ANALYSIS OF HOST AND HERPESVIRUS INTERACTIONS USING BIOINFORMATICS

Ria Holzerlandt

Submitted to the University of London for the degree of Doctor of Philosophy

July 2004

Viral Genomics and Bioinformatics Group
Department of Immunology and Molecular Pathology
Windeyer Institute of Medical Sciences
University College London

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Bioinformatics methods have become central to analysing and organising the sequence data continually produced by new and existing sequencing projects. The field of bioinformatics covers both the static aspects of organising and presenting these raw data, by compiling existing knowledge into accessible databases, ontologies, and libraries; and the more dynamic aspects of knowledge discovery informatics for interpreting and mining existing data. The aim of this thesis is to utilise such methods to analyse the herpesvirus-host relationship.

In Chapter 2 comparative host and herpesvirus genome analysis is used to compare the sequences of all currently sequenced herpesvirus open reading frames to the conceptually translated human genome with the aim of identifying herpesvirus-human (host) sequence homologues. Collating in one search all currently known host homologues provides the first complete assessment of herpesvirus-host homologues. This search identified 55 previously identified herpesvirus-host homologues, and 4 previously unknown herpesvirus-host homologues.

The work performed in Chapter 2 highlighted the need for consistent annotation of genomes and gene products to allow greater comparative genomics. It is not feasible to manually curate large numbers of genes whose relationships to each other are not immediately clear. Therefore, Chapters 3 and 4 focus upon the use of the Gene Ontology; a resource that is made publicly available for the purpose of annotating gene products with unified vocabulary derived from a structured directed acyclic graph. The Gene Ontology was extended to allow host-pathogen interaction annotation by a) adding 187 new terms relating specifically to virus function and structure (Chapter 3), and b) using both the new and existing terms to annotate the entire Human Herpesvirus 1 genome using references from the available literature (Chapter 4).

Finally, Chapter 5 examines the utility of the Gene Ontology when analysing such large-scale host and herpesvirus gene expression datasets as produced experimentally by DNA microarray studies. Using such automated annotation methods a cluster of 12 proteins were identified that increase mitochondrial function in HUVEC cells 24 hours post HCMV infection. A cluster of nine proteins that function in the MAPK pathway were also identified using the Gene Ontology that provide evidence for HCMV inhibition of the MAPK pathway.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AIHV	alcelaphine herpesvirus
ATF	activating transcription factor
BCL-2	B-cell lymphoma protein-2
BKS	bovine, KSHV, swinepox
BLAST	basic local alignment search tool
BLASTP	BLAST for proteins
BLOSUM	blocks substitution matrix
BoHV	bovine herpesvirus
BRCA1	breast cancer protein 1
BSH	biologically significant hit
CATH	class, architecture, topology, homologous superfamily
CCHV	channel catfish herpesvirus
CEA	carcinoembryonic antigen
CeHV	cercopithecine herpesvirus
CLN	ceroid liporfuscinosis
CNS	central nervous system
CRD	carbohydrate recognition domain
CRE	cAMP responsive element
CREB	CRE binding
CVS	Concurrent Versions System
CXCR4	chemokine (C-X-C motif) receptor 4
DAG	Directed Acyclic Graph
DDBJ	DNA Data Bank of Japan
DUSP (aka MKP)	dual specificity phosphatase
dUTP	2'-deoxyuridine 5'-triphosphate
EBI	European Bioinformatics Institute
EBV	Epstein-Barr virus
EC	enzyme classification
EHV	equine herpesvirus
EMBL	European Molecular Biology Laboratory
ERK	extracellular-signal regulated kinase
EST	expressed sequence tag
FADD	FAS associated death domain
FAS	fatty acid synthase
FGARAT	formylglycineamide ribonucleotide aminotransferase
FTP	file transfer protocol
GaHV	gallid herpesvirus
GO	Gene Ontology
GPS1	G protein pathway suppressor 1
GXD	Gene Expression Database
HCCS	holocytochrome c-type synthetase
HCMV	human cytomegalovirus
HHV	human herpesvirus
HMM	hidden markov model
HPF	Homologous Protein Family
HSP	high-scoring segment pair
HSV	herpes simplex virus
HUVEC	human umbilical vein endothelial cell
HVS	saimiriine herpesvirus
ICAM	intracellular adhesion molecule

ICTV	International Committee for Taxonomy of Viruses
IE	immediate early
JNK	Jun N-terminal kinase
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	K nearest neighbour
KSHV	Kaposi's sarcoma herpesvirus
MAPK	mitogen activated protein kinase
MCMV	murine cytomegalovirus
MDM	mutation data matrix
MeHV	meleagrid herpesvirus
MEKK/MAPKKK	mitogen activated protein kinase kinase kinase
MGD	Mouse Genome Database
MHC	major histocompatibility complex
MHV	murine herpesvirus
MKK/MEK/MAPKK	mitogen activated protein kinase kinase
MRP	mitochondrial ribosomal protein
MSP	maximal seqment pair
NADH	nicotinamide adenine dinucleotide
NCBI	National Center for Biotechnology Information
NFkB	nuclear factor of kappa light polypeptide gene enhancer in B-cells
NK	natural killer
NLM	National Library of Medicine
OMIM	Online Mendelian Inheritance in Man
ORF	open reading frame
PAM	point accepted mutation
PML	promyelocytic leukemia
PSG	pregnancy-specific glycoprotein
PSI-BLAST	position specific iterated BLAST
PSSM	position specific scoring matrix
PTM	post-translational modification
RaHV	ranid herpesvirus
RCMV	rat cytomegalovirus
SaHV	salmonid herpesvirus
SGD	*Saccharomyces* Genome Database
SMART	simple modular architecture research tool
SOM	self-organising map
TAIR	The *Aribidopsis* Information Resource
TK	thymidine kinase
TNFR	tumour necrosis factor receptor
TRAF	TNFR associated factor
TrEMBL	translated EMBL
UL	unique long
US	unique short
vFLIP	viral FLICE inhibitory protein
VIDA	virus database
vIL-10	viral interleukin 10
VZV	Varicella-Zoster virus
XML	extensible markup language

1.0 Introduction

1.1 Viral Bioinformatics

Virology has often been a driving force behind advances in biological understanding: for example, virology was the first discipline to become post-genomic with the full sequence of bacteriophage MS2 becoming available in 1976 (Fiers, Contreras et al. 1976), and the sequence of Simian Virus 40 being completed in 1978 (Fiers, Contreras et al. 1978). Since then 1278 viral genome sequences have been discerned (Entrez Viral Genomes: 30 April 2004), of which 32 are herpesviruses, including all eight human herpesviruses (Arrand, Rymo et al. 1981; McGeoch, Dolan et al. 1985; Chee, Bankier et al. 1990; Russo, Bohenzky et al. 1996; Dargan, Jamieson et al. 1997; Dolan, Jamieson et al. 1998; Davison, Dolan et al. 2003); (Davison and Scott 1986; Gompels, Nicholas et al. 1995) (NC_000898; NC_001716). Deposited viral genome sequences are also updated and corrections to existing records are made, often with new insights. Such was the case with the recent comparison of wild-type human herpesvirus 5 (human cytomegalovirus; HCMV) with the close relative chimpanzee cytomegalovirus (Davison, Dolan et al. 2003). This study discounted 51 previously putative HCMV proteins, modified 24 and proposed 10 novel genes. Such co-linear base-by-base genome comparison, even with genomes of relatively small sizes, was previously impossible without the use of computers. Therefore, even this level of computer based analysis produces defined changes in virology knowledge.

1.2 The Search for Homology

Searching for sequence or structural based homology is one of the areas of research that is most efficiently accomplished using bioinformatics tools. Homology is a measure of similarity linking ancestral conservation of structure, sequence and function. By searching for homology between two proteins, relationships between sequences/structures can be determined, and function inferred.

1.2.1 Homology, Homoplasy, Orthology, Paralogy, and Xenology

Sequence similarity searches of databases with a defined sequence are designed to identify sequence relatives and thereby infer homology. Many sequence alignment programs are currently available such as BLAST (Altschul, Gish et al. 1990), PSI-BLAST (Altschul, Madden et al. 1997), or IMPALA (Schaffer, Wolf et al. 1999), to search reference databases such as GenBank (Karsch-Mizrachi and Ouellette 2001).

Homology has become a ubiquitous definition when studying two or more sequences in reference to each other; and as a term, is often used inappropriately to indicate that two sequences share similar characteristics. However, two sequences can only be homologous if they both inherited their shared characteristics from the same common ancestor(Page and Holmes 1998).

Alternatively, homoplasy describes the independent acquisition of similar features by two unrelated sequences, otherwise known as convergent evolution. Homology is often misused because it is difficult to determine whether two similar sequences are homologous or homoplasious without prior knowledge of their common ancestor. Thus, homology is inferred when the two sequences are closely related, whereas two sequences that appear more distantly related, yet share common characteristics at the sequence level, are difficult to categorise.

There are three subcategories of homology, depending upon inheritence. Orthology is homology shared by two genes that resulted from a speciation event. Paralogy is the result of gene duplication after speciation, the two resulting genes being paralogous, The hemoglobin gene family is a classic example of paralogy (Gribaldo, Casane et al. 2003).

The final type of homology that is often seen in viral context, is xenology. This describes homology acquired by horizontal gene transfer. Thus, the viral oncogene, v-src, that was acquired from an avian ancestor, shares xenology with the avian equivalent, c-src.

In virology, homology detection has been used to infer functional properties of related viral genes to host genes and of host function to viral genes (Holzerlandt, Orengo et al. 2002).

1.3 Sequence Alignment

A protein is encoded by sequences of amino acids, their order effecting its molecular structure and function. Those amino acids vital to function are conserved creating recognisable unique patterns or motifs of amino acids that can be directly related to protein function. These motif patterns can be searched for within a sequence database using a query sequence as a template. There are a variety of programs available that can search for individual protein matches (pairwise alignments), multiple protein matches (multiple alignments), functional domain homology (local alignment) or protein homology (global alignment); each is equipped with a variety of algorithms to deal with such issues as amino acid insertion, deletion, or substitution.

1.3.1 Pairwise Alignments Algorithms

Two proteins can share homology across their entire length. This is a 'global' alignment and is usually indicative of a shared evolutionary history, as well as shared function, (i.e. homology). More commonly, regions of similarity within two sequences are detected. This 'local' alignment is indicative of two proteins that have similar functions (such as kinase activity) limited to areas or 'domains' of the protein sequence. Algorithms have been designed to look for global alignments (Needleman-Wunsch) and local alignments (Smith-Waterman) between two sequences.

1.3.1.1 Needleman-Wunsch Algorithm

The Needleman-Wunsch algorithm (Needleman and Wunsch 1970) is an algorithm that finds the optimal global alignment between two sequences by searching for the maximum number of residues from one sequence that can be matched to another while allowing for all possible deletions/insertions. Introducing a gap penalty (a penalty induced by the insertion of a gap into the alignment) into the scoring system inhibits arbitrary gap insertions. The two sequences (of lengths x and y) are aligned in a two-dimensional matrix with the beginning of the alignment being represented by cell (1,1)

and end of the alignment being represented by cell (x, y), signifying a global alignment. Scores are accumulated along the diagonals of the grid. All alignment algorithms can be used with a residue scoring substitution matrix (see below). The advantage of dynamic algorithms, such as the Needleman-Wunsch, is they calculate all possible permutations in order to guarantee that the alignment found is indeed optimal.

The disadvantage of this method is that it cannot detect local areas of similarity between multidomain proteins where only a subset of domains match. Three-dimensional structural analysis of proteins reveals the domain nature of proteins and suggests protein domains are an evolutionary unit (Ponting and Russell 2002; Teichmann 2002; Vogel, Berzuini et al. 2004). Such evolution of domains by processes of domain shuffling or swapping highlights a need for methods that identify regions of local similarity within and between proteins.

1.3.1.2 Smith-Waterman Algorithm

Smith and Waterman devised an algorithm that looks for regions of local sequence similarity between two sequences by reducing the accumulated score through long regions of dissimilarity to a negative score (Smith and Waterman 1981). Like Needleman-Wunsch, the Smith-Waterman algorithm uses a two-dimensional matrix, but in this case each cell potentially represents the beginning or end of a region of local alignment. In order to emphasize regions of similarity, matches are positively scored and accumulated through the matrix along diagonals, mismatches are given a negative score. Thus if any cell's score becomes negative, it is replaced with a default 0, indicating the end of an alignment. The advantage of this algorithm is that accumulation of scores can be calculated for every cell in the matrix to detect more than one region of local sequence similarity between two multidomain proteins (Figure 1.1).

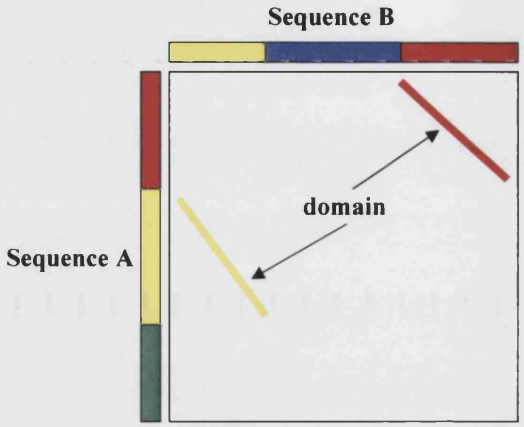


Figure 1.1 Multiple Local Alignments. Local alignment algorithms can be used to discover more than one region of local similarity between multidomained sequences. Sequence A and B share two regions of similarity (in red and yellow) which are found by tracing back all high scoring cells through the matrix until a 0 is met.

1.3.2 Substitution Matrices

Aligning two sequences (as described above) uses a scoring function that looks only for occurrences of residue identity (i.e. matching two identical sequences at similar positions). In reality, substitutions of one amino acid for a similar amino acid result in occurrences of residue similarity. Evolutionary pressures allow those substitutions that do not adversely affect the function of the protein to occur far more readily than those that do. Thus, the probability of a certain amino acid substitution occurring within a sequence is a direct result of its phenotypic affect upon the protein, and the genetic code itself (Higgins and Taylor 2000). These probabilities can be incorporated into substitution matrices that weight specific amino acid substitutions with scores that reflect their probability of occurrence in nature. The two most popular substitution matrices are the MD or PAM, and BLOSUM matrices.

1.3.2.1 Dayhoff Mutation Data (MD) Amino Acid Substitution Matrices

The mutation data matrix (MDM) is based upon the evolutionary unit of time the Point Accepted Mutation (PAM) (Dayhoff, Schwartz et al. 1978). One PAM is the amount of evolutionary time it would take to change, on average, 1% of the residues in a protein. To estimate the relatedness of two proteins, Dayhoff used the common ancestor method by taking two 85% identical present day sequences and deducing the sequence of the common ancestor by building phylogenetic trees and inferring the most likely ancestor at each node. This allows the calculation of the number of PAMs (substitutions) per sequence pair of either: the two present day sequences, or one present day sequence and its ancestor. Each substitution is counted twice as it can never really be known whether residue 1 mutated to residue 2 or vice versa. PAM matrices are symmetrical (ie Leucine (L)-> Valine (V) is given the same score as Valine (V)-> Leucine (L) as similarity is considered a symmetrical concept (Valdar and Jones 2003).

The PAM 1 matrix, therefore, indicates how likely one amino acid will be substituted to another over the period of 1 PAM. It is easy to calculate the scores for PAM (n) matrices by simply raising each score to the power of (n); this is necessary to compare sequences that do not share much sequence similarity (ie are quite distantly

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	W	Y
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
W	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	17	
Y	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	10

Figure 1.2 PAM250 Substitution Matrix. This substitution matrix can be used when aligning two sequences that share as little as 20% sequence homology. A positive score indicates a more likely replacement, a negative score a less likely replacement. Probability of such a replacement increases with score. The scores from each amino acid pair are summed together to produce a final score for the alignment. The higher the score the better the alignment.

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	W	Y
C	9																			
S	-1	4																		
T	-1	1	4																	
P	-3	-1	1	7																
A	0	1	-1	-1	4															
G	-3	0	1	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	1	-1	-2	-1	1	6												
E	-4	0	0	-1	-1	-2	0	2	5											
Q	-3	0	0	-1	-1	-2	0	0	2	5										
H	-3	-1	0	-2	-2	-2	1	1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	11	
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	2	7

Figure 1.3 BLOSUM62 Substitution Matrix. The most commonly used BLOSUM matrix based upon blocks of protein family multiple alignments that share an average of 62% sequence homology. The scores from each amino acid pair are summed together to produce a final score for the alignment. The higher the score the better the alignment.

related). The PAM 250 (Figure 1.2) matrix is frequently used in alignment algorithms as its scores can be applied to sequences that share as little as 20% similarity, i.e. sequences that are separated by 250 PAMs (250 substitutions per 100 amino acids).

1.3.2.2 Blocks Substitution Matrices (BLOSUM)

Henikoff and Henikoff developed the BLOcks SUbstitution Matrices (BLOSUM) in order to create more empirically scored matrices (Henikoff and Henikoff 1992). Unlike the MDMs, BLOSUM matrices do not rely upon an evolutionary model to derive substitution rates. Instead, BLOSUM matrices are derived from families of proteins that share a known function and thus sequence motifs (Blocks). This is achieved by counting the number of amino acid substitutions found in the Blocks taken from the BLOCKS database (Henikoff, Henikoff et al. 1999; Henikoff, Greene et al. 2000). The BLOCKS database is constructed by searching for regions of high amino acid conservation within protein family multiple alignments; these regions are stored as Blocks in the database. Each block is distinguished by the average percentage sequence identity shared between its members. Thus, the BLOSUM 62 (Figure 1.3) matrix is derived from a multiple alignment of sequences that share 62% sequence identity, giving the matrices an evolutionary context; it is the most commonly used of the BLOSUM matrices. Unlike the MDMs, it is not possible to multiply the BLOSUM (n) matrix to obtain the BLOSUM (x) matrix. Each matrix must be derived from a multiple alignment of sequences that share a defined percentage of residue identity.

1.3.3 Multiple Alignments

Pairwise alignments using algorithms such as the Smith-Waterman or the Needleman-Wunsch can only find homology between two sequences that are relatively well conserved at either a local or global level. When sequence similarity drops to below approximately 25% in proteins it becomes difficult to distinguish true negatives from distantly related family members (Orengo 2003). However, distantly related members of a protein family often retain certain structural or funtional features in their sequence that are evolutionarily significant, and explain their presence in the same protein family. Multiply-aligning a number of sequences simultaneously and statistically scoring their shared similarities can build a descriptive two-dimensional definition of a protein family. These definitions can then be used to identify more distantly related proteins.

One such multiple alignment method is the Position Specific Scoring Matrix (PSSM). Whereas Blocks based BLOSUM matrices condense the information within many multiple alignments into a given scoring matrix, PSSMs build a scoring matrix per multiple sequence alignment.

1.3.3.1 Position Specific Scoring Matrices (PSSMs)

A PSSM records the frequency and position of conserved residues within a multiple sequence alignment (Figure 1.4). A PSSM can be built from a multiple alignment of related sequences, usually found through a series of pairwise alignment matches, or as the result of a multiple alignment program.

A two-dimensional matrix: the length of the sequences by the number of sequences in the group, is built from the multiple alignment. Each position in the matrix records either an amino acid or a gap character. Gaps are often considered as a 21st letter of the amino acid alphabet to aid in the computation of statistics. A PSSM records the likelihood of a particular residue occurring at a particular position in the sequence. This can be calculated and weighted in conjunction with a substitution matrix, but is also dependent upon conditions such as: whether all of the residues in a column are identical, or if not, the number of independent observations that occur in each column (ie the number of different amino acids that occur in the column). Independent observations are based upon the assumption that each unique amino acid present in a column is the result of a point mutation at that position during the family's evolution. Each PSSM method differs slightly, but most will also take into account the global sequence similarity of each sequence to the others, in order to avoid a biased score matrix.

The PSSM scoring matrix that results (known as the PSSM) is of the dimensions: length of query sequence by number of amino acids + gap character (i.e. 21). Thus, instead of using a similarity matrix which simply records the probability of residue A being replaced by residue B over time (such as BLOSUM62), a PSSM calculates the probability of residue A being replaced by residue B at position N in a given multiple sequence alignment (1 protein family) over time.

1.3.4 Scoring Functions

When searching a database with an alignment program each result is accompanied by a score indicating its statistical significance. Every alignment program uses its own scoring function, and these will be described on a program by program basis.

1.4 Data Sources

With the increase in production of biological sequence data comes an increase in the necessity for efficient storage and search facilities. Avoiding redundancy and providing public access to such data are two important aspects that need to be considered when developing and maintaining storage facilities.

Available primary research data is stored in a variety of databases that range from literature resources such as Pubmed (NCBI), to biological sequence data stores such as GenBank (Karsch-Mizrachi and Ouellette 2001), EMBL (Kulikova, Aldebert et al. 2004), DDBJ (Miyazaki, Sugawara et al. 2004). Three principal methods exist for organising biological sequence data into primary, secondary or compilation databases. Primary databases store primary sequence information for nucleic and/or amino acid sequences. They receive their sequences from a number of sources from individual submitters to genome sequencing projects. Many primary databases perform low level curation, for example by classifying their entries according to their reliability (eg. annotated sequences vs. conceptual translations). Secondary databases analyse and organise primary sequence data using their own algorithms. There are a large number of secondary databases already in existence, the reliability of each database depending upon the methods used to build it, its quality of annotation, and the frequency of its sequence updates. Compilation databases collate data from a number of different primary and secondary sources for convenient user access.

```
...AKGHITQTPITYITPY.....FA...   species 1
...AKGDLTMTPLYGITPY.....FA...   species 2
...AKGYITQTPKERITPYPIIPLFA...   species 3
...AKGGITMTPP...TPY.....FA...   species 4
...AKGSLTMTPRIPITPY.....FA...   species 5
...AKGHITQTPIREITPY.....FA...   species 5
```

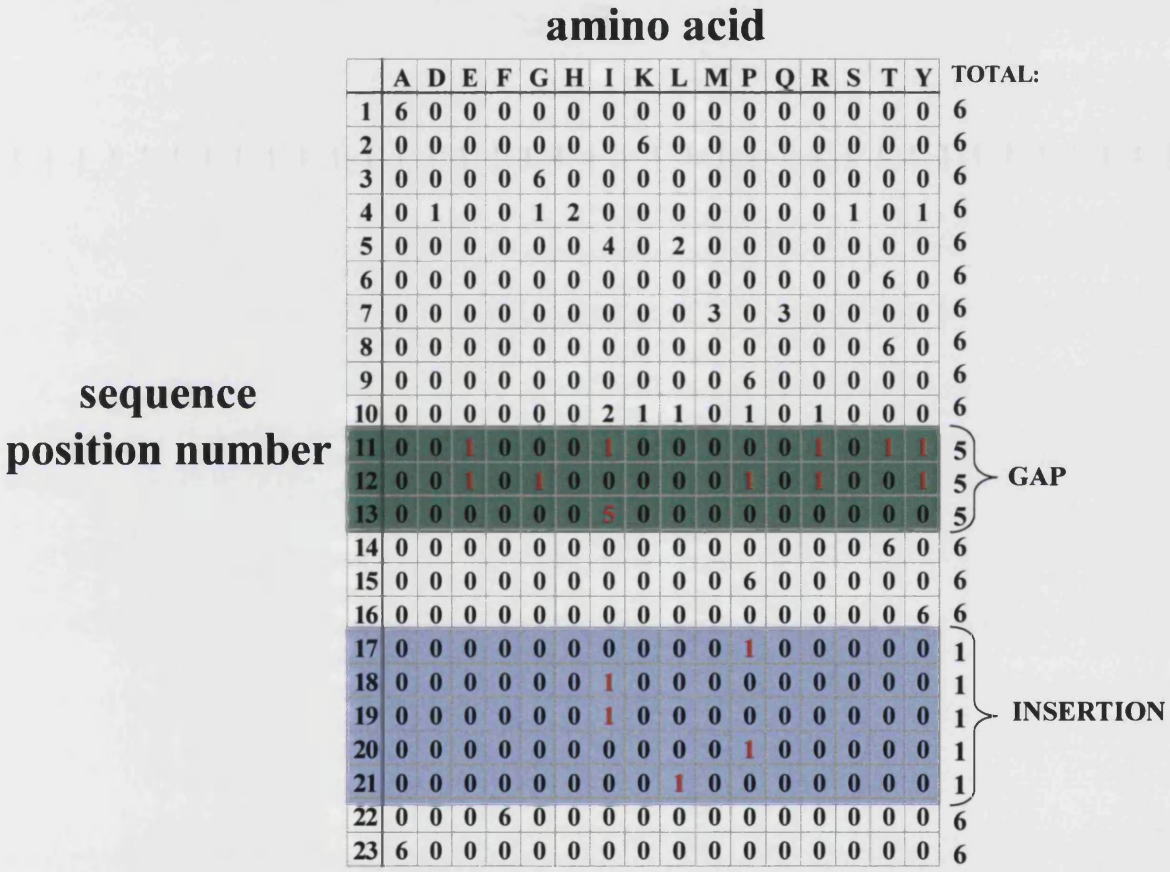


Figure 1.4 A Position Specific Scoring Matrix (PSSM). The multiple alignment above can also be represented in a simple PSSM by recording the frequency of residue occurrence at each position in the alignment. By looking at the totals of each position (to the right of the matrix) the positions of the gap in species 4 (highlighted in green), and the insertion in species 3 (highlighted in blue) can be identified. This is merely a simple representation of a PSSM; PSSM calculations most often include scores for each position based upon some form of similarity matrix and gap penalties.

1.4.1 Primary Sequence Database: GenBank

GenBank (Karsch-Mizrachi and Ouellette 2001), is a publicly accessible primary database that stores nucleic and amino acid sequences with accompanying bibliographical and biological annotation. It is maintained by the NCBI, a division of the National Library of Medicine, on the National Institute of Health campus in the USA. GenBank, in conjunction with DDBJ (DNA Data Bank of Japan) and EMBL (European Molecular Biology Laboratory; accessible from the British outstation EBI: European Bioinformatics Institute) maintains the largest repositories of biological sequence data in the world. The three databases consolidate daily by exchanging new sequence submissions thereby keeping all three synonymous and up-to-date.

The bulk of GenBank is populated from EST projects, but also by public submission from authors, individual labs, large sequence projects, genome survey sequencing (GSS), the US Patents and Trademarks Office (USPTO), and other high-throughput data produced by various sequencing centres.

GenBank is composed of 39 billion nucleotide bases from 33.7 million different sequences, approximately 33% of which are of human origin (25% of all sequences of human ESTs) (release 141, 15 April 2004). There are currently over 130,000 species represented in GenBank and new species sequences are being added at a rate of 1100 per month.

1.4.2 Compilation Sequence Databases

1.4.2.1 SWISS-PROT/TrEMBL

The SWISS-PROT protein database (Boeckmann, Bairoch et al. 2003) is maintained by the Swiss Institute of Bioinformatics (ExPASy) and the EBI, and is a compilation database that draws together amino acid sequence information with experimental results and information from the literature to provide a comprehensive overview of all relevant data relating to a protein entry. Unlike many compilation databases that are publicly available, SWISS-PROT supplies high quality manual annotation for each entry, using standardised nomenclature officiated by international committees where possible and controlled vocabularies elsewhere. In addition to compiling data into easily

decipherable records, SWISS-PROT provides direct links to specialised databases, including cross-references to such resources as the originating DNA sequences from GenBank/EMBL/DDBJ, 2D and 3D protein structure databases, a variety of protein domain and family characterisation databases, species-specific projects, variant and disease databases and post-translational modification (PTM) databases (Boeckmann, Bairoch et al. 2003). SWISS-PROT accommodates all species but focuses upon the human genome and other model organisms. SWISS-PROT contains over 149,914 sequence entries with information cross-referenced from 66 different databases (release 43.2, 24 April 2004).

It is difficult for a manual curation databases such as SWISS-PROT to keep pace with the high volume of sequence data being produced, therefore, the TrEMBL (translation of EMBL) has been developed as a computationally annotated counterpart to SWISS-PROT. It is derived from translations of all coding sequences in EMBL that are not already in SWISS-PROT, and like SWISS-PROT, then computationally maps to additional online resources. TrEMBL also includes new protein sequences from publication and sequences submitted by the public and currently contains over 1,065,889 sequences (release 26.2).

1.4.2.2 InterPro

The Integrated Documentation Resource of Protein Families, Domains, and Functional Sites (InterPro) (Mulder, Apweiler et al. 2003) is maintained by the European Bioinformatics Institute. InterPro is a compilation database that brings together into a single resource the information from ten protein signature databases: SWISS-PROT (Boeckmann, Bairoch et al. 2003), PRINTS (Attwood, Bradley et al. 2003), TrEMBL (Boeckmann, Bairoch et al. 2003), Pfam (Bateman, Coin et al. 2004), PROSITE (Hulo, Sigrist et al. 2004), ProDom (Servant, Bru et al. 2002), SMART (Letunic, Copley et al. 2004), TIGRFAMs (Haft, Selengut et al. 2003), PIR SuperFamily (Huang, Barker et al. 2003), and SUPERFAMILY (Gough, Karplus et al. 2001). InterPro contains information about protein families, conserved sequence domains, repeats and translational modification sites, providing an interface for protein pattern analysis.

InterPro contains approximately 10709 entries, with over 2411 domains, 8035 families, 197 repeats, 26 active sites, 20 binding sites, and 20 post-translational modifications

(release 7.2, 29 March 2004). Updates are scheduled every two months. The database is built by processing flat files from source databases and merging the information to form InterPro records. Overlapping domains, signatures, profiles or families are integrated into one entry and given a unique accession number (taking the form: IPR123456) (Mulder, Apweiler et al. 2003). Proteins that have no identifiable counterparts in other source databases are assigned their own InterPro entries and accession number.

1.4.3 Genome Sequencing Projects

Genome sequencing projects are providing an increasing number of complete genome and organelle sequences to public repositories, with virus genome sequences comprising 49% of the total (Table 1.1). As complete host genome sequences become available, such as the human genome, completed in April 2003 (NIH-Newsroom 2003), further research into the host-virus relationship can be undertaken.

1.4.4 Viral Databases

The large amount of viral data available has seen an increase in the number of websites and databases that specialise in their organisation (Table 1.2). These databases are a combination of secondary and compilation resources, providing access to sequence data using their own algorithms and content formats. Resources that supply viral data accompanied by initial sequence analysis can provide useful bases from which to conduct further virus research.

1.4.4.1 Secondary Sequence Database: VIDA – (VIrus DAtabase)

VIDA is a secondary database that processes animal virus open reading frames (ORFs) from GenBank for a given virus family and organises them by homology into homologous protein families (HPFs). The current release, VIDA 2.0, is populated with all viral open reading frames in GenBank release 130 (last VIDA update: 19 August, 2002) from Arteriviridae, Coronaviridae, Herpesviridae, Papillomaviridae, and Poxviridae. Users can search pre-compiled libraries for proteins of interest, by HPF number, virus name, protein function (or functional class as designated by VIDA), GenBank number, free text, or by using their input query sequence. Links to the

complete genome sequences for Arteriviruses, Coronaviruses, Herpesviruses, and Poxviruses are provided where available. VIDA is populated by a defined algorithm (Figure 1.5). Viral data are parsed from GenBank records, which are filtered to extract only the open reading frames (ORFs) from a given virus family. HPFs are constructed using the PSCBuilder algorithm and additional data pertaining to the viral genomes and ORFs are acquired from external databases: CATH (a Protein Structure Classification Database) and the ICTV (International Committee of the Taxonomy of Viruses). Each HPF is then represented by a table, which includes links to SWISS-PROT, TrEMBL, and each ORF's complete genome record where appropriate (Figure 1.8).

1.4.4.1.1 Homologous Protein Families (HPF)

The HPF (Figure 1.6) results from the VIDA algorithm. HPFs are constructed using MKDOM/XDOM (Gouzy, Eugene et al. 1997; Gouzy, Corpet et al. 1999), which identifies conserved amino acid motifs within a protein using the local multiple alignment program BLASTP (Altschul, Gish et al. 1990). Once these motifs are identified in all ORFs from a given virus family (for example, Herpesviruses), the ORFs are clustered into protein families according to shared motifs (Figure 1.7) by the VIDA specific program PSCBuilder. The HPF is the lowest common grouping possible within VIDA as subgroups within HPFs are not formed. HPFs can be defined by more than one motif as long as all members of the family contain the same defining motifs. Each HPF is annotated with a functional description and functional class derived from a combination of GenBank records and manual curation and can contain proteins from any or all of a virus family's subtaxonomies (i.e. subfamily).

In some instances, no homology to other ORFs within the chosen virus family can be found. These 'singleton' proteins are still included in the database and displayed in an identical fashion to the other HPFs, but contain only a single protein entry.

An example of an HPF (family 308) that has three members all involved in gene expression regulation is shown in Figure 1.8. The functional descriptions in VIDA include a representative gene name, in this case, *immediate early regulatory protein, HHV-1 US1*. Along with links to each ORF's FASTA format sequence and SWISS-PROT/TrEMBL entries, each HPF includes a link to the conserved motif alignment(s)

that define each HPF (Figure 1.8b). This alignment can be retrieved in FASTA sequence format (Figure 1.8c). In some instances HPFs contain several proteins from the same virus species. This is due to the existence of proteins from different strains or to the presence of more than one copy of the gene in the virus genome. Those sequences retrieved from GenBank that share 100% sequence identity are noted and only one is included in the HPF table, a link beneath the table indicates the details of the other redundant sequences. Some virus genes have determined 3D structures, which are also linked to the HPF table.

Table 1.1 Complete Genomes/Organelles in NCBI Entrez Genomes[1,2]

Organism	Type	Number of Complete Genomes
Eukaryota	Genome	20
	Organelles[3]	556
	Plastids	35
	Plasmids	14
Nucleomorph	Genome	1
Archaea	Genome	18
	Plasmids	23
Bacteria	Genome	154
	Plasmids	501
Viruses		1287

[1]Source: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
[2]30 April 2004
[3]Mitochondria & Chloroplasts

Table 1.2 Selection of Virus Data Databases[1]

Database	Contents	Location	Reference
HIV Database	Sequences, alignments, immunology, drug resistence,vaccines	Los Alamos National Laboratory, USA	(Kantor, Machekano et al. 2001)
Influenza Sequence Database	Sequences, 3-D tools	Los Alamos National Laboratory, USA	(Macken, Lu et al. 2001)
Hepatitis C Virus Database	Sequences, sequence analysis tools	Réseau National Hépatites, France	http://hepatitis.ibpc.fr
IAH Virus Pages	Sequences, alignments, documentation (Animal Pathogenic Viruses)	Institute for Animal Health, UK	http://www.iah.bbsrc.ac.uk/virus/
VIRGO (Viral Genome Office)	Complete genomes and protein sequences, orthologous clusters (Poxviruses, Herpesviruses)	University of Victoria, Canada	(Hiscock and Upton 2000)
VIDA (Virus Database)	Homologous protein families, alignments (Herpesviruses, Poxviruses, Papillomaviruses, Coronaviruses, Arteriviruses)	University College London, UK	(Alba, Lee et al. 2001)
Viroid and viroid-like Sequence Database	Sequence, RNA secondary structure prediction	Université de Sherbrooke, Canada	(Lafontaine, Mercure et al. 1997)

[1]Adapted from (Kellam and Alba 2002)

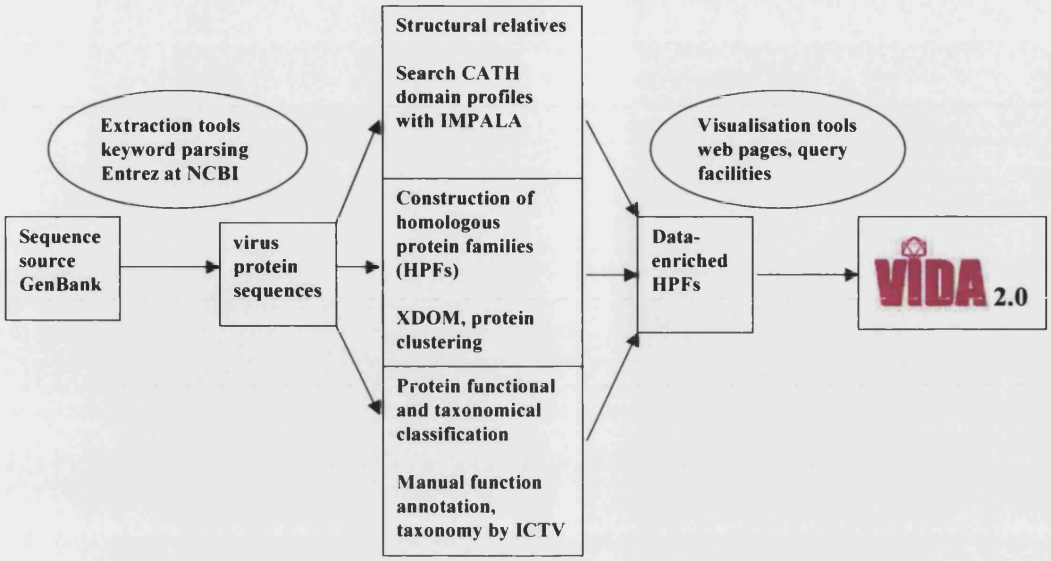


Figure 1.5 An overview of the VIDA algorithm. VIDA, as a secondary database retrieves its source data from a primary database, GenBank, and performs further computational processes on the new data. VIDA also integrates further details from other databases to maximise the quality of its annotation.

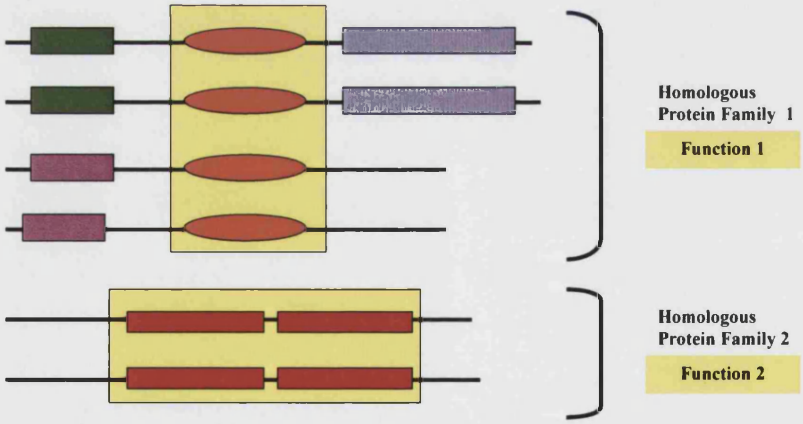


Figure 1.6 Building homologous protein families. Once MKDOM has identified all conserved regions for all relevant ORFs, PSCBuilder (Mar Alba) groups the proteins into HPFs according to conserved shared motifs. HPFs can be defined by more than one conserved motif, but each member of the family must contain all defining motifs as in the case of Homologous Protein Family 2 above which is defined by two conserved motifs.

