovirulent in both mice and rabbits and causes fatal encephalitis in approximately 50% of animals. In addition, it spontaneously reactivates with high frequency relative to other HSV-1 strains in rabbits. We sequenced the McKrae strain genome and compared its coding protein sequences with those of six other HSV-1 strains. Most of the 74 predicted protein sequences are conserved; only eleven are less than 98% conserved. Eight proteins were identified to be unique for McKrae based on sequence homology bit score ratio (BSR). These include five proteins showing significant variations (RL1, RS1, UL49A, US7, US11), two truncated proteins (UL36 and UL56) and one (US10) containing an extended open reading frame. The McKrae strain also has unique features in its 'a' sequence and non-coding sequences, such as LAT and miRNA. These data are indicative of strain variation but need further work to connect observed differences with phenotype effects.

Introduction:

Herpes simplex virus type 1 (HSV-1) is a widespread human pathogen that establishes a lifelong infection characterized by recurrent cycles of latency and reactivation. To date, there are 17 strains of HSV-1 that have been isolated (ICTV 2011 taxonomy). Among them, 7 genomes have been completed or partly completed. The reference strain 17 (NC_001806, X14112) is often used for genome sequence comparison. HSV-1 has a large double stranded DNA genome of approximately 150 kb that is composed of a unique long (UL) region (106.5 kb), a unique short (US) region (13.5 kb), two copies of a long inverted repeat (RL) (8.75 kb each), and two copies of a short inverted repeat (RS) (6.25 kb each).

During primary infection, HSV-1 undergoes productive replication in epithelial cells typically of the oral, nasal, or ocular mucosa. HSV-1 virions eventually infect the trigeminal ganglia by entering nerve endings at the site of infection, and a life-long latent infection is established in a subset of neurons (Hill et al., 1996). The HSV-1 genome is largely inactive during latency, except for expression of the latency associated transcript (LAT) (Rock et al., 1987; Stevens et al., 1987). HSV-1 latency may reactivate spontaneously or in response to stress resulting in production of infectious viruses/ particles. Infectious particles are transported to peripheral sites

where lytic replication ensues. Investigations into HSV-1 reactivation have implicated several viral gene products and genome regions including RS1 (ICP4) (a major transcriptional activator/repressor), RL1 (ICP34.5) (a neurovirulence factor), RL2 (ICP0) (a transcription regulator and ubiquitin ligase), and LAT (Halford et al., 2001; Halford and Schaffer, 2001; Perng et al., 1994). Ectopic expression of RS1 (ICP4) or RL2 (ICP0) induces viral replication from a quiescent state *in vitro*, and deletions in the LAT promoter or LAT transcript have been shown to severely diminish reactivation *in vivo* (Block et al., 1993; Halford et al., 2001; Halford and Schaffer, 2001; Hill et al., 1990; Leib et al., 1989; Perng et al., 1994; Perng et al., 1996; Perng et al., 1999). These genes all map to the inverted repeat regions of the viral genome. In addition, both micro RNAs (Umbach et al., 2008) and apoptosis have been implicated in latency reactivation (Jin et al., 2005; Perng et al., 2000). Although several factors influencing reactivation from latency have been identified, the study of HSV-1 reactivation is complicated by the fact that reactivation frequency and ability to reactivate are both strain and host specific *in vivo* (Hill et al., 1987). For example, in the rabbit model, HSV-1 strains McKrae, 17, and KOS each have a distinct reactivation phenotype (high, medium, and low frequency, respectively) (Hill et al., 1987; Perng et al., 2002).

In addition, HSV-1 strains show variation in virulence and pathogenesis (Perng et al., 2002; Stroop and Schaefer, 1987). Among laboratory strains (McKrae, 17, KOS, F, and H129), McKrae is relatively virulent and neuroinvasive in mice and rabbits, resulting in 50% mortality at 2×10^5 pfu per eye in animals infected by the ocular route without corneal scarification (Jin et al., 2007; Jin et al., 2005), whereas strain 17 and KOS infect the rabbit eye poorly without corneal scarification (Hill et al., 1987; Perng et al., 2002). RL1 (ICP34.5) is a neurovirulence factor, and the McKrae strain variant has been shown to confer the virulent phenotype in recombinants (Bower et al., 1999; Mao, 2002; Mao and Rosenthal, 2003; Perng et al., 2002). In this report we describe the use of high throughput sequencing to determine the HSV-1 McKrae genome sequence and employ comparative genomic analyses using previously sequenced HSV-1 genomes from strains F, H129, 17, HF10, and KOS sequences to locate protein and genomic sequence variability between strains that may contribute to the neurovirulence and high frequency reactivation associated with the McKrae strain.

Results and Discussion

The McKrae Genome Sequence

The HSV-1 genome has an architecture composed of a unique long (UL) element, a unique short (US) element, and two distinct inverted repeat elements which flank the UL and US termed the repeat long (RL) and repeat short (RS) (Fig. 1). A total of 243,155 sequence reads were generated from 454 GS FLX+ system with the sequence length ranging from 40 bp to 1209 bp with mean sequence length at 339 bp and mod read length at 429 bp. The N50 contig size is 22,626 bp. More than 99% of the sequences have an average base Phred quality score of greater than 20 (Supplemental Fig. 1). The total bases are 82,442,028 with the genome coverage of 589.6X. From 243,155 sequence reads, the initial *de novo* assembling using Newbler software generated nine large contigs homologous to the UL, US, RL, and RS of the strain 17 genome (Fig. 1). There are 12,808 sequence singlets that were not assembled and may be the host source sequences or low quality sequences. PCR sequencing was used to confirm contigs and gaps. Excluding one copy of each repeat element, the strain 17 reference genome (Accession No. X14112) is approximately 136,500 bp, indicating that the initial contigs covered nearly the entire genome. Contigs were ordered by alignment to the strain 17 genome. The McKrae genome was assembled in such a manner as to create one continuous contig beginning with viral gene UL1 at the left end and viral gene US12 at the right end. Terminal repeat elements were added after the above sequence was determined by inverting the respective long and short internal repeat sequences and appending them to the draft sequence ends. This resulted in a final sequence of 151,135 bp. Omitting tandem repeats present at the UL – internal repeat long (IRL) and US – internal repeat short (IRS) junctions, the HSV-1 McKrae genome regions are: UL 107,759 bp, US 13,514 bp, IRL 8,720 bp, IRS 6,200 bp. The length of internal repeats reported here include the viral ‘a’ sequence. The genome contains all 77 open reading frames (ORFs) known to encode proteins in the HSV-1 genome (this includes repeated coding sequences in the inverted repeat elements), as well as two ORFs corresponding to ORF-O and ORF-P in the long inverted repeat. ORFs shared by McKrae and strain 17 show synteny across the entire genome.

The HSV-1 genome is GC rich and contains many repeat elements, which can cause

complications when attempting to assemble a genome *de novo*. The Roche 454 sequencing platform was specifically chosen because it generates relatively long reads that were hoped to span tandem repeat elements. One drawback, however, is its inability to resolve mononucleotide repeats past seven or eight residues due to detection saturation. Several regions in the McKrae genome contain long mononucleotide repeats, and in an alignment to strain 17 we noted deletions in strain McKrae mononucleotide runs. It's unclear whether these deletions represent real variation or natural fluctuations in repeat size, or are simply sequence regions where saturation was reached. Discrepancies within coding regions were confirmed by PCR product sequencing to ensure accurate predicted protein sequences for comparative analyses. High throughput sequencing also had difficulty sequencing particular genome regions [specifically oriL and RS1 (ICP4)]. However, this same deficiency was seen using standard Sanger methods, suggesting the underlying sequence was responsible for the difficulties as opposed to the sequencing method.

Sequence Comparison of Predicted Proteins

Predicted protein sequences from HSV-1 strains McKrae, 17, F, H129, HF10, and KOS were aligned for comparative analyses using Geneious Software. Fully sequenced and annotated HSV-1 genome data is available for F strain (Accession No. GU734771), H129 (Accession No. GU734772), and strain 17 (Accession No. X14112). The strain HF10 (Accession No. DQ889502) sequence is not complete, but most coding regions are annotated. We included strain HF10 predicted protein sequences in our analysis, but omitted any non-coding comparisons due to missing sequence data and an uncharacterized reactivation phenotype. In addition, two fully sequenced and annotated KOS genomes (Accession No. JQ780693 and JQ673480) were published recently. Both of them were included in our analysis.

The percentage of identical amino acids was calculated in multiple alignment for each protein. Of the 74 unique predicted protein sequences found in HSV-1, 63 proteins are ≥ 98% conserved between strains, whereas 11 proteins are less than 98% conserved (Table 1). In an earlier comparative analysis of three HSV-1 strains (17, F, and H129), ten protein sequences were found to be completely conserved (Szpara et al., 2010). The addition of three strains (McKrae and KOS) in these analyses revealed only four proteins (UL16, UL20, UL35, and UL45) to be conserved between strains, and only two proteins (UL35 and UL45) were found to be completely

conserved when HF10 was included and all seven strain sequences were compared (Tables 1& 2). Not all of these perfectly conserved proteins are necessary for viral growth in culture, and it has been suggested that maintenance of coding sequence in these proteins is the result of selection and enhanced fitness (Szpara et al., 2010). The addition of McKrae, HF10, and KOS coding sequences has narrowed the focus of this list and has provided more evidence for an evolutionary advantage associated with sequence conservation in these proteins.

Identification of Protein Variants

The goal of this comparative analysis was to identify protein or sequence features unique to strain McKrae that might contribute to its characteristic pathogenesis and high frequency of spontaneous reactivation. We searched the homology of each protein of McKrae against six other strains by BLAST tool (blast.ncbi.nlm.nih.gov) and calculated the sequence homology bit score (BS). A bit score matrix was generated by searching each protein of McKrae against that of six other strains as well as that of McKrae itself (Supplemental Table 1). The homology bit score ratio (BRS) was the ratio between the average BS of each other strains against McKrae BS and the BS of McKrae against McKrae. BSR value then is a representation of how different (substitutions and indels) a McKrae protein is from a relevant protein of other strains. If there is no difference, the BSR will be 1; however, the BSR will be lower than 1 if there are variations between McKrae and other strains. Using this approach, we were able to generate a list of proteins containing variation comparing to those from McKrae (Table 2). Proteins UL56, RL1 (ICP34.5), US10, RS1 (ICP4), US11, US07 (glycoprotein I; gI), UL36 and UL49A are found to contain a relatively large amount of variations (BSR < 0.85), which include proteins that were previously identified to be associated with pathogenesis and latency reactivations, specifically RL1 and RS1.

Protein Variants

The identification of RL1 (ICP34.5) using the BSR method indicates it is capable of identifying protein variants that have been linked to a specific phenotype (Mao, 2002; Perng et al., 2002). The ICP34.5 protein of HSV-1 is a neurovirulence factor that plays critical roles in viral replication and anti-host responses (Bolovan et al., 1994; Li et al., 2011). Strain McKrae RL1 (ICP34.5) has a lower BSR (0.387) due to an extended P-A-T repeat between residues 159 and

160 that results in 8 iterations while other strains contain only 3-5 iterations. This extended repeat has been shown experimentally to control cellular localization of the RL1 (ICP34.5) protein, while a chimeric virus suggests this region in important for virulence phenotype when integrated into a non-virulent HSV-1 strain (Mao, 2002; Perng et al., 2002).

The RS1 (ICP4) of McKrae has been previously identified to be important for latency reactivation *in vivo* (Halford et al., 2001). RS1 is an immediate early transcriptional regulator and recognizes DNA motifs in several immediate early genes as well as in the LAT promoter (Kuddus et al., 1995; Shepard et al., 1989). The coding sequence contains a block of variation in an alanine rich region (AASAPDAADALAAA) between residues 707 and 720, which are present in the other six strains but not in McKrae (Fig. 2), where the alanine rich region is replaced by a serine rich sequence (GPRRSSSSSGVAA-) (Fig. 2). The serine rich block of substitutions present in McKrae is adjacent to the nuclear localization signal (NLS) (amino acid 728 - 734). A change in conformation of this region may alter the NLS and in turn affect localization of not only ICP4, but also other viral proteins (e.g. ICP0, ICP8) that are affected by ICP4 localization (Knipe and Smith, 1986). The fact that ICP4 is an immediate early transcriptional regulator and has been implicated in reactivation suggests that this variation may influence phenotype.

The low BSR for US07 (gI) (0.78) is also the result of an extended repeat element located within the coding sequence. Strain McKrae contains six perfect iterations of the internal tandem repeat STPSTTT. The STPSTTT repeat resides in the predicted extravirion domain of the US07 (gI) protein (UniProt Identifier: P06487). US07 (gI) interacts with host immunoglobulin G (IgG) through heterodimerization with US08 (gE), and alterations in the extravirion domain could affect recognition of either protein (Johnson et al., 1988). In a report investigating HSV-1 phylogeny, gI sequences were used to look at diversity of HSV-1 isolates, and therefore a large number of gI sequences are available (Norberg et al., 2004). An alignment of forty-five gI protein sequences showed considerable diversity in the number and composition of the STPSTTT repeat element (Fig. 3). Only one isolate (lcl|31811) contained the six perfect iterations found in strain McKrae, suggesting this large iteration number is relatively rare. HSV-1 glycoproteins are known to interact with host cell membrane proteins and mediate envelope fusion and viral entry (Browne et al., 2001; Turner et al., 1998). Disruption of the US07 (gI) gene results in attenuation *in vivo* and

a compromised spread-phenotype *in vitro*, which effectively suggests US07 (gI) contributes to the virus pathogenicity (Balan et al., 1994). It is tempting to speculate that the McKrae strain US07 variant may facilitate an increased cell-to-cell spread and pathogenesis.

Proteins predicted to contain large truncations (UL36, UL43, UL56) or extended ORFs (US10) were also found to have lower BSR (<0.85). McKrae UL36, the large tegument protein, contains a premature stop codon introduced due to a G nucleotide deletion in a mononucleotide string encoding amino acid residue 2453 (nt 72,535). As a consequence, UL36 encodes a unique and relatively short C-terminus of 321 amino acids relative to the mutation site as opposed to over 700 residues in the compared strains. The premature stop codon causes a loss of the 70-amino acid PQ repeat. It has been reported that residues 2430 to 2893 of wild type HSV-1 UL36 contain a binding site for the capsid protein UL25 (pUL25), which is required to recruit protein encoded by UL36 (pUL36) onto cytosolic capsids during assembly for secondary envelopment. In addition, the 167 residues of the very C terminus contain a second pUL25 binding site crucial to maintain pUL36 on incoming capsids during cell entry (Schipke et al., 2012). This suggests that McKrae pUL36 may have an altered interaction with pUL25. At a minimum, partial UL36 function is intact as McKrae does not have a UL36 null phenotype which would be characterized by a cytoplasmic build-up of unenveloped viral capsids (Desai, 2000).

McKrae strain gene UL56 (180 aa) contains a single base pair insertion at nucleotide 115,992 (amino acid 97). This results in a divergent and truncated C-terminus of only 83 aa (compared to 137 aa in other strains), which significantly reduced the BSR of UL56. UL56 has been implicated in pathogenicity (Kehm et al., 1996); restoring UL56 expression in the non-expressing, avirulent HSV-1 strain HFEM resulted in a pathogenic phenotype. In addition, C-terminal deletions in UL56 lead to a much reduced pathogenicity. The predicted C-terminal loss in the McKrae strain does not seem to agree with these previous studies. The authors acknowledge their observations may be specific to the infection route (intraperitoneally, IP); however, strain McKrae is pathogenic in mice by IP infection (Kehm et al., 1996; Weir et al., 1989). The presence of the UL56 mutation in the virulent McKrae strain may reflect phenotypic masking by a dominant protein variant in strain McKrae, possibly RL1 (ICP34.5) or US07 (gI).

US10 extended ORF is the result of a single bp insertion at nucleotide 143,416. The frameshift causes a stop codon loss in McKrae and a unique C-terminal protein sequence. Strain McKrae US10 is therefore predicted to be 317 amino acids long - five to seventeen residues longer than the compared strains (Table 1). The maximum length of US10 is the predicted sequence length for strain McKrae US10 from start codon to the first encountered stop codon. The revised length above takes into account an intact US10 transcription termination sequence (Rixon and McGeoch, 1984). The transcriptional machinery should terminate as usual, giving rise to a US10 transcript containing no translational stop codon. US10 is a capsid/tegument protein and localizes to the nucleus during infection (Yamada et al., 1997). It has not been fully characterized, and the effects of a predicted extended ORF are unknown.

UL49A (glycoprotein N) of strain McKrae was also found to have a lower BSR when compared to the other six strains. UL49A has two aa differences at residues 28 and 51 with histidine and threonine for McKrae, arginine and threonine for 17, and histidine and alanine for the other strains. Among all the strains, UL49A was also found to contain variability unique to F strain (Szpara et al., 2010). No UL49A was annotated in HF10 strain; therefore, BSR is not calculated when HF10 strain was included in the comparison. US11 is an RNA binding protein involved in translation regulation. This analysis also finds that US11 is unique for McKrae and H129 with 6, 9, and 12 aa deletion near the c-terminal when compared to HF10, 17, and KOS strains, respectively (Sup Fig. 2A).

Non-Coding Sequence Variation: Origin of Replication L (oriL)

The initial strain McKrae genome assembly did not appear to contain an oriL site. Standard PCR amplification of the region, followed by amplicon sequencing, confirmed this result. Although oriL is not required for genome replication and establishment of latency, previous research has associated oriL with enhanced neuronal replication *in vitro* and efficient reactivation *in vivo* (Balliet and Schaffer, 2006; Polvino-Bodnar et al., 1987). oriL also contains neuron specific protein binding domains, suggesting an important role for replication during active infection and adaptation to respond to neuronal signals during reactivation (Hardwicke and Schaffer, 1997). The presence of oriS in the initial assembly led to the conclusion that 454 sequencing is capable of sequencing through the hairpin structures formed by the replication origins. To further

examine the McKrae strain oriL deletion, the same region in strain 17 was amplified. However, it also produced a PCR product reflecting an oriL deletion (Fig. 4). As shown in Figure 4, the 409 bp band in lane A is an amplicon of the oriL site in a stable hairpin. PCR sequencing of this product indicated an oriL deletion. The 557 bp product in lane B is the relaxed oriL site and indicates the presence of oriL. Sequencing of the relaxed hairpin PCR product produced sequence ends that confirmed the product was from the oriL site. We were, however, unable to sequence through the entire hairpin using either standard or modified Sanger sequencing protocols. Although we were unable to sequence a 102 bp region (strain 17 bp 62,416 to 62,517) encompassing the apex of the oriL hairpin, our data is consistent with the presence of oriL, and it is reasonable to conclude that the McKrae oriL does not significantly differ from replication origins present in the compared strains, which are 100% conserved.