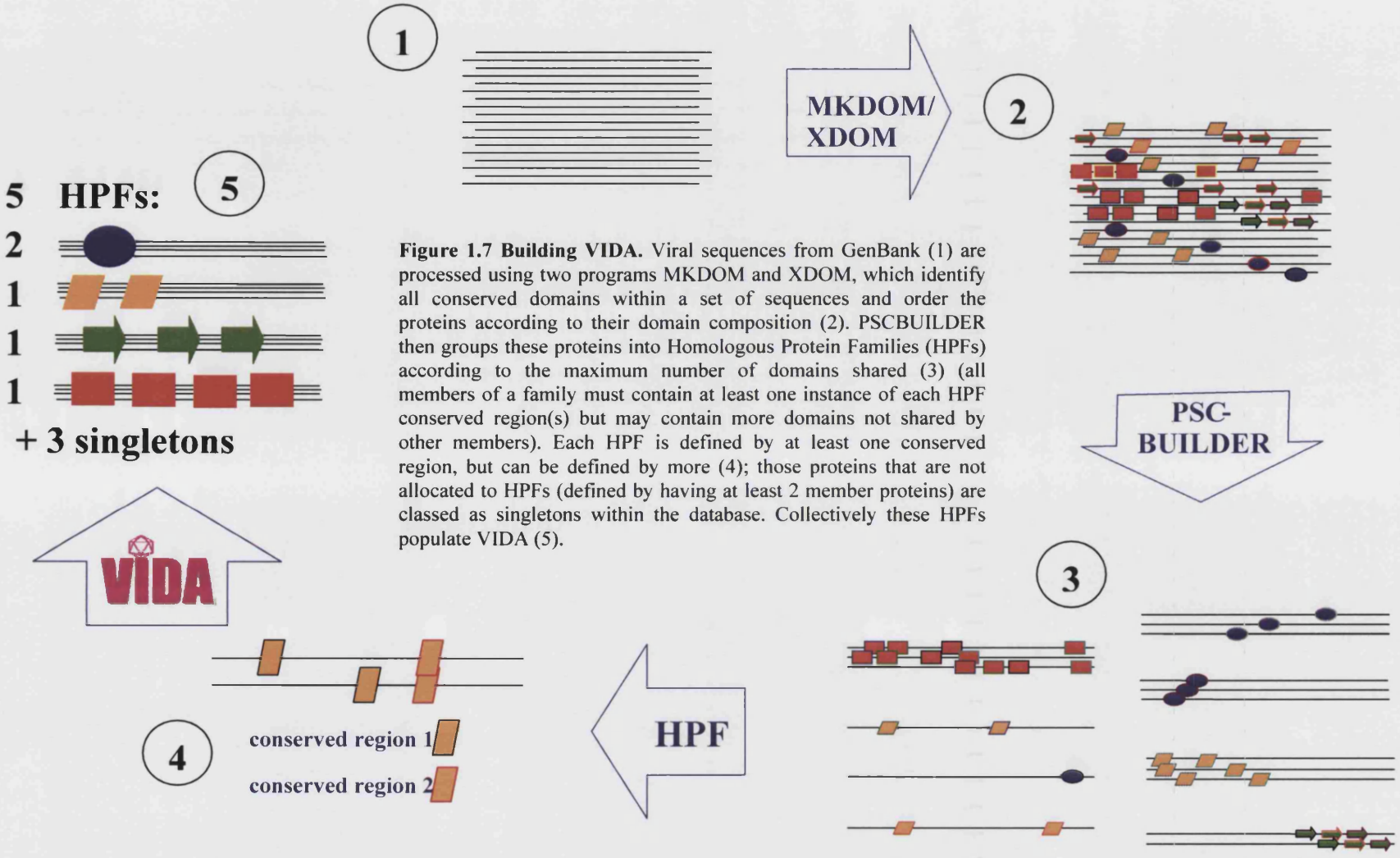


Figure 1.7 Building VIDA. Viral sequences from GenBank (1) are processed using two programs MKDOM and XDOM, which identify all conserved domains within a set of sequences and order the proteins according to their domain composition (2). PSCBUILDER then groups these proteins into Homologous Protein Families (HPFs) according to the maximum number of domains shared (3) (all members of a family must contain at least one instance of each HPF conserved region(s) but may contain more domains not shared by other members). Each HPF is defined by at least one conserved region, but can be defined by more (4); those proteins that are not allocated to HPFs (defined by having at least 2 member proteins) are classed as singletons within the database. Collectively these HPFs populate VIDA (5).

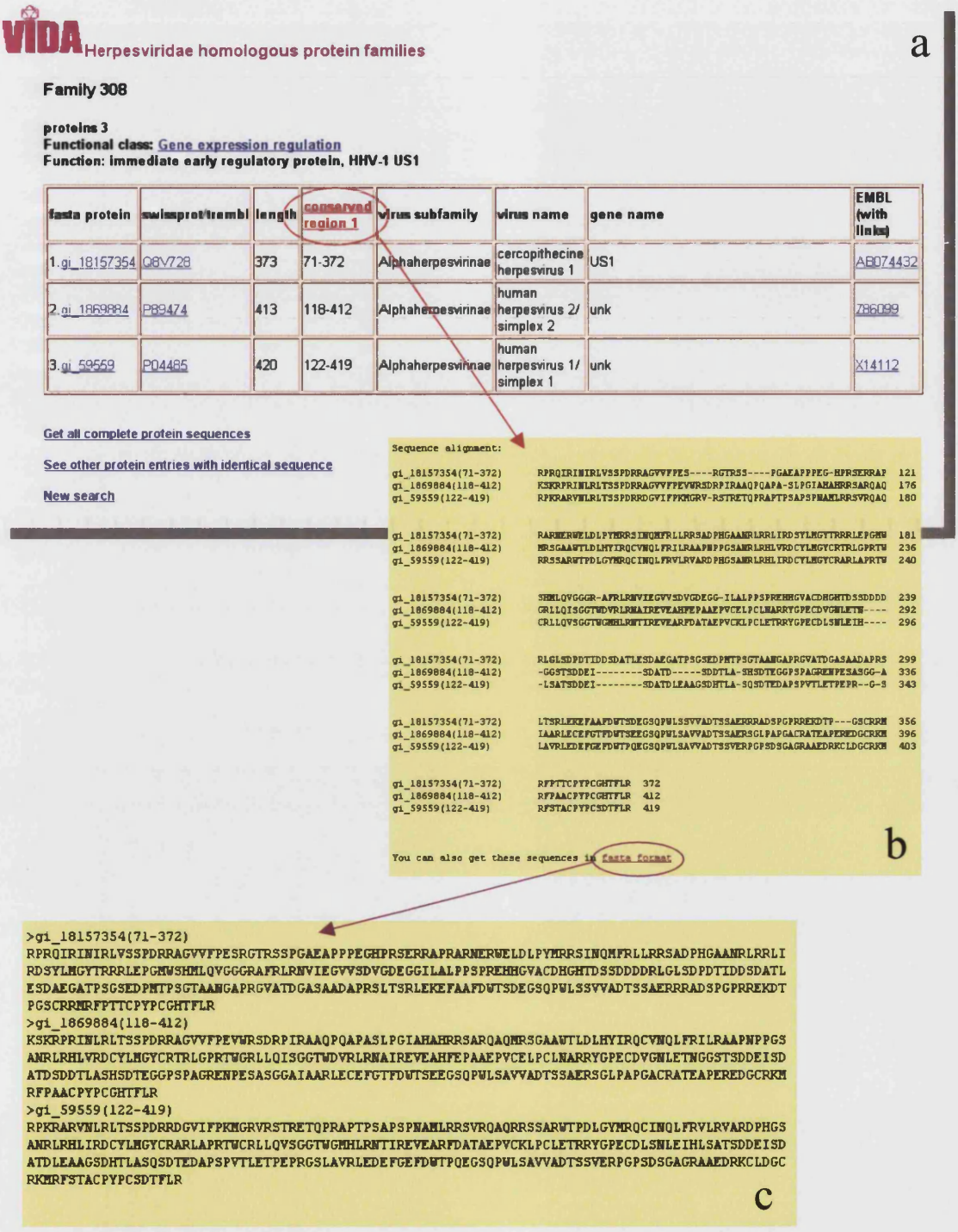


Figure 1.8 An online view of an HPF. a) Each HPF lists the functional class in which it is placed (by VIDA curators) along with the general function of the member proteins and an example protein. Each member protein is listed with links to its fasta sequence, SWISS-PROT/TrEMBL entry, and complete genome entry (where available). Beneath the HPF table are links to protein sequences that were not included in the HPF due to redundancy at the 100% sequence identity level with an existing member; a complete fasta formatted list of all member proteins; and where available, a link to any protein structures associated with the HPF. b) An alignment of the conserved regions that define the HPF can be obtained; including a fasta format version (c).

1.5 Gene and Genome Annotation

Using sequence alignment to confer knowledge from one protein to another can only be accomplished if there is sufficient structural/functional knowledge available about one of the two proteins. Thus, high-quality sequence data annotation is important for bioinformatics research, to ensure that the information conferred is appropriate, and to avoid the perpetuation of incomplete or inaccurate annotations. Bioinformatics is also capable of handling large datasets, containing information from multiple genes, pathways, organisms and species. In order to process these data most efficiently, uniform methods of annotation are necessary to ensure that data are organised and cross-comparable.

Such machine-readable annotation requires defined and delimited vocabularies, and controlled use of the vocabularies. A number of ontologies and vocabularies have been developed that aim to unify annotation. Databanks and Encyclopedias such as the ENZYME data bank (Bairoch 2000), or the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, Goto et al. 2004) collate and organise gene products and genomes, and assign, in the case of the ENZYME data bank, EC (Enzyme Classification) numbers to each enzyme function. These EC numbers can be applied to enzymes from any species, allowing genes from different species with the same enzymatic function to be annotated and easily compared. Similarly, the Gene Ontology (Harris, Clark et al. 2004); (Ashburner, Ball et al. 2000; Consortium 2001) organises three annotation ontologies that structure biological processes, molecular functions, and cellular components (locations) to allow gene products from any species to be appropriately annotated. The descriptional terms within the Gene Ontology are organised in an hierarchical manner into Directed Acyclic Graphs (DAGs), similar to trees, that allow for relationships between gene products to be studied.

The ability to handle large datasets has allowed the study of genomics to embrace technology that can capture the gene expression pattern (transcriptomics), or the protein content (proteomics) of a cell at a given moment in time. The development of microarrays allows for thousands of gene expression levels to be studied simultaneously. Some array studies simply characterise expression profiles, looking at the entire expression profile and listing genes that have been up- or downregulated at certain timepoints in the genome's replication cycle. It is also perfectly valid to study

only subdivisions of the dataset by using clustering methods to group together genes with similar expression patterns. Integration of functional genomics data with gene and protein functional meta data will undoubtedly increase the utility of microarrays.

1.6 Herpesviridae

The *Herpesviridae* virus family consists of large, enveloped, DNA virus members. Believed to co-evolve with their hosts, they appear to be ubiquitous, as most vertebrate animal species examined thus far have evidence of infection by at least one herpesvirus; 130 species have been identified to date (International Committee on Taxonomy of Viruses; ICTV). Herpesvirus family members have a viral core comprised of a linear, double-stranded DNA genome, surrounded by an icosadeltahedral capsid approximately 100-110nm in diameter comprising 162 capsomers. The capsid is in turn encompassed by an amorphorous material called the tegument, which contains a number of viral proteins. A host derived trilaminar-membrane envelope, embedded with glycoproteins, encloses the entire structure (Epstein 1962).

1.6.1 Genome Configuration

While linear in the virion, the herpesvirus genome immediately forms a closed circular viral-episome upon entry to the host cell nucleus. The herpesvirus genome codes for approximately 70-120 open reading frames (ORFs), except in the case of cytomegaloviruses where the genomes can encode for up to 220 ORFs. There are 26 genes (known as the herpesvirus core genes) that are conserved across all members of the herpesvirus family. These are contained within seven core genome blocks (blocks A-G) that are found in various orders and orientations in each herpesvirus species. Within the seven blocks, which contain between 2-12 genes, the order, function and polarity of the genes is conserved. While varying between subfamilies, the block order (and therefore gene order) is usually conserved at the level of the genera (Roizman and Pellet 2001).

1.6.2 Herpesvirus Life Cycle

Herpesviruses are differentiated from other viruses by their ability to remain latent in their natural hosts. When the virus enters latency, its genome adopts an episomal

configuration, and the period is characterised by a lack of infectious viral progeny. The latent genome, however, can be reactivated to produce infectious progeny at any time although it is not clear what triggers reactivation.

When reactivation occurs, the virus enters into the lytic cycle; the production of infectious progeny eventually leading to the destruction of the infected cell. During the lytic cycle, herpesvirus replication is recogniseable by the synthesis of viral DNA and capsids in the host cell nucleus. Capsids are then enveloped as they bud through the nuclear membrane (Vlazny, Kwong et al. 1982), although the immature particle is reported to undergo de-envelopment as it travels through the outer nuclear membrane and re-envelopment in the cytoplasm as it egresses the cell (Skepper, Whiteley et al. 2001).

It is clear that most viral ORFs are important in the activation and maintenance of the lytic replication cycle, however, a subset of viral ORFs are expressed and function predominantly during the latent cycle. Herpesviruses also, like other large DNA enveloped viruses (*Poxviridae*), encode their own enzymes responsible for protein processing, DNA synthesis, and nucleic acid metabolism. Numbers (and types) of enzymes may vary between species, but their presence in the genome sets herpesviruses apart from many other virus families (Roizman and Pellet 2001).

1.6.3 Herpesviruses and Their Host Genomes

Many of the genes encoded by herpesviruses are enzymes involved in nucleotide repair/metabolism, DNA synthesis, and protein processing, however, a number of proteins encoded by herpesvirus genomes are involved in host-virus interactions. Herpesviruses often attempt to overcome complex immune responses by interfering or mimicking host response pathways to effect viral immune evasion (Ploegh 1998). A number of homologues shared between herpesviruses and their designated hosts have already been discovered (Damania and Desrosiers 2001; Hughes 2002) (McGeoch 2001) (Raftery, Muller et al. 2000) (Davis-Poynter, Degli-Esposti et al. 1999) (Davis-Poynter and Farrell 1996). These genes are thought to have been acquired by herpesviruses from their hosts over the course of herpesvirus evolution for the purpose of genome replication and immune evasion, the prevalence of homologues among herpesvirus subfamilies indicating the approximate evolutionary time of acquisition. It has been estimated that immune evasion genes compose >50% of large DNA virus genomes (herpesviruses and poxviruses) (Alcami and Koszinowski 2000). Once acquired, these genes perform a wide range of functions including interfering with cellular and humoral host immune responses, infected cell apoptosis, and interferon, chemokine and cytokine activity. This is in addition to the regulatory control over cellular transcription and translation machinery exacted by the virus.

1.6.4 Herpesvirus Subfamilies

The herpesvirus family is divided into three subfamilies based upon biological properties determined before comprehensive genomic information was available. These subfamilies are further divided into genera according to more subtle genome differences. The three subfamilies, *Alphaherpesvirinae, Betaherpesvirinae,* and *Gammaherpesvirinae*, are defined by their differences in cellular tropism (for lytic and latent cycles), replication cycle length and efficiency, and disease manifestation (Armstrong, Pereira et al. 1961; Asher, Heller et al. 1969; Arvanitakis, Geras-Raaka et al. 1997). There also exist a number of herpesviruses isolated from ectothermic (cold-blooded) animals that have been classified within the family due to their highly conserved virion structure similarities; however, within the family they remain unclassified as further genomic similarities have yet to be established.

1.6.4.1 *Alphaherpesvirinae*

The Alphaherpesviruses are characterised by their variable host range, and latent infection occurring primarily in the sensory ganglia. They have short reproductive (lytic) cycles in vitro, and thus spread rapidly in culture, efficiently destroying infected cells. Genera: simplex viruses, Varicellovirus, Marek's disease-like virus, and infectious laryngotracheitis-like virus.

1.6.4.2 *Betaherpesvirinae*

The betaherpesviruses have a more restricted host range and longer reproductive cycle *ex vivo*, leading to a slower progress in culture. Latency can be maintained in cells from secretory glands, kidneys, lymphoreticular cells, and other tissues. Infected cells frequently become enlarged and form cytomegalia. Genera: cytomegalovirus, muromegalovirus, and roseolovirus.

1.6.4.3 *Gammaherpesvirinae*

The gammaherpesviruses have variable reproductive cycles in lymphocytes and in some instances also replicate in epithelial, endothelial, and fibroblastic cells. Usually T or B lymphocyte specific, latent virus infection can frequently be found in lymphoid tissues. Gammaherpesviruses are associated with lymphoproliferative disease and tumors in immunocompromised hosts. Genera: lymphocryptovirus, and rhadinovirus.

1.6.5 Human Herpesviruses

To date, eight human herpesviruses have been discovered (Table 1.3), three alpha herpesviruses, three betaherpesviruses, and two gammaherpesviruses.

Table 1.3 Human herpesviruses and their Disease Associations

Sub-family	Genus	ICTV name	Medical Name	Disease
alpha	Simplex virus	HHV-1†	Herpes-simplex virus (HSV-1)	Oropharangeal herpes (cold sores) Genital herpes
	Simplex virus	HHV-2	Herpes-simplex virus (HSV-2)	Genital herpes
	Varicello-virus	HHV-3	Varicella-Zoster virus (VZV)	Varicella (chickenpox) Zoster (shingles)
beta	Cytomegalo-virus	HHV-5	Human cytomegalovirus (HCMV)	CMV-mononucleosis CMV disease
	Roseolo-virus	HHV-6A		Exanthem subitum (sixth disease)
	Roseolo-virus	HHV-6B		Exanthem subitum (sixth disease)
	Roseolo-virus	HHV-7		Exanthem subitum (sixth disease)
gamma	Lymphocrypto-virus	HHV-4	Epstein-Barr virus (EBV)	Infectious mononucleosis Nasopharyngeal carcinoma Burkitt's lymphoma Classical Hodgkin's lymphoma
	Rhadino-virus	HHV-8	Kaposi's sarcoma-associated herpesvirus (KSHV)	Kaposi's sarcoma Primary effusion lymphoma Multicentric Castleman's disease

†HHV: human herpesvirus

1.7 Aims

As sequencing projects provide increasing numbers of complete genomes it is important to take advantage of the available information and conduct searches for viral-host homology. In order to be able to efficiently administer and analyse the large amount of data these searches manipulate, it is necessary to have consistent annotation to allow cross-comparison between different species. Such annotation could then be used to conduct further analysis upon existing host-viral interaction data. The aims of this thesis are to: 1) find and catalogue host-herpesvirus homologues in an entire host genome; 2) provide consistent and concise annotation of an entire viral genome; and 3) to use such an annotation system to study host-pathogen interaction microarray data.

2.0 Identification of new herpesvirus gene homologues in the human genome

2.1 Introduction

Large DNA viruses, such as herpesviruses, may contain up to a few hundred open reading frames (ORFs) and among the proteins they encode it is possible to distinguish between those that have essential viral functions, such as genome replication and capsid assembly, and those that are involved in direct interaction with the host, effecting immune evasion, cell proliferation, and apoptosis control (Ploegh 1998). Many of the latter genes are likely to have been acquired from the host, to mimic or block the corresponding normal cellular functions (Moore, Boshoff et al. 1996; Alcami and Koszinowski 2000; McFadden and Murphy 2000). Identifying and understanding the functions of such 'acquired' viral proteins, should lead to a greater understanding of the co-ordinate range of host pathogen interactions and could lead to the development of therapeutic strategies to combat persistent herpesvirus infection.

Herpesviruses persist and replicate their genomes in the nucleus, and over the course of their evolution have acquired host genes (Chaston and Lidbury 2001), possibly by retrotransposition (Brunovskis and Kung 1995). Most of these acquired genes are located in regions outside the five gene blocks common to all herpesvirus genomes. Previous work has identified a set of 26 open reading frames that are conserved across all herpesviruses (McGeoch and Davison 1999; Alba, Das et al. 2001a). The remaining herpesvirus genes are either present in all members of a virus subfamily, a subset of viruses in a subfamily, or are unique to a particular virus. Many of these potentially important proteins, however, remain functionally ill-defined.

One approach for the identification of virus proteins that interfere with the host system is to search for homologous ORFs (xenologues) in the host genome. Until recently, the fraction of host genome sequence data available for analysis, and the quality of annotation of such data, has limited the identification of such homologues. The publication of the draft of the human genome and conceptual translated products (IHGSC 2001) enables us to conduct, for the first time, a comprehensive assessment of homologous proteins between a host vertebrate genome and viral ORFs.

Herpesviruses have co-evolved with their hosts, and many of their genes were acquired from common ancestors before speciation events led to a diversification in herpesvirus host range. Those such as DNA polymerase, and uracil DNA-glycosylase, are found in every member of the family, and were therefore acquired early in herpesvirus evolution. Others, such as the the chemokine receptors in HHV-5 (UL33, UL78), and MCMV (M33, M78), and vIL-10 in HHV-4 (BCRF-1), CeHV-8 (cercopithecine herpesvirus 8; UL111A), and EHV-2 (IL-10 gene) demonstrate acquisition after subfamily division and speciation have occurred, perhaps in separate acquisition events (Spriggs 1996; Dairaghi, Greaves et al. 1998; Alcami and Koszinowski 2000; Kotenko, Saccani et al. 2000; Lalani, Barrett et al. 2000; Tortorella, Gewurz et al. 2000) (Hughes 2002).

The prevalence of these acquired genes among a number of herpesvirus species, and their often identifiable sequence similarity using various sequence similarity methods (Montague and Hutchison 2000; Alba, Das et al. 2001a) indicates that many of them are viral orthologues of each other. It should, therefore, be possible to search one representative host genome, such as the human genome, to find evidence of xenology across all herpesviruses (including non-human herpesviruses) that resulted from a horizontal gene transfer between a common host ancestor and a common virus ancestor. This indicates that it would be reasonable to search the human genome not only with human herpesvirus ORFs, but also those of other host specificity. Given the size and complexity of the human genome compared to the relative simplicity of the virus, such a search may well yield new information about ORFs from different herpesviruses; likewise, it could be possible to aid annotation of the newly sequenced human genome using knowledge from other studied herpesvirus species where gene products have a defined function.

There are two methods particularly applicable to mass analysis of sequence databases. The first involves searching of individual protein sequences against a database using pairwise sequence comparison algorithms. Viral proteins, however, are subject to high mutation rates and that may cloud or mask true homology. A second, more sensitive approach is to search databases with amino acid sequence motifs that are conserved between related proteins. Motifs can be defined as regions of amino acid sequence that are more highly conserved than the rest of the protein due to functional constraints. An accurate representation of such motifs can be obtained by constructing Position Specific

Scoring Matrices (PSSMs) that store the frequency of occurrence of different amino acids that comprise the motif.

The most rigorous methodology, therefore, is to use both methods, for a number of reasons. First, as pairwise comparisons have been used to discover most, if not all, of the known sequence-based viral-host homologues, the method should act as a good positive control: the minimum number of homologues this method is expected to yield would be all currently recognised human herpesvirus-human sequence-based homologues. Pairwise comparisons, however, examine the similarity of two sequences to each other, therefore, the probability that a pairwise alignment has achieved a significant score by chance increases with the length of either protein and the size of the database(s) being searched. Thus, the results from this method are expected to include far more false positives than those yielded from the more rigid PSSM method. By using both methods, the validity and accuracy of each method can be compared, possibly highlighting one method as superior for large global homology searches with divergent viral protein sequences.

2.1.1 The BLAST Suite

A suite of programs based upon the original BLAST program can be used to align two sequences of nucleic or amino acid (BLASTn or BLASTp); create PSSMs (PSI-BLAST) by iteratively aligning a chosen sequence against a selected database of sequences, taking advantage of the more sensitive nature of multiple alignments; or compare a chosen sequence against a database of PSSMs built using PSI-BLAST (IMPALA). All programs are available at the NCBI for searching GenBank online, or for downloading and using against a user defined database.

2.1.1.1 BLAST

BLAST forms the basis for all other BLAST suite programs. BLAST (Basic Local Alignment Search Tool) is a local alignment tool based upon the Smith-Waterman Algorithm (Altschul and Koonin 1998). BLAST initially had the advantage of speed over other methods. This was largely due to the heuristic nature of BLAST, which resulted in a reduction of computational time for identifying 'high scoring' local alignments at the expense of the completeness of the search. Another advantage of

BLAST over its predecessors was that it allowed for the analysis of the program's performance, i.e. the alignments generated could be statistically ranked. It had not been previously possible to statistically verify the reliability of alignments produced by similar programs. BLAST can be used for both protein and DNA alignment and database searches. For these reasons, and the simplicity of use, BLAST is extensively used today.

The BLAST algorithm uses a measure of Smith-Waterman based local alignment: the Maximal Segment Pair (MSP) (Altschul, Gish et al. 1990). An MSP is defined as the highest scoring pair of identical length, ungapped segments chosen from two aligned sequences; in essence, the best local ungapped alignment. MSPs are detected by first identifying lengths of aligned sequence between two sequences, for example, the query sequence and one sequence from a database that score above a minimal threshold 'T' (known as 'words'). These words are then extended in either direction to see if they can qualify as MSPs (i.e. score above a second threshold score 'S'). The boundaries of the MSP are not fixed to allow the score to be maximised; thus, the MSPs can be of any length, the final length being determined by the necessity of a gap insertion at either end. Scores are derived for the MSP using similarity matrices; for example, default BLAST parameters use the PAM-120 similarity matrix for amino acid alignments.

BLAST returns pairwise alignments for all high scoring segment pairs (HSPs) that score above a threshold score S, from the query sequence search against the database. BLAST is, therefore, designed to not only look for the MSP of the search, but for any locally maximal sequence segment (i.e. sequence segments that score above the cutoff S that cannot be improved by lengthening or shortening both segments) (Altschul, Gish et al. 1990). This allows the program to identify multiple occurrences of similarity such as matches to multiple incidences of the same domain, or matches to the multiple different domains in the query sequence.

As with other similarity programs, the length of time it takes to calculate the MSP scores is directly proportional to the product of the length of the query sequence and the number of residues in the database, i.e. the more residues to be aligned, the longer the search will take.

2.1.1.2 Gapped-BLAST

The disadvantage of the original BLAST program was its inability to generate gapped alignments. Thus, Gapped-BLAST was developed to improve the accuracy of pairwise alignments generated by the program. Based upon the original BLAST algorithm, Gapped-BLAST uses a 'two-hit' method to generate gapped alignments. Before the extension of words occurs to find MSPs that score ≥ 'S', two words must be found within the Smith-Waterman 2D matrix that are on the same diagonal within a distance of 'A' from each other (i.e. not overlapping). In theory, the value of 'T' from the original BLAST should be lowered to increase sensitivity. Although this would not compromise run time – as the number of extensions that are made using gapped BLAST is lower than BLAST as only word pairs are extended and not all HSPs – this is not necessary as explained later. Once all word pairs are found, extension occurs in both directions to find all regions (words) of homology.

Unlike the original BLAST program, once a word with a score over 'T' is found, the entire matched sequence is searched for similarity to the query sequence and the entire sequence returned as one alignment; i.e., if global similarity exists it is reported as such, otherwise the two sequences are returned as a local alignment. This is in contrast to BLAST, which returned each MSP found as a separate alignment – if two proteins shared global alignment across 3 domains, three separate local alignments would be reported, not one global alignment. Therefore, if only one word pair that scores above T needs to be found per database sequence (as the remaining MSPs would be automatically searched for after identification of the first), the value of T need not be decreased, as mentioned earlier. If the completeness of the search is no longer compromised by a higher value of T, this results in a faster run time environment.

2.1.1.3 PSI-BLAST

BLAST takes as input a query sequence and an appropriate substitution matrix (such as PAM120). The program PSI-BLAST (Position Specific Iterated-BLAST) (Altschul, Madden et al. 1997) is an adapted version of BLAST that takes as input a position-specific score matrix. PSI-BLAST functions by creating a position-specific scoring matrix from an initial BLAST search of a database with a query sequence. This matrix

is used to more sensitively search the same database to find more distantly related sequences than would otherwise have been possible by a simple pairwise search.

The results from the first iteration (i) can be used as input for the second iteration (i+1) an infinite number of times, or until no further statistically significant sequences are found within the database. New significantly matched sequences to the given matrix are used to recalculate the PSSM which is then used in the next iterative round of database searching. More distantly related sequences can be identified by collecting sequences in a matrix after each iteration, because of the greater variation of residue conservation recorded in the matrix. Because PSI-BLAST uses gapped-BLAST to perform its initial pairwise searches of the database, PSI-BLAST also looks for local alignments within the database.

Matrices are built using each of the results that scored beneath a certain cut-off (0.01 E-value is default; see Statistics below) from an initial gapped-BLAST search of a database. The query sequence is used as a template and the two-dimensional matrix is the size of: the length of the query by the number of significant database sequences found. Any sequences that are identical to the query are removed and only one sequence is retained from groups of sequences that share more than 98% identity across their length.

In order to maintain a matrix that is always the same size as the query sequence, any column that requires the insertion of a gap character is ignored. Thus, each of the columns may contain different totals of residues. This is due to either the presence of gaps in the database sequences (or insertions in the query sequence), or because the algorithm is based upon local alignment, and will not necessarily find high scoring residue matches for the entire length of the query. Position-specific gap scores are not calculated for each PSSM, so the gap scores from the original gapped BLAST search are used for each iteration instead.

Once the matrix is constructed each sequence is weighted to avoid any bias towards a certain pattern due to the over-representation of similar sequences that might otherwise bias away from more divergent sequences that are not as numerously represented in the matrix. This is the only adjustment made to the matrix before the score matrix is calculated and distinguishes PSI-BLAST from other true multiple alignment algorithms.

For example, other methods may perform a number of further calculations, such as gap score determination, before calculating final scores.

One disadvantage this method has is that, due to its iterative nature, PSI-BLAST can amplify any errors that BLAST encounters, such as false positives due to query compositional bias (Altschul and Koonin 1998). Therefore, PSI-BLAST masks out these regions of the query sequence using the SEG program before performing a database search. PSI-BLAST was developed as an alternative to other more complicated multiple alignment tools, and aimed for speed and simplicity though automation of the entire process. By using multiple query sequences, a library of PSSMs can be built, and retained, for future querying.

2.1.1.4 IMPALA

Databases or collections of PSSMs can represent extensive sequence variation information for carefully grouped sets of proteins. These databases, when searched for similarity with a query sequence, allow statistical matching of protein group-specific conserved and divergent amino acid positions, rather than simple pairwise alignments using generalised substitution matrices.

The IMPALA (Schaffer, Wolf et al. 1999) algorithm can be used to search a database of PSSMs (of the type constructed by PSI-BLAST) with a single query sequence. In addition, IMPALA provides a more refined analysis of the matches, and by using the Smith-Waterman algorithm, guarantees an optimal alignment. Due to the nature of a PSSM library (which has non-similar proteins removed during its creation), an IMPALA query also takes less time to complete than either a PSI-BLAST or a BLAST search would against an entire non-redundant database.

The IMPALA program works by seeking an optimal alignment between the query sequence and each PSSM in the library using the Smith-Waterman alignment (extended for gap costs (Gotoh 1982)). If an alignment between a query sequence and a PSSM proves to have local alignment with a significant score, then any additional local alignments between the two are sought. This is done using the multiple match extension of the Smith-Waterman algorithm (Waterman and Eggert 1987) and searching in either direction along the sequences from the site of initial significance. If an alignment of the

query sequence to a PSSM scores beneath a user-defined threshold, then the alignment is reported. Each PSSM has a consensus sequence that is used to present the results of the search in a pairwise alignment similar to BLAST or PSI-BLAST.

2.1.2 Scoring Functions

The reliability of the results from any algorithm is dependent upon the probability of its significance. This is a measure of the possibility that a similar result could be obtained by running random simulations of the program. By calculating the probability of an alignment being a 'true' hit, the program can verify its accuracy. The calculation of an alignment score is dependent upon the residue composition and the size of both the query sequence and the database being searched. An increase in residue bias or size of database increases the possibility of random alignments being identified by the algorithm being scored.

2.1.2.1 BLAST Statistics

All BLAST programs display the results of a query as a list with each hit having statistical validation in the form of a score (measured in bits) and an E-value. Beneath the list of hits is the graphical representation of all the alignments in the lists. The Score (bits) is the score of the HSP normalised to account for the particular scoring system used. Normalisation occurs to allow the score to be statistically comparable to the scores from other alignments within the search. The E value looks at the probability of a match occurring by chance, thus E represents the number of HSPs with a score $\geq$ S that are expected to occur by chance in a database of a similar size to the one being searched. The E-value of an alignment is calculated as a function of the normalised score, the residue composition and lengths of the query sequence and search database. Each of the four BLAST algorithms listed here calculates the E value in a slightly different way.

The basic BLAST program uses the adjusted lengths of each sequence in the database, while using the actual residue composition of each query sequence, and an average residue composition of the database to calculate E. Gapped-BLAST calculates a random simulation using two theoretical proteins before the query is run. This random

simulation uses a standard residue composition that does not necessarily reflect the composition of either the query sequence or the database.

PSI-BLAST calculates two simulated comparisons: of a given PSSM against a query of average residue composition, and of two protein sequences of standard residue composition using a standard substitution matrix. The PSSM is then re-scored accordingly so that the two comparisons have similar scores, and the gap scores calculated for the pairwise sequence comparison can be conferred upon the PSSM. Using the method above, PSI-BLAST rescales a PSSM only once and assumes that the proteins in the database have an average residue composition.

IMPALA initially uses the same scaling method as PSI-BLAST by assuming the query has a standard residue composition; however, when a significant query-PSSM alignment is found the PSSM is rescaled based upon the actual residue composition of the query and the pairwise alignment between the two is recalculated. This makes its statistical analysis of produced alignments more reliable.

2.1.3 The Human Genome Project

The final version of the human genome (completed: 14th April, 2003) contains the sequence for 99% of its euchromatic regions; the entire sequence is highly accurate (99.999%), and highly continguous. The only regions that aren't currently sequenced are contained in less than 400 defined gaps that are found in areas of the genome that cannot be accurately sequenced with the technology available to date (NIH-Newsroom 2003).

Although substantially sequenced, the exact number of genes encoded by the human genome is still unknown; latest estimates by gene-prediction programs suggest as few as 24,500 genes are present (Pennisi 2003). There are two peptide translation projects publicly available that differ in their protein prediction methods: the Human Genome Resources (NCBI) and the Ensembl Project (EMBL-EBI/Wellcome Trust Sanger Institute). Ensembl bases its gene predictions on experimental evidence from Swiss-Prot (Boeckmann, Bairoch et al. 2003), SPTrEMBL (Boeckmann, Bairoch et al. 2003), RefSeq (Wheeler, Church et al. 2004), and cDNA entries from EMBL (Kulikova, Aldebert et al. 2004). Transcripts that map to existing human genes are named 'known

genes', those based on similarity to closely related species are termed 'novel genes'. The NCBI initially aligns reference sequences from RefSeq to the genome sequence. The heuristic gene prediction program Gnomon (NCBI) is then used to annotate those regions not transcribed by RefSeq alignments.

2.2 Methods

2.2.1 Initial data sets

All complete herpesvirus open reading frames (ORFs) from GenBank release 124 were available through the viral database VIDA (Alba, Lee et al. 2001). A total of 393 homologous multi-protein families (HPFs) and 494 singleton proteins were used in the analysis. This comprises all herpesvirus ORFs from VIDA including all eight human herpesviruses. A total of 4740 herpesvirus ORFs that include strain variant ORF sequences with <100% sequence identity, of which 4054 were non-redundant proteins, were used in this study.

The conceptual protein translations of two human genome databases were searched in this study: the collection of human genome gene products sequence build 23 at the National Centre for Biotechnology Information (NCBI), and the Ensembl Project beta release at the European Bioinformatics Institute (EBI). Both databases were downloaded by anonymous FTP and stored locally. The two databases were also concatenated into a single library and low-complexity protein segments masked using the SEG program with default parameters (Wootton and Federehen 1993).

2.2.2 Construction of motif position specific scoring matrices

Herpesvirus homologous protein families (HPFs) containing two or more proteins are defined by one or more amino acid motifs conserved across all members of the family (Alba, Lee et al. 2001). The large majority of HPFs are identified by a single motif (371 out of 393). However, there are 11 HPFs that contain 2 conserved motifs, 8 HPFs that contain 3 conserved motifs and 3 HPFs that share 4 motifs (see Figure 2.2). The motifs, in the form of multiple amino acid sequence alignments, were used to construct PSSMs using the programme PSI-BLAST (Altschul, Madden et al. 1997). Taking into account that some families contain more than one motif, we constructed a total of 429 PSSMs to be used in conjunction with the multiple alignment programme IMPALA.

2.2.3 Construction of a herpesvirus protein dataset at the 95% identity level

To complement the PSSM search method and to allow singleton herpesvirus proteins to be analysed, a dataset was derived for use by pairwise database search methods. A dataset of all individual herpesvirus proteins with $\leq$ 95% sequence identity was constructed for use with the pairwise alignment program BLAST. This dataset was constructed to save computational time by avoiding searching with nearly identical proteins, as proteins that share > 95% sequence identity would not provide different results. These representative proteins (termed an N95-rep) were selected by computing the global amino acid identity between all non-redundant herpesvirus proteins and grouping these proteins into subsets that shared at least 35%, 60%, 90% and 95% sequence identity using the programmes HOMOL and SEQCLUSTER respectively (Orengo, Michie et al. 1997) (Figure 2.1). The total protein population (black circle) is subdivided into groups of proteins that share 35% sequence identity, then into groups that share 60% (green circles), 90% (yellow circles), and finally 95% identity (red circles). An ORF was then selected at random from each 95% subset and used to perform pairwise sequence similarity searches of the human protein databases. For example, nine proteins from HPF 13 (protein kinase, HHV-1 UL13) were selected that represent (at the 95% sequence identity level) the thirty-three proteins in the HPF. In other words, the HPF13 could be subdivided into 9 subsets (of varying number) that have $\geq$95% sequence identity from which 9 representative sequences are selected, one from each subset. A total of 3986 N95-reps were derived.

2.2.4 Singleton Proteins

There exist a number of proteins in VIDA that do not share sufficient homology with any other herpesvirus protein to be placed in an HPF. These proteins are termed singleton proteins and exist as individual records within VIDA. In these cases the proteins cannot be defined by conserved motifs, and there are no multiple alignments from which to derive PSSMs. In this context, therefore, they are represented as N95-reps, although there are no other ORFs in their 95% sequence identity subgroup.

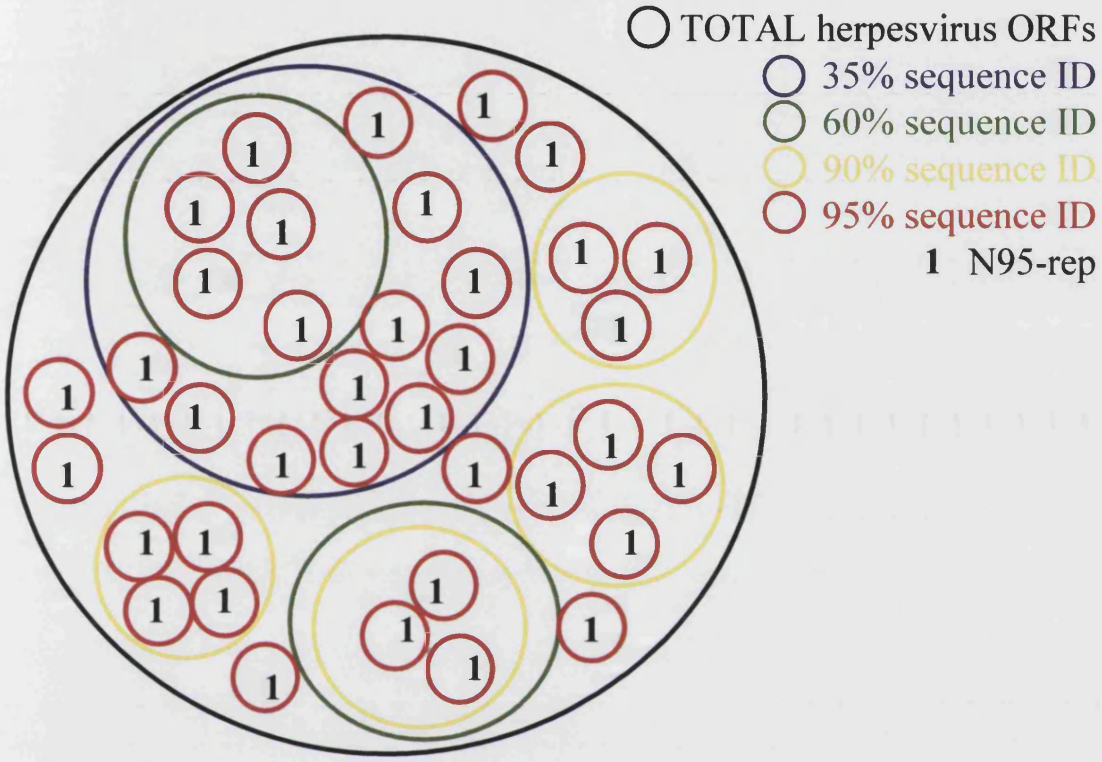


Figure 2.1 A Schematic Representation of N-reps. The black encompassing circle represents the total data set (4054 herpesvirus ORFs). Within the data set the proteins are divided into groups of 35% shared sequence identity (ID) (in blue), 60% (green), 90% (yellow), and 95% (red). One representative protein is then chosen at the 95% homology level to represent that group in future work.

2.2.5 Database searches and sequence analysis

The IMPALA program (Schaffer, Wolf et al. 1999) was used to perform searches using the two separate, unconcatenated human genome libraries against the 429 PSSMs derived from the motifs in VIDA. An E-value cut-off score of 0.01 and default parameters were used. The collection of N95-reps (which include all 494 singleton protein sequences) were searched using the programs BLASTP (Altschul, Gish et al. 1990) and PSI-BLAST (Altschul, Madden et al. 1997) against the concatenated Human Genome library, with default parameters and an E-value cut-off of 0.01. The total procedure is summarised in Figure 2.2.

All significant database hits were examined and curated manually based on detailed sequence alignments, conserved domain regions, functional annotation and reference to the literature. The manual inspection of putative homologues led to the removal of some of the initial hits, which appeared to be due to amino acid compositional bias (i.e. proteins such as collagen whose sequence is comprised of a high percentage of proline and glycine) rather than true homology. Where appropriate, additional proteins from different organisms were retrieved from GenBank for further multiple sequence alignment construction. These alignments were produced by the program MULTALIN (Corpet 1988) and, where necessary, manually edited using JALVIEW (http://www2.ebi.ac.uk/~michele/jalview/contents.html/) followed by visualisation using BOXSHADE (http://bioweb.pasteur.fr/seqanal/interfaces/boxshade.html/). Analysis of novel homologous families also included searching the domain database at the NCBI, which is linked to the Pfam (Bateman, Coin et al. 2004) and SMART (Letunic, Copley et al. 2004) domain databases.

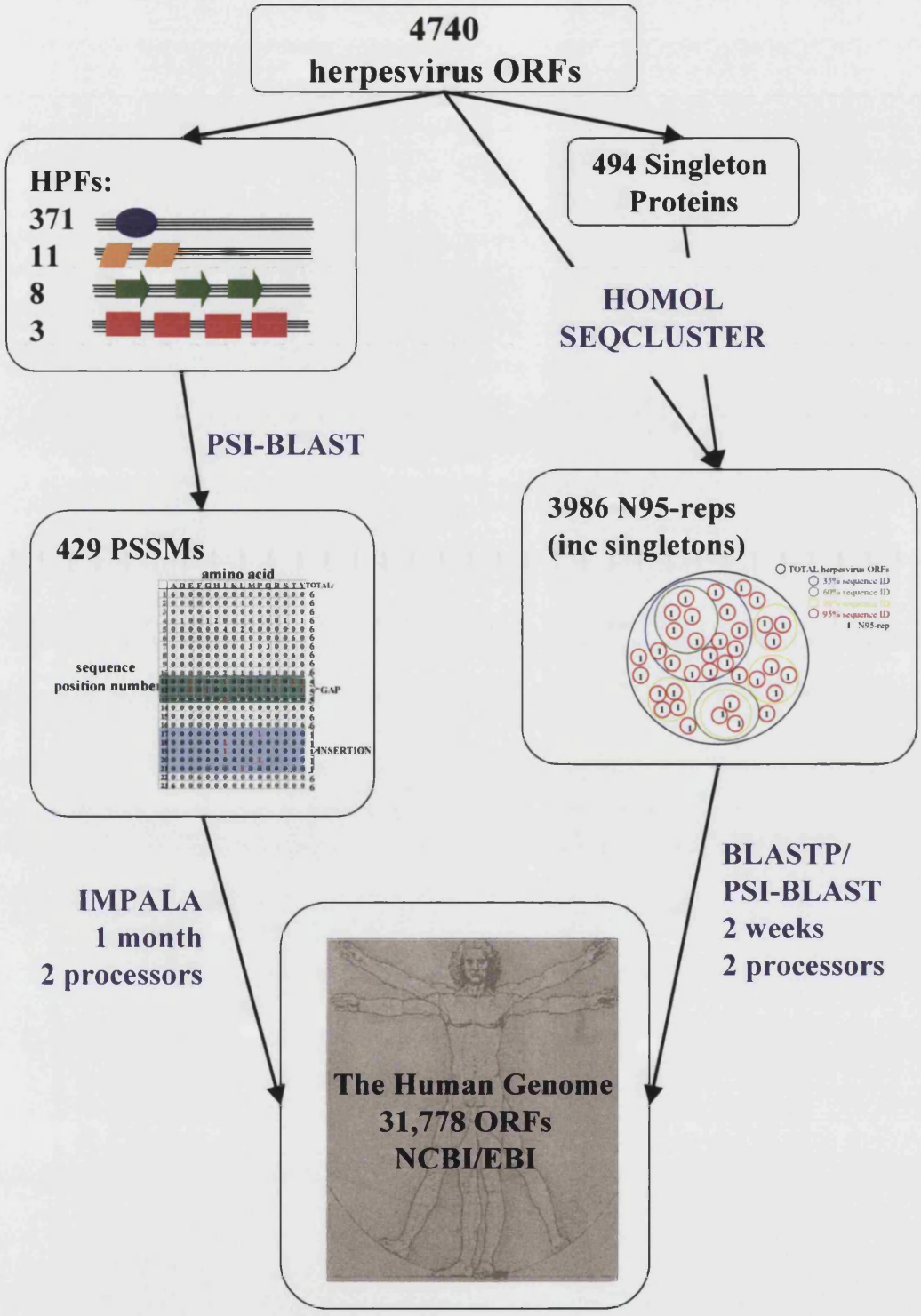


Figure 2.2 A Summary of the Human-Herpesvirus Homologue Search. The 4054 non-redundant herpesvirus open reading frames (ORFs) were organised into Homologous Protein Families (HPFs) defined by at least one shared sequence motif per family. Each motif formed the basis of a Position Specific Scoring Matrix (PSSM), which was then used to search the Human Genome Conceptual Protein Translation using IMPALA. All 4054 herpesvirus proteins were also compared sequentially using the programs HOMOL and SEQCLUSTER, which allowed for the selection of a representative protein (N95-reps) per cluster of proteins that shared at least 95% sequence identity. These N95-reps were used to search the Human Genome Conceptual Translation using the programs PSI-BLAST and BLASTP.

2.3 Results

2.3.1 Validating the Results

The raw data generated from both database search strategies were extensive and needed to be filtered for true 'hits' before further investigation to validate the matches was possible. A script NO_HIT_FILTER (C++) was written to extract the N95-reps that showed significant scores to human proteins from the N95-reps that failed to match any human protein and was used for BLASTP and PSI-BLAST results. The program BLASTP_PARSER (C++) was written to parse the results files from the BLASTP search by GI numbers, collating these with their HPF name and function, and the first match in the BLASTP results list. A suite of programmes known as FILTER (PERL/C++), was produced to filter out the human proteins that comprise each PSSM built by iterative rounds of PSI-BLAST (FILTER can also parse proteins from the viral PSSMs used when searching redundant databases). IMPALA results required no additional parsing before analysis was conducted. The new results files were then manually selected based upon e-value, with the e-value range of selected hits to investigate being 0.009 to e-149.

2.3.1.1 ENSEMBL Hits

The EMSEMBL human protein translation release used for this study had the disadvantage of coded annotation for each open reading frame (Figure 2.3). It was possible in most cases to make informed assumptions as to the nature of the ENSEMBL hit, i.e. the functions of proteins A and B (Figure 2.3) are easily deduced from the information provided by the NCBI human protein translation. Sequence A is most probably a ribosomal protein S6 kinase, and sequence B a phosphorylase kinase. Thus, it was possible to tentatively conclude that the viral search protein GI:3374481 was a protein kinase. Further research, however, would be required to verify this annotation, and to determine the exact kinase mechanism used by the viral protein. There were certain viral proteins that found matches only within the ENSEMBL human protein database. The functions of those matches with significant scores could not be ascertained from ENSEMBL's annotation, therefore, those matches were

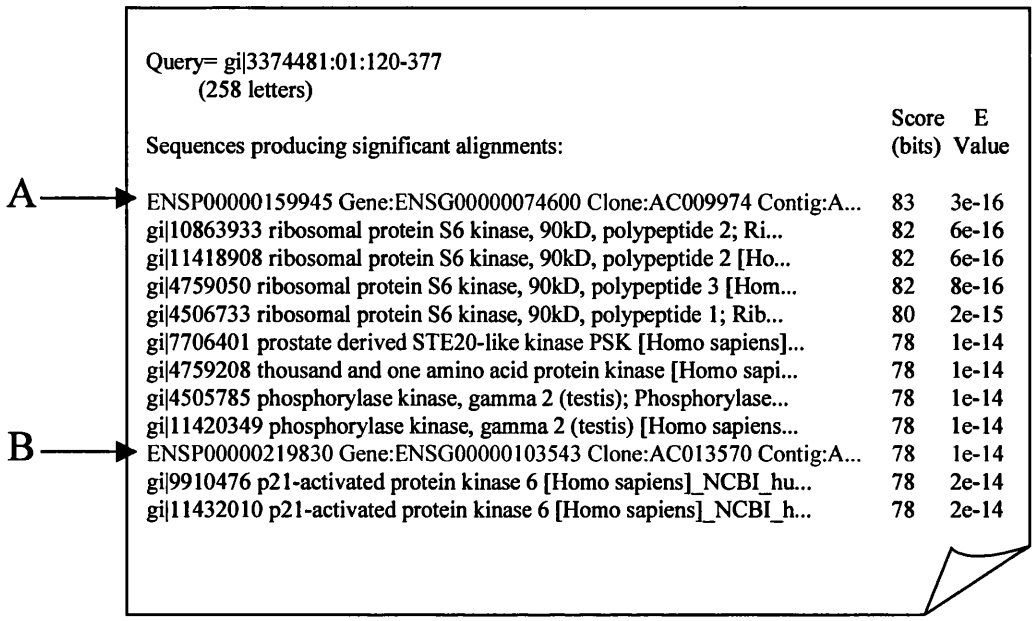


Figure 2.3 An example of ENSEMBL proteins from BLASTP output. The ENSEMBL human genome project proteins were not well annotated. Thus, in the case of A, it is necessary to query the sequence of ENSP00000159945 against GenBank or the NCBI human genome project in order to determine its name and function (if known); in the case of B, it can be hypothesized that ENSP00000219830 is a phosphorylase kinase, gamma2 (testis) given its identical score and e-value with NCBI proteins. Overall viral protein GI:3374481 could be putatively annotated as a protein kinase.

identified by searching non-redundant GenBank with the ENSEMBL sequence using BLASTP. All ENSEMBL proteins identified using this method matched records in GenBank with E-values of 0.

2.3.1.2 Initial Results

The initial analysis of VIDA herpesvirus protein and HPF statistics used for the xenologue searches are outlined in Table 2.1. A total of 14 complete human-herpesvirus genomes were included in this study (including strain variants), plus a number of human-herpesvirus ORFs that have been individually sequenced.

The 228 HPFs/singletons with known annotated viral functions were used as positive controls for the analysis. It is not expected, however, to find human homologues for all 228 for a number of reasons:

a) not all HPFs have functions that have correlating homologues (or even homoplasts) in their hosts' genomes, i.e. these are virus specific functions;
b) not all HPFs have functions that have been elucidated (or can be discerned) by sequence similarity i.e. these are experimentally derived;
c) not all HPFs contain open reading frames from human herpesviruses. This decreases the chances of a match being found between members of these HPFs and proteins within the human genome.
d) not all herpesviruses infect hosts which are closely related to humans. They therefore, may encode functions acquired from their respective hosts, or are designed to combat host specific responses not found in humans.

There are 622 HPFs/singletons of unknown function, therefore our hypothesis and previous data suggest that, should any of these significantly match proteins in the human genome, the HPF/singleton can be initially annotated with a function, or functional group, depending on the quality of annotation of the corresponding host-homologue. Likewise, should any viral proteins of known function significantly align to a host protein of unknown origin, functional annotation can again be conferred.

Table 2.1 Initial VIDA Statistics

VIDA Statistics	Total
total number of herpesvirus ORFs in database	4740*
of which non-redundant	4054
inclusive complete genomes	35
Alphaherpesviridae (of which human)	14(4)
Betaherpesviridae (of which human)	8(6)
Gammaherpesviridae (of which human)	12(4)
Unclassified	1
total genomes (partial/complete) in VIDA	52
total number of HPFs and Singletons	887
number of unknowns**	622/887 (70.1%)
number of singletons	494/887 (55.7%)

* including strain variants and redundancies

**HPF/singletons with unknown function

2.3.1.3 Search Statistics

Table 2.2 summarises the raw data statistics following parsing by BLASTP_PARSER, NO_HIT_FILTER, and FILTER gathered before each raw hit was analysed for accuracy. Due to the similarity in results format, PSI-BLAST data and statistics are considered synonymous with BLASTP results. Raw Data consists of each HPF and singleton that matched a human ORF below an E-value of 0.009 using either of the two search methods (BLASTP and IMPALA) regardless of its biological integrity (collectively termed 'raw hits'). A total of 135 raw hits out of a possible 887 (15.2%) matched at least one human ORF from the human genome. Raw hits were allocated to one of six result groupings (Table 2.3):

1. True hits: recognised (already documented) viral-cellular homologues;
2. Non-hits: matches due to random amino acid sequence devoid of functional motif or active site structure;
3. Repeat hits: matches biased by high individual peptide sequence repeats;
4. Domain Hits: matches that correctly identify functional domains (such as 7 transmembrane domains or immunoglobulin domains) within search proteins (i.e. identify regions of local similarity);
5. ID hits: matches of known viral proteins to unknown human proteins that allow for the possible annotation of the human ORF.
6. New hits: matches of known human proteins to unknown viral proteins that allow for the possible annotation of the viral ORF.

Table 2.2 Raw Data Search Statistics

Raw Data	Total	%
Number of HPFs/singletons (hereafter termed 'raw hits') that matched a human ORF in both searches* (out of total HPFs; see table 1)	135/887	15.2
Number of raw hits of known viral function	81/135	60
Number of raw hits of unknown viral function	56/135	41.5
Number of raw hits using BLASTP	89/135	65.9
Number of raw hits using IMPALA	92/135	68.1
Number of raw hits found by both methods (BLASTP and IMPALA)	43/135	31.9
Number of raw hits with human-herpesvirus proteins	70/135	51.9

*Both Searches: IMPALA of HPF PSSMs & BLASTP of N95-reps

Table 2.3 Breakdown of Raw Hits by Subfamily Composition

Subfamily composition	Raw Hits Total	True Hits* (% of total)	Non-Hits	Repeat Hits	Domain Hits	ID Hits*	New Hits*
γβα	10	5 (9.3)	3		2		
γβ	3	3 (5.6)					
γα	4	2 (3.7)	1	1			
α/unclassified	1	1 (1.9)					
α	30	6 (11.1)	11	8	3		2
β	31	8 (14.8)	5	14	1	1	2
γ	42	21 (38.9)	5	13	2	1	
unclassified	14	8 (14.8)	5	1			
TOTAL	135	54	30	37	8	2	4

*Significant hits can be further subdivided into True hits, ID hits, and New hits.

The quality of these raw hits is not yet determined by undertaking further analysis, therefore, this number does not have biological significance. The raw hit scores for the two different methods do not initially indicate a method-based bias. BLASTP found 89/135 (65.9%), while IMPALA found 92/135 (68.1%), however, only an approximate 50% of the raw hits from the two methods overlap, suggesting that each method is able to detect subtly different sequence similarities. These data also indicate that only by using both methods can the full extent of sequence similarity be assessed. This hypothesis, however, can only be verified by further analysis of the total 135 raw hits for biological significance based upon the following criteria: a) accuracy of the initial sequence alignment made by BLASTP or IMPALA, b) percentage of sequence similarity (including location/function of conserved residues), and c) any existing laboratory data available to confirm significance from the literature.

Hits were determined to be biologically significant by alignment of the human protein hit to the viral representative and analysis of conserved domains in both. Table 2.4 outlines the biologically significant hit (BSH) statistics gathered from the analysis of the raw hits. A total of 60 hits from the 135 raw hits were assigned biological significance (43.7%). These BSHs can be further divided into those hits involving viral proteins of known function (54/81 [66.7%]; defined as true hits, see Table 2.3), viral proteins of unknown function (4/56 [7.1%]; defined as new hits, see Table 2.3), and human proteins of unknown function (2/59 [3.4%]; defined as Identifying (ID) hits, see Table 2.3). Having analysed the significance of each hit, the comparison of the two methods, BLASTP and IMPALA, reveals that the combined use of both methods yields highest coverage of the searched databases, with the two methods yielding 86.7% (BLASTP) and 70% (IMPALA) of the BSH results, but only 34/60 BSHs (56.7%) were identified by both methods (Figure 2.4). This confirms that whilst both methods can identify a consistent set of BSHs, each method also identifies additional BSHs justifying their use in this study.

Over half of the BSHs identified match proteins from human-herpesviruses (33/60, 55%), therefore, 45% of the BSHs match proteins from herpesviruses of non-human hosts. This demonstrates the ability of this method to identify xenologues in all herpesviruses and not just human-herpesviruses. Xenologues, such as DNA polymerase, which was acquired early in herpesvirus evolution (evident from its ubiquity in the family), are easily identifiable in sequence similarity searches due to their highly

conserved functional motifs. It is, therefore, not necessary to search every herpesvirus host genome to confirm that DNA polymerase's presence in every herpesvirus can be attributed to a common ancestor. This method has demonstrated that xenologues from herpesviruses of varying host specificity can be found by searching one representative host genome.

VIDA classifies all HPFs and singletons into functional groups. Just over half of the BSHs found have host-interaction functions (31/60, 51.7%). The remainder of the hits (29/60) are functionally classified into the following groups: DNA replication, Nucleotide metabolism/repair, Enzymatic, Gene expression regulation, Glycoprotein, and Unknown. This is not surprising and reflects the observed tendency of herpesviruses to acquire host genes into their own genomes to manipulate their host's immune system (Alcami and Koszinowski 2000).

Table 2.3 breaks down raw hits according to their herpesvirus subfamily composition. HPFs can contain herpesvirus proteins from species in one or more of the three subfamilies, or from one of the unclassified herpesvirus species in VIDA. HPFs with members from multiple subfamilies demonstrate fewer BSHs 11/135 (8.1%) than HPFs/singletons from single subfamilies 49/135 (36.3%), indicating that many of the homologues that exist in herpesvirus genomes were acquired after subfamily divergence. Collectively, the Gammaherpesviruses exhibit the most homologues, followed by the betaherpesviruses, and finally the alphaherpesviruses.

2.3.1.4 IMPALA versus BLASTP

As shown, neither search method can be identified as outperforming the other for this type of search (Figure 2.4). As expected, most of the hits (33/59, 56.7%) were found by both methods. BLASTP appeared better at detecting non-human herpesvirus homologue hits, finding eleven non-human herpesvirus homologues, compared to IMPALA's two, although interestingly, both methods detected similar number of the human-herpesvirus homologue hits (BLASTP: 27/33; IMPALA 26/33). This observation most likely reflects the relative sensitivity and specificity of each method. By forcing the formation of PSSMs using a motif region from a predefined group of proteins (HPFs), the resulting matrices can only encode as much motif variation as exists in its the herpesvirus proteins contained in the HPF. This can possibly reduce the sensitivity of

the PSSMs to more distantly related proteins whose motifs are too dissimilar to the herpesvirus proteins in the HPF. In these cases, a pairwise alignment program, such as BLASTP, could detect more distant relatives due to the less stringent nature of pairwise comparison. In addition, BLASTP uses each full-length herpesvirus protein and may detect regions of sequence similarity outside the HPF motif; although the ability of BLASTP to detect sequences that are more distantly related is accompanied by the increased tendency of BLASTP over IMPALA to identify false positives. Overall, BLASTP (52/60, 86.7%) identified 10 more homologues than IMPALA (42/60, 70%).

2.3.2 Herpesvirus proteins with human homologues

Careful examination of putative host/virus sequence homologues showed that 39 herpesvirus HPFs and 20 singleton proteins had significant sequence similarity to human gene products (Table 2.5). One of the singleton proteins, HHV-5 US21, of unknown function, matched a human protein of known function, as well as two human proteins of unknown function, making it both a 'new hit' and an 'ID hit' bringing the total number of BSHs to 60. The 39 HPFs contain 483 proteins giving a total of 483+20 herpesvirus proteins that match a human sequence. This represented 12.4% of all herpesvirus ORFs in GenBank. Sequence similarity between herpesvirus and human proteins is clearly related to functional similarity, where function is known, based upon previous experimental data.

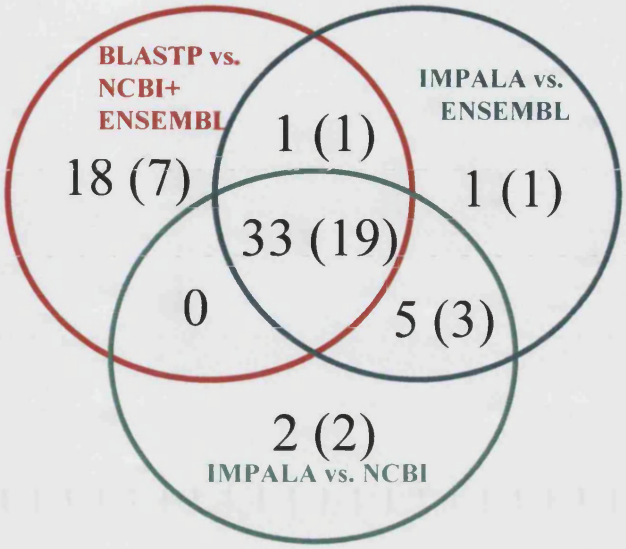


Figure 2.4 Distribution of BSHs Found per Method. Each circle represents one of the methods (IMPALA was run against each human genome library separately; whereas the two were concatenated for the BLASTP search). The numbers indicate the number of BSHs found by each method; the numbers in brackets is the number of BSHs that contain human-herpeviruses found by each method.

Table 2.4 Biologically Significant Hits Statistics

Biologically Significant Hits	Total	%
Number of biologically significant hits (BSHs; out of total number of raw hits; see table 2a)	60/135	43.7
Number of BSHs of known viral function (out of total number of raw hits of known function; see table 2a)	54/81	66.7
Number of BSHs of unknown viral function (out of total number of raw hits of unknown viral function; see table 2a)	4/56	7.1
Number of BSHs that matched human ORF(s) of unknown function (out of total number of BSHs)	2/60	3.3
Number of BSHs using BLASTP	52/60	86.7
Number of BSHs using IMPALA	42/60	70
Number of BSHs found by both methods (BLASTP and IMPALA)	34/60	56.7
Number of BSHs with human-herpesvirus proteins (out of total BSHs)	33/60	55
Number of BSHs with host-virus interaction functions (out of total BSHs)	31/60	51.7
Number of BSHs with host-virus interaction functions that contain human-herpsevirus proteins	19/33	57.6

Viral xenologues continue to participate in similar enzymatic functions (such as DNA and protein binding, kinase activity, chemoattraction) to their homologue counterparts, although their goal, pathway, and regulation are often not identical. The human herpesvirus 8 viral cyclin, for example, participates in the cell cycle as a cyclin D homologue but unlike the host cyclin D is not negatively regulated (Swanton, Mann et al. 1997). Likewise, cellular chemokine receptors invoke a second messenger system when bound to by a soluble chemokine (New and Wong 2003; Thomsen, Nansen et al. 2003). Viral chemokine receptor equivalents, however, usually serve one of two functions when present on the cell surface: either to deplete the surrounding area of host chemokines, or to redirect the cellular response by initiating an alternative secondary messaging system (Dairaghi, Greaves et al. 1998; Alcami and Koszinowski 2000; Lalani, Barrett et al. 2000). Therefore, the virus maintains accurate function of some enzymes but expresses them when the virus requires their function by removing transcriptional control and mutating regulatory mechanisms of some proteins.

Thus, it is not unusual that approximately 54% of the combined HPF and singleton hits corresponded to proteins classified in VIDA as being involved in 'host-virus interaction', primarily effecting immune and/or apoptosis controls. Of the remaining homologues, 32% have functions that can be generally termed 'metabolic': being 'enzymes' or involved in 'DNA replication' or 'nucleotide repair/metabolism'. These are notably more highly conserved, probably due to their earlier acquisition in herpesvirus evolution. Homologues to capsid constituents or capsid assembly proteins were not detected. This is not necessarily surprising as these are specific viral functions that have no obvious equivalents in cellular organisms.

In addition, approximately 42% of the HPFs and singletons that showed homology to human proteins did not contain any human herpesvirus ORF members. Many of these viral xenologues have been shown to share functional homology with their host equivalents through experimentation verifying that the method outlined in this chapter can be used to annotate gene products from non-human herpesviruses for which the relevant host genome sequence information is still unavailable.

Table 2.5 Herpesvirus-Human Xenologues. [1] HPF: homologous protein family number, S: singleton. HPF details can be visualised by searching VIDA by HPF number at http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA.html (Herpesviridae link). [2] a: alphaherpesvirus; b: betaherpesvirus; g: gammaherpesvirus; o: other; '–': indicates that only a subset of subfamily members are represented. For singletons, virus abbreviation and gene name is given: CCHV: channel catfish herpesvirus; SaHV-1: salmonid herpesvirus 1; RaHV-1: ranid herpesvirus 1; BoHV-4: bovine herpesvirus 4; HHV-8: human herpesvirus 8; EHV-2: equine herpesvirus 2; HVS-2: saimiriine herpesvirus 2; MeHV-1: meleagrid herpesvirus 1; HHV-5: human herpesvirus 5; HHV-4: human herpesvirus 4; RCMV: rat cytomegalovirus; AIHV-1: alcelaphine herpesvirus 1; GaHV-1: gallid herpesvirus 1. [3] GenBank: GenBank protein accession number (GI number). Only the human protein that hit with the lowest E-value is shown.

Table 2.5 Herpesvirus-Human Xenologues

Function Class	Viral Function (VIDA)	HPF[1]	Virus[2]	GenBank[3]	Human Function
DNA Replication	DNA polymerase	1	a,b,g	8393995	polymerase (DNA-directed), alpha
		293	O	15303524	polymerase (DNA directed), delta 1
	helicase/primase	16	A,b,g	5523990	DNA helicase
Nucleotide repair/metabolism	uracil-DNA glycosylase	8	a,b,g	6224979	uracil-DNA glycosylase
	ribonucleotide reduct. large sub.	24	a,b,g	4506749	ribonucleotide reductase M1 polypeptide
	ribonucleotide reduct. small sub.	33	a,g	4557845	ribonucleotide reductase M2 polypeptide
	thymidylate synthase	92	a-, g-	15297069	thymidylate synthetase
	dihydrofolate reductase	141	g-,b-	15297069	dihydrofolate reductase
	dUTP pyrophosphatase	S	CCHV ORF49	4503423	dUTP pyrophosphatase
		S	SaHV-1 ORF49	14756895	dUTP pyrophosphatase
	thymidine kinase	S	CCHV ORF5	11430716	thymidine kinase 2, mitochondrial
	DNA methyltransferase	S	RaHV-1 54_21	4503351	DNA (cytosine-5-)-methyltransferase 1
Enzyme	protein kinase	29	a,b,g-	14746991	serine/threonine-protein kinase PRP4
		40	a,o	4505649	protein kinase cdc2-related PCTAIRE-2
		214	o	9994197	G protein-coupled receptor kinase 7
		S	RaHV-1 54_2	14741902	CamKI-like protein kinase
	phospholipase-like protein	328	a-	5174497	endothelial cell-derived lipase precursor
	b-1,6-N-acetylglucosaminyltransf.	S	BoHV-4 ORF3-4	11431963	glucosaminyl (N-acetyl) transferase 3
	serine protease	S	CCHV ORF47	4505577	paired basic amino acid cleaving system 4
Gene Expression Regulation	transcriptional activator	74	a	5174653	ring finger protein (C3H2C3 type) 6
	bZIP domain	174	a-	4504809	jun B proto-oncogene
Glycoprotein	glycoprotein OX-2 like	242	g-	730246	OX-2 membrane glycoprotein precursor
	glycoprotein OX-2 like	194	b-	730246	OX-2 membrane glycoprotein precursor
Host-Virus Interaction	TNFR receptor homologue	13	HHV-5, UL144	4507571	tumor necrosis factor receptor, member 14
	virion-assoc. host shutoff factor	48	a	14738228	flap structure-specific endonuclease 1
	viral interferon regulatory factor	89	g-	4504723	interferon regulatory factor 2
		243	g-	13629153	interferon consensus seq. binding prot. 1
		S	HHV-8 vIRF-3	4505287	interferon regulatory factor 4
	G protein-coupled receptor	27	b,g-	13643500	chemokine (C-C motif) receptor 2
		248	b-	4758468	G protein-coupled receptor 50
		S	EHV-2, ORF74	4502639	chemokine (C-C motif) receptor 5
	complement binding protein	10	g-	10835143	decay accelerating factor for complement
	viral cyclin	102	g-	14767736	cyclin D1
	viral interleukin 10	140	g-	10835141	interleukin 10

	viral interleukin 6	315	g-	10834984	interleukin 6 (interferon, beta 2)
	viral interleukin 17	S	HVS-2 ORF13	4504651	interleukin 17
		161	g-	4502363	BCL2-antagonist/killer 1
	vBcl-2	259	g-	4557355	B-cell lymphoma protein 2 alpha
		S	MeHV-1 ORF1	11433559	BCL2-like 10 (apoptosis facilitator)
	MHC I downregulation	150	g-	8923613	hypothetical protein FLJ20668
	viral FLICE-inhibitory protein	256	g-	14731507	CASP8 and FADD-like apoptosis regulator
		S	EHV-2 E8	4505229	Fas (TNFRSF6)-associated via death domain
	CxC chemokine vIL8	531	a-	10834978	interleukin 8
	vMIP-I	225	g-	5174671	small inducible cytokine subf. A, member 26
	alpha chemokine	321	b-	4885589	small inducible cytokine subf. B, member 9B
	beta-chemokine	387	b-	5174671	small inducible cytokine subf. A, memer 26
	vMIP-III	S	HHV-8 K4.1	13628199	small inducible cytokine subf. A, member 17
	signal transduction protein	316	g-	12056967	Fc fragment of IgG, receptor for (CD16)
	CARD-like apoptotic protein	355	g-	4502379	CARD-like apoptotic protein
	U-PAR antigen CD59	352	g-	13639271	CD59 antigen p18-20
	natural killer (NK) cell decoy pr.	S	HHV-5 UL18	5031745	major histocompatibility complex, class I, E
	colony-stimulating factor 1	S	HHV-4 BARF1	4885123	CD80 antigen
	C-type lectin-like protein	S	RCMV lectin	4504883	killer cell lectin-like receptor subf. C, member 2
	semaphorin homolog	S	AIHV-1 A3	4504237	sema domain, Ig domain, GPI memb. anchor
	MHC1 heavy chain	S	RCMV R144	9665232	major histocompatibility complex, class I
Unknown	unknown	258	a-	4504883	killer cell lectin-like receptor subf. C, member 2
	Unknown	S	GaHV-1 UL45	4504883	killer cell lectin-like receptor subf. C, member 2
	Unknown	S	HHV-5 UL1	14764567	pregnancy specific beta-1-glycoprotein 5
	Unknown	S	HHV-5 US21	6912468	Lifeguard

2.3.2.1 Human Xenologous proteins of human herpesvirus proteins

This analysis provides an estimate of the number of xenologous proteins between the eight different human herpesviruses and the translated products from the human genome. A total of 33 different human herpesvirus proteins, including HPFs and singletons, showed significant homology to human proteins (Figure 2.5). This probably represents only a minimum estimate for a number of reasons. First, some proteins may still be functionally homologous but not show significant sequence similarity, because of the high rate of viral sequence divergence. These may be detectable when searching using different parameter set for IMPALA or BLASTP, or when using alternative alignment methods such as Hidden Markov Models (HMMs). Alternatively, searching by protein structure matching or experimental determination of viral gene function may elucidate other xenologues, such as the cellular survivin xenolog KSHV K7 (Wang, Sharp et al. 2002). Second, the total number of genes in the human genome is still uncertain. Thus, new genes that are homologous to viral ORFs may yet be discovered, and currently recognised putative genes may turn out to be pseudogenes. Nevertheless, this study provides the first detailed catalogue of all putative herpesvirus protein homologues in the host genome.

If the total number of human homologues relative to the total viral genome gene content in the different human herpesviruses is examined differences in subfamily homolog accumulation become apparent (Figure 2.6). The distribution of xenologs is 11-16% of the genes in human alphaherpesviruses, 11% in the betaherpesviruses, 10%-30% of the genes in gammaherpesviruses. As HHV-5 has a significantly larger genome than its fellow-subfamily members, the percentage indicates that more homologues were discovered for HHV-5 than the rest of the betaherpesvirus subfamily.