Inverted Repeats

The Inverted Repeat consists of the long and short inverted (or internal) repeats (Fig. 5A). An inverted copy of each repeat, of course, also flanks the opposite end of the unique long and unique short regions, respectively. The inverted repeats surrounding UL and US are of particular interest because they are the only significantly active sites of the viral genome during latency, and emergence from the latent state is affected by proteins and other genome features in this region (Fig. 5) (Block et al., 1993; Halford et al., 2001; Halford and Schaffer, 2001; Hill et al., 1990; Leib et al., 1989; Perng et al., 1996; Perng et al., 1995; Stevens et al., 1987). The inverted repeats contain the viral protein coding sequences for RL1 (ICP34.5), RL2 (ICP0), and RS1 (ICP4), the non-coding latency associated transcript (LAT), open reading frames ORF-O and ORF-P, the ‘a’ sequence, oriS, several tandem repeat elements, and several micro-RNAs (miRNAs) (Fig. 5a). Tandem repeats found at the UL-RL junction are related but distinct, and they are reduced in strain McKrae (49 bp in McKrae as opposed to 181 bp in strain 17 and KOS). Immediately following the UL-RL junction repeat, strain McKrae is missing a significant amount of sequence compared to other strains (approximately 330 nt). The RS-US junction is nearly identical between strains, and the oriS sequence just upstream is perfectly conserved.

The ‘a’ Sequence

HSV-1 McKrae strain has a unique ‘a’ sequence compared to the other well-defined lab strains

included in these analyses (Fig. 5b). The HSV-1 'a' sequence is present at the internal RL - RS junction and at the ends of the linear genome. Variability in the 'a' sequence is common both between strains and within strain isolates, but the subunit pattern is conserved (Umene, 2001; Umene et al., 2008). The McKrae sequence contains unique variation within the 'a' sequence direct repeat 2 (DR2) array. Instead of consisting of a series of unbroken tandem repeats, the McKrae DR2 repeats are interrupted twice by identical guanine-rich sequences. The viral 'a' sequence has been implicated in genome stability and recombination rate (Umene, 1991, 1993). The effect of interrupting the DR2 repeat with divergent sequence strings may have an effect on recombination, but could also affect RL1 (ICP34.5) expression (Chou and Roizman, 1986; Martin and Weber, 1998). Upstream promoter elements influencing RL1 (ICP34.5) expression have been found in the DR2 array and have been shown to alter expression *in vitro*. Differences in repeat length, especially in the DR2 array, are present between strains, but this may not be as significant as the structural variation described here for strain McKrae.

The LAT region

Previous work has shown HSV-1 LAT to be required for spontaneous reactivation, and we hypothesized that a multiple sequence alignment of the LAT region would identify sequence features unique to strain McKrae (Fig. 5a). In these analyses, the full LAT transcript was inferred from existing McKrae sequence data, annotations, and sequence markers present in the compared strains. The LAT TATA promoter sequence is 27 nt upstream of the proposed transcription start site at nt 117,769 in strain McKrae, and the 3' transcript end is assumed to terminate near a polyA signal ending at nt 125,965. Therefore, the strain McKrae LAT is predicted to be 8,170 nt with a transcription start site at nt 117,796. The LAT promoter is well conserved between strains, and the little variability that does exist is unlikely to be responsible for increased reactivation in strain McKrae. A previous study showed that substitution of the McKrae LAT promoter for the non-spontaneously reactivating KOS strain promoter in a KOS background did not impart increased reactivation (Strelow et al., 1994). HSV-1 KOS contains an extended repeat of 15 T residues beginning at base 118,292 (McKrae) that is not present in other strains. Downstream of this site is the predicted 5' end of the McKrae stable LAT intron. The McKrae LAT intron is 2,006 bp in length (genomic coordinates 118,458 to 120,463) and is 95.9% conserved between strains. The 5' half of the LAT intron does not overlap protein coding sequences, whereas the 3' end overlaps

RL2 (ICP0). Major variation within the LAT intron is due to differences in a repeat element (GCACCCCCACTCCCAC) that varies in iteration number beginning at nucleotide 119,482 in McKrae strain (F, H129, and 17: 9 repeats; KOS: 15 repeats; McKrae: 13 repeats). The 3' end of the LAT transcript is well conserved between strains likely due to overlapping RL2 (ICP0) coding sequence on the opposing strand.

A region of increased sequence variation is present between the 'a' sequence and the 3' end of the LAT transcript (McKrae coordinates 125,298 to 125,965). It is the result of tandem repeats of distinct composition and number. Tandem repeat variation between strains is found beginning at strain McKrae base 125,520. McKrae repeat elements include twelve iterations of CCCCAGCCCTCCCCAG and eight iterations of CCCCTCGCCCCCTCCCG. The first repeat unit is unique from other strains in that it contains a G to A transition, and strain McKrae contains three iterations more than any other strain. The McKrae strain second repeat element is collapsed, missing 188 nucleotides relative to all other strains, and separated from the upstream repeat by a 100% conserved sequence of 105 bp containing miR-H5.

Micro-RNAs

The role of miRNAs in translation regulation has recently become of interest in HSV-1 biology. Several miRNAs have been discovered in the inverted repeat regions that are capable of interfering with RS1 (ICP4) and RL2 (ICP0) translation (Umbach et al., 2008). Deep sequencing was used to identify six miRNAs, miR-H1 through miR-H6, from the IR region (Umbach et al., 2008). miR-H1 and miR-H6 are anti-parallel on opposing strands at McKrae nucleotide 117,312. A KOS-specific single nucleotide deletion occurs at base 117,351, affecting only the mature miR-H1 miRNA. The site giving rise to miR-H1 and H6 is upstream of the LAT region. miR-H2 and miR-H5 are 100% conserved between all strains (McKrae nt 120,871). miR-H3, located at McKrae nucleotide 124,768, contains two sites of H129 strain specific variation that fall outside the mature miRNA. miR-H4, just downstream of miR-H3 at nucleotide 124,916, contains a 6 nucleotide deletion at the 3' end of the proposed miRNA precursor in strain H129.

The proposed region of miRNA interference in RL2 (ICP0) is 100% conserved between strains, whereas the RS1 (ICP4) transcript contains two sites of variation, one of which is specific to

strain McKrae. Umbach et al. demonstrated that the accumulation of several point mutations in the RS1 (ICP4) transcript abrogated translational suppression, and one of these mutations (T to G at bp 126,118) is present in McKrae. If mutations in the miR-H6 recognition site decrease or abolish duplex formation, it seems logical that de-repression of RS1 (ICP4) translation could facilitate viral reactivation by allowing transactivation of viral early gene transcription.

This sequence analysis identified at least 8 genes and several non-coding sequences unique for McKrae strain. In addition to those previously identified to be associated with pathogenesis and latency reactivations, such as RL1, RS1, and RL2, three UL genes (UL36, UL49A, UL56) and three US genes (US7, US10, US11) were found to be unique for McKrae strain. In addition to gene variations, non-coding sequences such as LAT, 'a' sequence, and miRNAs, were also identified to contain variations unique for McKrae. These differences may contribute to observed pathogenic and reactivation phenotype variability among different strains.

Materials and Methods:

Virus and Cell Culture

HSV-1 strain McKrae was propagated in Vero cells maintained in Eagle Minimal Essential Medium (EMEM) with 2 mM L-glutamine, 0.1 mM nonessential amino acids, 1 mM sodium pyruvate, 10% fetal bovine serum (Promega Scientific), penicillin (100 U/ml), and streptomycin (100 µg/ml) (Sigma, St. Louis, MO).

Purification of viral DNA

Viral DNA was extracted either from virions purified as previously described (Jin et al., 2008) or from purified intracellular nucleocapsids. For nucleocapsid purification, infected cells were harvested when 70-90% of the monolayer showed CPE and processed as described previously (Scherba et al., 1992). Briefly, flasks were freeze-thawed twice and pelleted at 5000 rpm at 4°C for 20 min. The cell pellet was then washed in PBS and resuspended in 18 ml hypotonic buffer (10 mM Tris-HCl, 10 mM KCl, 5 mM EDTA, pH 8.0) and 2 ml Triton X-100 and incubated on ice for 10 minutes. The cell debris was removed by low speed centrifugation at 5000 rpm at 4°C for 10 minutes. The viral nucleocapsids were then centrifuged through a 60% sucrose cushion for

virion purification. Purified virions or nucleocapsids were digested overnight at 50°C in 10 mM Tris-HCl (pH 8.0), 100 mM EDTA, 1% *N*-lauroyl sarcosine, and 200 µg/ml proteinase K. The viral DNA was extracted twice with an equal volume of phenol:chloroform (1:1) and then precipitated with 2 volumes of ethanol and 1/10 volume of 3.0 M sodium acetate. The precipitate was washed once in 70% ethanol and resuspended in TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0).

McKrae Genome Sequencing

Sample preparations for 454 sequencing were carried out using protocols provided by the manufacturer. The viral genome, total 5µg of purified viral DNA, was nebulized to produce fragments less than 800bp before sequencing. DNA sequencing was done using the GS FLX+ System from 454 Life Technology (Roche). *De novo* assembly of the McKrae genome was assembled using 454 Newbler *de novo* assembler (version 2.5).

PCR DNA Sequencing

Primers were designed using Primer3 software to generate PCR products spanning genome gaps (Table 3). PCR products were either gel purified using the IBI Scientific Gel/PCR DNA Fragments Extraction Kit or by ChargeSwitch PCR Clean-Up Kit (Invitrogen) and sequenced by Sanger Sequencing using standard protocols by the Center for Genome Research and Biocomputing (CGRB) at Oregon State University using an ABI Prism®3730 Genetic Analyzer with a BigDye® Terminator v. 3.1 Cycle Sequencing Kit and employing ABI Prism®3730 Data Collection Software v. 3.0 and ABI Prism® DNA Sequencing Analysis Software v. 5.2. All major sequence variants in the McKrae draft genome were confirmed by PCR amplification and sequencing as described above. Invitrogen PCR Enhancer was used as needed in PCR sequencing reactions.

oriL PCR amplification

PCR amplification with oriL specific primers was performed as follows: a 25 µl reaction solution containing 1X Pfx amplification buffer (Invitrogen), 1X or 2X PCR Enhancer solution (Invitrogen), 0.5 µM $MgSO_4$, 0.4 µM dNTP's, 0.4 µM primers (Forward and Reverse), 1.0 U of Platinum Pfx DNA polymerase (Invitrogen), and 0.01-0.1 µg of viral DNA, was subjected to

94°C for 2 min, 30 cycles of 94°C for 30 s, 66°C for 30 s, and 72°C for 30 s, followed by a 5 min elongation reaction at 72°C after the final cycle.

Comparative Sequence Analyses

Primary sequence data and PCR sequence data were assembled using Geneious software. Annotations of the McKrae genome and comparative analyses were also done using Geneious software. Protein domain and structure information was gathered from NCBI and UniProt databases (http://www.ncbi.nlm.nih.gov/; http://www.ebi.ac.uk/uniprot/).

Strain 17: Accession No. X14112

Strain F: Accession No. GU734771

Strain H129: Accession No. GU734772

Strain McKrae: Accession No. JQ730035

Strain HF10: Accession No. DQ889502

Strain KOS: Accession No. JQ780693

Strain KOS: Accession No. JQ673480

Acknowledgments

G. Watson and W. Xu made equal contributions to this paper.

This work was supported by the College of Veterinary Medicine at Oregon State University. We thank Claire Ostertag-Hill and Liang Fang for the technical supports.

Disclosure statement

The authors declare that they have no conflict of interest.

Figure Legends

Figure 1. Construction of the McKrae genome. All sequence contigs and primer sequenced regions are aligned vertically. A: HSV-1 genome regions. The unique long (UL) and unique short (US) regions are flanked by distinct inverted repeats termed the repeat long (RL) and repeat short (RS), respectively. B: Initial de novo assembled contigs (C1 – C9) from 454 Sequencing and their approximate location determined by alignment to HSV-1 strain 17. Staggering of

contigs indicates a gap in the assembly. PCR-sequencing was used to bridge the contigs. C: Gene coding regions confirmed by PCR sequencing and their location (triangles) in the McKrae genome.

Figure 2. Variation in the HSV-1 RS1 (ICP4) sequence. A: Schematic structure of RS1 (ICP4) protein with functional domains. NLS: nuclear localization signal. B: Multiple sequence alignment within the expanded region near NLS between AA 686-739. The McKrae unique substitution is noted. Multiple sequence alignment and the expanded regions were produced using Geneious Software.

Figure 3. HSV-1 US07. A: Diagram of US07 (gI) extravirion and intravirion domains separated by the transmembrane (TM) domain. B: Alignment of the US07 (gI) STPSTTT tandem repeat from 44 protein sequences. Tandem repeats are sorted by increasing number of perfect repeats. Strain McKrae and isolate lcl|31811 contain the largest number of perfect repeats. Graphic from Geneious Software.

Figure 4. oriL PCR amplification products. The oriL hairpin is able to be amplified only in the presence of PCR Enhancer (Invitrogen). Lane A: oriL PCR supplemented with enhancer to 1X concentration. Lane B: oriL PCR supplemented with enhancer to 2X concentration. The 409 bp band in lane A is an amplicon of the oriL site in a stable hairpin. The 557 bp band in lane B is the relaxed oriL PCR product. L: 1 kb Plus DNA Ladder (Invitrogen). Template is genomic DNA.

Figure 5. Internal repeat (IR) region multiple sequence alignment and HSV-1 'a' sequence. A: alignment of six sequences covering the internal repeat (IR) region: McKrae strain, F strain, H129, strain 17, KOS promoter sequence, and KOS sequence covering the LAT intron. The histogram above the alignment represents percent identity between strains using a sliding window of 99 base pairs; dark green indicates 100% identity; red indicates below 30% identity. Gaps between the black boxes in the alignment of 5 strains shown correspond to gaps in the alignment. Annotations below the multiple sequence alignment are relative to strain McKrae and cover regions of interest in these analyses. RL: Repeat Long; UL: Unique Long; RS: Repeat

Short; US: Unique Short; SI: StyI region; dLAT2903: 5' LAT transcript region; LP: LAT Promoter; oriS: origin of replication S; RL2: ICP0; RL1: ICP34.5; RS1: ICP4: 'a': 'a' sequence. Base numbers are relative to strain McKrae. B: subunit structure of the HSV-1 'a' sequence and the composition of the DR2 array by strain. Strain McKrae contains two identical guanine rich insertions (McKrae nt 125,114 and nt 125,170). DR1: direct repeat 1 (19-20 bp); Ub: unique sequence (65-77 bp); DR2 array: direct repeat 2 array (108-235 bp); DR4; direct repeat 4 (19-20 bp); Uc: unique sequence (58 bp). The y-axis is the scale of sequence alignment identity. Image produced using Geneious Software.

Supplemental Fig. 1. Average quality score per sequence. Reads with low scores were filtered out of initial drafts at arbitrary scores through several contig draft iterations.

Supplemental Fig. 2. Alignment of US11 amino acids between selected strains. Lane 1: McKrae (Accession No. JQ730035), 2:17 (Accession No. X14112), 3: H129 (Accession No. GU734772), 4: F (Accession No. GU734771), 5: HF10 9 Accession No. DQ889502), 6: KOS (Accession No. JQ673480), KOS (Accession No. JQ780693).

References:

Balan, P., Davis-Poynter, N., Bell, S., Atkinson, H., Browne, H., Minson, T., 1994, An analysis of the in vitro and in vivo phenotypes of mutants of herpes simplex virus type 1 lacking glycoproteins gG, gE, gI or the putative gJ. J Gen Virol 75 (Pt 6), 1245-1258.

Balliet, J.W., Schaffer, P.A., 2006, Point mutations in herpes simplex virus type 1 oriL, but not in oriS, reduce pathogenesis during acute infection of mice and impair reactivation from latency, In: J Virol.United States, pp. 440-450.

Block, T.M., Deshmane, S., Masonis, J., Maggioncalda, J., Valyi-Nagi, T., Fraser, N.W., 1993, An HSV LAT null mutant reactivates slowly from latent infection and makes small plaques on CV-1 monolayers. Virology 192, 618-630.

Bolovan, C.A., Sawtell, N.M., Thompson, R.L., 1994, ICP34.5 mutants of herpes simplex virus type 1 strain 17syn+ are attenuated for neurovirulence in mice and for replication in confluent primary mouse embryo cell cultures. J Virol 68, 48-55.

Bower, J.R., Mao, H., Durishin, C., Rozenbom, E., Detwiler, M., Rempinski, D., Karban, T.L., Rosenthal, K.S., 1999, Intrastrain variants of herpes simplex virus type 1 isolated from a neonate with fatal disseminated infection differ in the ICP34.5 gene, glycoprotein processing, and neuroinvasiveness. J Virol 73, 3843-3853.

Browne, H., Bruun, B., Minson, T., 2001, Plasma membrane requirements for cell fusion induced by herpes simplex virus type 1 glycoproteins gB, gD, gH and gL. J Gen Virol 82, 1419-1422.

Chou, J., Roizman, B., 1986, The terminal a sequence of the herpes simplex virus genome contains the promoter of a gene located in the repeat sequences of the L component. Journal of virology 57, 629-629.

Desai, P.J., 2000, A null mutation in the UL36 gene of herpes simplex virus type 1 results in accumulation of unenveloped DNA-filled capsids in the cytoplasm of infected cells. Journal of Virology 74, 11608-11608.

Halford, W.P., Kemp, C.D., Isler, J.A., Davido, D.J., Schaffer, P.A., 2001, ICP0, ICP4, or VP16 expressed from adenovirus vectors induces reactivation of latent herpes simplex virus type 1 in primary cultures of latently infected trigeminal ganglion cells. J Virol 75, 6143-6153.

Halford, W.P., Schaffer, P.A., 2001, ICP0 is required for efficient reactivation of herpes simplex virus type 1 from neuronal latency. J Virol 75, 3240-3249.

Hardwicke, M.A., Schaffer, P.A., 1997, Differential effects of nerve growth factor and dexamethasone on herpes simplex virus type 1 oriL-and oriS-dependent DNA replication in PC12 cells. Journal of virology 71, 3580-3580.

Hill, J.M., Gebhardt, B.M., Wen, R., Bouterie, A.M., Thompson, H.W., O'Callaghan, R.J., Halford, W.P., Kaufman, H.E., 1996, Quantitation of herpes simplex virus type 1 DNA and latency-associated transcripts in rabbit trigeminal ganglia demonstrates a stable reservoir of viral nucleic acids during latency. J Virol 70, 3137-3141.

Hill, J.M., Rayfield, M.A., Haruta, Y., 1987, Strain specificity of spontaneous and adrenergically induced HSV-1 ocular reactivation in latently infected rabbits. Curr Eye Res 6, 91-97.

Hill, J.M., Sedarati, F., Javier, R.T., Wagner, E.K., Stevens, J.G., 1990, Herpes simplex virus latent phase transcription facilitates in vivo reactivation. Virology 174, 117-125.

Jin, L., Perng, G.C., Carpenter, D., Mott, K.R., Osorio, N., Naito, J., Brick, D.J., Jones, C., Wechsler, S.L., 2007, Reactivation phenotype in rabbits of a herpes simplex virus type 1 mutant containing an unrelated antiapoptosis gene in place of latency-associated transcript. J Neurovirol 13, 78-84.