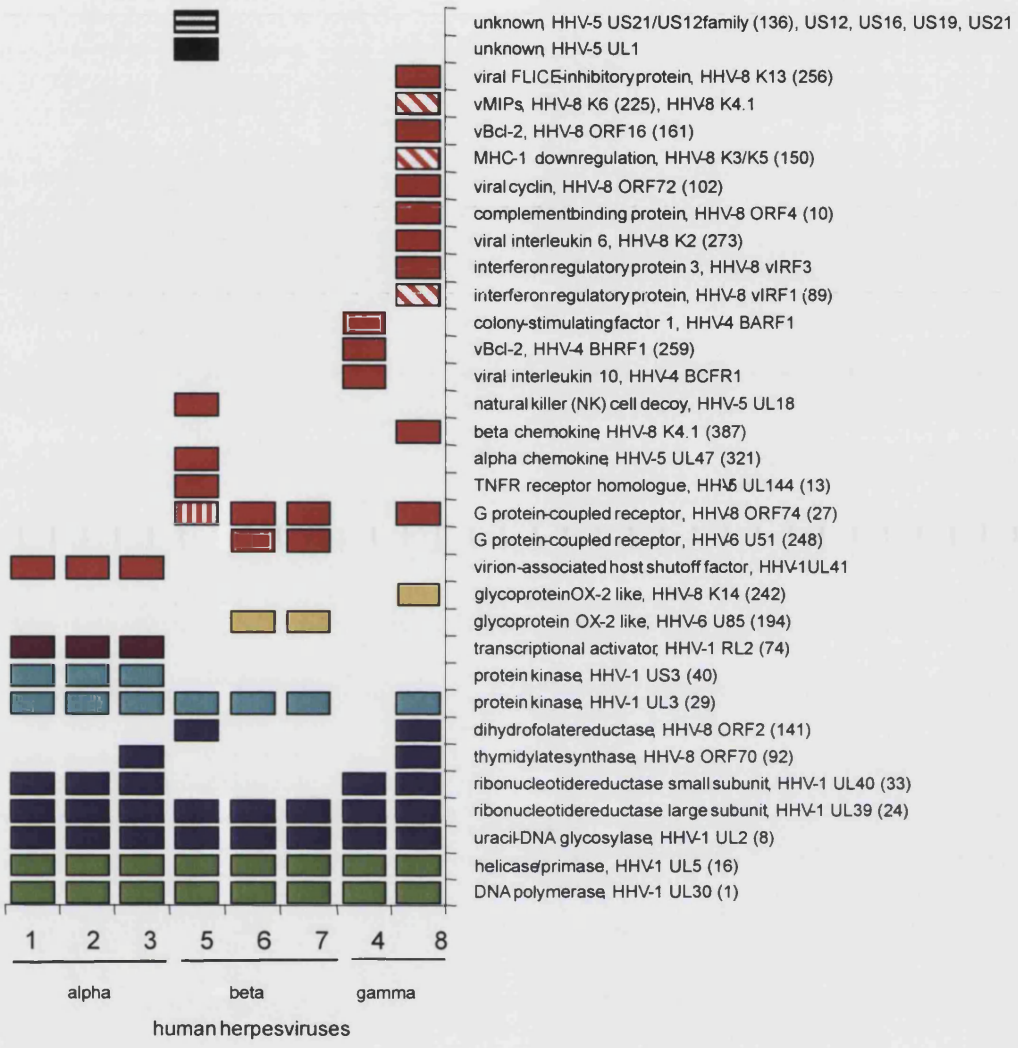


Figure 2.5 Human herpesvirus proteins with human homologues. Labels show the virus protein function, the name of a member of the HPF (homologous protein family) or singleton, and, for HPFs, the corresponding number in brackets. All the annotations and HPF numbers are taken from VIDA. The graph is color coded according to functional class: light green: DNA replication; dark blue: nucleotide repair/metabolism; light blue: enzyme; purple: gene expression regulation; yellow: glycoprotein; red: host-virus interaction; black: unknown. Diagonal lines within a box indicate 2 gene copies (per viral genome), vertical lines indicate 3 copies, and horizontal lines indicate 10 copies.

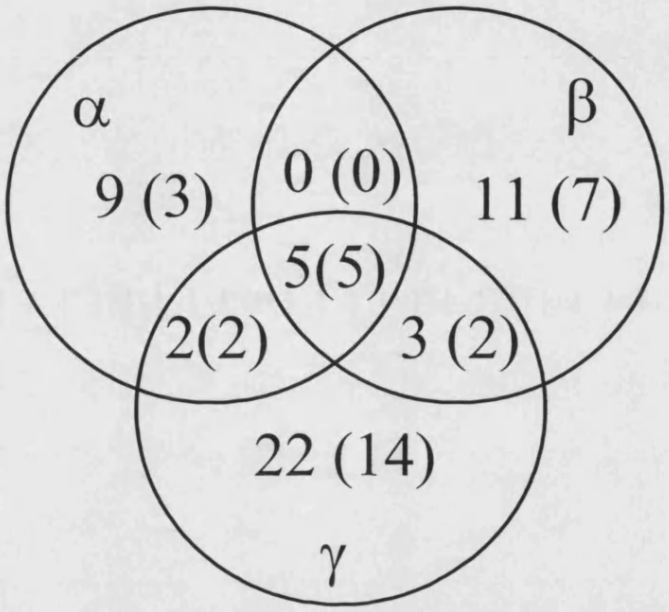


Figure 2.6 BSH distribution between the three Human Herpesvirus Subfamilies. The distribution of the 52 biologically significant hits that are classified in one of the three subfamilies. The number in brackets indicates the number of hits containing human herpesvirus proteins. A further eight BSHs are to unclassified herpesvirus proteins. One hit contained members from the alphaherpesvirus subfamily and unclassified proteins and is counted here as an alphaherpesvirus BSH.

The number of homologues discovered differed between the three subfamilies, directly reflecting the differences in host-virus interaction between members of each subfamily. The alphaherpesviruses have the smallest number of homologues. They exhibit a limited cellular tropism remaining latent in neurones, which exist in the immune privileged CNS and exhibit less anti-immune function. Thus, the alphaherpesviruses appear to require fewer proteins to inhibit apoptosis and cellular defense mechanisms. However, they do require core viral replication proteins such as transcriptional activator (HHV-1 RL1), thymidylate synthase (HHV-3 gene 13), and an additional protein kinase (HHV-1 US3), as the neuron no longer actively replicates its genome.

The beta- and gammaherpesviruses, on the other hand, have a larger number of homologues. HHV-5's cellular tropism, in contrast to the alphaherpesviruses, is much more varied. HHV-5 has been isolated from numerous tissues including liver, brain, thyroid, myeloid cells and leukocytes (Pass 2001). HHV-8 has been associated with a number of diseases including Kaposi's sarcoma, multicentric Castleman's disease, and primary effusion lymphoma (PEL) (Moore and Chang 2001). Both HHV-5 and HHV-8 infect cells of the immune system, and are more frequently targeted by host responses, thereby explaining their necessity for a larger number of proteins that can interact with their host environment. It is interesting to note that while HHV-4, HHV-6, and HHV-7 also infect cells of the immune system, they have fewer xenologues than HHV-5 and HHV-8 (Figure 2.5).

The five virus-host xenologues shared between the three human herpesvirus subfamilies are those found in most herpesviruses. Four of these are known to be present in all human herpesviruses, namely: DNA dependent DNA polymerase, helicase/primase, uracil-DNA glycosylase and ribonucleotide reductase large subunit, and these were all correctly identified by our methods. An additional protein family, protein kinase HHV-1 UL13, is present in all human herpesvirus except in HHV-4, suggesting it was lost from this genome during HHV-4 evolution.

It is known that the gammaherpesviruses share a common evolutionary branch with the betaherpesviruses and that the alphaherpesviruses form a separate lineage (McGeoch and Davison 1999; Alba, Das et al. 2001a). It is interesting, therefore, that one of the human homologues, ribonucleotide reductase small subunit, is found in the alpha- and gammaherpesviruses, but not in the betaherpesviruses. This could be due to a loss in the

latter lineage after the divergence of the three subfamilies. However, while the small subunit is encoded next to the large subunit in both HHV-1 and HHV-8 (both encoded at the end of genome block A), their orientations and positions within the genome blocks are varied, indicating that an independent acquisition event is also a strong possibility.

The HPFs have been used previously to construct phylogenetic trees of herpesvirus lineage (Alba, Das et al. 2001a). In this work there are two virus-human xenologues of particular interest as they appear in disparate positions in the herpesvirus evolutionary tree: thymidylate synthase in HHV-3 (varicella zoster virus) and in HHV-8 (Kaposi's sarcoma associated herpesvirus) and, dihydrofolate reductase in HHV-5 (human cytomegalovirus) and HHV-8. These could be explained by either independent acquisition of these genes from the host genome by each virus, or multiple gene loss events in different herpesvirus lineages. The second is a plausible explanation for genes such as protein kinase (HHV-1, UL13) where the gene is missing from only one of the eight human herpesviruses, but not in this situation where the gene appears at almost random positions in two disparate genomes. A third possibility is horizontal transfer between virus genomes, which, in the case of dihydrofolate reductase, could reasonably have occurred between HHV-5 and HHV-8 during co-infection as HHV-5 has been shown to enhance HHV-8 lytic infection in endothelial and keratinocytes (Vieira, O'Hearn et al. 2001). This theory is less likely, however, for HHV-3 and HHV-8 (in the case of thymidylate synthase) whose cellular tropisms differ quite drastically.

Sequence similarity alone revealed a minimum estimate of human homologues in different human herpesvirus genomes to be about 9-16% of virus genes, with the exception of human herpesvirus 8, which is approximately 30% of viral genes. The reason for a higher percentage of homologues in this virus, and in gammaherpesviruses in general, is unclear. Most of the herpesvirus/human homologues identified correspond to proteins involved in immune modulation and apoptotic control. These proteins are normally specific to one or a few viruses and they often show a complex distribution across the herpesvirus phylogeny tree indicating a lack of evolutionary pattern of acquisition. They are, therefore, likely to contribute to the virus's adaptation to different hosts or different cellular tropisms on an 'as and when needed' acquisition basis. This is in contrast to a more stable group of homologues, composed of proteins involved in

DNA replication and nucleotide metabolism, components of the well-conserved virus (and host) DNA genome replication machinery.

For proteins with viral structural functions, such as capsid constituents and capsid assembly proteins, which make a large proportion of herpesvirus genome coding capacity (20% of the genes of HHV-1), no resemblance to any human protein could be found. This is perhaps not surprising, as these have 'viral-only' functions. Recently, however, another method of formulating functional hypotheses of viral proteins, *in silico* protein structure prediction using threading techniques, has been applied to herpesvirus proteins. This was performed for all proteins of HHV-5, yielding complete structural identifications for 36 viral proteins, only eight of which were previously known (Novotny, Rigoutsos et al. 2001). These included some HHV-5 structural proteins indicating a possibility that viral derivations from the host may extend much further than currently estimated, although, the relationship between structural homology and functional similarity indicates that function cannot always be conferred upon structural homologues (Todd, Orengo et al. 2001).

2.3.3 Identification of new virus-human homologues

Of special interest for this study was the identification of human homologues for herpesvirus protein families and singletons of unknown function. The new homologues may provide putative functional annotations for several herpesvirus and/or human proteins. New herpesvirus/human protein homologues were found for the US12 (Unique Short) human cytomegalovirus protein family, the UL1 (Unique Long) human cytomegalovirus protein, the gallid/meleagrid herpesvirus UL45 protein family and, the K3/K5 human herpesvirus 8 family (Table 2.5).

2.3.3.1 HHV-5 US21

HHV-5 US21 is a distant member of a larger HHV-5 protein family, the US12 protein family, encompassing gene products US12 to US21 (Chee, Satchwell et al. 1990a). US21 showed significant overall sequence similarity to three human proteins: lifeguard, CGI-119 and PP1201. Other members of the US12 protein family, including an HPF which groups 6 of them in VIDA, did not initially hit any human proteins but multiple sequence alignments revealed the extent of amino acid similarity between all these

proteins (Figure 2.7). All of the members in the US12 family shown in Figure 2.7 have been shown to be non-essential for *in vitro* HCMV AD169 replication (Yu, Silva et al. 2003), and for *in vitro* HCMV Towne strain replication in fibroblasts (Dunn, Chou et al. 2003), except for US13 deletion mutants which were observed to cause moderate growth defects by Dunn *et al.* The herpesvirus and human proteins also contain a putative seven transmembrane domain, UPF0005, from the Pfam database. Pfam is a secondary databases that constructs protein families using Hidden Markov Models (HMMs) (Bateman, Birney et al. 2000).

Lifeguard is the human homologue of the rat protein neuromembrane protein 35 (Schweitzer, Taylor et al. 1998), proposed to protect against Fas-mediated apoptosis (Somia, Schmitt et al. 1999) without interfering with Fas associated death domain (FADD) binding to fatty acid synthase (FAS), or the tumour necrosis factor α (TNFα) apoptotic signal; therefore, the related HHV-5 proteins may also have an anti-apoptotic role. Viral-FLIPs (FLICE inhibitory protein) that interfere with Fas-mediated caspase-8 (FLICE) activated apoptosis have already been described in gammaherpesviruses (Belanger, Gravel et al. 2001), and the UL36 gene in HCMV has been designated a viral inhibitor of caspase-8-induced apoptosis (vICA) (Skaletskaya, Bartle et al. 2001), although the two (viral-FLICE and vICA) demonstrate little sequence similarity. From our analysis HHV-5 potentially encodes a number of anti-Fas apoptosis homologues, distinct from vICA, and the gammaherpesvirus FLIP homologues. Interestingly, in the cowpox virus, a member of the poxviridae family, a gene termed SR1, of unknown function but similar to the CGI-119 protein, was also identified (Shchelkunov, Safronov et al. 1998).

This hit is of dual interest as it allows the oportunity to functionally annotate not only viral ORFs using knowledge gathered from their human homologues, but also newly discovered human proteins based upon search results. This 'backward' annotation from 'usurper' to 'usurped' has occurred before, with the initial discovery of viral FLICE inhibitory protein (vFLIP), present in several gammaherpesviruses including HHV-8. vFLIP was initially identified using a flexible motif search profile (Bucher, Karplus et al. 1996) constructed from death-effector domains from human and murine FADD, FLICE and Mch4 (Thome, Schneider et al. 1997). This was followed by the subsequent identification of cellular equivalents using a similar bioinformatics method, based upon

a profile constructed from six known members of the vFLIP family (Irmler, Thome et al. 1997). Both identifications were confirmed with experimental evidence.

2.3.3.2 HHV-5 UL1

Homology was found between the HHV-5 UL1 gene product, a member of the RL11 family, and the CEA/PSG human protein family. The region of sequence similarity covers about two thirds of the UL1 protein and the N-terminal region of PSG and CEA subgroup proteins (Figure 2.8). UL1 was also shown to be non-essential for *in vitro* HCMV AD169 replication in fibroblasts, along with the other members of the RL11 family (Yu, Silva et al. 2003); Dunn *et al* also found the majority of the RL11 family to be non-essential for *in vitro* replication of HCMV Towne strain in fibroblasts, apart from UL11 (deletion mutant caused moderately defective replication), UL9 (enhanced replication), and UL1, for which no deletion mutant was tested.

HHV-5 UL1 showed particular similarity to the pregnancy-specific glycoprotein 5 (PSβG-5) and other members of the human carcinoembryonic antigen (CEA) protein family. The CEA (carcinoembryonic antigen) family is a member of the immuglobulin superfamily and contains three subgroups: the CEA subfamily, the pregnancy-specific glycoprotein (PSG) subfamily, and a remaining subfamily composed of six proteins (Teglund, Olsen et al. 1994). Known functions for the CEA family include involvement in cell adhesion, signal transduction, and possible innate immunity (Hammarstrom 1999). They are also utilised as clinical tumour markers since their discovery in tumour tissue in 1965 by Gold and Freedman (Maxwell 1999). Analysis of the CEA and PSG subfamilies indicates that the CEA subgroup are cell surface membrane proteins, while the PSG subgroup members are secreted from the cell.

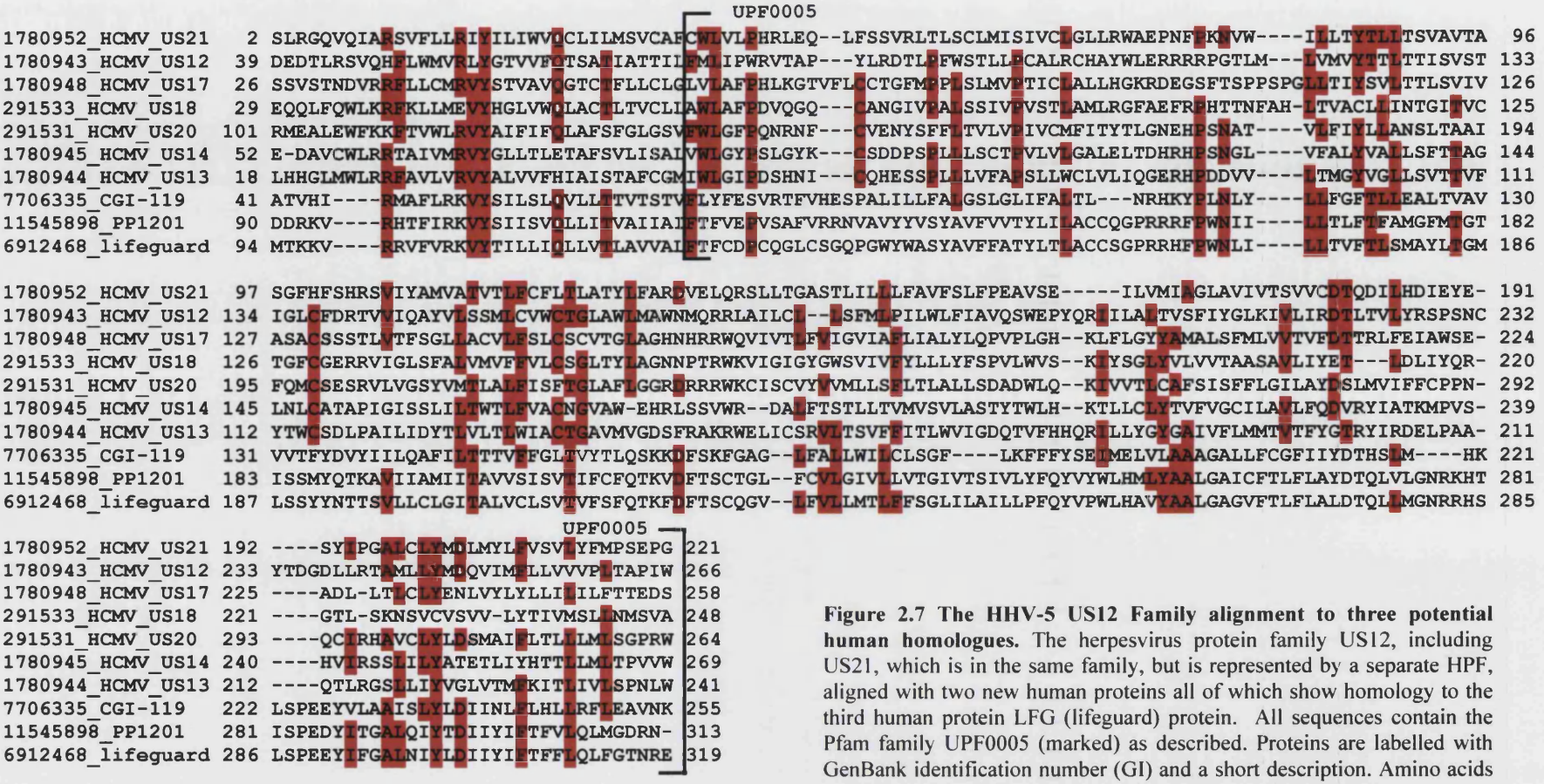


Figure 2.7 The HHV-5 US12 Family alignment to three potential human homologues. The herpesvirus protein family US12, including US21, which is in the same family, but is represented by a separate HPF, aligned with two new human proteins all of which show homology to the third human protein LFG (lifeguard) protein. All sequences contain the Pfam family UPF0005 (marked) as described. Proteins are labelled with GenBank identification number (GI) and a short description. Amino acids shaded red share identity across 50% or more of the alignment.

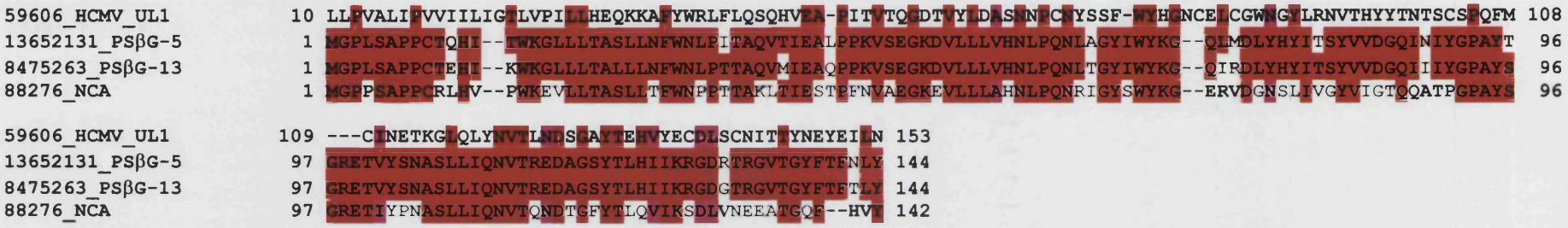


Figure 2.8 Alignment of HHV-5 UL1 to Members of the CEA Family. The alignment shows HHV-5 UL1, two PSG proteins (PSβG 5 and 13) and one member of the CEA subfamily (NCA, non-reacting antigen). Proteins are labelled with GenBank identification number (GI) and a short description. Amino acids shaded red share identity across 50% or more of the alignment; amino acids shaded pink share alternative 50% identity.

The 11 members of the PSG subgroup still have no known function, their expression predominantly occurring in the syncytiotrophoblast during pregnancy (Zhou, Baranov et al. 1997) possibly regulating immune system responses, although cDNA clones have been isolated from other tissues including fetal liver, salivary glands, testis, and myeloid cells (Hammarstrom 1999). HHV-5 infection, which is usually benign in immunocompetent individuals, can have catastrophic consequences during pregnancy (Fisher, Genbacev et al. 2000). Infection of the placenta has a 30 to 40% risk of intrauterine virus transmission to the foetus. Similarity of UL1 to pregnancy-specific glycoproteins (PSG) could subsequently be related to the pathology of HHV-5 during pregnancy, or to general immune modulation in the host, although there is a marked lack of HHV-5 infection of the syncytiotrophblast, the primary PSG expressor (Wagener and Ergun 2000).

2.3.3.3 GaHV-1 UL45

The protein family represented by UL45 in gallid (includes Marek's disease herpesvirus) and meleagrid herpesviruses shows homology to human C-type (calcium-dependent) lectin domain containing natural killer (NK) cell receptor proteins. Two other herpesvirus proteins, from rat cytomegalovirus (RCMV) and from a different gallid herpesvirus strain (GenBank accession Y14300), also demonstrate significant sequence similarity to C-type lectin domain containing NK cell receptors (Figure 2.9). The presence of C-type lectin domain in the RCMV protein was recently reported (Voigt, Sandford et al. 2001); now clearly this potential functionality extends to homologues in avian herpesviruses. Interestingly, proteins with C-type lectin-like domains are also found in the poxviruses (Bugert and Darai 2000).

NK cell receptors interact with HLA class I antigens and facilitate triggering or inhibition of natural killer cell-mediated cytotoxicity (Biassoni, Cantoni et al. 2001). C-type lectins contain a carbohydrate recognition domain (CRD), which includes four conserved cysteine residues forming two di-sulphide bonds and is responsible for carbohydrate ligand recognition activity in cell surface lectins (Day 1994; Cebo, Vergoten et al. 2002). These conserved cysteines are also present in the herpesvirus C-type lectin-like homologues.

The accuracy of this hit demonstrates that this method can be used to identify functional homologues between disparate virus-host pairings. In this case, none of the viral genes

come from human herpesviruses. The functional homology, however, has been demonstrated experimentally, in the case of RCMV, thereby increasing the probable accuracy of these host and viral xenologues findings. Clearly the presence of homology between the human genome and avian and rat herpesviruses suggests that either a) there also exist C-type lectins in human herpesviruses that have yet to be identified (perhaps by other experimental means that do not rely upon sequence similarity); b) human herpesvirus equivalents existed but have since been lost through evolution; or c) that the avian and rat herpesviruses independently acquired similar C-type lectin proteins from their respective hosts at some point during their co-evolution. When the completed host genomes for these viruses become available it will be possible to try and find, through similar methods, their direct homologues.

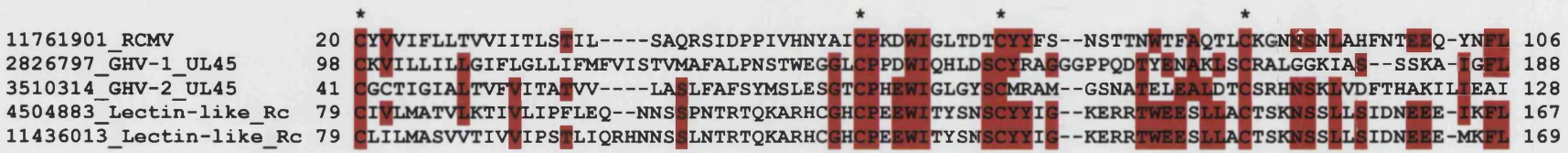


Figure 2.9 Alignment of GaHV-1/2 UL45 with RCMV, human and GaHV-2 equivalents. A representative from each of the herpesvirus protein families found to contain C-type lectin domains and two natural killer receptors (NKG2-A). The four conserved cysteines, important for di-sulphide bond formation in the CRD, are indicated. Proteins are labelled with GenBank identification number (GI) and a short description. Amino acids shaded red share identity across 50% or more of the alignment.

2.3.3.4 HHV-8 K3/K5

RING finger motifs bind zinc, but are distinct from other zinc-binding motifs, and are found in a wide variety of proteins in species ranging from the piroplasmic protozoa *Babesia microti*, and *Arabidopsis thaliana*, to viruses, yeast, and humans (Saurin, Borden et al. 1996). Proteins that contain RING finger motifs range in function from transcription factors (ICP0, HSV1), to photomorphogenesis (COP1, *Arabidopsis thaliana*). It is also interesting to note that a number of proto-oncogenes and oncogenes also contain RING finger motifs such as BRCA1 (breast cancer gene 1), human PML (cause of acute promyelocytic leukemia), and the human proto-oncogene CBL (Interpro family IPR001841). RING fingers are referred to as C3HC4 (3 cysteines, 1 histidine, 4 cysteines) motifs (Figure 2.10a), in contrast to the related C4HC3 (4 cysteines, 1 histidine, 3 cysteines) (Figure 2.10b) of the PHD/LAP finger motif.

PHD/LAP fingers are RING finger variants found in similar ranges of species and function. A highly conserved PHD/LAP finger has been identified in the proteins K5 and K3 from HHV8, IE1 in BHV-4 (bovine herpesvirus), and ORF12 in MHV-68 (murine gamma herpesvirus). These proteins are grouped together in VIDA as HPF 150. An additional gene, ORF 12 in saimiriine herpesvirus 2 (HVS-2), a singleton in VIDA that also contains the PHD/LAP finger motif, should also be considered a member of the family (Nicholas, Ruvolo et al. 1997). Figure 2.10c demonstrates the spatial difference between PHD/LAP fingers and the K5/K3 finger, which led Nicholas et al to classify the herpesvirus finger as a subclass of PHD/LAP finger known as the BKS (BHV-4, KSHV, and swinepox) subfamily (Nicholas, Ruvolo et al. 1997).

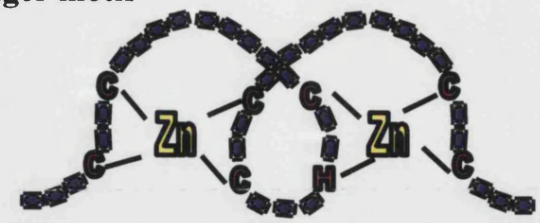
While this BKS motif has been found in a number of different species, most of these peptides have no known function; the mechanisms used by RING/PHD/LAP/BKS fingers during protein-protein interaction are also as yet unknown. The only function associated with this motif is in yeast. The ssm4 coiled-coil protein is believed to contain a microtubule-binding motif at its N-terminus (Yamashita, Watanabe et al. 1997), which is essential for its association with microtubules during meiotic division.

We identified six unannotated human proteins, including three identified by pairwise searches (Jenner and Boshoff 2002), that contain this highly conserved BKS finger motif (Figure 2.11). In four of the human peptides, the motif can be found at the N-

terminus, as in its viral counterparts. In the fifth human protein (gi:7243179; KIAA1399) the motif is in the middle of the peptide, and in the sixth (gi:12383066; similar to axotrophin) the motif is embedded in the C-terminus, as it is in murine axotrophin (not shown; gi:10181210).

Members of this family have been demonstrated to downregulate cell surface molecules. K3, K5 (HHV-8) and ORF 12 (MHV68) downregulate major histocompatability complex (MHC) class I proteins by ubiquitination to facilitate endocytosis (K3: HLA-A, -B, -C, -E; K5: HLA-A, -B) (Lorenzo, Jung et al. 2002; Means, Ishido et al. 2002), or by binding to the proteins in the endoplasmic reticulum (ORF12: H-2D) (Boname and Stevenson 2001), before targeting the internalised proteins for degradation. In addition to HLA-A and -B, K5 also reduces the levels of intercellular adhesion molecule 1 (ICAM1) and the costimulatory molecule B7-2 (Ishido, Wang et al. 2000; Coscoy, Sanchez et al. 2001). This is one of the first positive indications of a function for the BKS motif, coupled with the fact that the BKS motif has not been found previously in mammals, indicates a possible function for the host protein.

a. RING finger motif



C-2x-C-(9-39)x-C-(1-3)x-H-(2-3)x-C-2x-C-(4-48)x-C-2x-C

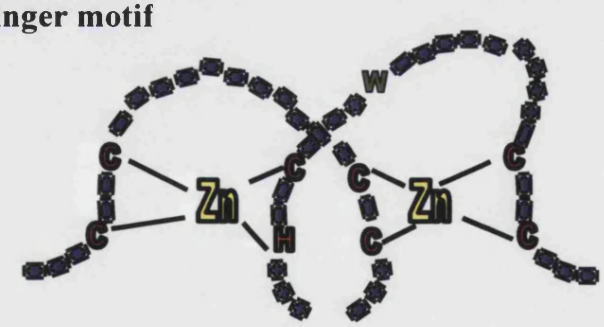
C3 H C4

b. PHD/LAP finger motif

C-2x-C-(4-11)x-C-2x-C-4x-H-2x-C-(9-14)x-C-2x-C

C4 H C3

c. BKS finger motif

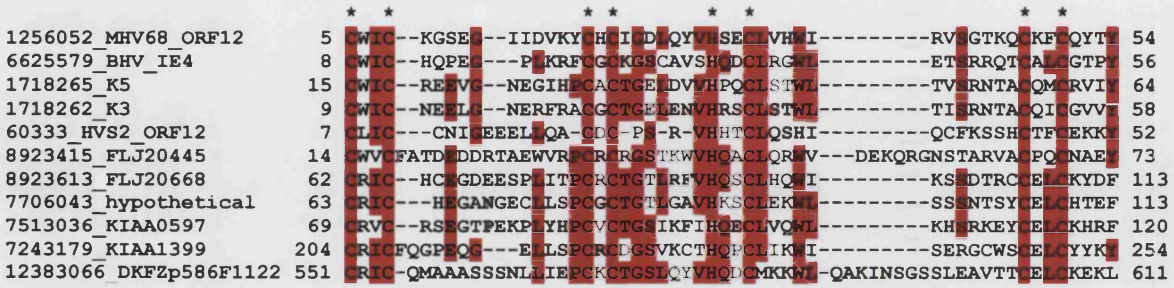


C-2x-C-(10-15)x-C-x-C-7x-H-2x-C-3x-W-(8-21)x-C-2x-C

C4 H C3

Figure 2.10 The Spatial Differences between RING, PHD/LAP, and BKS Finger Motifs. a) The RING zinc finger motif defined by its C3HC4 configuration in contrast to b) the PHD/LAP zinc finger motif defined by C4HC3; and c) the BKS zinc finger motif which is also C4HC3, designating it to the PHD/LAP family, but in its own class due to the less rigid configuration that includes a conserved tryptophan between the 5th and 6th cysteine. C=cysteine; H=histidine; W=tryptophan; x=any amino acid; 0-9=number of residues.

a.



b.

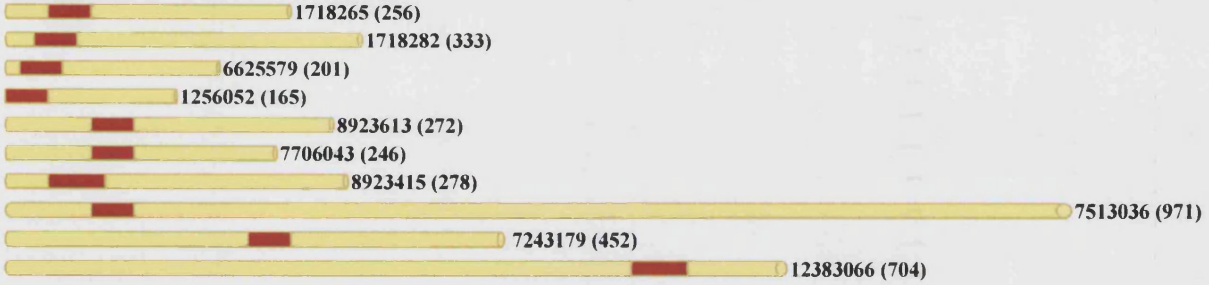


Figure 2.11 The Alignment and Positioning of the BKS Motif in Viral and Human Proteins. a) K3/K5 herpesvirus protein family with six human homologues. Cysteine/histidine conserved residues in the BKS motif are indicated. Proteins are labelled with their GenBank identification number (GI) and a short description. Amino acids shaded red share identity across 50% or more of the alignment. b) The position of the BKS motif is indicated in red along each protein; each protein is labelled with its GI number and its total length in brackets.

2.4 Conclusion

The publication of the human genome has provided the opportunity to analyse host-parasite interactions using new methods. Herpesviruses capture genes from their host and use them during their infection cycle. This work has analysed virus-host protein homology using consistent cross-comparative methods for herpesvirus proteins and gene products of the human genome. The study has allowed us to derive a global picture of cellular functions captured by herpesviruses.

There are a variety of pairwise and multiple alignment tools available for use in such studies. Those used here, BLASTP, PSI-BLAST, and IMPALA (all members of the BLAST family of programs) were chosen for their speed and automation. This is of particular concern when analysing large datasets. Using members of the same program family assures the possibility of cross-program results comparison, as the statistical methods used to measure significance are related.

We have detected sequence homology to human proteins for approximately 12.5% of all known herpesvirus proteins. The question remains whether the remaining 87.5% can be considered exclusively viral. It is likely that a fraction may still be functional homologues with global sequence similarity too limited to be detectable by the methods used here. In addition, our methods will not detect very small sequence motifs such as phosphorylation and protein binding sites. Therefore, viral proteins such as HHV8 K15, which contains a tumour necrosis factor receptor associated factor (TRAF) binding domain (Glenn, Rainbow et al. 1999), or EBV LMP-2A, which contains immunoreceptor tyrosine-based activation motif (ITAM) sequences (Fruehling and Longnecker 1997), are not detected here, although these proteins are known to provide signals/functions similar to their related human counterparts in a given cellular context. Attempts have been made to identify these motifs in HHV-5 (Rigoutsos, Novotny et al. 2003). This is difficult to do, as short sequences segments are too readily detected in large datasets. Their short length produces results with high statistical scores, but with an increased rate of false positives. The subsequent alignments produced by Rigoutsos *et al* were tenuous and full of gaps indicating a lack of functional significance in the results (Rigoutsos, Novotny et al. 2003).

A further confounding factor for detection of viral homologues is the rapid evolution of some viral sequences. It has been estimated that herpesvirus proteins typically evolve one or two orders of magnitude more rapidly than host proteins (McGeoch and Cook 1994), and this may quickly mask any common sequence identifiable ancestry of two proteins. For example, one known human/herpesvirus homologue, thymidine kinase (TK), is present in all known herpesviruses but, due to very limited sequence similarity, could not be identified using our methods; although a human TK mitochondrial homologue of the channel catfish herpesvirus (CCHV) TK protein was detected. Human homologues of the MHV-68 serpin (serine protease inhibitor), M1, were similarly not identified using sequence similarity searches.

The relative number of homologues between herpesviruses and the human genome may also increase as the prediction methods and number of human gene products from the human genome become more accurate. This is highlighted by an initial failure to detect the sequence based homology between human and herpesvirus α-N-formylglycineamide ribonucleotide aminotransferase (FGARAT). Neither of the human predicted protein datasets contained FGARAT even though a human FGARAT gene was recently reported (Patterson, Bleskan et al. 1999). Additional homologues for non-human herpesviruses may also be identified when their host genome sequences becomes available.

The reverse of this argument applies equally to herpesvirus proteins. Many of the open reading frames in the herpesvirus genomes are only conceptual translations from the virus genome sequence and are, therefore, predicted hypothetical proteins. Most of the hypothetical proteins are singletons, only 4% of which showed homology to human proteins, in contrast to 10% of the herpesvirus HPFs. The lack of supporting information surrounding such conceptual translations are highlighted by Davison et al, in their recent overhaul of the wild-type HHV-5 genome. By comparing the wild-type HHV-5 genome with the chimpanzee equivalent (CCMV), they were able to discount 51 hypothetical proteins, reinterpret 24 proteins, and propose 10 novel ORFs (Davison, Dolan et al. 2003). Therefore, careful re-analysis of sequenced herpesvirus genomes may identify more authentic viral proteins.

The analysis of the expression of all open reading frames using methods such as DNA array-based profiling (Chambers, Angulo et al. 1999; Stingley, Ramirez et al. 2000;

Hill, Lukiw et al. 2001; Jenner, Alba et al. 2001; Moses, Jarvis et al. 2002; Polson, Wang et al. 2002; Wagner, Ramirez et al. 2002; Jones and Arvin 2003) will establish if these potential products are expressed during the virus cycle. Overall, the continued, virus-focused searching of constantly growing protein databases using cross-comparable methods is likely to increase our understanding of the relationship between virus and host.

There also appears to be a direct correlation between genome size (in case of HHV5) and number of host homologues (HHV8) and the flexibility of infection in terms of cell tropism and host-virus interaction/avoidance. Beta- and gammaherpesviruses infect lymphocytes and encode a number of proteins that prevent apoptosis and antigen presentation, possibly to avoid viral detection and immune elimination. The human alphaherpesviruses, on the other hand, encode fewer host-interaction homologues than their beta- and gammaherpesvirus counterparts. Alphaherpesviruses infect neurones present in the immune privileged CNS and, therefore, potentially do not face such a vigorous immune responses as met by beta- and gammaherpesviruses. They do encode a number of additional homologues not found in the other two subfamilies, such as kinases and transactivators, likely to assist alphaherpesvirus genome replication in a cellular environment otherwise unsuited to DNA replication. This, however, is not an absolute rule. Many alphaherpesvirus proteins have extensive immune evasion capabilities. For example, the HHV-1 protein kinase US3 has been recently found to inhibit apoptosis by phosphoylation of the protein Bad, a member of the bcl-2 family (Cartier, Komai et al. 2003).

3.0 New viral additions to the Gene Ontology

3.1 Introduction

3.1.1 The Gene Ontology

As available sequence data increases, it is important that consistent annotation and organisation of information does not fall behind. Cross comparison between species is particularly vulnerable to misinterpretation if there is no universal understanding of functional definition for homologous gene functions. The importance of creating and maintaining a universally accepted terminology is magnified when the gene product title and gene product function, while often sharing the same name, are confused with each other. Additionally, confusion is increasingly common when there are multiple names for the same gene product or multiple, unrelated genes sharing the same name (Pearson 2001). It is for this reason that the Gene Ontology (GO) Consortium (Consortium 2001) developed a universal vocabulary (ontology) of biological process, cellular component, and functional definitions. The aim of the ontology is to provide a shared, structured vocabulary adequate for the annotation of molecular characteristics across organisms (Ashburner, Ball et al. 2000).