Jin, L., Perng, G.C., Mott, K.R., Osorio, N., Naito, J., Brick, D.J., Carpenter, D., Jones, C., Wechsler, S.L., 2005, A herpes simplex virus type 1 mutant expressing a baculovirus inhibitor of apoptosis gene in place of latency-associated transcript has a wild-type reactivation phenotype in the mouse. J Virol 79, 12286-12295.

Jin, L., Valentine, B.A., Baker, R.J., Lohr, C.V., Gerlach, R.F., Bildfell, R.J., Moerdyk-Schauwecker, M., 2008, An outbreak of fatal herpesvirus infection in domestic rabbits in Alaska. Vet Pathol 45, 369-374.

Johnson, D.C., Frame, M.C., Ligas, M.W., Cross, A.M., Stow, N.D., 1988, Herpes simplex virus immunoglobulin G Fc receptor activity depends on a complex of two viral glycoproteins, gE and gI. J Virol 62, 1347-1354.

Kehm, R., Rosen-Wolff, A., Darai, G., 1996, Restitution of the UL56 gene expression of HSV-1 HFEM led to restoration of virulent phenotype; deletion of the amino acids 217 to 234 of the UL56 protein abrogates the virulent phenotype. Virus Res 40, 17-31.

Knipe, D.M., Smith, J.L., 1986, A mutant herpesvirus protein leads to a block in nuclear localization of other viral proteins. Mol Cell Biol 6, 2371-2381.

Kuddus, R., Gu, B., DeLuca, N.A., 1995, Relationship between TATA-binding protein and herpes simplex virus type 1 ICP4 DNA-binding sites in complex formation and repression of transcription. J Virol 69, 5568-5575.

Leib, D.A., Bogard, C.L., Kosz-Vnenchak, M., Hicks, K.A., Coen, D.M., Knipe, D.M., Schaffer, P.A., 1989, A deletion mutant of the latency-associated transcript of herpes simplex virus type 1 reactivates from the latent state with reduced frequency. J Virol 63, 2893-2900.

Li, Y., Zhang, C., Chen, X., Yu, J., Wang, Y., Yang, Y., Du, M., Jin, H., Ma, Y., He, B., Cao, Y., 2011, ICP34.5 protein of herpes simplex virus facilitates the initiation of protein translation by bridging eukaryotic initiation factor 2alpha (eIF2alpha) and protein phosphatase 1. J Biol Chem 286, 24785-24792.

Mao, H., 2002, An N-terminal Arginine-rich Cluster and a Proline-Alanine-Threonine Repeat Region Determine the Cellular Localization of the Herpes Simplex Virus Type 1 ICP34.5 Protein and Its Ligand, Protein Phosphatase 1. Journal of Biological Chemistry 277, 11423-11431.

Mao, H., Rosenthal, K.S., 2003, Strain-dependent structural variants of herpes simplex virus type 1 ICP34.5 determine viral plaque size, efficiency of glycoprotein processing, and viral release and neuroinvasive disease potential. J Virol 77, 3409-3417.

Martin, D.W., Weber, P.C., 1998, Role of the DR2 repeat array in the regulation of the ICP34.5 gene promoter of herpes simplex virus type 1 during productive infection. J Gen Virol 79 (Pt 3), 517-523.

Norberg, P., Bergstrom, T., Rekabdar, E., Lindh, M., Liljeqvist, J.A., 2004, Phylogenetic analysis of clinical herpes simplex virus type 1 isolates identified three genetic groups and recombinant viruses. J Virol 78, 10755-10764.

Perng, G.-C., Mott, K.R., Osorio, N., Yukht, A., Salina, S., Nguyen, Q.-H., Nesburn, A.B., Wechsler, S.L., 2002, Herpes simplex virus type 1 mutants containing the KOS strain ICP34.5 gene in place of the McKrae ICP34.5 gene have McKrae-like spontaneous reactivation but non-McKrae-like virulence. Journal of General Virology 83, 2933-2942.

Perng, G.C., Dunkel, E.C., Geary, P.A., Slanina, S.M., Ghiasi, H., Kaiwar, R., Nesburn, A.B., Wechsler, S.L., 1994, The latency-associated transcript gene of herpes simplex virus type 1 (HSV-1) is required for efficient in vivo spontaneous reactivation of HSV-1 from latency. J Virol 68, 8045-8055.

Perng, G.C., Ghiasi, H., Slanina, S.M., Nesburn, A.B., Wechsler, S.L., 1996, The spontaneous reactivation function of the herpes simplex virus type 1 LAT gene resides completely within the first 1.5 kilobases of the 8.3-kilobase primary transcript. J Virol 70, 976-984.

Perng, G.C., Jones, C., Ciacci-Zanella, J., Stone, M., Henderson, G., Yukht, A., Slanina, S.M., Hofman, F.M., Ghiasi, H., Nesburn, A.B., Wechsler, S.L., 2000, Virus-induced neuronal apoptosis blocked by the herpes simplex virus latency-associated transcript. Science 287, 1500-1503.

Perng, G.C., Slanina, S.M., Yukht, A., Drolet, B.S., Keleher, W., Jr., Ghiasi, H., Nesburn, A.B., Wechsler, S.L., 1999, A herpes simplex virus type 1 latency-associated transcript mutant with increased virulence and reduced spontaneous reactivation. J Virol 73, 920-929.

Perng, G.C., Thompson, R.L., Sawtell, N.M., Taylor, W.E., Slanina, S.M., Ghiasi, H., Kaiwar, R., Nesburn, A.B., Wechsler, S.L., 1995, An avirulent ICP34.5 deletion mutant of herpes simplex virus type 1 is capable of in vivo spontaneous reactivation. J Virol 69, 3033-3041.

Polvino-Bodnar, M., Orberg, P.K., Schaffer, P.A., 1987, Herpes simplex virus type 1 oriL is not required for virus replication or for the establishment and reactivation of latent infection in mice. J Virol 61, 3528-3535.

Rixon, F.J., McGeoch, D.J., 1984, A 3' co-terminal family of mRNAs from the herpes simplex virus type 1 short region: two overlapping reading frames encode unrelated polypeptide one of which has highly reiterated amino acid sequence. Nucleic Acids Res 12, 2473-2487.

Rock, D.L., Nesburn, A.B., Ghiasi, H., Ong, J., Lewis, T.L., Lokensgard, J.R., Wechsler, S.L., 1987, Detection of latency-related viral RNAs in trigeminal ganglia of rabbits latently infected with herpes simplex virus type 1. J Virol 61, 3820-3826.

Scherba, G., Jin, L., Schnitzlein, W.M., Vodkin, M.H., 1992, Differential polymerase chain reaction for detection of wild-type and a vaccine strain of Aujeszky's disease (pseudorabies) virus. J Virol Methods 38, 131-143.

Schipke, J., Pohlmann, A., Diestel, R., Binz, A., Rudolph, K., Nagel, C.H., Bauerfeind, R., Sodeik, B., 2012, The C terminus of the large tegument protein pUL36 contains multiple capsid binding sites that function differently during assembly and cell entry of herpes simplex virus. J Virol 86, 3682-3700.

Shepard, A.A., Imbalzano, A.N., DeLuca, N.A., 1989, Separation of primary structural components conferring autoregulation, transactivation, and DNA-binding properties to the herpes simplex virus transcriptional regulatory protein ICP4. J Virol 63, 3714-3728.

Stevens, J.G., Wagner, E.K., Devi-Rao, G.B., Cook, M.L., Feldman, L.T., 1987, RNA complementary to a herpesvirus alpha gene mRNA is prominent in latently infected neurons. Science 235, 1056-1059.

Strelow, L.I., Laycock, K.A., Jun, P.Y., Rader, K.A., Brady, R.H., Miller, J.K., Pepose, J.S., Leib, D.A., 1994, A structural and functional comparison of the latency-associated transcript promoters of herpes simplex virus type 1 strains KOS and McKrae. Journal of General Virology 75, 2475-2480.

Stroop, W.G., Schaefer, D.C., 1987, Severity of experimentally reactivated herpetic eye disease is related to the neurovirulence of the latent virus. Invest Ophthalmol Vis Sci 28, 229-237.

Szpara, M.L., Parsons, L., Enquist, L.W., 2010, Sequence variability in clinical and laboratory isolates of herpes simplex virus 1 reveals new mutations. J Virol 84, 5303-5313.

Turner, A., Bruun, B., Minson, T., Browne, H., 1998, Glycoproteins gB, gD, and gHgL of herpes simplex virus type 1 are necessary and sufficient to mediate membrane fusion in a Cos cell transfection system. J Virol 72, 873-875.

Umbach, J.L., Kramer, M.F., Jurak, I., Karnowski, H.W., Coen, D.M., Cullen, B.R., 2008, MicroRNAs expressed by herpes simplex virus 1 during latent infection regulate viral mRNAs. Nature 454, 780-783.

Umene, K., 1991, Recombination of the internal direct repeat element DR2 responsible for the fluidity of the a sequence of herpes simplex virus type 1. J Virol 65, 5410-5416.

Umene, K., 1993, Herpes simplex virus type 1 variant a sequence generated by recombination and breakage of the a sequence in defined regions, including the one involved in recombination. J Virol 67, 5685-5691.

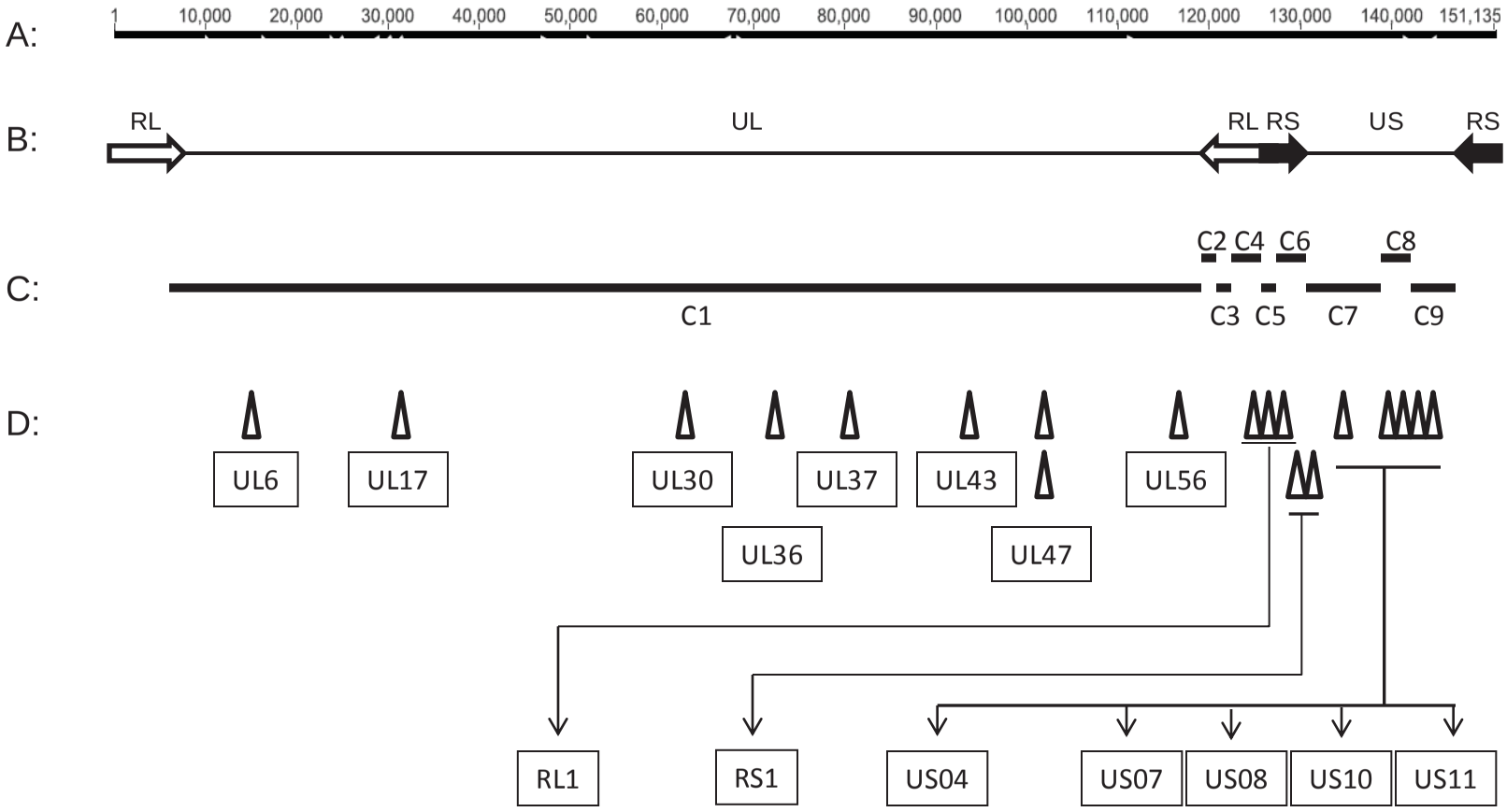
Umene, K., 2001, Cleavage in and around the DR1 Element of the a Sequence of Herpes Simplex Virus Type 1 Relevant to the Excision of DNA Fragments with Length

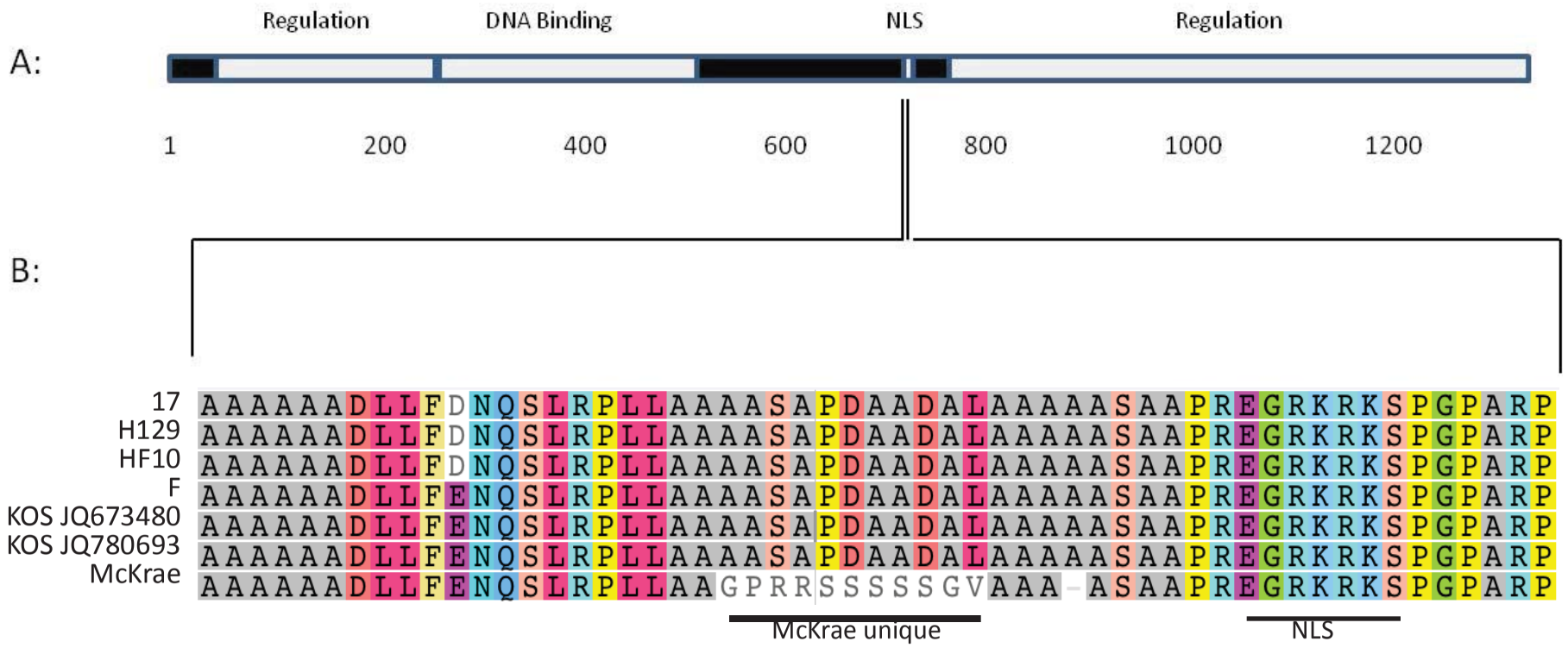
Corresponding to One and Two Units of the a Sequence. Journal of Virology 75, 5870-5870.

Umene, K., Oohashi, S., Yoshida, M., Fukumaki, Y., 2008, Diversity of the a sequence of herpes simplex virus type 1 developed during evolution. Journal of General Virology 89, 841-841.

Weir, J.P., Bennett, M., Allen, E.M., Elkins, K.L., Martin, S., Rouse, B.T., 1989, Recombinant vaccinia virus expressing the herpes simplex virus type 1 glycoprotein C protects mice against herpes simplex virus challenge. J Gen Virol 70 (Pt 10), 2587-2594.

Yamada, H., Daikoku, T., Yamashita, Y., Jiang, Y.M., Tsurumi, T., Nishiyama, Y., 1997, The product of the US10 gene of herpes simplex virus type 1 is a capsid/tegument-associated phosphoprotein which copurifies with the nuclear matrix. J Gen Virol 78 (Pt 11), 2923-2931.