The GO Consortium was established in 1998 as a collaboration between three independent model organism databases: FlyBase (*Drosophila*) (Consortium 2003), SGD (the *Saccharomyces* Genome Database) (Ball, Dolinski et al. 2000), and MGI (the integrated database comprising Mouse Genome Database (MGD) (Bult, Blake et al. 2004), and the Gene Expression Database (GXD) (Hill, Begley et al. 2004). In 2000, two more model organism databases joined: The *Arabidopsis* Information Resource (TAIR) (Garcia-Hernandez, Berardini et al. 2002), and the *Caenorhabditis elegans* group, WormBase (Harris, Chen et al. 2004). The collaborators provide three services: a) the creation and maintenance of the ontologies, b) the association of genes and gene products from the contributing databases, and c) the development of tools to aid usage of the ontologies and their continued maintenance and growth.

The Gene Ontology project is divided into three species-independent ontologies: molecular function, biological process, and cellular component. The molecular function

ontology describes protein functions at the biochemical level without designating their time, place, or context. The biological process ontology links functions together in an ordered assembly. There is still no designation of time, place or specific context, thus biological processes do not represent biological pathways. In contrast, the third ontology, cellular component, does outline the place or structure in a cell where a particular gene product can be found. This ontology will refer to the site where a gene product is active, so 'place' can refer to a location such as the cell membrane, or a structure such as a proteasome (Ashburner, Ball et al. 2000).

Biological process and molecular functions are often confused; to help differentiate, biological processes usually have more than one distinct step (examples of processes are: molecular transport, metabolism, translation or replication). Molecular function terms, on the other hand, often end with the words 'activity' or 'binding' (examples are: DNA polymerase activity, or DNA binding). The necessity for these words at the end of enzymatic activity such as DNA polymerase is to help distinguish between the physical gene product that performs such a function (and is therefore named after it), and the name of the function it performs, as these two also are often confused. Therefore, the protein known as 'DNA polymerase' has (among others) the function 'DNA polymerase activity', which is a term name found in the molecular function ontology.

Each of these ontologies is represented by directed acyclic graphs (DAGs). As in a strict hierarchy, each parent can have multiple children (more specific terms), but in a DAG each child can also have more than one parent (more general terms) (Figure 3.1a). The relationship between a child and a parent can be either of the "is a" type, where the child is an instance of the parent, or of the "part of" type, where the child is a component of the parent (Figure 3.1b). The advantage of the DAG structure is that a child can have different relationships with each of its different parents (Consortium 2001), allowing for a more realistic depiction of biological systems than the classic hierarchy.

The DAGs must also follow the "True Path Rule" (Consortium 2001): the path from each child to each parent must be true to the top level parent/s. Should any new child not follow the rule, a new node in the DAG must be constructed (or removed) and the links reformed such that the new paths are true.

Each level or term of the DAG is designated a number and this number is then conferred upon any gene products associated with that level (term) of the ontology (Figure 3.1b). Each term is also accompanied by a definition that should apply to all gene products at that level, and information concerning any synonyms.

The GO Consortium makes very clear what should not be expected of GO. The Gene Ontology is an ordered description of the behaviour of gene products in a cellular context. It is not, therefore, a gene product database, and terms within the ontologies do not describe individual gene products – but functions, processes, or components that can be conducted, partaken in, or comprised of gene products. It therefore is not a unifying solution, it does not dictate a minimum standard or nomenclature that should be universally adopted. Finally, GO does not attempt to describe all aspects of gene products. Details of domains, 3D structure, protein-protein interactions, splicing parameters, disease associations, tissue or cellular tropisms, or developmental stages are not included in GO. Many of the above listed, however, can be found in other ontologies that are currently being developed simultaneously with GO.

GO can be accessed from its website and browsed online, downloaded by anonymous FTP, or accessed by CVS (Concurrent Versions System). Files are available in three different formats: flat file (updated daily), XML, and MySQL (both updated monthly).

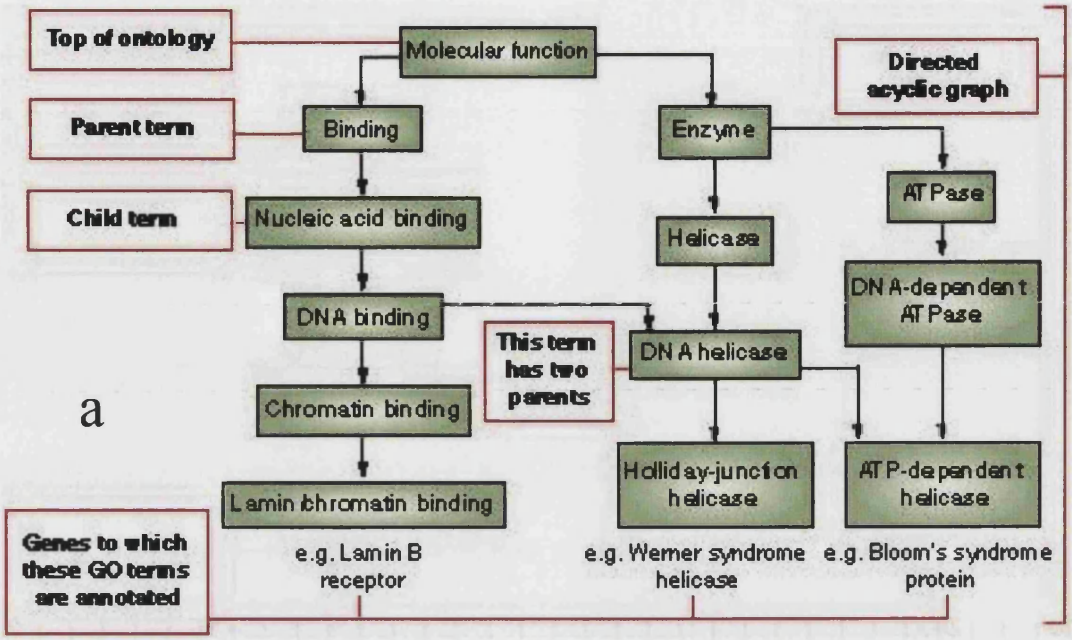
3.1.2 Adding new virus-related terms to the gene ontology

Analysis of host-viral interaction, especially with high throughput methods such as DNA microarrays, can be achieved without a formalised method of regulated gene annotation. The Gene Ontology, however, provides the ideal framework necessary to compare two different species at the genomic level, utilising a common language and allowing comprehensive analysis to be conducted more easily. For this to be possible, though, both species require their gene products to be annotated with the appropriate GO numbers.

The current custom of the Gene Ontology consortium is to include terms in the various ontologies that apply to more than one taxonomic Kingdom. Occasionally, however, the same term has two different meanings in two different species. When this occurs, the term 'sensu' is utilised to distinguish between definitions, as in the example:

GO:0016065:humoral defense mechanism (sensu Invertebrata), and **GO:0016064:humoral defense mechanism (sensu Vertebrata)**. Unfortunately, this annotation technique is not always sufficient to distinguish viral from organism-based definitions. While a few can be shared, such as DNA polymerase and protein kinase, there are a large number of viral gene products that do not share functions, processes or components with other taxonomical groups of organisms, for example, **GO:0046773:viral inhibition of host cell protein biosynthesis shutoff**, or **GO:0046740:viral spread within host, cell to cell**.

An initial set of viral terms was available within GO, but these required extensive consistency checking and placement to make them usable by virologists. This was undertaken here and any necessary new terms were also created. These terms, in accordance with GO guidelines, are fully integrated into the existing GO DAGs and can be accessed alongside cellular terms at http://www.geneontology.org .



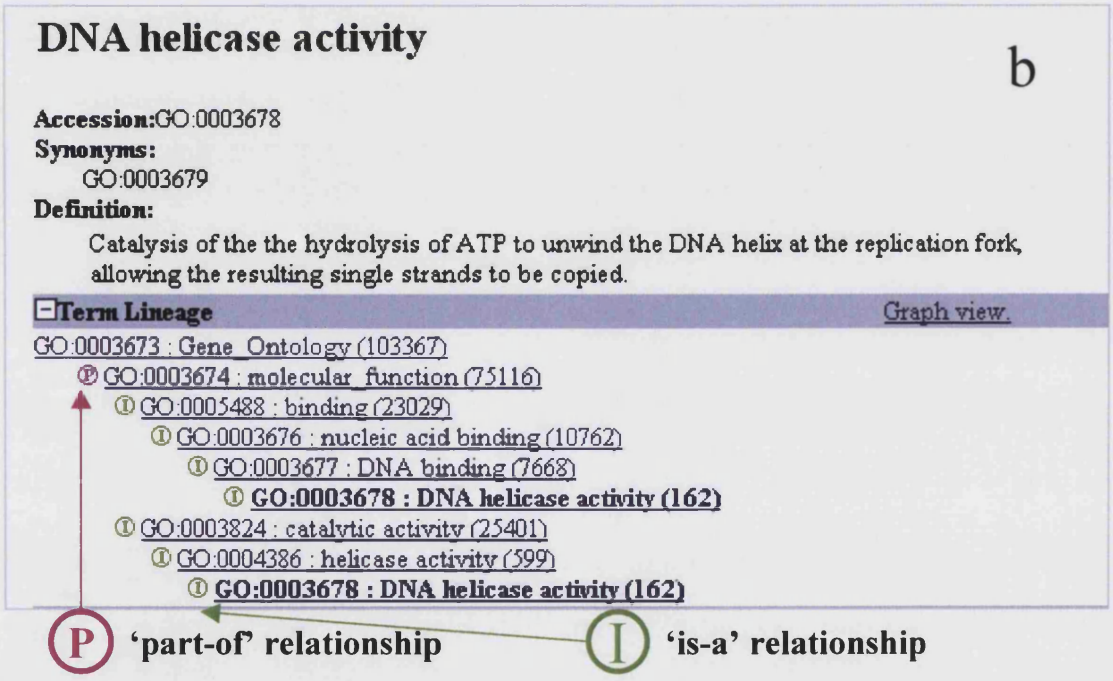


Figure 3.1 The Structure of the Gene Ontology and its Terms. a) The Gene Ontology is organised into Directed Acyclic Graphs (DAGs) characterised by the fact that parents have multiple children (as with any hierarchy) but children can also have multiple parents such as the term DNA helicase. Terms are then annotated to gene products to aid in unifying annotation between species. This picture was taken from the Gene Ontology website at www.geneontology.org. b) Each term in the DAG is identified by its name, accession number and definition. It can have either a 'part-of' relationship, or an 'is-a' relationship, relationship with each of its parents.

3.2 Methods

3.2.1 New Viral GO Terms

Examination of GO revealed that there were a number of instances where it was necessary to create new viral GO terms. All terms were created in accordance with GO guidelines available from www.geneontology.org/. New terms were devised by documenting and breaking down into component parts the various stages of viral infections using standard references and expert opinion. Each component was defined as a specific, new GO term and annotated with a definition and supporting reference, where available. The presence of synonymous terms was determined manually. Where possible, Fields Virology (Knipe, Howley et al. 2001) and VIDA GenBank derived annotations were used for new GO term assignments. All terms, definitions and references are curated manually by the Gene Ontology Consortium before a unique GO accession number is assigned to each term and they are integrated into the existing Gene Ontology DAGs.

3.2.2 Visualisation

Schematic diagrams of DAGs are retrieved from the QuickGO browser, available from http://www.ebi.ac.uk/ego/. The DAGs in Figure 3.5 were manually produced. Figure 3.3 DAGs are produced by the AmiGO brower, available from www.geneontology.org/.

3.2.3 Data Availability

All new viral terms, definitions, and references are listed in Appendix A, and have been integrated into the ontologies, which are accessible by searching the database at www.geneontology.org using any of the available search engines.

3.3 Results

3.3.1 Assigning New GO Terms (placing them in the ontologies)

Gene Ontology guidelines for creating new terms require each term to be named, defined and referenced. Each term is then curated by the Consortium before being added to the existing ontologies. The current ontologies are not complete; rather, new terms are constantly being created. When adding to an existing ontology it is important that accuracy, redundancy, overlap, and placement are all taken into consideration. Because of the enormity of the project the Gene Ontology is not without inconsistencies and nor is it static, with such inconsistencies being constantly corrected and terms updated or revised. Figure 3.2 demonstrates the necessity for such close attention to detail, which is especially important when using the ontology to decipher host-pathogen relations, as many terms will refer directly to various interactions involving products from both species. Before assignments can be made the term and its definition must be universally understood.

3.3.1.1 Accuracies

The need for accuracy is exemplified by the initial viral term **GO:0019054:virus-host cell process manipulation**, (Figure 3.2a), whose original definition is "*defined **cellular** processes that are disturbed by **viral** products*". This definition implies that *host* products should be annotated to this term and to any of its child terms, however, both of the two child terms clearly relate to *viral* products that interfere with *host* processes (Figure 3.2b), **GO:0019056:viral pertubation of host cell transcription** and **GO:0019057:viral pertubation of host cell mRNA translation**. Indeed, a viral product that manipulates host processes, such as HHV-1 UL41 (virion host shut-off protein), should be assigned to a child of this GO term, however, it is not a defined *cellular* process. A more accurate definition for the parent term would be: *The manipulation of host cell processes by viral products.* Thus, the term **GO:0019054:virus-host cell process manipulation**, whose original definition was misleading was changed to: *Alteration of defined cellular processes that viruses target during replication,* a more accurate description. This illustrates the necessity for accurate, semantic use of language in GO term creation to prevent annotation inaccuracies.

3.3.1.2 Redundancy

Given the hierarchical levels of the DAGs, redundancy was not commonly found. Some examples were initially identified for further investigation, as in Figure 3.2b. The term **GO:0019048:virus-host interaction**, is defined as: *interactions directly with the host cell macromolecular machinery, to allow virus replication.* This can be interpreted as encompassing every viral gene presently known to function during the viral life cycle, as viral replication cannot occur without interaction with cellular macromolecular machinery. This creates confusion when a viral product of unknown function is being annotated. The product could be assigned **GO:0000004:unknown**, but could also be more informatively annotated with the general term **GO:0019048:virus-host interaction**. The need for the term **GO:0019048** at the given DAG level becomes apparent as it is the only logical step between higher, less specific terms, such as: **GO:0030383:host-pathogen interaction**, and its children **GO:0019049:viral-host defense evasion**, and **GO:0019054:virus-host cell process manipulation**. Therefore, the term **GO:0019048** is not necessarily designed for annotation use, but primarily to structure the DAGs. The definition is all encompassing, however, because children terms have more specific definitions, and their placement is logical beneath the general parent term. Therefore, the correct annotation for an ORF involved in unknown viral processes would be **GO:0000004:unknown**, until more information is known and the ORF can be more accurately described, but the term **GO:0019048** is not redundant.

3.3.1.3 Overlapping

Examination of the existing Gene Ontology identified some overlapping terms (Figure 3.2c). The terms 'cytoplasm' and 'nucleus' are two examples of terms easily named, defined, and placed within the cellular component ontology. The creation of the terms **GO:0042025:host cell nucleus**, and **GO:0030430:host cytoplasm**, clearly overlaps with the pre-existing terms. There are three immediate concerns that arise from this duality. First, should new terms relating to both healthy and infected cell cytoplasms be placed appropriately as children of both terms (**GO:0005623:cell** and **GO:0018995:host**)? While there are certain cellular reactions, and thus certain gene products, that would only function during infection, these discrepancies in product function can be recorded by creating appropriate biological process and molecular

function terms. Second, since 'nucleus' and 'cytoplasm' have been designated separate GO numbers to distinguish between 'cell' and 'host', new terms would have to be created to represent every structure in the cell. Third, the term **GO:0018995:host** and all of its children have been placed as children of the **GO:0005576:extracellular** term. It would, therefore, be inaccurate to place any virus-related GO terms within the 'host' term DAG as that would place them in the extracellular compartment – an inappropriate assignation for any viral gene product operating inside the cell.

This apparent 'overlap' in the ontologies, however, serves a necessary purpose. The existence of 'host' terms in the Gene Ontology, is used to distinguish between cellular based pathogens that, as an integral part of their life cycles, are internalised by other cells (i.e. *Plasmodium falciparum*). Indeed, the term **GO:0018995:host** has been defined as *Any organism in which another organism, especially a parasite or symbiot, spends part or all of its life cycle and from which it obtains nourishment and/or protection.* Therefore, some of these pathogenic gene products may function in the pathogen's nucleus (**GO:0005634:nucleus**), and others in the host cell's nucleus (**GO:0042025:host cell nucleus**), thus requiring the two terms in order to distinguish between the two cells. As this dichotomy cannot occur during the viral life cycle, the terms relating to a cell (**GO:00055623:cell** and all its children) and not those relating to a host cell (**GO:0018995:host** and all its children) were used when integrating new virus-related GO terms (and eventually to annotate viral products). This procedure was agreed with the GO Consortium for two reasons: i) The DAG beneath **GO:00055623:cell** is more complete in it's complexity and depth of terms, and ii) this structure creates a distinction, allowing for the annotation of cellular based intracellular pathogens.

3.3.1.4 Placement Errors

Some examples of placement errors were identified in the existing viral terms (Figure 3.2d) such as **GO:0019036:viral transcriptional complex**, which was originally assigned as a child of **GO:0042025:host cell nucleus**. This error constitutes a true-path rule violation (see below) as transcription, whether viral or cellular, cannot occur extracellularly. In a similar example, **GO:0030430:host cytoplasm** (Figure 3.2c) was originally a child of **GO:0042025:host cell nucleus**, instead of being its sibling. These

placements were easily adjusted in accordance with current GO guidelines because the fluid nature of the ontologies allows for the modification of the DAGs when necessary.

Finally, certain distinctions between cellular and viral functions were made where, despite the presence of similarity in function/process, there was a clear distinction between the two that sufficiently warranted a new GO term. This was the case for terms such as **GO:0019039:viral-cell fusion molecule**, and **GO:0019083:viral transcription**. The process of cell membrane fusion involves a number of cell surface proteins; however, the process of viral-induced cell membrane fusion will also involve a number of fusion proteins of viral origin, which therefore require distinction. The process of cellular transcription and viral transcription are also distinct processes during viral infection, which in the case of viral transcription, can be controlled by viral specific processes. In this situation it is important to distinguish which processes of transcribing a viral gene are unique and linked to viral infection, especially when the viral and cellular process compete.

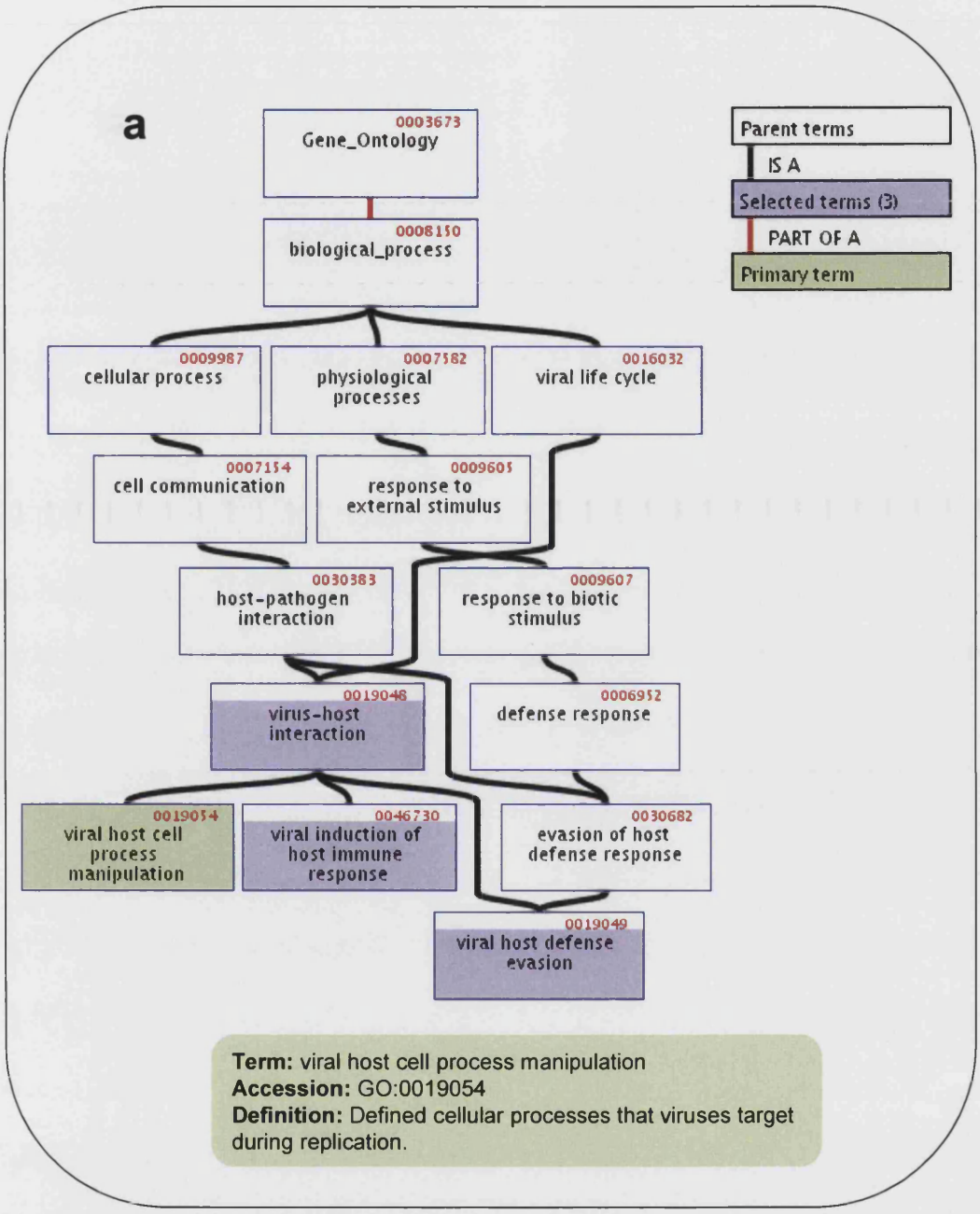
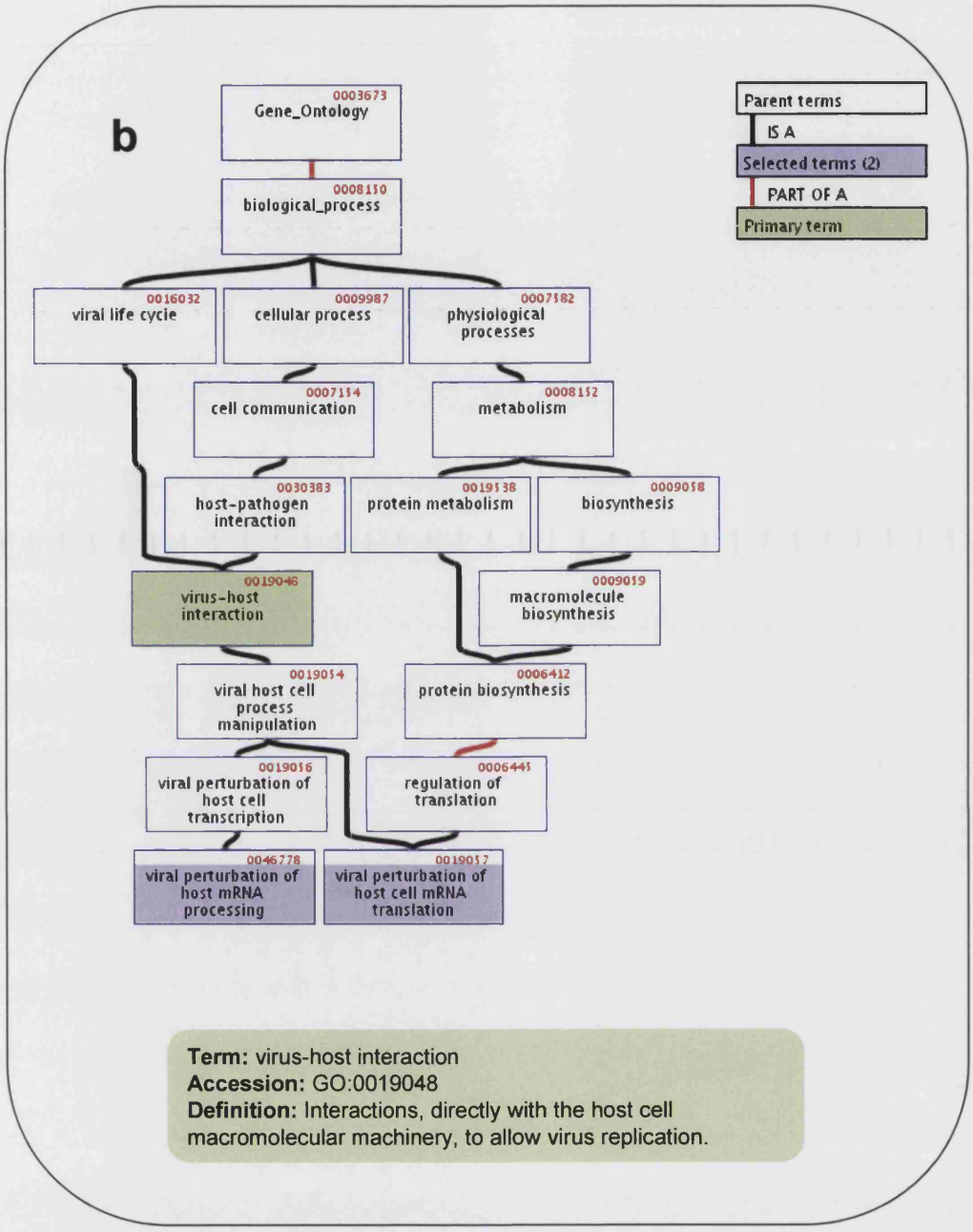
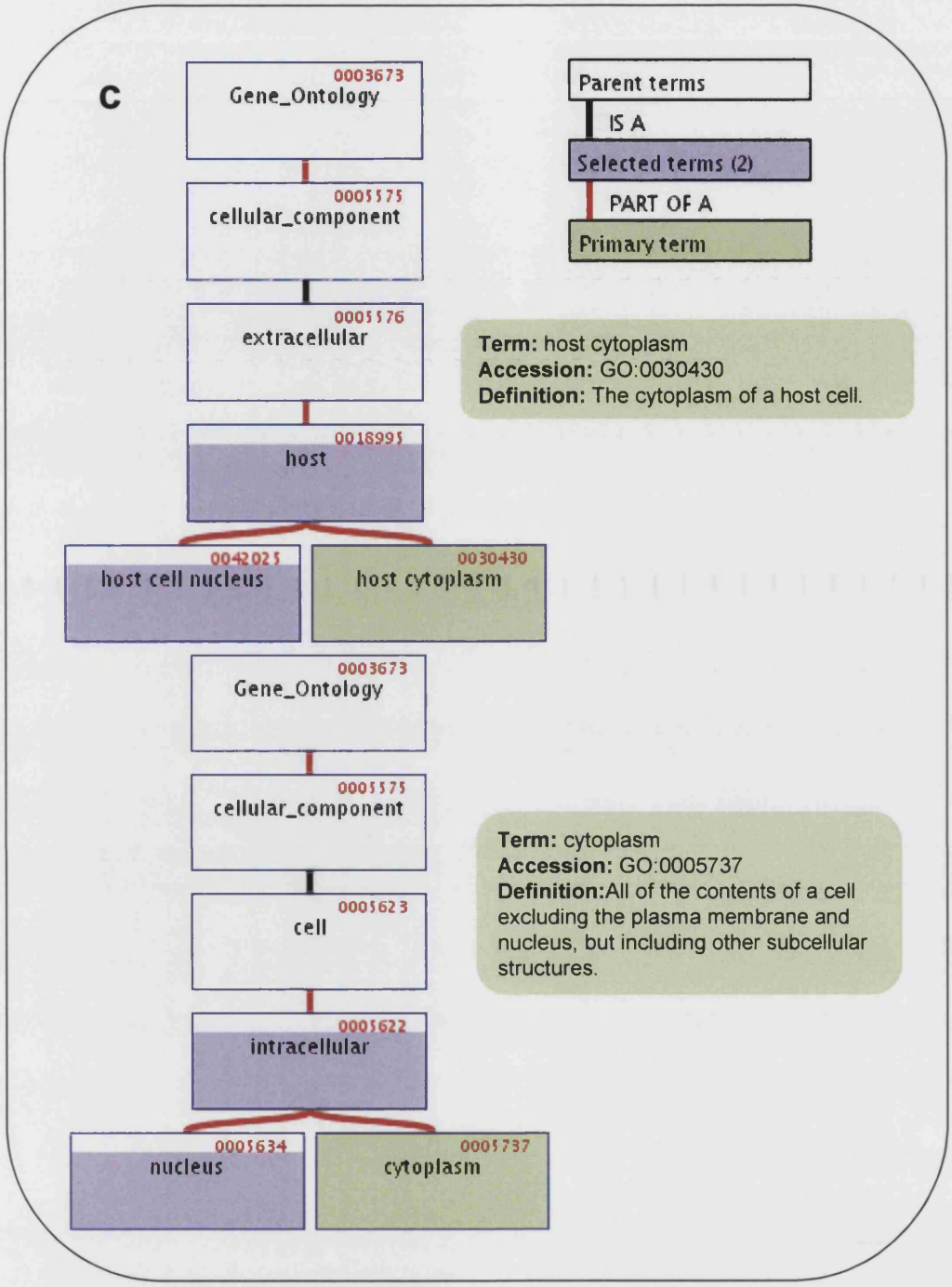
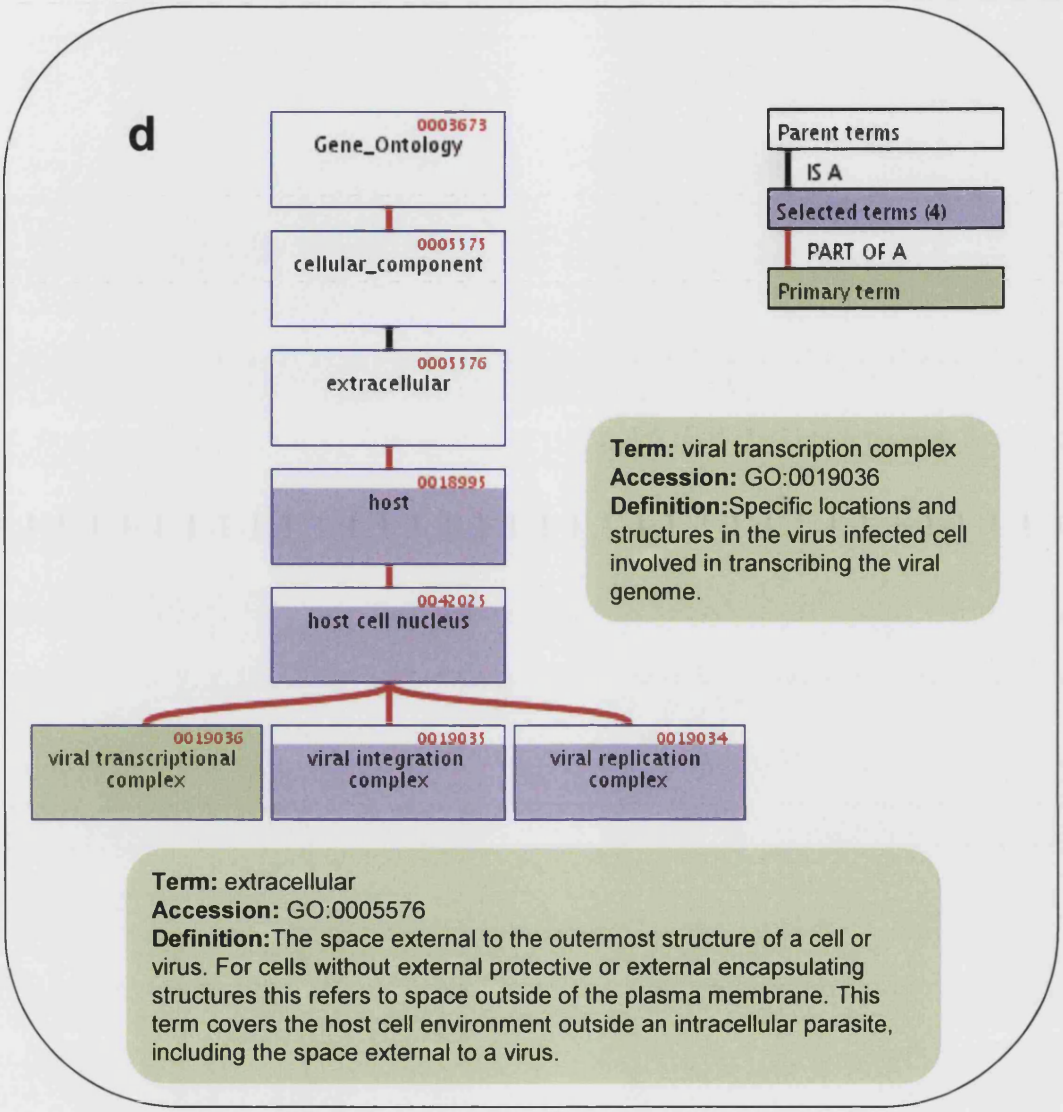


Figure 3.2 Accuracy, Redundancy, Placement, & Overlap within the Gene Ontology. There are a number of potential problems within an Ontology that must be constantly considered when creating new terms such as a) Accuracy of term names and definitions; b) Redundancy of new terms with respect to existing terms; c) Placement of new terms; and d) Overlapping of old and new terms. These were all addressed during creation of viral GO terms.


b
0003673
Gene_Ontology
0008150
biological_process
0016032
viral life cycle
0009987
cellular process
0007582
physiological processes
0007154
cell communication
0008152
metabolism
0030383
host-pathogen interaction
0019538
protein metabolism
0009058
biosynthesis
0019048
virus-host interaction
0009059
macromolecule biosynthesis
0019054
viral host cell process manipulation
0006412
protein biosynthesis
0019016
viral perturbation of host cell transcription
0006445
regulation of translation
0046778
viral perturbation of host mRNA processing
0019057
viral perturbation of host cell mRNA translation
Parent terms
IS A
Selected terms (2)
PART OF A
Primary term
Term: virus-host interaction
Accession: GO:0019048
Definition: Interactions, directly with the host cell macromolecular machinery, to allow virus replication.


C
0003673
Gene_Ontology
0005575
cellular_component
0005576
extracellular
0018995
host
0042025
host cell nucleus
0030430
host cytoplasm
0003673
Gene_Ontology
0005575
cellular_component
0005623
cell
0005622
intracellular
0005634
nucleus
0005737
cytoplasm
Parent terms
IS A
Selected terms (2)
PART OF A
Primary term
Term: host cytoplasm
Accession: GO:0030430
Definition: The cytoplasm of a host cell.
Term: cytoplasm
Accession: GO:0005737
Definition:All of the contents of a cell excluding the plasma membrane and nucleus, but including other subcellular structures.


d
0003673
Gene_Ontology
0005575
cellular_component
0005576
extracellular
0018995
host
0042025
host cell nucleus
0019036
viral transcriptional complex
0019035
viral integration complex
0019034
viral replication complex
Parent terms
IS A
Selected terms (4)
PART OF A
Primary term
Term: viral transcription complex
Accession: GO:0019036
Definition:Specific locations and structures in the virus infected cell involved in transcribing the viral genome.
Term: extracellular
Accession: GO:0005576
Definition:The space external to the outermost structure of a cell or virus. For cells without external protective or external encapsulating structures this refers to space outside of the plasma membrane. This term covers the host cell environment outside an intracellular parasite, including the space external to a virus.

3.3.1.5 Maintaining the true path rule

The "True Path Rule" is integral to GO and states that the path from each lower level node ('child') up to the top level node ('parent') must be true, i.e. every intermediate parent must make sense in the context of the process, function, or component in question. It is important when placing terms in the ontology that this rule is not violated. Violations can occur through placement error, where a term is not logically, or correctly placed within the ontologies (Figure 3.3a). In the case of the term **GO:0019038:provirus**, which has been placed correctly beneath **GO:0019015:viral genome**, with an "IS-A" relationship, the violation occurs in the next few parents, which are not true: the viral genome is not always a part of the nucleocapsid, and by definition cannot always be extracellular. In particular, the provirus is only ever found in the nucleus of a host-cell, as it is the term given to integrated viral DNA, it is never found in the nucleocapsid or extracellularly. This, therefore, constitutes a true-path rule violation. Violations can also occur when a term is placed in multiple positions within the ontology as it is possible that one of the placements would not be accurate for an open reading frame assigned to that GO term (Figure 3.3b). The term **GO:0007323:pheromone processing**, for instance, was originally found in three different places with the biological process ontology and any gene assigned to this term would find that the term follows the True Path Rule for the first and third placements only (as child to **GO:0019236:pheromone response**, and **GO:0016485:protein processing**) (Figure 3.3b). The problem occurs, however, with the second placement beneath **GO:0007322:mating (sensu Saccharomyces)**, as any assignment of this term to a species other than Saccharomyces would immediately violate the True Path Rule. Both of these situations were rectified by altering the shape of the DAGs and removing inappropriate placements (Figure 3.3c). In addition, the term **GO:0007323** has evolved in name and definition to further avoid confusion.

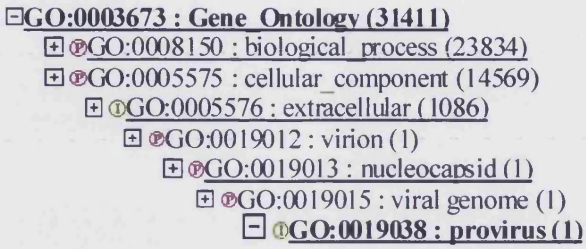
3.3.1.6 Use of sensu

The use of the terminology 'sensu' in the case of the term **GO:0005618:cell wall**, is an example of a situation where the term 'cell wall' has a number of different meanings across a wide variety of species. Thus, 'sensu' is used to distinguish each 'cell wall' term according to its specific definition; each definition relating to a different species. In

this case there exist five different cell wall terms: **GO:0009274:cell wall (sensu Bacteria)**, with two children gram negative (**GO:0009276**) and gram positive (**GO:0009275**), **GO:0009277:cell wall (sensu Fungi)**, and **GO:0009505:cell wall (sensu Magnoliophyta)**, each term has its own number and its own definition. The use of sensu is a particular concern when annotating viral gene products as there are a number of homologous gene products which may share the same name between virus and host, but whose function or structure is sufficiently different to warrant an alternative term.

Within viruses, there are certain terminology distinctions that need to be made in order to refine the Gene Ontology. An example of such a situation is the creation of the term **GO:0046728:viral capsid (sensu Retroviridae)** (Figure 3.4). This was done to allow viral annotators to distinguish between the viral capsid that immediately surrounds the viral genome (**GO:0019028:viral capsid**) and the capsid that surrounds the nucleocapsid (as is the case in retroviruses). The nucleocapsid is defined as the capsid structure that immediately surrounds the viral genome in viruses that have more than one capsid. The new *sensu* term was placed in the ontology as another child of the term **GO:0019028:viral capsid**, and the term **GO:0019013:viral nucleocapsid**, previously a direct child of **GO:0019012:virion**, was moved to be a child of **GO:0046728:viral capsid (sensu Retroviridae)**, reflecting that its relevance to viral annotation relies upon the existence of an additional capsid surrounding the viral genome (Figure 3.4).

a. **Term:** provirus
Accession:GO:0019038
Synonyms: None.
Definition: The name given to a viral genome after it has been integrated into the host genome; particularly applies to retroviruses and is a required part of the retroviral replication cycle.



b. **Term:** pheromone processing
Accession:GO:0007323
Synonyms: None.
Definition: None.

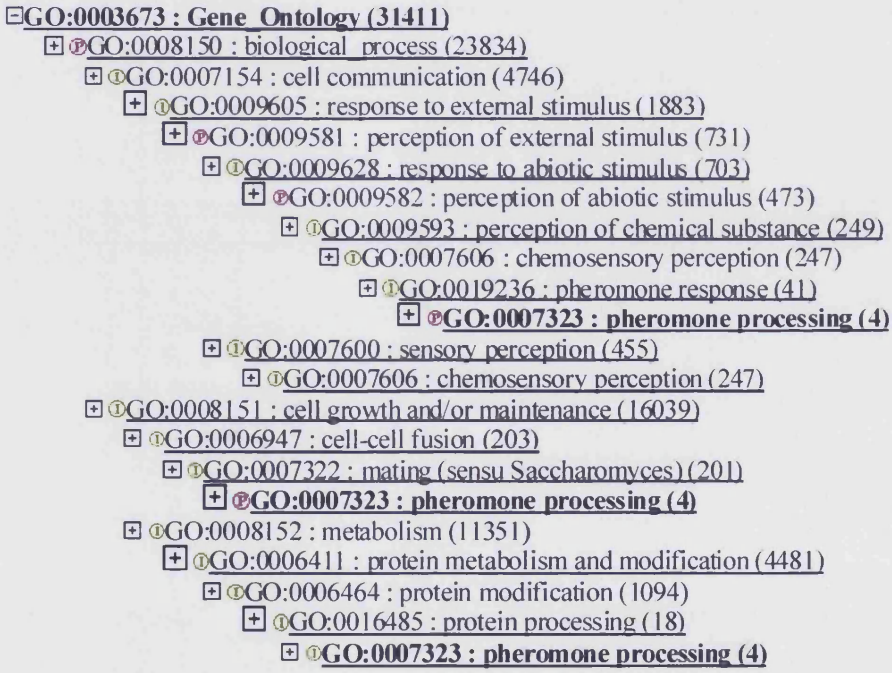
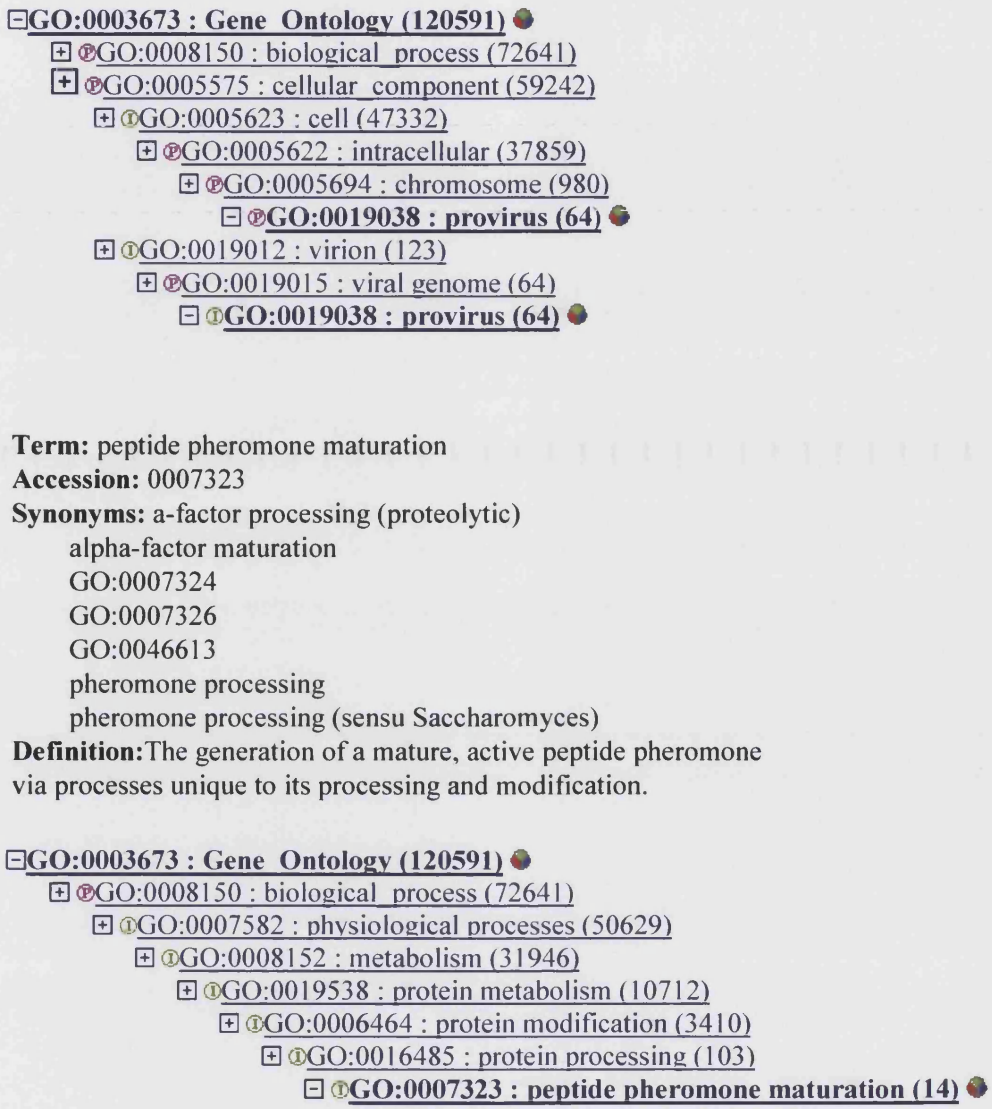


Figure 3.3 Examples of True Path Rule Violations. a) an example of placement error leading to the term **GO:0019038:provirus** violating the true path rule; b) the term **GO:0007323:pheromone processing** is found in three different places in the DAGs, however, not all placements are appropriate for every gene product that could be assigned to this term; c) both the violation and potential violation in a) and b) were easily rectified by rearranging the DAG, in the case of a) and b), and by amending the original term in the case of the term **GO:0007323**.

c.



Term: peptide pheromone maturation
Accession: 0007323
Synonyms: a-factor processing (proteolytic)
alpha-factor maturation
GO:0007324
GO:0007326
GO:0046613
pheromone processing
pheromone processing (sensu Saccharomyces)
Definition: The generation of a mature, active peptide pheromone via processes unique to its processing and modification.

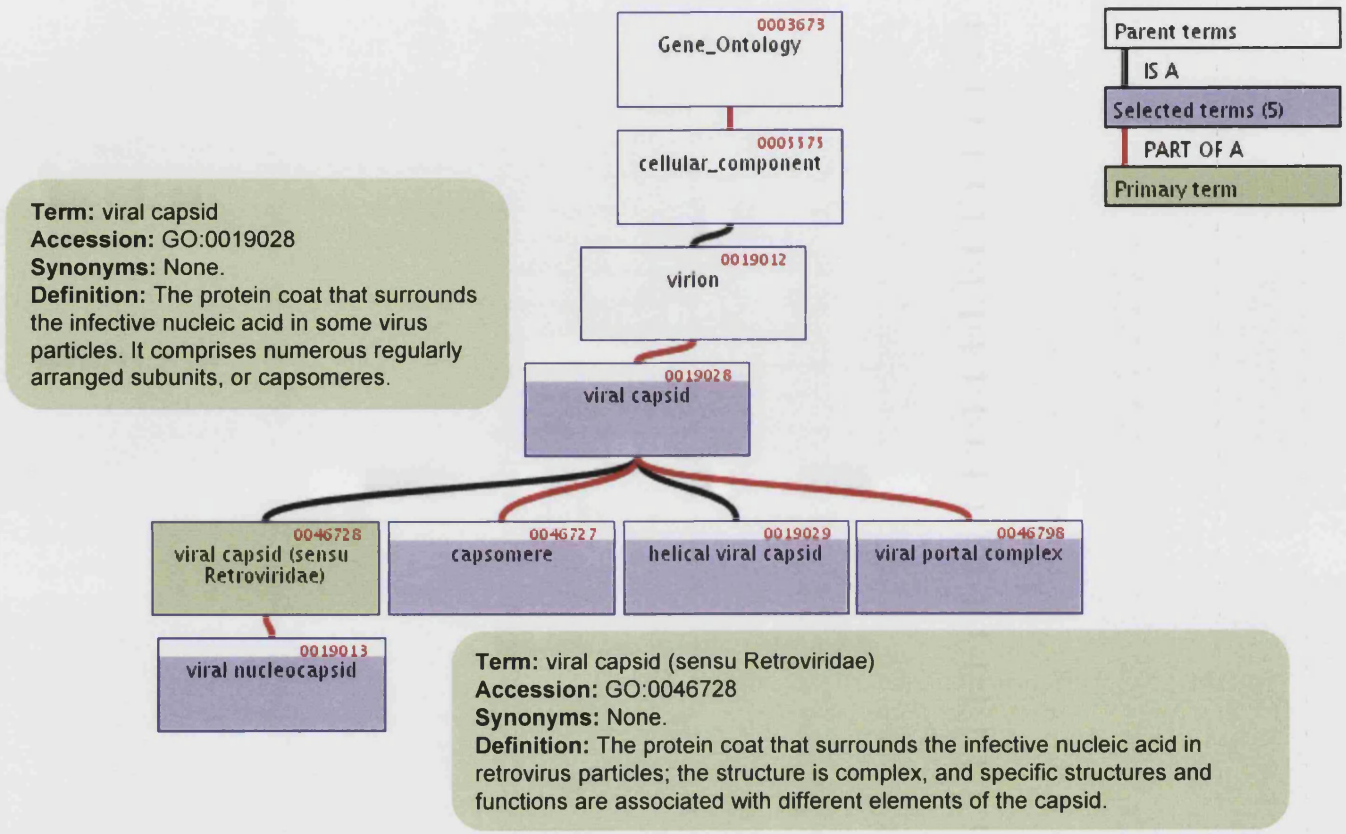


Figure 3.4 Example of *sensu* Usage in viral terms. When the same term has two different meanings between species the word *'sensu'* is used to determine the sense in which the term is intended; here it is used to distinguish between the viral capsid that directly surrounds the viral genome (GO:0019028), and the capsid found in retroviral virions that surrounds a nucleocapsid, which in turn directly surrounds the viral genome.

3.3.1.7 Refining terms

When new terms are added to the existing GO certain knockon effects result in a number of terms that require altering in order to preserve DAG structure. These adjustments include moving term placements within the ontologies as new terms are inserted into the existing DAGs, altering or adding definitions as the ontologies are refined, and in a few cases changing the term name to reflect alterations to the DAGs. All new terms assigned to the ontologies were carefully checked for True Path Rule violations. In addition, any existing terms that were encountered that violated the rule were adjusted by either: removal of the term from the incorrect placement, movement of the term to a more appropriate placement, or creation of a new term using the terminology 'sensu' to distinguish between the different definitions of the terms.

In total, 92 new terms relating to viral function, process, and component were created and incorporated into the ontologies to complement the 95 pre-existing terms, thereby comprising a total of 187 virus-related gene ontology terms (Appendix A). This process was undertaken with respect to all the annotation and placement criteria described in this chapter. Integration of the terms into the existing DAGs (Figure 3.5) was accomplished, often utilising the acyclic structure of DAGs to give certain terms more than one parent (i.e. **GO:0046773:viral inhibition of host cell protein biosynthesis shutoff** is a child of both **GO:0019049:viral-host defense evasion,** and **GO:0019054:virus host-cell process manipulation.**) (Figure 3.5d). All of the terms can also be seen incorporated into the ontologies by using an online browser at www.geneontology.org.

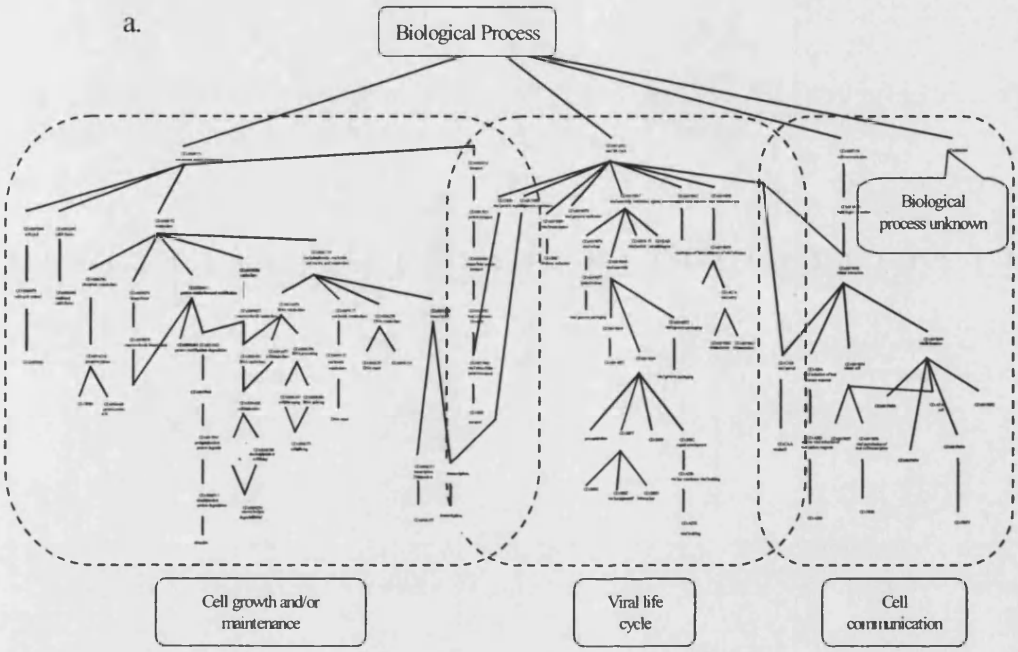
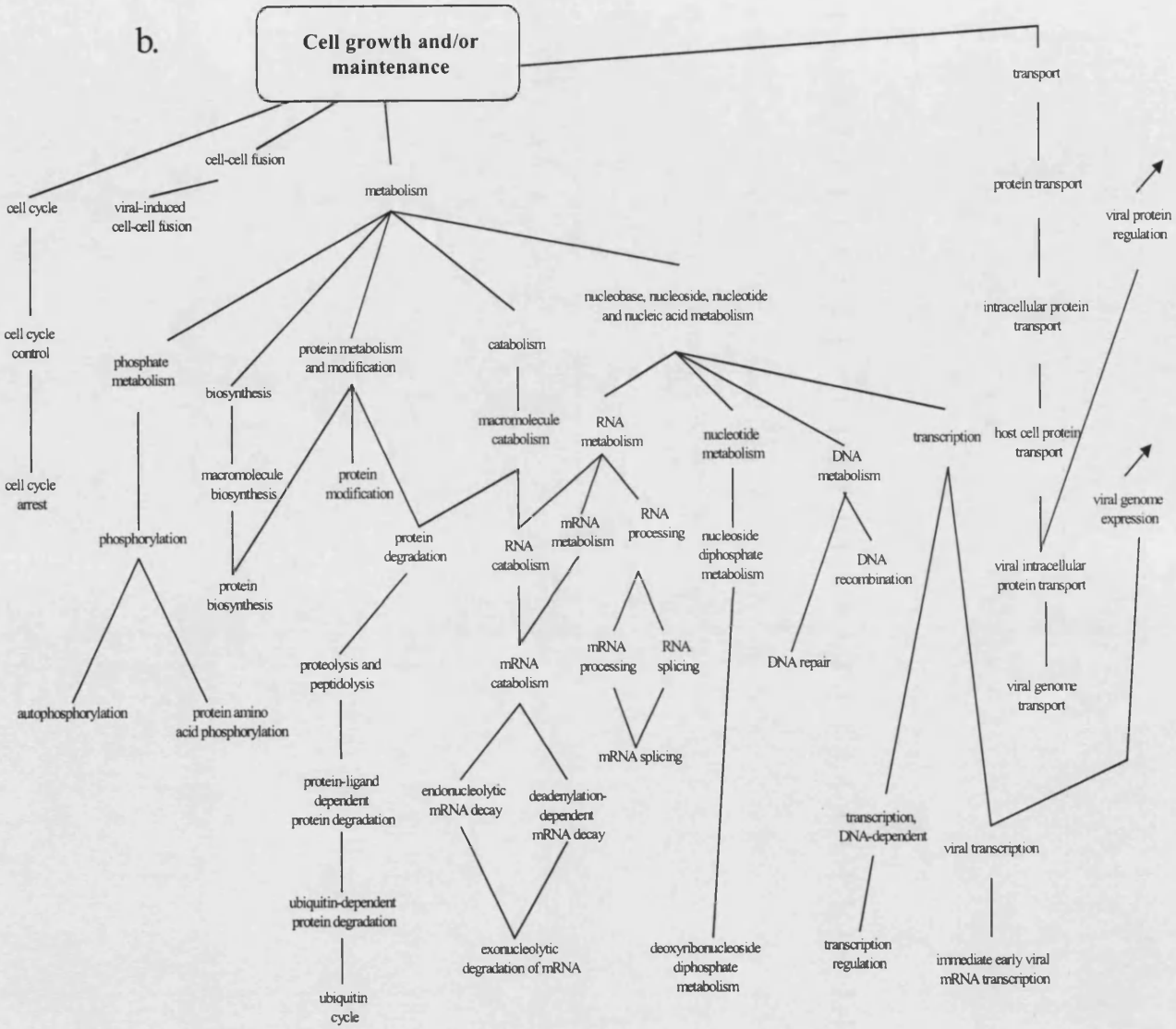
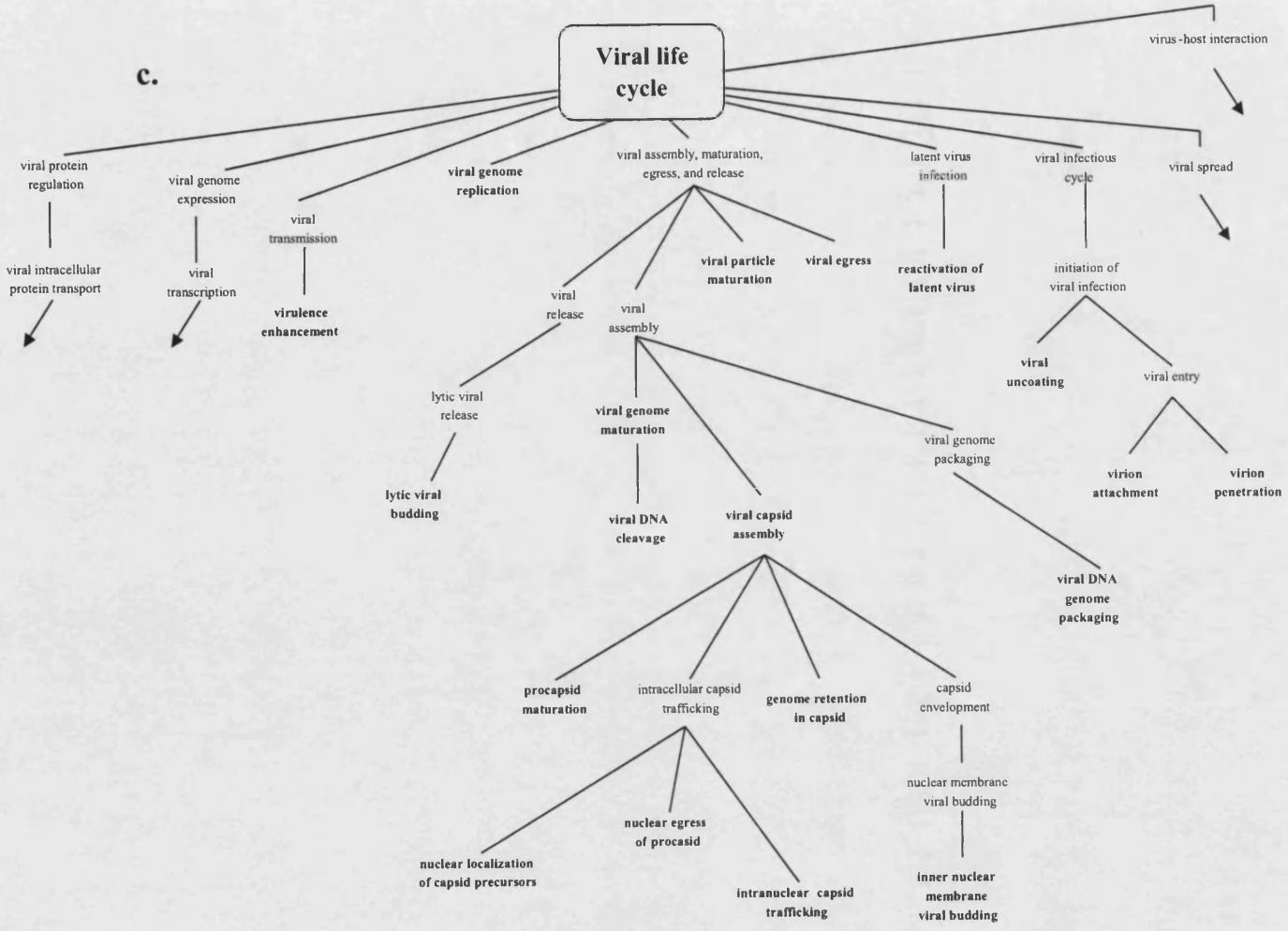


Figure 3.5 Subsections of the Biological Process DAGs with integrated viral terms (above and opposite). All new GO terms are integrated into the existing DAGs. a) shows how the three sections fit together within the ontology; terms are depicted here fully integrated into the b) cell growth and/or maintenance, c) viral life cycle, and d) cell communication sections of the existing biological process ontologies.

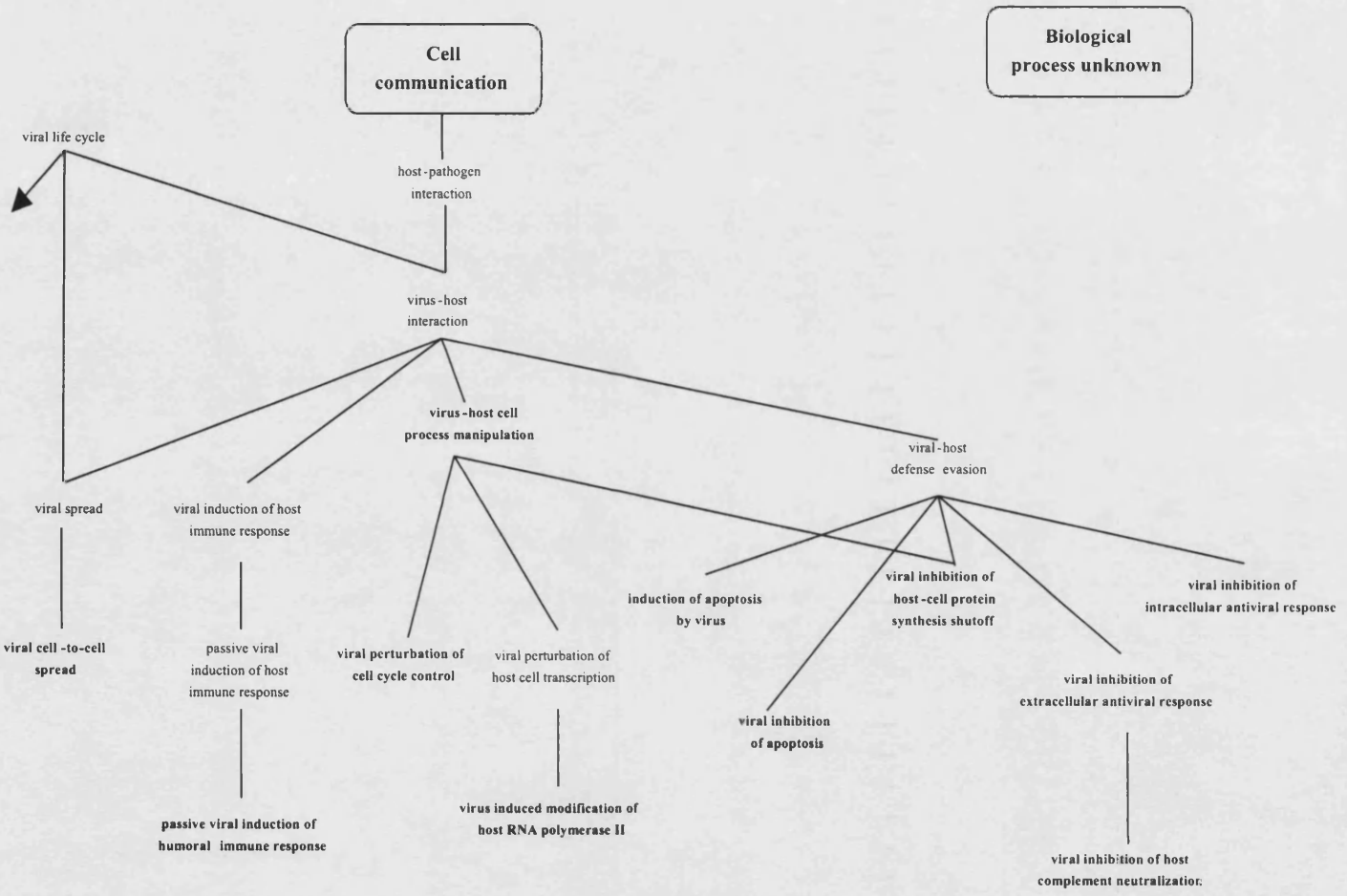
b.


Cell growth and/or maintenance
cell cycle
cell-cell fusion
viral-induced cell-cell fusion
metabolism
transport
cell cycle control
cell cycle arrest
phosphate metabolism
phosphorylation
autophosphorylation
protein amino acid phosphorylation
biosynthesis
macromolecule biosynthesis
protein biosynthesis
protein metabolism and modification
protein modification
protein degradation
proteolysis and peptidolysis
protein-ligand dependent protein degradation
ubiquitin-dependent protein degradation
ubiquitin cycle
catabolism
macromolecule catabolism
RNA catabolism
mRNA catabolism
endonucleolytic mRNA decay
deadenylation-dependent mRNA decay
exonucleolytic degradation of mRNA
nucleobase, nucleoside, nucleotide and nucleic acid metabolism
RNA metabolism
mRNA metabolism
RNA processing
mRNA processing
RNA splicing
mRNA splicing
nucleotide metabolism
nucleoside diphosphate metabolism
deoxyribonucleoside diphosphate metabolism
DNA metabolism
DNA recombination
DNA repair
transcription
transcription, DNA-dependent
transcription regulation
protein transport
intracellular protein transport
host cell protein transport
viral intracellular protein transport
viral genome transport
viral protein regulation
viral genome expression
viral transcription
immediate early viral mRNA transcription

c.


Viral life cycle
virus-host interaction
viral protein regulation
viral genome expression
viral transmission
viral genome replication
viral assembly, maturation, egress, and release
latent virus infection
viral infectious cycle
viral spread
viral intracellular protein transport
viral transcription
virulence enhancement
viral release
viral assembly
viral particle maturation
viral egress
reactivation of latent virus
initiation of viral infection
viral uncoating
viral entry
virion attachment
virion penetration
lytic viral release
viral genome maturation
viral genome packaging
lytic viral budding
viral DNA cleavage
viral capsid assembly
viral DNA genome packaging
procapsid maturation
intracellular capsid trafficking
genome retention in capsid
capsid envelopment
nuclear membrane viral budding
nuclear egress of procasid
nuclear localization of capsid precursors
intranuclear capsid trafficking
inner nuclear membrane viral budding

d.


Cell communication
Biological process unknown
viral life cycle
host-pathogen interaction
virus-host interaction
virus-host cell process manipulation
viral-host defense evasion
viral spread
viral induction of host immune response
viral cell-to-cell spread
passive viral induction of host immune response
passive viral induction of humoral immune response
viral perturbation of cell cycle control
viral perturbation of host cell transcription
virus induced modification of host RNA polymerase II
induction of apoptosis by virus
viral inhibition of host-cell protein synthesis shutoff
viral inhibition of intracellular antiviral response
viral inhibition of apoptosis
viral inhibition of extracellular antiviral response
viral inhibition of host complement neutralizatior

3.4 Conclusion

The development and maintenance of scientific databases, repositories, and ontologies does not necessarily produce conclusive scientific insight and 'evidence'; however, it does produce results upon which further research can be based (Guettler, Jackson et al. 2003; Lord, Stevens et al. 2003; McCarter, Mitreva et al. 2003; Palmer, O'Shaughnessy et al. 2003). Since its conception in 1998, the Gene Ontology has become the focus of a number of new studies (Schug, Diskin et al. 2002; Jensen, Gupta et al. 2003; King, Foulger et al. 2003; Lagreid, Hvidsten et al. 2003), numerous primary and secondary databases and analysis tools have integrated GO into their systems as a base requirement for annotation (Biswas, O'Rourke et al. 2002; Hodges, Carrico et al. 2002; Camon, Magrane et al. 2003; Dennis, Sherman et al. 2003; Doniger, Salomonis et al. 2003; Garavelli 2003; Rhee, Beavis et al. 2003; Sprague, Clements et al. 2003; Tulipano, Millar et al. 2003), and a number of secondary tools have been independently developed to make GO more accessible to the biological science community (Tanoue, Yoshikawa et al. 2002; Berriz, White et al. 2003; Yeh, Karp et al. 2003; Zeeberg, Feng et al. 2003).

The fluidity of the ontologies allows them to adapt to the addition of new terms, the alteration of existing terms, and thus, the constant restructuring of the DAGs. There is also no rigidity or inflexibility when using GO, as any biological discrepancies can be immediately addressed. The ontologies, which were originally designed for cellular organisms, have shown the ability to accommodate terms that were created specifically for plant organisms (*Arabidopsis thaliana*), intracellular parasites (*Plasmodium falciparum*), and, from the work here, viruses. However, before any research can be conducted that is based upon the new viral additions to the Gene Ontology, it is necessary to annotate individual viral gene products with GO terms.

4.0 Annotation of herpesvirus gene products using the Gene Ontology

4.1 Introduction

4.1.1 Annotating HHV-1 with Gene Ontology terms

The Gene Ontology provides a network of terms that, when annotated to gene products from different biological systems, can provide a global view of system interactions within and between organisms. There are three stages to maximising usage of GO's resources: creation of terms (where necessary) to annotate to a chosen organism(s), annotation of all available gene products from the organism, and analysis of research results using new annotations. The first stage encompasses Chapter 3, the second two stages, Chapters 4 and 5 of this thesis.

The completion of a basic viral ontologies framework allows for the assignation of GO numbers to viral gene products. The first step in this process, however, was to determine which viral Open Reading Frames (ORFs) have already been assigned GO numbers by other resources, such as InterPro, before then assigning the remaining ORFs with the appropriate terms. GO annotations, like the ontologies, are not static: they are easily changed in accordance with the emergence of new data.

To fully examine the practicality of the newly created viral GO terms, Human Herpesvirus 1 (HHV-1; Herpes Simplex Virus, HSV-1) was annotated. HHV-1 is a widely studied virus and is thus an ideal candidate for such a project. As HHV-1 is highly characterised, the homologous protein families built by VIDA (Alba, Lee et al. 2001) could then be used to automatically annotate a number of less well studied viruses with GO terms. The annotation of viruses, and their hosts, with GO terms opens the possibility of integrating viral-host gene function analysis utilising the expression data from microarrays (Chapter 5), and other high through-put functional genomics methods.

4.1.2 Human Herpesvirus 1 (HHV-1; Herpes Simplex Virus 1, HSV-1)

Human Herpesvirus 1, or Herpes Simplex Virus 1, is an alphaherpesvirus whose DNA genome is approximately 152kbp in size. The HHV-1 genome encodes for approximately 90 transcriptional units, of which at least 84 encode proteins (Roizman and Knipe 2001). HHV-1 infection occurs most commonly in oral mucosal tissue after direct contact with infectious agent. Primary infection includes replication at the site of infection, but also the infection of sensory neurons that supply the area. The virus is then transported via retrograde axonal transport to the dorsal root ganglion where latency is established.

HHV-1 is uniquely characterised by its neurovirulence, as it not only infects neurones from peripheral sites, but also is able to replicate in the non-dividing neuronal cell. When the virus is reactivated, it replicates and travels back to the site of initial infection (and surrounding area) by axonal transport and replicates in the epithelial tissue at the peripheral site. HHV-1 has been linked to such diseases as oropharyngeal herpes, recurrent labialis, encephalitis, keratitis, and mucocutaneous diseases in immunocompromised hosts.

4.2 Methods

4.2.1 HSV-1 annotation dataset

The complete list of ORFs (open reading frames) from HSV-1 was obtained from VIDA (VIrus DAtabase) (Alba, Lee et al. 2001). VIDA 1.0 was used, which is derived from GenBank release 124.0. ORFs in VIDA are identified by their GI numbers (as assigned by GenBank at the NCBI), and their SWISS-PROT or TrEMBL number (as assigned by SWISS-PROT) (Boeckmann, Bairoch et al. 2003). Existing functional information pertaining to the HPF of each HSV-1 ORF was also extracted from VIDA. A total of 4054 non-redundant, non-fragmented herpesvirus ORFs representing 887 HPFs/Singletons and 237 HHV-1 ORFs (including strain variants) were used.

4.2.2 GO FINDER

The only existing compilation database to include GO annotations and viral ORFs, at the time of this work, was InterPro (Mulder, Apweiler et al. 2003). An algorithm, GO FINDER, was designed to match ORFs from VIDA to InterPro families, and to extract any pre-existing GO annotations. To complete this step for herpesviruses, the herpesvirus ORFs (identified by their GI numbers) must be linked via SWISS-PROT to their IPR (InterPro) family numbers, some of which have been assigned GO numbers. All of this information is then collated into a table (Figure 4.1).

The information from five different files was collated and corresponding entries extracted. The first three files: Herpesviridae_124_functions_updated.txt, Herpesviridae_124_mkpsc_updated.txt, and Herpesviridae_124_gene_table.txt are subsidiary flatfiles produced when updating VIDA with new releases of GenBank. Each contains lists of information pertaining to the ORFs being incorporated into VIDA. The last two files, protein2ipr.dat and ipro2go can be downloaded by anonymous FTP from InterPro and parsed down to the essential information using the scripts SWPRO_EDIT and IPRO_EDIT. After obtaining existing GO annotations from InterPro, herpesviruses ORFs were then mapped onto their HPFs in VIDA to aid in computational annotation of other, related herpesvirus ORFs. These computational first pass annotations identified by GO FINDER were used as a general guideline, with every GO assignment being

subsequently manually curated and reconfirmed according to the most up-to-date information from the Gene Ontology, available literature, and VIDA.

4.2.3 Literature Based Curation

Each ORF in the HSV-1 genome was individually researched in the literature. Papers dating from the original HSV-1 full-length genome sequencing publication (McGeoch, Dolan et al. 1985; McGeoch, Dolan et al. 1986; McGeoch, Dalrymple et al. 1988; Perry and McGeoch 1988) were considered. Papers were searched for functional information, based upon laboratory confirmation, in accordance with Gene Ontology assignment criteria.

4.2.4 GO Term Assignments

Each HSV-1 ORF was assigned one GO number per function/location. There is no limit to the number of GO numbers each ORF can be designated, thereby allowing annotation of multifunctional proteins. As GO is not time specific, some viral terms also have more than one location GO number according to their movement in the cell during viral infection. Each term assignment is accompanied by a documented evidence code (Appendix B), as outlined by the Gene Ontology Consortium, giving weight to the assignment.

4.2.5 Data Availability

The complete HSV-1 GO term annotations (Table 4.2) can also be viewed at the VIDA website: http://www.biochem.ucl.ac.uk/bsm/virus_database/Table2_VIDA_linked.html, and can be accessed at the Gene Ontology website: www.geneontology.org via CVS, FTP and HTTP. All terms have been integrated into the ontologies and can be accessed by searching the database at www.geneontology.org using any of the available search engines.