□i□□e□□
□ i□□ □e□e to □own□o□□ □i□□e□□i□□□ □□□□□□ppt

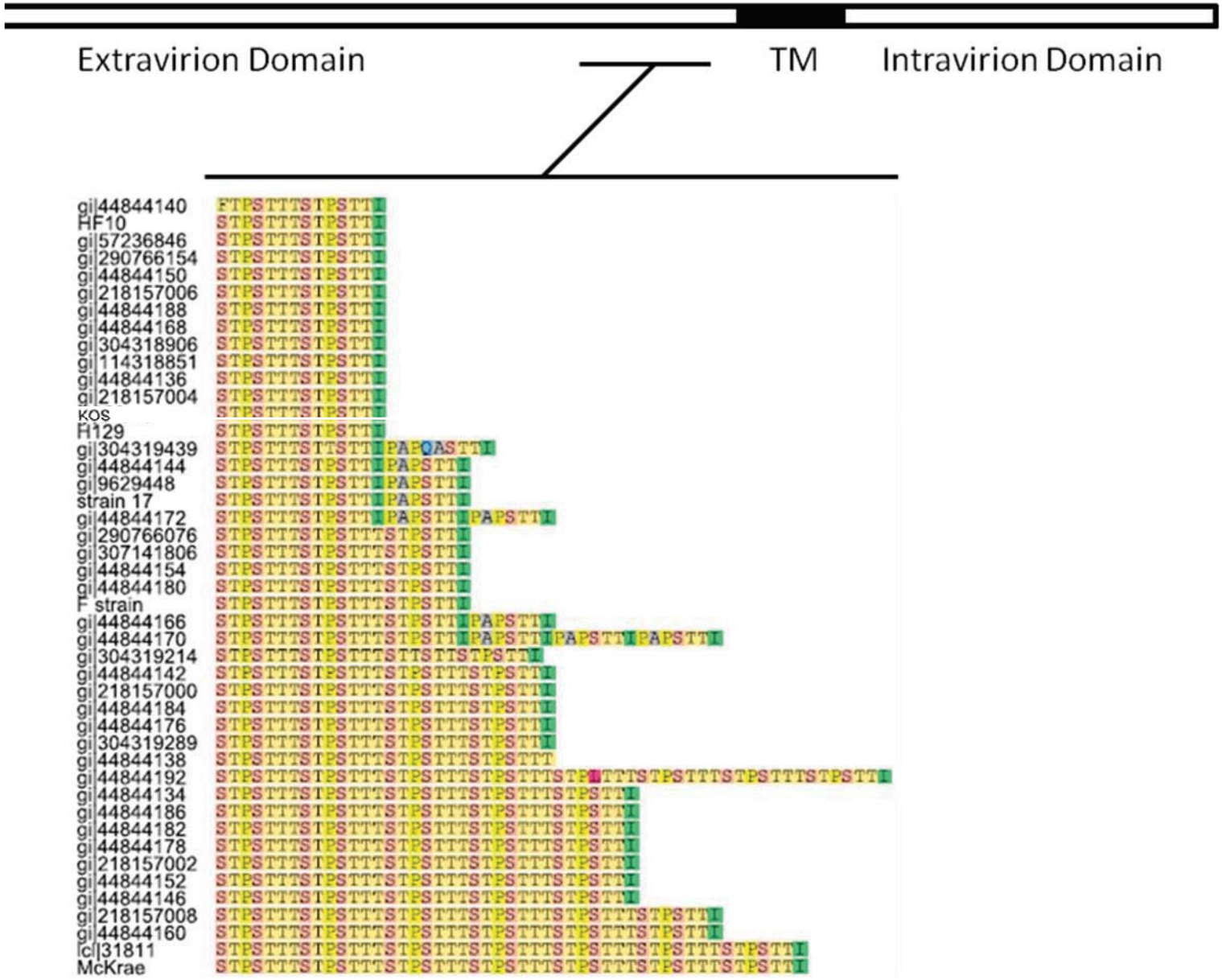


Figure-4

Click here to download Figure: Fig. 4 053012.ppt

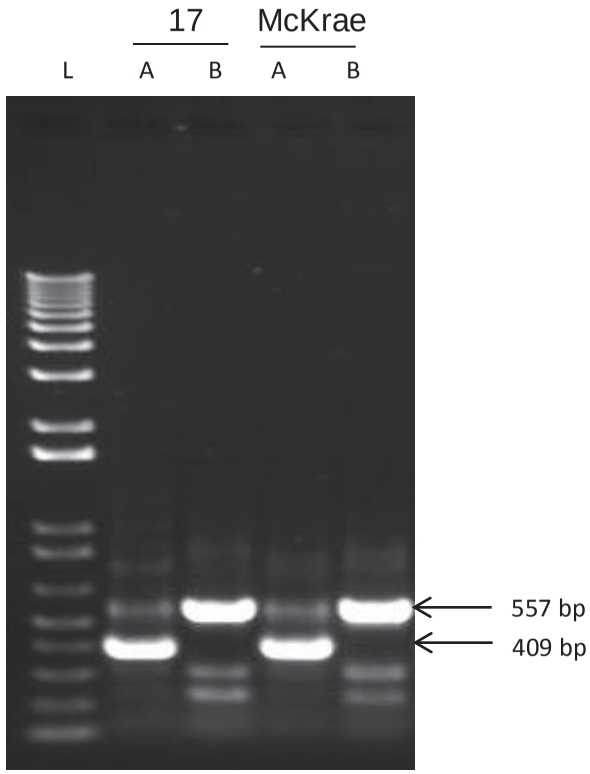
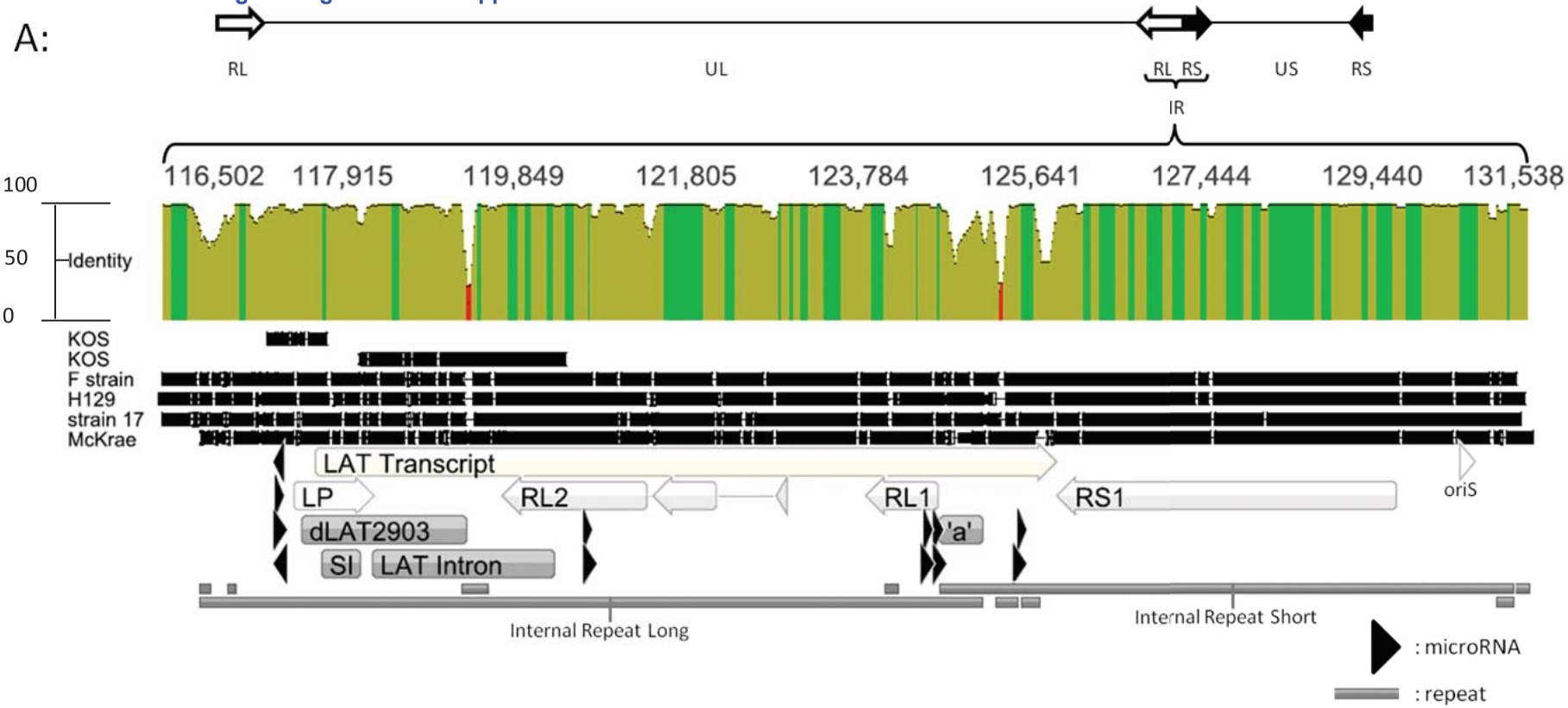


Figure 5

Click here to download Figure: Figure 5 0□2012.ppt



Strain	DR2
17	$(CGCTCCTCCCCC)_{18}$
KOS	$(CGCTCCTCCCCC)_{5}(CGCTCCTCCCC)_{4}$
H129	$(CGCTCCTCCCCC)_{19}$
F	$(CGCTCCTCCCCC)_{17}$
McKrae	$(CGCTCCTCCCC)_{2}(GTCTGTGGGTGGGGCTCCTCCCC)(CGCTCCTCCCC)_{3}(GTCTGTGGGTGGGGCTCCTCCCC)$

□a□le 1

Table 1. Conservation of genes between HSV-1 strains: 17, McKrae, F strain, H129, HF10, and two KOS

Name	% [a] Pairwise Identity	No.[b] Sequences	Max Sequence Length	Min Sequence Length	Sequence Length
RL1	95.00%	7	257	240	258
RL2	98.10%	7	777	768	783
RS1	98.30%	7	1303	1294	1304
UL1	97.30%	7	224	224	224
UL2	97.70%	7	334	334	334
UL3	99.70%	7	235	235	235
UL4	99.00%	7	199	198	199
UL5	99.50%	7	882	882	882
UL6	99.40%	7	676	676	676
UL7	99.60%	7	296	296	296
UL8	99.20%	7	750	750	750
UL9	97.80%	7	851	793	851
UL10	99.30%	7	473	473	473
UL11	97.60%	7	96	96	96
UL12	99.40%	7	626	626	626
UL13	99.00%	7	518	518	518
UL14	98.50%	7	219	219	219
UL15	99.90%	7	735	735	735
UL16	99.90%	7	373	373	373
UL17	99.10%	7	703	702	703
UL18	99.70%	7	318	318	318
UL19	99.60%	7	1374	1374	1374
UL20	99.70%	7	222	222	222
UL21	99.80%	7	535	535	535
UL22	99.50%	7	838	838	838
UL23	99.10%	7	376	376	376
UL24	99.10%	7	269	269	269
UL25	99.80%	7	580	580	580
UL 26	99.40%	7	635	635	635
UL26.5	99.30%	6	329	329	329
UL27	99.10%	7	904	904	904
UL28	99.90%	7	785	785	785
UL29	99.70%	7	1196	1196	1196
UL30	99.50%	7	1235	1235	1235
UL31	99.70%	7	306	306	306
UL32	99.60%	7	596	595	596
UL33	99.80%	7	130	130	130