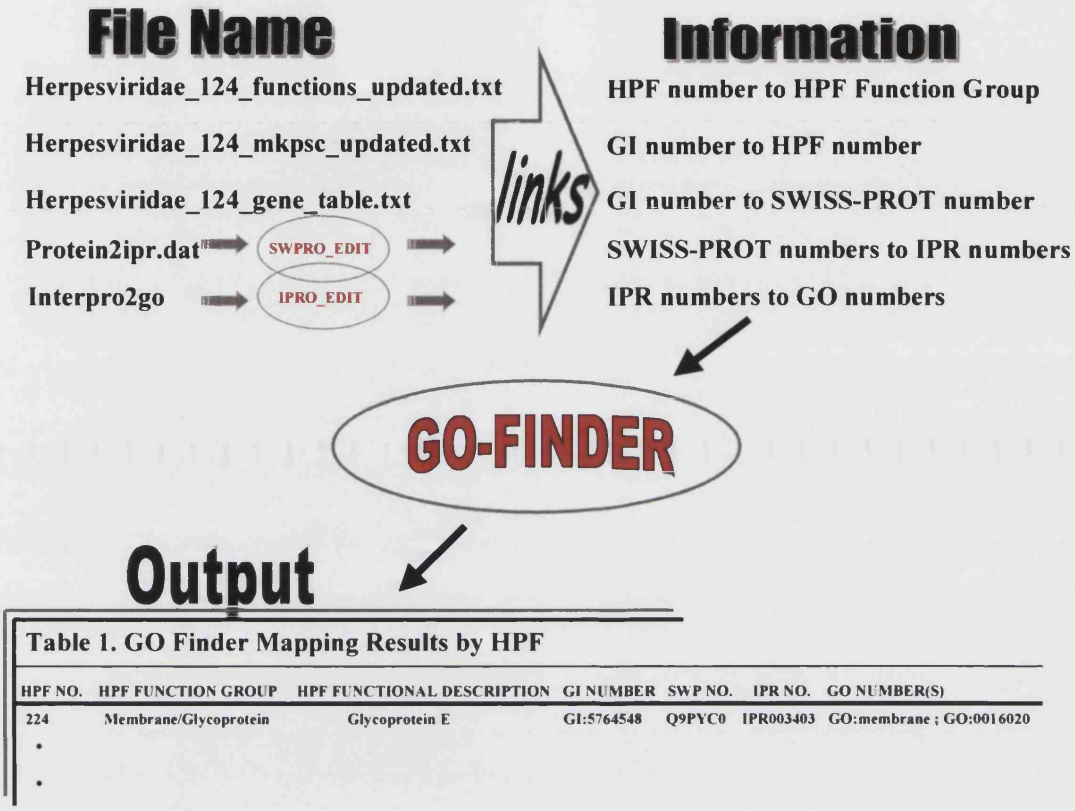


Table 1. GO Finder Mapping Results by HPF

HPF NO.	HPF FUNCTION GROUP	HPF FUNCTIONAL DESCRIPTION	GI NUMBER	SWP NO.	IPR NO.	GO NUMBER(S)
224	Membrane/Glycoprotein	Glycoprotein E	GI:5764548	Q9PYC0	IPR003403	GO:membrane ; GO:0016020
•						
•						

Figure 4.1. GO FINDER. GO FINDER collates the information required from the five files listed above and outputs them in a summarised table organised by VIDA HPF numbers. Two smaller programmes, SWPRO_EDIT and IPRO_EDIT, are run prior to GO FINDER in order to parse Protein2ipr.dat and Interpro2go thereby decreasing GO FINDER's runtime. A full version of the resulting table is included in the text (Table 4.2).

4.3 Results

4.3.1 Annotating HHV-1 using GO terms

4.3.1.1 GO FINDER

Automated GO annotation of as many of the 4504 herpesvirus proteins as possible was performed by the program GO FINDER. By extracting and collating the necessary information from five input files (see Methods) provided by VIDA and InterPro, 37.1% of HSV-1 ORFs, and 32.3% of all herpesvirus ORFs, were assigned GO numbers (Table 4.1). The latter number is determined by inheriting GO assignments to all members of an HPF from which at least one ORF has been previously annotated. All annotations determined using GO FINDER are still subject to further manual confirmation from the literature.

Table 4.2 summarises the results from the program GO FINDER. The table is organized by the HPF number to which each corresponding ORF belongs; these results do not include those GI's that did not have corresponding GO annotations. A number of HPFs are represented by more than one line in the table, demonstrating that within one HPF there are ORFs that have been assigned different GO numbers by InterPro. HPF27, for example, contains 7-transmembrane G-protein coupled receptors (GPCR), two of which were assigned different GO numbers due to their being members of different InterPro families (**IPR000276:Rhodopsin-like GPCR superfamily** and **IPR000355:Chemokine receptor**; Figure 4.2a). The two InterPro families have a parent-child relationship indicating that all of the members of HPF27 can be annotated with GO numbers pertaining to the rhodopsin-like GPCR superfamily (**IPR000276**), but some may also be annotated with the GO numbers pertaining to chemokine receptors (**IPR000355**). Therefore, the functional annotation of HPF27 as 'G-protein coupled receptors' is accurate.

Similarly, three members of HPF29 are also placed in three different, but related, InterPro families (**IPR000719:Protein kinase**, **IPR001245:Tyrosine protein kinase**, and **IPR002290:Serine/threonine protein kinase**; Figure 4.2b). Again, all of the proteins in HPF29 can be annotated with GO numbers associated with the parent family protein kinase (**IPR000719**); however, only those within HPF29 whose specific kinase

mechanism has been identified can be further annotated with the information from the two child families (**IPR001245** and **IPR002290**).

GO FINDER also found instances of HPFs that are assigned different functional groups in VIDA, such as HPF27 and HPF711 (*'host-virus interaction'* and *'unknown'*), but are assigned identical GO numbers. This possibly indicates that results have been published concerning the function of ORFs in HPF 711 (*'unknown'*) since the building of VIDA 2.0 using GenBank release 124.

Many of the unknown ORFs from VIDA have been automatically assigned to the term **GO:0008166:viral replication**. This annotation is an artefact from before the creation of numerous GO terms relating to viral function. Previously, the Gene Ontology contained only the one term relating to viruses causing it to be annotated to all viral proteins. This has been rectified (Chapter 3), and the term **GO:0008166** has since been made obsolete in GO, but is still present in InterPro annotation.

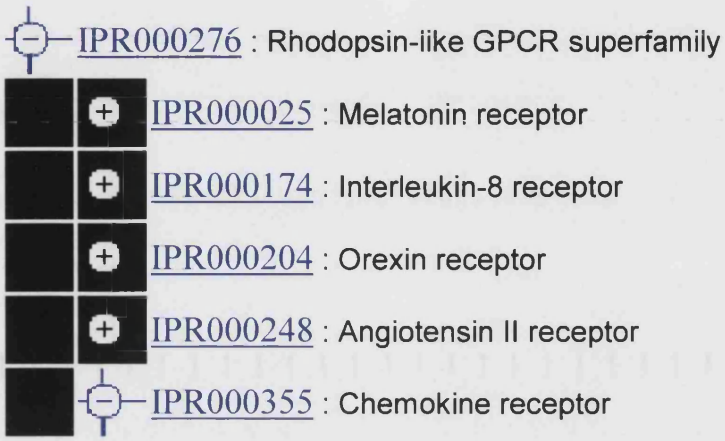
Because of the inferred annotation process through database cross-referencing, and because some errors were immediately obvious, all HSV-1 ORFs were then manually annotated from the literature. This was also necessary to annotate the remaining 63% of HSV-1 ORFs that were not annotated by GO FINDER.

Table 4.1 GO FINDER Based Annotation Statistics.

	VIDA	GO Finder Results	% Annotated
Homologous Protein Families	**985**	**99**	**10.1**
Number of Herpesvirus ORFs	**4504**	**1456**	**32.3**
Number of HSV-1 ORFs	**237†**	**88**	**37.1**

† **Includes strain variant ORFs**

a.



b.

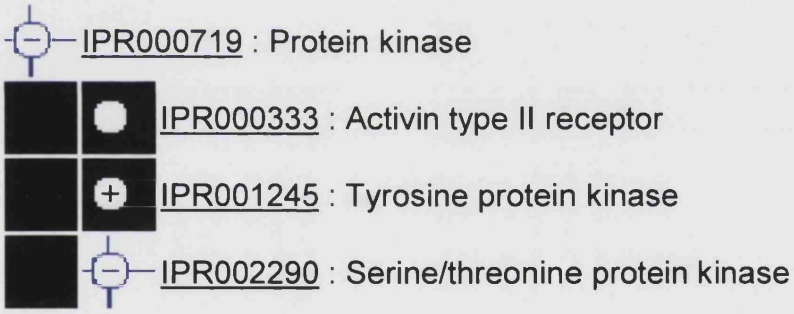


Figure 4.2 The parent-child relationship of InterPro families. a) The tree indicates a parent-child relationship between the two InterPro families (**IPR000276** and **IPR000355**) assigned to HPF27. b) The tree indicates a parent-child relationship between **IPR000719** and families **IPR001245** and **IPR002290**, all three have been assigned to ORFs from HPF29. Tree taken from InterPro resources.

Table 4.2 GO FINDER Results

HPF NO.[1]	HPF FUNCTION GROUP	GI NUMBER[2]	SWP NO.[3]	IPR NO.[4]	GO NUMBER(S)
1	DNA Replication	GI:59530	P04293	IPR002064	GO:DNA binding ; GO:0003677 GO:DNA-directed DNA polymerase ; GO:0003887 GO:3'-5' exonuclease ; GO:0008408 GO:DNA replication ; GO:0006260
2	Nucleotide and nucleic acid metabolism	GI:7384849	Q9IR31	IPR001889	GO:thymidine kinase ; GO:0004797 GO:ATP binding ; GO:0005524 GO:TMP biosynthesis ; GO:0006230
5	DNA Replication	GI:7385024	Q9J3N7	IPR003450	GO:DNA replication origin binding ; GO:0003688 GO:ATP binding ; GO:0005524 GO:viral replication ; GO:0008166
5	DNA Replication	GI:10180713	Q9E6Q7	IPR003593	GO:nucleotide binding ; GO:0000166
8	Nucleotide and nucleic acid metabolism	GI:405185	P52447	IPR003249	GO:uracil-DNA glycosylase ; GO:0004844 GO:DNA repair ; GO:0006281
9	Glycoprotein	GI:804969	Q65587	IPR003600	GO:defense/immunity protein ; GO:0003793 GO:immune response ; GO:0006955
11	Nucleotide and nucleic acid metabolism	GI:695210	Q66641	IPR001616	GO:DNA binding ; GO:0003677 GO:exonuclease ; GO:0004527
13	other	GI:6456008	Q9PWY0	IPR001368	GO:receptor ; GO:0004872
14	Virus structural_protein	GI:221900	P23984	IPR001847	GO:serine-type endopeptidase ; GO:0004252 GO:proteolysis and peptidolysis ; GO:0006508
15	DNA Replication	GI:60018	P09246	IPR000635	GO:single-stranded DNA binding ; GO:0003697 GO:DNA replication ; GO:0006260 GO:nucleus ; GO:0005634
16	DNA Replication	GI:7385028	Q9J3N5	IPR003840	GO:helicase ; GO:0004386 GO:ATP binding ; GO:0005524 GO:viral replication ; GO:0008166
18	other	GI:330792	P28969	IPR003499	GO:DNA packaging ; GO:0006323
20	Glycoprotein	GI:405175	P52449	IPR000785	GO:molecular_function unknown ; GO:0005554 GO:membrane ; GO:0016020
22	other	GI:971317	Q65567	IPR003499	GO:DNA packaging ; GO:0006323
24	Nucleotide and nucleic acid metabolism	GI:437736	P50643	IPR000788	GO:ribonucleoside-diphosphate reductase ; GO:0004748 GO:DNA replication ; GO:0006260 GO:ribonucleoside-diphosphate reductase ; GO:0005971
26	Glycoprotein	GI:459194	Q65530	IPR003404	GO:membrane ; GO:0016020
27	other	GI:5929959	Q9QEV2	IPR000276	GO:G-protein coupled receptor ; GO:0004930

					GO:membrane ; GO:0016020
27	other	GI:8671563	Q9IP69	IPR000355	GO:chemokine receptor ; GO:0004950 GO:membrane ; GO:0016020
29	other	GI:1209026	Q67670	IPR001245	GO:protein tyrosine kinase ; GO:0004713 GO:ATP binding ; GO:0005524 GO:protein phosphorylation ; GO:0006468
29	other	GI:1718289	P88924	IPR000719	GO:protein kinase ; GO:0004672 GO:ATP binding ; GO:0005524 GO:protein phosphorylation ; GO:0006468
29	other	GI:1869835	P89436	IPR002290	GO:protein serine/threonine kinase ; GO:0004674 GO:ATP binding ; GO:0005524 GO:protein phosphorylation ; GO:0006468
33	Nucleotide and nucleic acid metabolism	GI:703073	Q69279	IPR000358	GO:ribonucleoside-diphosphate reductase ; GO:0004748 GO:deoxyribonucleoside diphosphate metabolism ; GO:0009186
36	Glycoprotein	GI:540201	Q69472	IPR002567	GO:cell adhesion molecule ; GO:0005194 GO:cell adhesion ; GO:0007155 GO:membrane ; GO:0016020
40	other	GI:331282	Q00095	IPR000719	GO:protein kinase ; GO:0004672 GO:ATP binding ; GO:0005524 GO:protein phosphorylation ; GO:0006468
40	other	GI:331281	Q00094	IPR002290	GO:protein serine/threonine kinase ; GO:0004674 GO:ATP binding ; GO:0005524 GO:protein phosphorylation ; GO:0006468
43	Nucleotide and nucleic acid metabolism	GI:1718307	P88942	IPR001428	GO:pseudouridylate synthase ; GO:0004730 GO:tRNA metabolism ; GO:0006399
44	Virus structural_protein	GI:1718332	P88964	IPR000728	GO:enzyme ; GO:0003824
48	other	GI:1483517	Q82171	IPR002927	GO:transcription regulation ; GO:0006355
51	Nucleotide and nucleic acid metabolism	GI:695210	Q66641	IPR001616	GO:DNA binding ; GO:0003677 GO:exonuclease ; GO:0004527
52	Transcription	GI:330320	Q69113	IPR003174	GO:DNA binding ; GO:0003677 GO:transcription activating factor ; GO:0003710 GO:transcription regulation ; GO:0006355
55	Unknown	GI:9800360	Q9DW56	IPR003360	GO:viral replication ; GO:0008166
59	DNA Replication	GI:7673121	Q9IBW5	IPR003450	GO:DNA replication origin binding ; GO:0003688 GO:ATP binding ; GO:0005524 GO:viral replication ; GO:0008166
59	DNA Replication	GI:10180713	Q9E6Q7	IPR003593	GO:nucleotide binding ; GO:0000166
82	Unknown	GI:2647982	O39921	IPR003360	GO:viral replication ; GO:0008166
85	other	GI:2149636	O12000	IPR000276	GO:G-protein coupled receptor ; GO:0004930 GO:membrane ; GO:0016020
89	other	GI:4494967	Q9WRN9	IPR001346	GO:transcription factor ; GO:0003700

					GO:transcription regulation ; GO:0006355 GO:nucleus ; GO:0005634
92	Nucleotide and nucleic acid metabolism	GI:695245	Q89940	IPR000398	GO:thymidylate synthase ; GO:0004799 GO:dTMP biosynthesis ; GO:0006231
95	other	GI:59182	Q89558	IPR001204	GO:inorganic phosphate transporter ; GO:0005315 GO:phosphate transport ; GO:0006817 GO:membrane ; GO:0016020
99	Unknown	GI:2746235	O57302	IPR003360	GO:viral replication ; GO:0008166
104	DNA Replication	GI:1869865	P89463	IPR003202	GO:DNA binding ; GO:0003677 GO:DNA replication ; GO:0006260
107	Virus structural_protein	GI:2370241	O40637	IPR001847	GO:serine-type endopeptidase ; GO:0004252 GO:proteolysis and peptidolysis ; GO:0006508
109	Transcription	GI:4377600	Q69551	IPR003360	GO:viral replication ; GO:0008166
121	Unknown	GI:1780954	P09701	IPR003360	GO:viral replication ; GO:0008166
140	other	GI:7542407	Q9J4B8	IPR000098	GO:cytokine ; GO:0005125 GO:immune response ; GO:0006955
141	Nucleotide and nucleic acid metabolism	GI:60322	P09503	IPR001796	GO:dihydrofolate reductase ; GO:0004146 GO:glycine biosynthesis ; GO:0006545 GO:nucleotide biosynthesis ; GO:0009165
145	Transcription	GI:5733532	Q9QJ47	IPR003360	GO:viral replication ; GO:0008166
146	Nucleotide and nucleic acid metabolism	GI:235434	P30007	IPR001428	GO:pseudouridylate synthase ; GO:0004730 GO:tRNA metabolism ; GO:0006399
156	Unknown	GI:2746235	O57302	IPR003360	GO:viral replication ; GO:0008166
159	Unknown	GI:2647982	O39921	IPR003360	GO:viral replication ; GO:0008166
167	other	GI:529230	Q69087	IPR001847	GO:serine-type endopeptidase ; GO:0004252 GO:proteolysis and peptidolysis ; GO:0006508
171	Virus structural_protein	GI:5733553	Q9QJ30	IPR001847	GO:serine-type endopeptidase ; GO:0004252 GO:proteolysis and peptidolysis ; GO:0006508
172	Unknown	GI:330827	P28936	IPR001847	GO:serine-type endopeptidase ; GO:0004252 GO:proteolysis and peptidolysis ; GO:0006508
174	Transcription	GI:1185442	Q67633	IPR001871	GO:transcription factor ; GO:0003700 GO:transcription regulation ; GO:0006355
176	Transcription	GI:7158288	Q9JE49	IPR001083	GO:transcription factor ; GO:0003700 GO:transcription regulation ; GO:0006355 GO:nucleus ; GO:0005634
177	Unknown	GI:221458	Q01350	IPR003360	GO:viral replication ; GO:0008166
185	Unknown	GI:1139610	P52523	IPR003360	GO:viral replication ; GO:0008166
191	Glycoprotein	GI:4996008	Q9WT44	IPR003600	GO:defense/immunity protein ; GO:0003793 GO:Immune response ; GO:0006955
194	Glycoprotein	GI:1139686	Q69512	IPR003600	GO:defense/immunity protein ; GO:0003793 GO:immune response ; GO:0006955

195	other	GI:7107124	Q9J7C4	IPR000098	GO:cytokine ; GO:0005125 GO:immune response ; GO:0006955
214	other	GI:331224	P15443	IPR002290	GO:protein serine/threonine kinase ; GO:0004674 GO:ATP binding ; GO:0005524 GO:protein phosphorylation ; GO:0006468
219	Unknown	GI:2746316	O56303	IPR003360	GO:viral replication ; GO:0008166
224	Glycoprotein	GI:5764548	Q9PYC0	IPR003404	GO:membrane ; GO:0016020
225	other	GI:7330003	Q9J2M1	IPR001811	GO:chemokine ; GO:0008009 GO:immune response ; GO:0006955
226	Virus structural_protein	GI:1419024	Q83417	IPR001847	GO:serine-type endopeptidase ; GO:0004252 GO:proteolysis and peptidolysis ; GO:0006508
231	Unknown	GI:6552718	Q9Q6Z7	IPR000276	GO:G-protein coupled receptor ; GO:0004930 GO:membrane ; GO:0016020
238	other	GI:1167494	P16046	IPR001847	GO:serine-type endopeptidase ; GO:0004252 GO:proteolysis and peptidolysis ; GO:0006508
242	Glycoprotein	GI:2246507	O40948	IPR003599	GO:defense/immunity protein ; GO:0003793 GO:immune response ; GO:0006955
243	other	GI:7330050	Q9J2J5	IPR001346	GO:transcription factor ; GO:0003700 GO:transcription regulation ; GO:0006355 GO:nucleus ; GO:0005634
248	other	GI:854030	P52382	IPR000276	GO:G-protein coupled receptor ; GO:0004930 GO:membrane ; GO:0016020
250	Unknown	GI:469956	Q69581	IPR000634	GO:amino acid metabolism ; GO:0006520
253	Unknown	GI:853967	Q89660	IPR000564	GO:iron-sulfur electron transfer carrier ; GO:0008042 GO:electron transport ; GO:0006118
257	Transcription	GI:808657	Q69127	IPR001871	GO:transcription factor ; GO:0003700 GO:transcription regulation ; GO:0006355
273	Unknown	GI:9800260	Q9DWF5	IPR003360	GO:viral replication ; GO:0008166
287	Unknown	GI:4219031	Q9YQZ6	IPR002064	GO:DNA binding ; GO:0003677 GO:DNA-directed DNA polymerase ; GO:0003887 GO:3'-5' exonuclease ; GO:0008408 GO:DNA replication ; GO:0006260
293	DNA replication	GI:4219031	Q9YQZ6	IPR002064	GO:DNA binding ; GO:0003677 GO:DNA-directed DNA polymerase ; GO:0003887 GO:3'-5' exonuclease ; GO:0008408 GO:DNA replication ; GO:0006260
315	other	GI:1562494	Q98823	IPR003573	GO:cytokine ; GO:0005125 GO:immune response ; GO:0006955
315	other	GI:2246551	O40918	IPR003574	GO:interleukin-6 receptor ligand ; GO:0005138 GO:immune response ; GO:0006955
316	other	GI:4494908	Q9WRU4	IPR003006	GO:immunoglobulin ; GO:0003823
317	Unknown	GI:4996078	Q9WSZ6	IPR001257	GO:viral replication ; GO:0008166
321	Unknown	GI:1167932	Q68399	IPR001811	GO:chemokine ; GO:0008009

					GO:immune response ; GO:0006955
324	Virus structural_protein	GI:7673137	Q9IBU9	IPR001847	GO:serine-type endopeptidase ; GO:0004252 GO:proteolysis and peptidolysis ; GO:0006508
328	other	GI:11095831	Q9E1J0	IPR000379	GO:enzyme ; GO:0003824
344	Unknown	GI:1167926	Q68393	IPR003599	GO:defense/immunity protein ; GO:0003793 GO:immune response ; GO:0006955
345	Virus structural_protein	GI:535659	Q69223	IPR001847	GO:serine-type endopeptidase ; GO:0004252 GO:proteolysis and peptidolysis ; GO:0006508
367	Unknown	GI:695174	Q66607	IPR003599	GO:defense/immunity protein ; GO:0003793 GO:immune response ; GO:0006955
370	other	GI:1778606	P88968	IPR001811	GO:chemokine ; GO:0008009 GO:immune response ; GO:0006955
387	other	GI:606851	Q83145	IPR001811	GO:chemokine ; GO:0008009 GO:immune response ; GO:0006955
402	Glycoprotein	GI:6435841	Q9QC01	IPR003599	GO:defense/immunity protein ; GO:0003793 GO:immune response ; GO:0006955
424	other	GI:695246	Q66673	IPR000276	GO:G-protein coupled receptor ; GO:0004930 GO:membrane ; GO:0016020
473	Unknown	GI:4219046	Q9YQY1	IPR001525	GO:DNA binding ; GO:0003677 GO:DNA (cytosine-5-)-methyltransferase ; GO:0003886 GO:DNA methylation ; GO:0006306
486	other	GI:4877816	Q9XR29	IPR003600	GO:defense/immunity protein ; GO:0003793 GO:immune response ; GO:0006955
496	other	GI:3152729	O71294	IPR001346	GO:transcription factor ; GO:0003700 GO:transcription regulation ; GO:0006355 GO:nucleus ; GO:0005634
520	Unknown	GI:331280	Q00103	IPR000822	GO:transcription factor ; GO:0003700 GO:transcription regulation ; GO:0006355 GO:nucleus ; GO:0005634
531	other	GI:3873223	Q9YVA9	IPR002473	GO:cytokine ; GO:0005125 GO:immune response ; GO:0006955
539	other	GI:331257	Q00139	IPR002884	GO:subtilase ; GO:0004289 GO:proteolysis and peptidolysis ; GO:0006508
569	Nucleotide and nucleic acid metabolism	GI:331259	P28893	IPR001428	GO:pseudouridylate synthase ; GO:0004730 GO:tRNA metabolism ; GO:0006399
573	Unknown	GI:4219027	Q9YR00	IPR000719	GO:protein kinase ; GO:0004672 GO:protein phosphorylation ; GO:0006468 GO:ATP binding ; GO:0005524
587	other	GI:59457	P08560	IPR003600	GO:defense/immunity protein ; GO:0003793 GO:immune response ; GO:0006955
595	Unknown	GI:2625047	O39237	IPR003573	GO:cytokine ; GO:0005125 GO:immune response ; GO:0006955

616	Unknown	GI:2558898	O38018	IPR001428	GO:pseudouridylate synthase ; GO:0004730 GO:tRNA metabolism ; GO:0006399
626	Unknown	GI:330546	P24909	IPR003360	GO:viral replication ; GO:0008166
711	Unknown	GI:1780855	P16751	IPR000276	GO:G-protein coupled receptor ; GO:0004930 GO:membrane ; GO:0016020
732	Glycoprotein	GI:330353	P03218	IPR003599	GO:defense/immunity protein ; GO:0003793 GO:immune response ; GO:0006955
734	other	GI:1334917	P03228	IPR003600	GO:defense/immunity protein ; GO:0003793 GO:immune response ; GO:0006955
755	other	GI:8096689	Q9IZK2	IPR003406	GO:acetylglucosaminyltransferase ; GO:0008375 GO:membrane ; GO:0016020
762	Unknown	GI:59629	P16760	IPR003360	GO:viral replication ; GO:0008166
868	Unknown	GI:808657	Q69127	IPR001871	GO:transcription factor ; GO:0003700 GO:transcription regulation ; GO:0006355

1. Homologous protein family (HPF) number
2. GenBank Identifier (GI) number
3. SWISS-PROT number
4. Interpro number

4.3.1.2 Manual Annotation

The GO FINDER results were used only as initial guidelines for annotation of HSV-1 gene products. Annotation of the genome was achieved by manually mining the literature for relevant information pertaining to each ORF. While providing a useful base from which to proceed, GO FINDER results could not solely be used in individual gene product annotation as they are based upon the annotation of families of proteins: both the HPFs in VIDA and the IPRs in InterPro. When annotating individual gene products it is important not to limit their functional descriptions to those of their family. As seen with HPF27 and HPF29, there can exist within one family certain distinctions (such as tyrosine vs. serine/threonine kinase mechanisms) that would make it necessary to annotate different members with different, more specific, GO terms. It is also preferable to utilise, where necessary, the newly created virus related GO terms in this exercise, which are not incorporated into GO FINDER's results as they were not available for annotation by InterPro at the time.

All HSV-1 ORFs have been annotated using existing viral, cellular, and newly created viral GO terms (Table 4.3). Each ORF has at least three GO numbers: one from each of the ontologies: biological process, molecular function, cellular component, even if the term is 'unknown' (i.e. UL7). Where the literature indicated that a viral ORF functioned in different cellular compartments, the ORF was annotated with all relevant terms, as often there was insufficient evidence available to determine the exact functional location. In the case of multifunctional proteins, however, the protein may have more than one site of functional activation. In these cases, multiple general annotations may not prove to be incorrect. Nevertheless, the dynamic nature of GO allows all annotations to be altered as new knowledge emerges concerning each ORF.

The annotations in Table 4.3 are conservative; if an attribute was not found mentioned in the literature, it was not annotated to the gene product. This is to avoid incorrect 'assumptions' being perpetuated through GO annotation. Therefore, DNA polymerase (UL30) was not annotated with the GO term **GO:0003677:DNA binding** as the two were not mentioned together in the literature, although it is commonly accepted that DNA polymerase binds to DNA in order to function.

4.3.1.3 Using the 'Unknown' GO Term

The GO Consortium is very clear about the usage of the three 'unknown' GO terms **(GO:0008372:cellular component unknown, GO:0005554:molecular function unknown, GO:0000004:biological function unknown)**. Annotation of an ORF with the 'unknown' term in any of the ontologies must only occur *after* the literature (and any other resources) has been researched and it becomes clear that the function/process/location of the ORF is unknown not only to the annotator, but to the scientific community at large. This is to ensure that all gene product assignments to the 'unknown' terms are truly unknown, establishing a base threshold of knowledge up from which to work. Therefore, all terms in Table 4.3 annotated with any of the three 'unknown' terms are unknown in respect to that ontology in the literature.

4.3.1.4 Evidence Codes

The Gene Ontology annotation process includes a quality control check upon each GO term annotation made to a gene product. Each GO term assignment to a gene product must be accompanied by a reference, which may come from the literature, another database, or computational analysis; and an evidence code, which indicates how the assignment was determined (Table 4.4). As all annotation in this study was completed by searching the available literature, the only three evidence codes used were Traceable Author Statement (TAS), Non-traceable Author Statement (NAS), and No Biological Data Available (ND). All evidence codes and references for the annotations in Table 4.3 can be found in Appendix B.

Table 4.3 HSV-1 Genome GO Annotation

ORF	PROTEIN PRODUCT	BIOLOGICAL PROCESS	MOLECULAR FUNCTION	CELLULAR COMPONENT
RL1	ICP34.5, g134.5	BP:viral inhibition of host cell protein biosynthesis shutoff ; 0046773 BP:viral inhibition of cell cycle arrest ; 0046792 BP:phosphatase regulator ; 0019208	MF:protein binding ; 0005515	CC:unknown ; 0008372
RL2	a0, ICP0	BP:viral perturbation of cell cycle control ; 0019055 BP:viral inhibition of extracellular antiviral response ; 0019053 BP:ubiquitin cycle ; 0006512 BP:virus-host cell process manipulation ; 0019054 BP:viral transcription ; 0019083 BP:histone deacetylase inhibitor ; 0046811	MF:protein binding ; 0005515	CC:dense nuclear body ; 0046818
UL1	gL	BP:virion penetration ; 0019063 BP:viral-induced cell-cell fusion ; 0019064	MF:unknown ; 0005554	CC:viral envelope ; 0019031
UL2		BP:DNA repair ; 0006281	MF:uracil-DNA glycosylase ; 0004844	CC:nucleus ; 0005634
UL3		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:dense nuclear body ; 0046818
UL4		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:virion ; 0019012 CC:dense nuclear body ; 0046818
UL5		BP:viral genome replication ; 0019079	MF:DNA helicase ; 0003678 MF:ATPase ; 0016887 MF:DNA binding ; 0003677 MF:ATP binding ; 0005524	CC:replication compartment ; 0046809 CC:viral replication complex ; 0019034
UL6		BP:viral DNA genome packaging ; 0019073	MF:unknown ; 0005554	CC:nucleus ; 0005634 CC:viral portal complex ; 0046798
UL7		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:unknown ; 0008372
UL8		BP:viral intracellular protein transport ; 0019060 BP:positive regulation of DNA replication ; 0045740	MF:protein binding ; 0005515	CC:viral replication complex ; 0019034
UL8.5		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:unknown ; 0008372
UL9		BP:viral genome replication ; 0019079	MF:ATPase ; 0016887 MF:ATP dependent DNA helicase ; 0004003 MF:DNA replication origin binding ; 0003688	CC:replication compartment ; 0046809 CC:nucleus ; 0005634

			MF:ATP binding ; 0005524 MF:DNA binding ; 0003677	
UL9.5		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:unknown ; 0008372
UL10	gM	BP:viral spread within host, cell to cell; 0046740	MF:unknown ; 0005554	CC:plasma membrane ; 0005886 CC:viral envelope ; 0019031
UL10.5		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:unknown ; 0008372
UL11		BP:viral capsid envelopment ; 0046744 BP:viral egress ; 0046788 BP:viral capsid re-envelopment ; 0046745 BP:viral intracellular protein transport ; 0019060	MF:unknown ; 0005554	CC:nucleus ; 0005634 CC:dense nuclear body ; 0046818 CC:nuclear inner membrane ; 0005637 CC:endomembrane system ; 0012505 CC:viral tegument ; 0019033
UL12		BP:viral genome maturation ; 0019070 BP:nuclear egress of viral procapsid ; 0046802	MF:exonuclease ; 0004527 MF:DNA binding ; 0003677 MF:endonuclease ; 0004519	CC:nucleus ; 0005634
UL12.5		BP: exonucleolytic degradation of mRNA ; 0000291 BP:endonucleolytic mRNA decay ; 0000294	MF:endonuclease ; 0004519 MF:exonuclease ; 0004527	CC:isohedral viral capsid ; 0019030
UL13		BP:induction of apoptosis by virus ; 0019051 BP:protein phosphorylation ; 0006468	MF:protein kinase ; 0004672	CC:virion ; 0019012
UL14		BP:viral spread within host, cell to cell; 0046740	MF:unknown ; 0005554	CC:cytoplasm ; 0005737 CC:virion tegument ; 0019033 CC:dense nuclear body ; 0046818
UL15		BP:viral DNA genome packaging ; 0019073 BP:viral DNA cleavage ; 0019071	MF:ATP binding ; 0005524	CC: viral procapsid ; 0046729 CC:nucleus ; 0005634 CC:replication compartment ; 0046809
UL16		BP:viral DNA genome packaging ; 0019073 BP:viral DNA cleavage ; 0019071	MF:unknown ; 0005554	CC:assemblon ; 0046808 CC:cytoplasm ; 0005737 CC:nucleus ; 0005634 CC:virion ; 0019012 CC:replication compartment ; 0046809
UL17		BP:viral DNA cleavage ; 0019071 BP:viral DNA genome packaging ; 0019073 BP:nuclear viral capsid transport ; 0046742	MF:unknown ; 0005554	CC:nucleus ; 0005634 CC:viral tegument ; 0019033
UL15.5		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:unknown ; 0008372

UL18	VP23	BP:viral DNA cleavage ; 0019071 BP:viral DNA genome packaging ; 0019073	MF:unknown ; 0005554	CC:nucleus ; 0005634 CC:isohedral viral capsid ; 0019030
UL19	VP5; ICP5	BP:unknown ; 0000004	MF:unknown ; 0005554	CC:nucleus ; 0005634 CC:isohedral viral capsid ; 0019030 CC:capsomere ; 0046727
UL20		BP:viral intracellular protein traffic ; 0019060 BP:viral egress ; 0046788	MF:unknown ; 0005554	CC:Golgi stack ; 0005795 CC:viral envelope ; 0019031 CC:nuclear membrane ; 0005635 CC:virion transport vesicle ; 0046816
UL20.5		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:dense nuclear body ; 0046818
UL21		BP:microtubule cytoskeleton organization and biogenesis ; 0000226 BP:intracellular virion transport ; 0046795 BP:intracellular viral capsid transport ; 0046801 BP:microtubule polymerization ; 0046785	MF:microtubule associated protein ; 0005875 MF:microtubule binding ; 0008017	CC:cytoplasm ; 0005737 CC:viral tegument ; 0019033
UL22	gH	BP:viral egress ; 0046788 BP:viral spread within host, cell to cell; 0046740 BP:virion penetration ; 0019063 BP:viral-induced cell-cell fusion ; 0006948	MF:unknown ; 0005554	CC:viral envelope ; 0019031
UL23	ICP36	BP:reactivation of latent virus ; 0019046	MF:nucleoside kinase ; 0019206	CC:unknown ; 0008372
UL24		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:membrane ; 0016020
UL25		BP:viral DNA genome packaging ; 0019073 BP:genome retention in viral capsid ; 0046815 BP:virion penetration ; 0019063	MF:DNA binding ; 0003677	CC:nucleus ; 0005634 CC:isohedral viral capsid ; 0019030 CC:cytoplasm ; 0005737 CC: viral procapsid ; 0046729
UL26	VP24 VP21	BP:proteolysis and peptidolysis ; 0006508 BP:viral scaffold assembly and maintenance ; 0046807	MF:serine-type endopeptidase ; 0004252	CC:nucleus ; 0005634 CC:viral scaffold ; 00464806
UL26.5	ICP35 (VP22a)	BP:nuclear localisation of viral capsid precursors ; 0046752 BP:viral scaffold assembly and maintenance ; 0046807	MF:unknown ; 0005554	CC:nucleus ; 0005634 CC:cytoplasm ; 0005737 CC:viral scaffold ; 00464806
UL27	gB, VP7	BP:viral-induced cell-cell fusion ; 0019064	MF:host cell extracellular matrix binding ; 0046810 MF:viral-cell fusion molecule ; 0019039	CC:viral envelope ; 0019031
UL27.5		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:cytoplasm ; 0005737

UL28	ICP18.5	BP:viral DNA genome packaging ; 0019073 BP:viral DNA cleavage ; 0019071	MF:unknown ; 0005554	CC:cytoplasm ; 0005737 CC:nucleus ; 0005634
UL29	ICP8	BP:viral genome replication ; 0019079 BP:recruitment of helicase-primase complex to DNA lesions ; 0046799 BP:viral replication complex formation and maintenance ; 0046786	MF:single-stranded DNA binding ; 0003697	CC:viral replication complex ; 0019034 CC:replication compartment ; 0046809
UL30		BP:viral genome replication ; 0019079	MF:DNA-directed DNA polymerase ; 0003887 MF:3'-5' exonuclease ; 0008408	CC:replication compartment ; 0046809
UL31		BP:viral DNA cleavage ; 0019071 BP:viral DNA genome packaging ; 0019073 BP:inner nuclear membrane viral budding during viral capsid envelopment ; 0046771	MF:unknown ; 0005554	CC:nucleus ; 0005634
UL32		BP:viral DNA cleavage ; 0019071 BP:viral DNA genome packaging ; 0019073 BP:intranuclear viral capsid transport ; 0046742	MF:unknown ; 0005554	CC:cytoplasm ; 0005737 CC:nucleus ; 0005634 CC:replication compartment ; 0046809
UL33		BP:viral DNA cleavage ; 0019071 BP:viral DNA genome packaging ; 0019073	MF:unknown ; 0005554	CC:cytoplasm ; 0005737 CC:replication compartment ; 0046809 CC:nucleus ; 0005634
UL34		BP:inner nuclear membrane viral budding during viral capsid envelopment ; 0046771	MF:unknown ; 0005554	CC:virion ; 0019012 CC:nuclear membrane ; 0005635 CC:nuclear membrane lumen ; 0005641
UL35	VP26	BP:viral procapsid maturation ; 0046797	MF:unknown ; 0005554	CC:isohedral viral capsid ; 0019030 CC:nucleus ; 0005634
UL36	ICP1-2	BP:viral egress ; 0046788 BP:viral budding ; 0019078 BP:viral uncoating ; 0019061 BP:viral particle maturation ; 0019075	MF:unknown ; 0005554	CC:nucleus ; 0005634 CC:cytoplasm ; 0005737 CC:viral tegument ; 0019033
UL37		BP:nuclear membrane viral budding during viral capsid envelopment ; 0046749 BP:viral particle maturation ; 0019075 BP:cytoplasmic viral capsid transport ; 0046743 BP:viral egress ; 0046788 BP:viral capsid re-envelopment ; 0046745	MF:unknown ; 0005554	CC:viral tegument ; 0019033 CC:cytoplasm ; 0005737 CC:nucleus ; 0005634
UL38	ICP32/VP19C	BP:nuclear localisation of viral capsid precursors ; 0046752 BP:viral capsid assembly ; 0019069	MF:DNA binding ; 0003677	CC:isohedral viral capsid ; 0019030 CC:nucleus ; 0005634