UL34	99.70%	7	275	275	275
UL35	100.00%	7	112	112	112
UL36	96.40%	7	3164	2775	3169
UL37	99.50%	7	1123	1123	1123
UL38	99.70%	7	465	465	465
UL39	99.60%	7	1137	1137	1137
UL40	99.40%	6	340	340	340
UL41	99.50%	7	489	489	489
UL42	99.20%	7	488	488	488
UL43	96.90%	6	434	357	434
UL44	98.60%	7	511	511	511
UL45	100.00%	6	172	172	172
UL46	98.60%	7	719	717	719
UL47	99.80%	7	693	693	693
UL48	99.40%	7	490	489	490
UL49	98.80%	7	301	301	301
UL49A	99.00%	6	91	91	91
UL50	99.10%	7	371	370	371
UL51	99.30%	7	244	244	244
UL52	99.70%	7	1058	1056	1058
UL53	99.60%	7	338	338	338
UL54	99.40%	7	512	512	512
UL55	99.60%	6	186	186	186
UL56	88.70%	6	234	180	237
US1	98.70%	7	420	420	420
US2	98.80%	7	291	291	291
US3	99.30%	7	481	481	481
US4	96.80%	7	239	238	239
US5	99.50%	7	92	92	92
US6	99.40%	7	394	394	394
US7	95.30%	7	418	383	418
US8	99.20%	7	552	550	552
US8A	98.80%	7	190	159	190
US9	99.30%	7	90	57	90
US10	92.30%	7	507	300	516
US11	95.30%	7	161	149	161
US12	98.50%	7	88	88	88

[a] percentage of amino acid identity

[b] number of strains; 7=all strains ; 6= all stains but HF10

Black highlight <98% conserved protein

Table 2. Protein variants from McKrae sorted by bit score ratio (BSR)

McKrae	No. of other strains	BSR[a]	No. of other strains	BSR[b]
UL56	5	0.353	5	0.353
RL1	5	0.387	6	0.387
US10	5	0.479	6	0.479
RS1	5	0.669	6	0.669
US11	5	0.719	6	0.705
US7	5	0.787	6	0.784
UL49A	5	0.790	5	0.790
RL2	5	0.810	6	0.803
UL36	5	0.829	6	0.830
US9	5	0.852	6	0.875
UL43	5	0.885	5	0.885
US4	5	0.902	6	0.892
UL46	5	0.909	6	0.920
US8	5	0.920	6	0.916
UL1	5	0.925	6	0.938
UL40	5	0.935	6	0.944
UL2	5	0.955	6	0.963
UL17	5	0.964	6	0.968
UL9	5	0.976	6	0.979
UL11	5	0.979	6	0.980
UL4	5	0.980	6	0.980
UL32	5	0.981	6	0.972
UL50	5	0.982	6	0.984
UL55	5	0.985	5	0.985
UL14	5	0.987	6	0.975
UL42	5	0.987	6	0.987
UL23	5	0.988	6	0.986
US12	5	0.988	6	0.985
UL44	5	0.988	6	0.979
US3	5	0.989	6	0.990
UL6	5	0.990	6	0.990
US1	5	0.990	6	0.991
UL49	5	0.990	6	0.990
UL27	5	0.991	6	0.991
US2	5	0.991	6	0.991
UL37	5	0.992	6	0.992
UL13	5	0.992	6	0.991
UL8	5	0.993	6	0.992
UL12	5	0.993	6	0.993

UL26	5	0.993	6	0.993
UL51	5	0.993	6	0.994
UL48	5	0.994	6	0.994
UL38	5	0.994	6	0.995
UL5	5	0.994	6	0.995
UL7	5	0.994	6	0.994
UL24	5	0.994	6	0.994
UL52	5	0.994	6	0.995
UL10	5	0.995	6	0.994
UL26.5	5	0.995	5	0.995
UL30	5	0.996	6	0.995
US5	5	0.996	6	0.996
UL54	5	0.996	6	0.996
UL39	5	0.996	6	0.997
UL31	5	0.996	6	0.997
UL22	5	0.996	6	0.996
UL34	5	0.997	6	0.997
US8A	5	0.997	6	0.994
US6	5	0.997	6	0.997
UL41	5	0.997	6	0.997
UL29	5	0.997	6	0.997
UL3	5	0.998	6	0.998
UL53	5	0.998	6	0.997
UL47	5	0.998	6	0.998
UL19	5	0.998	6	0.998
UL33	5	0.998	6	0.999
UL18	5	0.999	6	0.998
UL21	5	0.999	6	0.998
UL25	5	0.999	6	0.999
UL15	5	0.999	6	0.999
UL28	5	0.999	6	0.999
UL16	5	1.000	6	0.999
UL20	5	1.000	6	0.999
UL35	5	1.000	6	1.000
UL45	5	1.000	6	1.000

[a] The BS mean of KOS (JQ673480), KOS (JQ780693), F(GU734771), H129(GU734772), and 17 (X14112), was divided by the BS of McKrae (JQ730035);

[b] The BS mean of KOS (JQ673480), KOS (JQ780693), F(GU734771), H129(GU734772), 17(X14112), and HF10 (DQ889502), was divided by the BS of McKrae (JQ730035). When a protein is not available from HF10, the numerator is the mean of five proteins available. The BS is generated by blast search against McKrae protein sequence using each relevant sequence of each strain including McKrae itself.

Table 3. Oligonucleotide primers for strain McKrae sequence confirmation.

Region	Coordinate	Forward Primer	Coordinate	Reverse Primer
Gap	116367	AGTTAACGGGCTACGCCTTC		
Gap	119377	CACTGTGGTTCTGGCTCCAT	119774	TCAGGGGATTTTTGCTGTCT
Gap	120557	AGGAAGAGGCAGAGGAGGAA	121176	CCCCATAGTGATCAGCGACT
Gap	125395	CAACACAACTCACAGCGACA	126001	ACGGCAACGACCTGATTAAG
Gap	131169	CTTTTTGCACGGGTAAGCAC	131670	TCGGTTTCGACCTCAGACTC
Gap	139181	GAAGGGACTCTGGCGTACAA	139502	GAGGTTGGGGTTTTGGGTCT
Gap	142633	TTTGCACCCAGCCTTTAACT	143050	GTACGGGGTGGTTTGTTCAT
Gap			121334	GGCTCCCGCCAGTCGCGAGCCGCGGCGCCGCGGGGGGCGT
Gap	142663	GGGTTCGATTGGCAATGTTGTCTCCCGGTTGATTTTTGGGT	143312	GGCTGTACTCGACCCACTGCCTGCATCTGTTTGGTGCGTT
ICP34.5	554	GGTGGGCCCCCGCCTTCTTGTT	1182	CGGGTGTAACGTTAGACCGAGTT
ICP4	127397	AACAGCGGGTGGTCCGTGAGC	127962	GCCGTGAGCCTGGTCGCC
ICP4	128022	GCATGAGCACCAGCGCGTCG	128594	CCATGAGCCGCCGATACGACC
ICP4	129010	CTCCCCGCTGACGTACCCGT	129381	TCCGCATCCTCTTCGTCCTCGT
oriL	61711	CGGTCAAGGGGAGGGTGGGA	62267	TTACCGGACCCCCAGGCTCG
UL17	32020	GGGGTCAGGGTTAGAACGTCGTT	33172	GCACGCACGCCTCCGTTGGCGAATA
UL30	63160	TGTTTTACCGCGTCTACGTC	63640	AGCTCGTTCAGGTGGGATT
UL36	71455	GGGTGAAGGGAGGGATTCCCGCGATT	72603	GCCGTGTCTGGTGCTCGTGGACAT
UL37	80910	CCAGGTGCGCCGTGGTTTCGGTGAT	81382	CCCCACTATGCAGATGGCCGACAA
UL43	93821	CGGAAGCCGTTACGCGTGTTACTTT	94322	GCGACATAACGACCGATCCCGCAAA
UL47	101926	GCGTCCCTCCAGAACCTCCACCTT	102638	CCTTCGCGGGAGATACTGCGTTTTT
UL47	100371	GGCGCTTTTTATCGGGAGGAGCTTAT	100604	GGACGGCAGACCCCGCCTTTCTTCT
UL56	115572	GGGGCGCTTACCGCCACAGGAATA	116190	GGAACGAGGGGCCGTTTGTCGTTATT
UL6	16504	GCCTGCAAACGTACCTGACAGACAT	16818	GCGGAGCAGCGTCTGCTGGCTAAT
US10	143780	CGCTCCATCTTGTGGGGAGAA	144137	GCGGCAGGGCAATGTGGAGATT
US11	143770	CCCGAGCGTACGCTCCATCTTGT	144258	CGACCCAGATGTTTACTTAAAAGG
US4	135763	CCCTCGCATGAAGCCCCCAACAT	136022	GGGACTACGGGACGGTTGGGTTTGT
US7	139132	GGACGCACCAAACGCCAGCCTGTTT	139603	CGCGGGGATGGCTATCTGGATTAT
US8	140646	CCCTGTCCGTGGGCGACATAAA	140969	GGATGGAGACGTTCGTGCTAAA

□upple□ ental Figure 1

Click here to download □upple□ entar□ □aterial □□o □e □u□li□hed□: □upple□ ental Fig. 1.ppt

□upple□ ental Figure 2

Click here to download □upple□ entar□ □aterial □□o □e □u□li□hed□: □up Fig. 2.ppt

□upple□ ental □a□le 1

Click here to download □upple□ entar□ □aterial □□o □e □u□li□hed□: □up □a□le 1.□l□