UL39	ICP6	BP:deoxyribonucleoside diphosphate metabolism ; 0009186 BP:viral genome replication ; 0019079 BP: passive viral induction of humoral immune response ; 0046733 BP:autophosphorylation ; 0046777	MF:protein kinase ; 0004672 MF:ribonucleoside-diphosphate reductase ; 0004748	CC:ribonucleoside-diphosphate reductase complex ; 0005971
UL40		BP:deoxyribonucleoside diphosphate metabolism ; 0009186 BP:viral genome replication ; 0019079	MF:ribonucleoside-diphosphate reductase ; 0004748	CC:ribonucleoside-diphosphate reductase complex ; 0005971
UL41	vhs	BP:transcription regulation ; 0006355 BP:viral inhibition of MHC class 1 cell surface presentation ; 0046776 BP:viral inhibition of host cytokine production ; 0046775 BP:viral inhibition of intracellular interferon activity ; 0046774 BP:viral perturbation of polysomes ; 0046783 BP:viral host cell process manipulation ; 0019054	MF:mRNA catabolism, endonucleolytic ; 0000294	CC:viral tegument ; 0019033 CC:cytoplasm ; 0005737
UL42		BP:viral genome replication ; 0019079	MF:DNA polymerase processivity factor ; 0030337 MF:DNA binding ; 0003677	CC:replication compartment ; 0046809
UL43		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:unknown ; 0008372
UL43.5		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:assemblon ; 0046808
UL44	gC, VP7.5	BP:enhancement of virulence ; 0046800 BP:viral inhibition of host complement neutralisation ; 0046791 BP:virion attachment ; 0019062	MF:unknown ; 0005554	CC:viral envelope ; 0019031
UL45		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:viral envelope ; 0019031
UL46	VP11/12	BP:unknown ; 0000004	MF:unknown ; 0005554	CC:nuclear membrane lumen ; 0005641 CC:cytoplasm ; 0005737 CC:viral tegument ; 0019033 CC:plasma membrane ; 0005886
UL47	VP13/14	BP:unknown ; 0000004	MF:unknown ; 0005554	CC:viral tegument ; 0019033 CC:nucleus ; 0005634 CC:cytoplasm ; 0005737
UL48	VP16;ICP25, aTIF	BP:immediate early viral mRNA transcription ; 0019085 BP:regulation of viral transcription ; 0046782 BP:viral egress ; 0046788	MF:transcription activator ; 0016563 MF:protein binding ; 0005515	CC:viral tegument ; 0019033 CC:nucleus ; 0005634 CC:cytoplasm ; 0005737

UL49	VP22	BP:viral spread within host, cell to cell; 0046740	MF:chromatin binding ; 0003682	CC:viral tegument ; 0019033 CC:nucleus ; 0005634 CC:cytoplasm ; 0005737
UL49.5		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:viral envelope ; 0019031
UL50		BP:tRNA metabolism ; 0006399 BP:nucleotide metabolism ; 0009117	MF:dUTPase pyrophosphatase ; 0004170	CC:unknown ; 0008372
UL51		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:nuclear membrane lumen ; 0005641 CC:nucleus ; 0005634 CC:virion ; 0019012
UL52		BP:viral genome replication ; 0019079	MF:DNA binding ; 0003677 MF:DNA helicase ; 0003678 MF:DNA primase ; 0003896	CC:viral replication complex ; 0019034 CC:replication compartment ; 0046809
UL53	gK	BP:viral egress ; 0046788 BP:intracellular viral capsid transport ; 0046801	MF:unknown ; 0005554	CC:nuclear membrane lumen ; 0005641 CC:nucleus ; 0005634 CC:Golgi apparatus ; 0005794 CC:endoplasmic reticulum membrane ; 0005789 CC:viral envelope ; 0019031 CC:nuclear membrane ; 0005635
UL54	a27;ICP27	BP:viral perturbation of host cell transcription ; 0019056 BP:intronless viral mRNA-nucleus export ; 0046784 BP:regulation of viral transcription ; 0046782 BP:negative regulation of viral genome replication ; 0045071 BP:viral inhibition of expression of host genes with introns ; 0046779 BP:positive regulation of viral genome replication ; 0045070 BP:viral dispersion of host splicing factors ; 0046781 BP:viral inhibition of host mRNA splicing ; 0046780	MF:transcription repressor ; 0016564 MF:RNA binding ; 0003723	CC:cytoplasm ; 0005737 CC:nucleus ; 0005634 CC:replication compartment ; 0046809
UL55		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:assemblon ; 0046808 CC:nucleus ; 0005634
UL56		BP:reduction of virulence ; 0046803	MF:unknown ; 0005554	CC:virion ; 0019012 CC:nucleus ; 0005634
RS1	a4, ICP4	BP: viral transcription ; 0019083 BP:cell cycle arrest ; 0007050 BP:viral perturbation of cell cycle control ; 0019055 BP:regulation of viral transcription ; 0046782	MF:DNA binding ; 0003677 MF:transcription regulator ; 0030528	CC:replication compartment ; 0046809 CC:nucleus ; 0005634

Us1	a22, ICP22	BP:virus induced modification of host RNA polymerase II ; 0046793	MF:unknown ; 0005554	CC:dense nuclear body ; 0046818 CC:nucleus ; 0005634 CC:cytoplasm ; 0005737
Us1.5		BP:induction of apoptosis by virus ; 0019051	MF:unknown ; 0005554	CC:unknown ; 0008372
Us2		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:nucleus ; 0005634 CC:viral tegument ; 0019033
Us3		BP:protein amino acid phosphorylation ; 0006468 BP:viral inhibition of apoptosis ; 0019050	MF:protein kinase ; 0004672	CC:unknown ; 0008372
Us4	gG	BP:virion attachment ; 0019062	MF:unknown ; 0005554	CC:viral envelope ; 0019031
Us5	gJ	BP:viral inhibition of apoptosis ; 0019050	MF:unknown ; 0005554	CC:unknown ; 0008372
Us6	gD, VP17/18	BP:viral-induced cell-cell fusion ; 0019064 BP:negative regulation of apoptosis ; 0043066	MF:host cell surface receptor binding ; 0046789 MF:virion attachment, binding of host cell surface co-receptor ; 0046814	CC:viral envelope ; 0019031
Us7	gI	BP:viral spread within host, cell to cell; 0046740	MF:unknown ; 0005554	CC:Golgi apparatus ; 0005794 CC:viral envelope ; 0019031
Us8	gE	BP:viral spread within host, cell to cell; 0046740	MF:unknown ; 0005554	CC:Golgi apparatus ; 0005794 CC:viral envelope ; 0019031
Us8.5		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:nucleolus ; 0005730
Us9		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:viral tegument ; 0019033 CC:viral envelope ; 0019031
Us10		BP:unknown ; 0000004	MF:unknown ; 0005554	CC:nucleus ; 0005634 CC:isohedral viral capsid ; 0019030 CC:viral tegument ; 0019033
Us11		BP:viral inhibition of intracellular anti-viral response ; 0019052 BP:viral inhibition of host-cell protein synthesis shutoff ; 0046773	MF:RNA binding ; 0003723	CC:small ribosomal subunit ; 0015935 CC:viral tegument ; 0019033 CC:nucleolus ; 0005730
Us12	a47, ICP47	BP:negative regulation of extracellular antiviral response by virus ; 0019053	MF:receptor antagonist activity ; 0048019	CC:endoplasmic reticulum ; 0005783

Table 4.4 Gene Ontology Evidence Codes

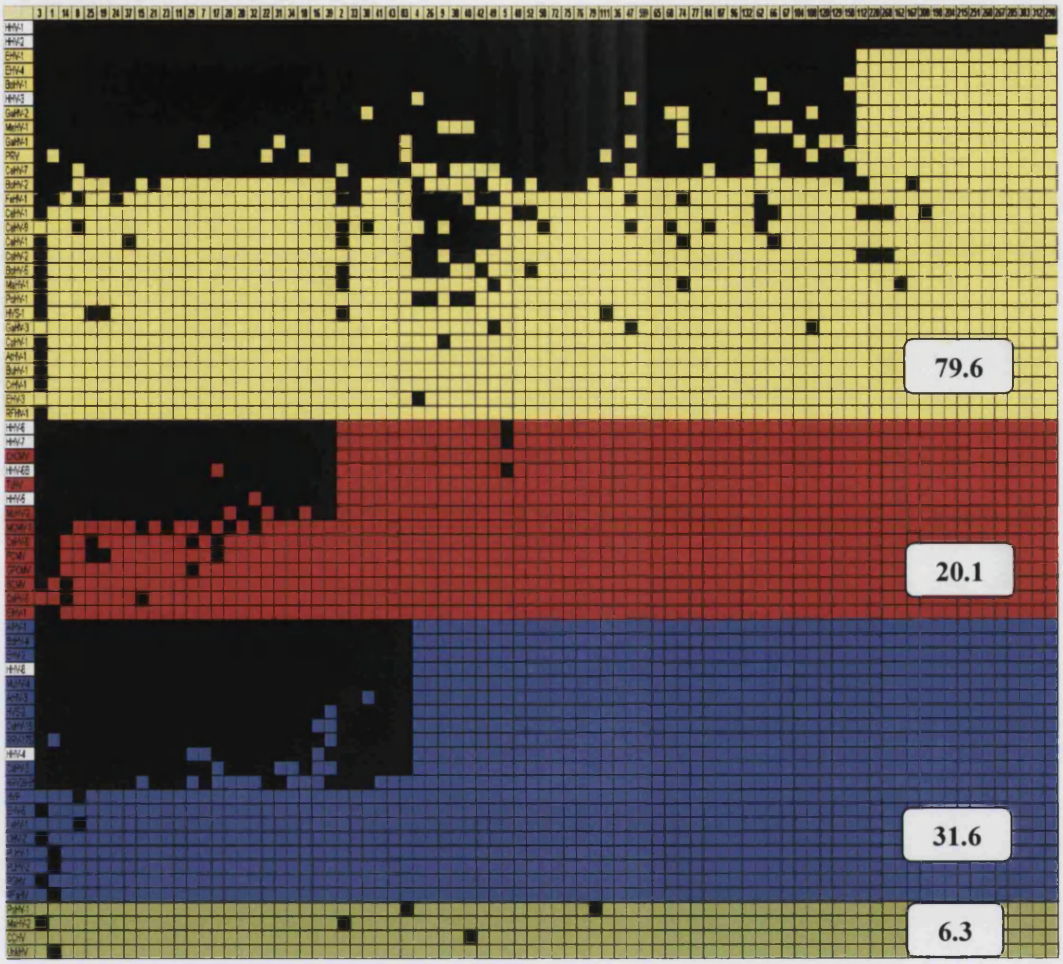
Evidence Code	Brief Description
IC	Inferred by Curator
IDA	Inferred by Direct Assay
IEA	Inferred from Electronic Annotation
IEP	Inferred from Expression Pattern
IGI	Inferred from Genetic Interaction
IMP	Inferred from Mutant Phenotype
IPI	Inferred from Physical Interaction
ISS	Inferred from Sequence or Structural Similarity
NAS	Non-traceable Author Statement; used for database entries that don't cite a paper and statements in papers that cannot be traced to another publication.
ND	No Biological Data Available
TAS	Traceable Author Statement; used for anything in the literature where the original experiments are traceable through that article and anything found in a text book or dictionary.

4.3.2 Conferring GO annotations to other Herpesviruses using VIDA's HPF structure

4.3.2.1 Annotating Herpesviruses with GO Numbers by sequence homology using VIDA

Having annotated the complete HSV-1 genome, it was possible to confer some of the more general GO annotations to other herpesvirus ORFs using sequence similarity. This is made easier by the existing HPF structure in VIDA. Each Homologous Protein Family is defined by a number of amino acid sequence motifs shared by every member of the family denoting potential functional similarity. It is therefore possible to confer GO annotations to all members of HPFs that contain one or more HSV-1 proteins. ORFs annotated by computational homology are initially assigned either the ISS (inferred from sequence or structural similarity), or the NAS (non-traceable author statement) evidence code. As there are different 'levels' of annotation possible with GO, in the form of evidence codes, the quality associated with any annotation is explicit. If new evidence regarding an annotation arises, i.e. a published 'traceable author statement', then the annotation evidence can be updated.

GO number annotations were thus computationally inferred across 81 herpesvirus HPFs that contain at least one annotated HSV-1 protein (Figure 4.3). These 81 HPFs comprise 2029 herpesvirus ORFs, from 66 different herpesviruses species/strains. Table 4.5 outlines the percentages of completely sequenced herpesvirus genomes that were annotated with GO numbers. Using this method, 45% of all 4504 known herpesvirus ORFs were annotated, corresponding to 79.6% of alphaherpesvirus ORFs, 31.6% of gammaherpesvirus ORFs, 20.1% of betaherpesvirus ORFs, and 6.3% of unclassified herpesvirus ORFs. What is immediately clear is that herpesviruses share much homology between species in each subfamily, as it was possible to annotate almost 80% of the alphaherpesvirus subfamily (Figure 4.3) based on one genome. Thus, an estimated further 70-80% of herpesvirus ORFs can be annotated simply by annotating two more genomes, namely one beta- and one gamma-herpesvirus.



Number of species partially annotated	66
Number of species not annotated	13
Total number of ORFs annotated	2029

Figure 4.3 Number of Herpesvirus ORFs annotated by Homology. The entire herpesvirus family is represented by genome down the side, divided into subfamilies: alpha (yellow), beta (red), gamma (blue), and unclassified (green). Along the top are the 81 HPFs that contain at least one HSV-1 ORF. Presence of an ORF in each HPF is denoted by a black box, thus HSV-1 is the first genome in the list, identifiable by a continuous line of black boxes. The genomes of human herpesviruses are highlighted by white boxes and occur in the order: HHV-1, HHV-2, HHV-3, HHV-6, HHV-7, HHV-6B, HHV-5, HHV-8, and HHV-4. The percentage of each subfamily annotated is in listed in each subfamily.

Table 4.5 Percentages of Complete Herpesvirus Genomes annotated with GO Terms

Herpesvirus Subfamily	Complete Herpesvirus Genome	% annotated
alphaherpesvirus	Bovine herpesvirus type 1.1	85
	Cercopithecine herpesvirus 7	90.6
	Equine herpesvirus 1	70.3
	Equine herpesvirus 4	78.6
	Gallid herpesvirus 2	64.6
	Gallid herpesvirus 3	43.8
	Human herpesvirus 1	100
	Human herpesvirus 2	95.9
	Human herpesvirus 3	86.9
	Meleagrid herpesvirus 1	74
betaherpesvirus	Chimpanzee cytomegalovirus	14.6
	Human herpesvirus 5	21.5
	Human herpesvirus 6	21.2
	Human herpesvirus 6B	27.1
	Human herpesvirus 7	26.8
	Mouse cytomegalovirus 1	24.4
	Rat cytomegalovirus Maastricht	-†
	Tupaia herpesvirus	15.8
gammaherpesvirus	Alcelaphine herpesvirus 1	39
	Bovine herpesvirus 4	37.2
	Callitrichine herpesvirus 3	41.7
	Cercopithicine herpesvirus 15	34.1
	Equine herpesvirus 2	38
	Human herpesvirus 4	25.7
	Human herpesvirus 8	17.3
	Macaca mulatta rhadinovirus 17577	36.8
	Macaca mulatta rhadinovirus 26-95	29.2
	Murid herpesvirus 4	39.8
	Saimiriine herpesvirus 2	33
unclassified	Ateline herpesvirus 3	-
	Ictalurid herpesvirus 1	2.5

† (-) indicates 0% of the genome was annotated.

4.4 Conclusion

The complete HHV-1 genome (84 ORFs) was annotated with Gene Ontology terms describing their Biological Process (es) involvement, Molecular Function(s), and Cellular Component(s)/Location(s). This was done using a combination of existing Gene Ontology terms and newly created terms relating to the viral life cycle in the host. Having annotated an entire herpesvirus genome, it was possible, utilising the Homologous Protein Family structure of VIDA, to annotate a further 1,945 herpesvirus ORFs by conferring annotations to the other members of HPFs that contained one or more HHV-1 proteins. By using a combination of manual and automated methods, 79.6% of all alphaherpesvirus ORFs have been assigned GO numbers. Of the 79 herpesvirus genomes represented in VIDA by one or more ORFs, only 13 were sufficiently dissimilar to HHV-1 to have none of their proteins annotated with GO number by correspondance to an HPF. By annotating a representative genome from each of the other two subfamilies, beta- and gammaherpesviruses, it is estimated that similar coverage can be achieved in these two subfamilies using the same combination of techniques, thereby annotating proteins not previously characterised due to their dissimilarity to HHV-1.

The program GO FINDER was used initially to ascertain the number of current annotations that existed in the database InterPro. This search highlighted the necessity for creating a number of viral function specific terms, as carried out in Chapter 2, as InterPro's annotations were often limited by the scope of the Gene Ontology, as was the case for HPFs 109 and 145 (Table 4.2). It also emphasized the benefits of a hierarchical family database, as some of InterPro's initial annotations were more specific than could be handled by VIDA; for example, the distinction between tyrosine and serine/threonine protein kinases (two separate families in InterPro), within the one HPF in VIDA (HPF 29). This made it necessary to confer only general annotations across HPFs in the last automated step of this annotation methodology to avoid inaccurate assignment. As each protein in the family is individually examined, its annotation can be enhanced by terms that apply more specifically to it, e.g. using **GO:0004950:chemokine receptor** instead of **GO:0004930:G-protein coupled receptor** in the case of certain proteins in HPF 27 (Table 4.2).

The process of defining the terms' vocabulary and DAGs is separate from the process of assigning ORFs to different GO terms; however, the annotation of a genome often requires the creation of a number of new terms in order to realistically, and accurately, complete the task. Many of these terms may not be directly related to herpesvirus function, but either pertain to aspects of other viral family infections in preparation for future viral annotation, or, are more general parent terms that are required for accurate DAG structure but are not intended for direct annotation to an ORF. Any terms created for these reasons were included and discussed in work outlined in Chapter 3.

Annotation of gene products with Gene Ontology terms has been the basis for a number of further studies into gene function and identification (Schug, Diskin et al. 2002; Bono, Nikaido et al. 2003) (Harhay and Keele 2003; King, Foulger et al. 2003; King, Lee et al. 2003). The necessity to characterise gene products as completely as possible has led to the development of intermediate tools that bridge between GO and other existing vocabularies (Hill, Blake et al. 2002) (Cantor, Sarkar et al. 2003), various technologies (Doniger, Salomonis et al. 2003) (Blaschke and Valencia 2002), and text manipulation approaches (Jenssen, Laegreid et al. 2001; McCray, Browne et al. 2002; Wren and Garner 2004). Tools have also been developed that aim to fully automate the annotation process (Khan, Situ et al. 2003) (Raychaudhuri, Chang et al. 2002; Xie, Wasserman et al. 2002; Hennig, Groth et al. 2003), removing the manual element undertaken here, in order to increase the quantity of existing annotations. All of this demonstrates that annotating gene products is not the final phase in a static process, but the necessary initial step that, once completed, produces results that can be manipulated to a variety of purposes and should be continually updated.

5.0 Analysis of host-virus interaction microarray data using the Gene Ontology

5.1 Introduction

5.1.1 Microarrays

The advantage of computational analysis of a biological system is the ability to study large data sets concurrently. It is finally possible to study cellular expression on the genomic rather than the single gene product or pathway level, thereby inferring interactions from expression pattern similarities and the co-ordinated functions of many genes.

The microarray experiment was developed to view an entire cell's expression phenotype, or transcriptome. The central dogma of information flow within a cell is from DNA to mRNA and then to protein. While there are factors, such as differing rates of degradation and post-transcription/translation modification that exclude the three from being synonymous with each other, it is becoming widely accepted that the levels of mRNA within a cell at a given point of time are a good indication of its protein levels (Kellam and Liu 2003). The microarray determines the levels of a cell's mRNA transcripts at a moment in time; the results provide a new level of cellular classification, the transcriptome.

There exists a wide variety of microarrays types, but they all rely upon the same principles of labelling RNA (as cDNA) isolated from cells and hybridising it to a slide or chip of gene specific PCR products or oligonucleotides. Microarray technology uses glass slides upon which the DNA products are robotically spotted. Arrays now routinely cover the predicted mRNA coding capacity of entire sequenced genomes.

For two coloured, spotted microarrays the RNA is isolated from cells and fluorescently labelled with either Cy3 (green fluorescent) or Cy5 (red fluorescent) using either reverse transcription or amino-allyl ligation (incorporation of an aminoallyl group to the cDNA before attaching a fluorphore). Two samples can then be compared to each other by being competitively hybridised to the same array, each RNA transcript binding to its

own unique probe spotted on the array. The hybridised array is scanned with a fluorescent detection scanner at two wavelengths (corresponding to each of the fluorophores). The combined fluorescent intensity in each spot of the array is used to calculate the ratio of expression between the two samples.

Microarrays can be used to analyse multiple samples, such as timecourses, by using a common reference sample in the place of sample two in each array experiment. Reference RNA (usually labelled with Cy3) is a pool of RNA taken from a plentiful source, such as cell lines, that serves as a control across all experiments. This becomes statistically significant when analysing the samples as it enables standardisation across an array experiment, eliminating discrepancies that can occur from differences in labelling, detection, or binding; while absolute levels of a transcript may vary between samples due to these systematic errors, the reference ensures that the overall ratio will be the same. Once the results are normalised, the direct comparison of the multiple samples being studied can be undertaken.

Here we examine the use of host-herpesvirus GO in the context of HSV-1 lytic cell gene expression from the microarray dataset of Stingley et al (Stingley, Ramirez et al. 2000), and the effect of HCMV infection of host gene expression from the microarray dataset of Eva Gramoustianou et al. The biology of HSV-1 has been previously described (Chapter 4).

5.1.2 Human Herpesvirus 5 (HHV-5; Human Cytomegalovirus, HCMV)

HHV-5 infection is characterised by a slow replication cycle *in vitro* and a distinct focal cytopathic effect in culture. Cells will often enlarge and round during infection (forming cytomegalia in some cases) (McGeoch, Cook et al. 1995; Pass 2001). The HHV-5 genome, as with all cytomegalovirus genomes, is much larger than other herpesviruses with 200-240kbp. Different strains code for different combinations of genes, with the average being approximately 200 open reading frames per genome, but a total of 213 unique HHV-5 genes are currently recognised (Mocarski and Courcelle 2001). This was, however, recently disputed in a new study by Davison *et al* (Davison, Dolan et al. 2003), which discounted 51 putative HCMV proteins, and proposed 10 new ones.

Characteristic of HHV-5 (and all β and γ herpesviruses) is the indefinite persistence in the host that can produce infectious progeny for months, even years, in the presence of an active host immune response. It is believed that the bone marrow acts as a reservoir of latent virus for HHV-5, as it is known to remain latent in myeloid cells that can develop into macrophages, dendritic cells, and granulocytes (Sinclair and Sissons 1996; Hahn, Jores et al. 1998; Soderberg-Naucler and Nelson 1999).

Congenital infection by HHV-5 is extremely detrimental and can be fatal to the unborn child, often with signs of involvement of multiple organ systems (Boppana, Pass et al. 1992), which can include the CNS. HHV-5 is also thought to be responsible for approximately 8% of mononucleosis cases in children and adults (Ljungman 1996).

5.1.3 Mapping Microarray Data onto the Gene Ontologies

Our laboratory currently uses microarrays containing 5000 human genes and a number of HHV-5 genes to study host-virus gene expression, and cellular reaction to viral infection. A major barrier to extensive datamining of such microarray data is the ability to map multiple functional annotations. The next logical step involving GO (both virus and host) would be to map the genes from microarrays onto the gene ontologies. The only information currently available *en masse* about the microarray genes is their GI numbers; and by using GO FINDER in conjunction with the resource LocusLink (Wheeler, Church et al. 2004), the GO numbers of as many of the 5000 genes as possible can be identified. This allows the coupling of microarray expression data and patterns of ORF expression to organised ontologies of function, process, and component/structure. The utility of this approach is investigated here by re-analysis of HCMV induced gene expression in fibroblasts and endothelial cells.

5.2 Methods

5.2.1 HHV-1 Microarray Data Presentation

Gene expression data for HSV-1, based on oligonucleotide microarrays, were obtained from the literature (Stingley, Ramirez et al. 2000). Gene expression values from 2 hours (early transcription) and 8 hours (late transcription) post HSV-1 infection were used. All transcripts detected with RNA signal intensities of >8 (arb units) from the two time intervals were mapped onto the biological process ontology. DAGs that have microarray data superimposed upon them were manually produced.

5.2.2 Statistical Preparation of Microarray Data

5.2.2.1 Data Source

Fibroblasts were infected with 2 HCMV strains at a Multiplicity of Infection (MOI) of 1. Strain AD169 is lab-adapted and highly-passaged, strain Toledo (endothelial-tropic) is low-passage, similar to wild type HCMV. RNA was extracted at 1,6,12,24,48,72 and 96h after infection and hybridised to glass slide microarrays with 5428 human probes (Clark, Edwards et al. 2002) and 23 HCMV probes (genes: UL18, gB, UL130, and UL132-UL151). Controls of uninfected cells were also undertaken to give a total of 16 microarrays (2 controls, 14 timepoints). These experiments are referred to as the '*viruses*' experiments.

In addition, two different cell types, fibroblasts and HUVECs (human umbilical vein endothelial cells) were infected with Toledo virus at MOI of 1. RNA was extracted at 1,6,12,24,48,72 and 96h after infection and hybridised to the same microarray type. Controls of uninfected cells were also undertaken to result in a total of 16 microarrays (2 controls, 14 timepoints). These experiments are referred to as the '*cells*' experiments. All data was kindly provided by Eva Gramoustianou.

Both the *cells* and the *viruses* timecourses utilise a common reference RNA sample in the Cy3 fluorophore channel to allow interarray normalisation and cross-array analyses. In order to ensure a signal intensity that is distinct from background noise for as many of the probes as possible, a combination of infected and uninfected cell types were used.

RNA was purified from HCMV ToledoE infected endothelial cells, uninfected peripheral blood mononuclear cells (PBMC) and MRC-5 fibroblasts. Batches of each cell type were grown to ensure the same reference samples were used across all timecourses.

The data from these two timecourses were used for GO assignment and analysis.

5.2.2.2 Assigning GO Numbers to Array Genes

The script GO FINDER was used in conjunction with the resource LocusLink to assign GO numbers to 3684 of the human probes on the microarrays. HCMV probes were manually assigned GO numbers using the same criteria outlined for HHV-1 (Chapters 3 and 4).

5.2.2.3 Log Transforming Data

The data provided were $\log_2$ expression ratios of sample (Cy5) divided by the reference (Cy3). The data are log transformed to the base 2 ($\log_2$) to allow all fold changes in regulation to be represented by the same magnitudes. For example, a gene upregulated by a factor of 2 has a $\log_2$ ratio of 1, and a downregulated gene by a factor of 2 has a $\log_2$ ratio of -1, a constant gene expression (where Cy5=Cy3) has a ratio of 1 and thus a $\log_2$ ratio of 0 (Kellam and Liu 2003).

5.2.2.4 Filling Missing Data Points

In a microarray experiment there are frequently missing values due to one of a number of factors such as insufficient resolution/intensity compared to the background of the slide, experimental variables (such as dust or scratches on the slide), or image corruption. It is possible to compute values to replace those missing rather than discount the gene spots that are affected (Troyanskaya, Cantor et al. 2001).

A program known as KNNImpute devised by Troyanskaya (2001) calculates missing values within a matrix of data based upon the K-nearest-neighbour method. By selecting K number of genes with a similar expression profile to the gene with the missing value (K being set by the user), the missing value is taken to be a weighted average of the

values present in the same experiment of those chosen genes. Each value is weighted according to the gene's expression profile similarity to the gene of interest (i.e. the gene with the missing value). KNNImpute is extremely accurate, able to predict missing values in datasets with up to 20% of the data missing with only a 10% drop in accuracy, and can be used with matrices that have as few as six columns of data.

KNNImpute was used to calculate the missing data points from all genes that were missing 1 or 2 data points (87.5% or higher data present in the datasets used here). Data from all genes that had 0,1, or 2 datapoints missing were used to calibrate the program. Following KNNImpute, a total of 1629 probes (human and HCMV) were included from the *cells* experiments, and 1185 probes (human and HCMV) were included from the *viruses* experiments for further analysis.

5.2.2.5 CLUSTER and TREEVIEW

Once missing values have been computed the data are presented in a matrix with the genes represented as rows and each experiment (for example timepoints in a timecourse) represented as a column. There are a number of programs available for preparing microarray data for analysis. Here the programs used were CLUSTER (Eisen, Spellman et al. 1998), which provides a range of tools for processing data, and TREEVIEW, which allows clusters to be viewed graphically.

5.2.2.6 Normalising the Data

It is necessary to normalise the data after they have been log transformed. This is performed by CLUSTER. Normalisation allows data, such as the gene expression pattern of one gene across a number of experiments (between columns of a matrix), to be compared by removing (normalising) systematic errors produced across all the array data. This helps to eradicate differences that occur between experiments such as varying mRNA levels, or labelling and detection inefficiencies between the two fluorophores (Quackenbush 2001). Normalisation of the data can be achieved by median centring the data on a common value, usually 0. This is achieved by subtracting the row-wise/column-wise median value of the data from each value in the row/column. This method assumes that the median gene in a given experiment (and the median

experiment for a given gene) has a ratio of one, and a log transformed ratio of 0 (Kellam and Liu 2003). The normalised data can then be cross compared.

5.2.2.7 Organising the Data in Self-Organising Maps (SOMs)

CLUSTER also allows simple Self-Organising Maps (SOM) to be constructed. Devised by Teuvo Kohonen in 1981 (Kohonen 1995), an SOM is a neural network-based approach to clustering that assigns genes to a series of predefined nodes according to their expression pattern. CLUSTER allows one-dimensional SOMs to be constructed, with the user determining which axis to organise (rows or columns), the number of nodes to be produced (usually the square root of the number of genes/experiments available for reorganisation), and the number of iterations to be run. By running an SOM for the gene axis, the genes can be roughly organised into clusters of similar expression patterns that are themselves ordered according to the similarity of one cluster to another. Each gene is then assigned a number according to the node it comprised.

5.2.2.8 Hierarchical Clustering of Data

The construction of the SOM provides a guideline for the final processing of the data by CLUSTER. The data are now hierarchically clustered according to similarities in gene expression between individual genes. The method used here was *Average-Linkage Hierarchical Clustering* and the aim is to organise all genes (in this case; it is also possible to do the same across experiments) into a tree structure where similarity of expression is represented by the length of branch connecting the genes. This is calculated by first finding the correlation co-efficient between every pair of genes in the matrix. CLUSTER uses the Pearson's correlation co-efficient which, if plotted, would detail the best fit line on a scatterplot; or in vector space, describes the angle between two vectors that both pass through the origin. Once calculated, those gene pairs with co-efficients closest to 1 (1=correlation; 0=no correlation,-1=anti-correlation) are clustered together and an average vector for the two is calculated. The entire process is then repeated using the average vector for each cluster to derive new correlation co-efficients until the entire dataset is mapped in the tree. Due to the nature of the algorithm, the order of the genes within clusters is not recorded since an average of the cluster is taken for each iteration This can affect the final visualisation of the tree. To facilitate

visualisation the node numbers assigned by the SOM are used to determine any final gene placement within the hierarchical clusters.

TREEVIEW allows the files from CLUSTER to be visualised. The data can be easily studied for changes in cellular transcription patterns over time (in a timecourse), or differences in response to similar stimuli between two cell lines over time. The scale used in TREEVIEW figures is a colour intensity bar (from red to green) representing positive fold-magnitude (red) expression and negative fold-magnitude expression in relation to the reference sample expression. The exact fold magnitudes are labelled on the individual figures.

5.2.3 Biological Pathway Visualisation

Biological pathway diagrams are taken, and if necessary, amended, from the pathway database of Kyoto Encyclopedia of Genes and Genomes (KEGG) http://www.genome.ad.jp/kegg/kegg2.html.

5.3 Results

5.3.1 Using GO's DAG framework to analyse microarray data

GO can be used in a number of ways to aide in microarray data analysis. By annotating entire microarrays with GO numbers, similar gene expression levels can be used in conjunction with the DAG structures to analyse the processes (or functions) occurring at different points in the time course, or even the prevalent locations of activity. For example, by mapping all highly expressed genes at a given timepoint against the DAGs, insight into which particular cellular processes are active at that time can be gained. Alternatively, by using the inherent DAG structure to cluster genes of similar function/process, patterns in expression of these genes can be explored.

5.3.1.1 Time dependent expression of HSV-1 using the Gene Ontology

We therefore analysed the data from two HSV-1 microarray experiments (2 hours and 8 hours post infection, correlating to early and late infection respectively) (Stingley, Ramirez et al. 2000). These data were juxtaposed with the GO term assignment data (Chapter 3) and the biological process DAG in a schematic representation (Figure 5.1). The microarray data effectively allows the integration of a dimension of time into the existing GO data, revealing changing global functional patterns created by the progress of viral infection. By looking only at the number of GO processes viral products are involved in at the two time points (Figure 5.1a), it is possible to see that late infection (8h; grey boxes) involves many more cellular processes, than early infection (2h; white boxes). Looking more closely at the three main GO parents that all the viral products of early and late infection are children of: cell growth and/or maintenance (Figure 5.1b), viral life cycle (Figure 5.1c), and cell communication (Figure 5.1d), the DAG demonstrates quite clearly the related, yet temporally distinct roles the viral ORFs play over an infectious timecourse.

Cell growth and/or maintenance processes are normally involved in the activation and systematic clearing of proteins and amino acids, metabolism and catabolism, general maintenance and repair of the cell. When the cell is damaged these processes progress to cell cycle arrest, and eventual cell death, if repair is not possible. Early in HSV-1 infection of the host cell, the virus co-ordinates the interruption of cellular metabolism

and cell protein production through the degradation of host cell mRNA (UL12.5), while altering protein phosphorylation patterns (US3, UL13, UL39) and increasing nucleotide metabolism (UL39) (Figure 5.1b, white boxes). In contrast, during late infection, when viral genome replication occurs (Figure 5.1b, grey boxes), viral proteins active in viral processes such as protein biosynthesis and DNA repair and recombination are evident, in accordance with virion production.

Similar contrasts can be seen in the viral life cycle and sections of the DAG (Figure 5.1c). During late infection the virus is entrenched in the process of replicating (viral DNA cleavage, UL17, UL31, UL32, UL33) and packaging (viral DNA genome packaging, UL6, UL17, UL25, UL31, UL32, UL33, UL36) its genome before egressing (UL36, UL48) the cell, as evident from the large number of proteins present and active in related processes at that time.

From examining the cell communication section of the DAG, it is apparent that an equal number of proteins active in both early and late infection are involved in some form of host defense evasion (Figure 5.1d). Although involved in a similar number of processes, the specific defences the virus manipulates to avoid the host's immune system are different. Early infection is concerned with controlling cellular apoptosis (UL13, US3, US5) and inhibition of intracellular viral response (US11); while late infection shifts to combat extracellular immune responses (US12, ICP0) such as the humoral immune response (UL39), or complement neutralization (UL44).

5.3.1.2 Expanding Microarray Data Analysis

These schematic representations of the biological process DAG of HSV-1 microarray data show the advantage of combining the Gene Ontology with a very limited Time Dependent method such as Stingley *et al*'s microarrays. By contrasting the two different time points (early vs. late infection) the progress of viral infection through a cell can be visualised process by process – even in the absence of host cell expression data. This analysis would be greatly expanded with increased time sampling and the use of inhibitors of viral and cellular functions. The advantage of the DAGs, in this setting, is the ability to see not only which processes the virus is involved with, but also to identify those closely related processes not affected by the virus. By adding the dimension of time to the analysis certain cellular location issues can be resolved; for

example, the proteins involved in DNA repair and recombination are found in the nucleus during late infection (Table 4.3). The aspect of location of protein activation is not addressed directly by microarray experiments; however, by mapping the expression data onto the cellular component DAG, it can be taken into consideration when analysing microarray results.

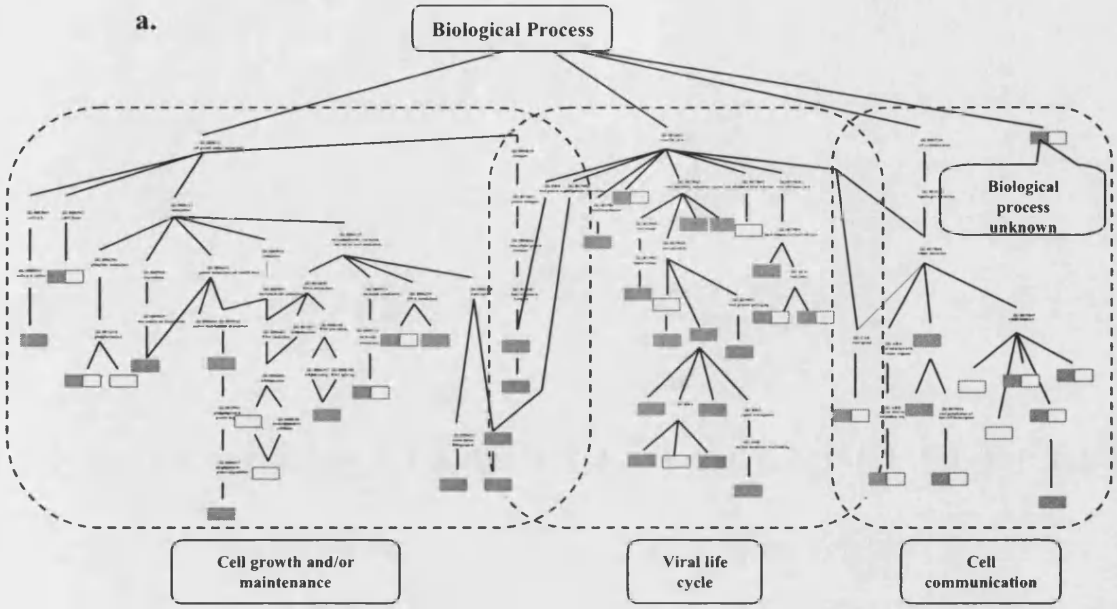
5.3.1.3 Contradictions in the Microarray Data

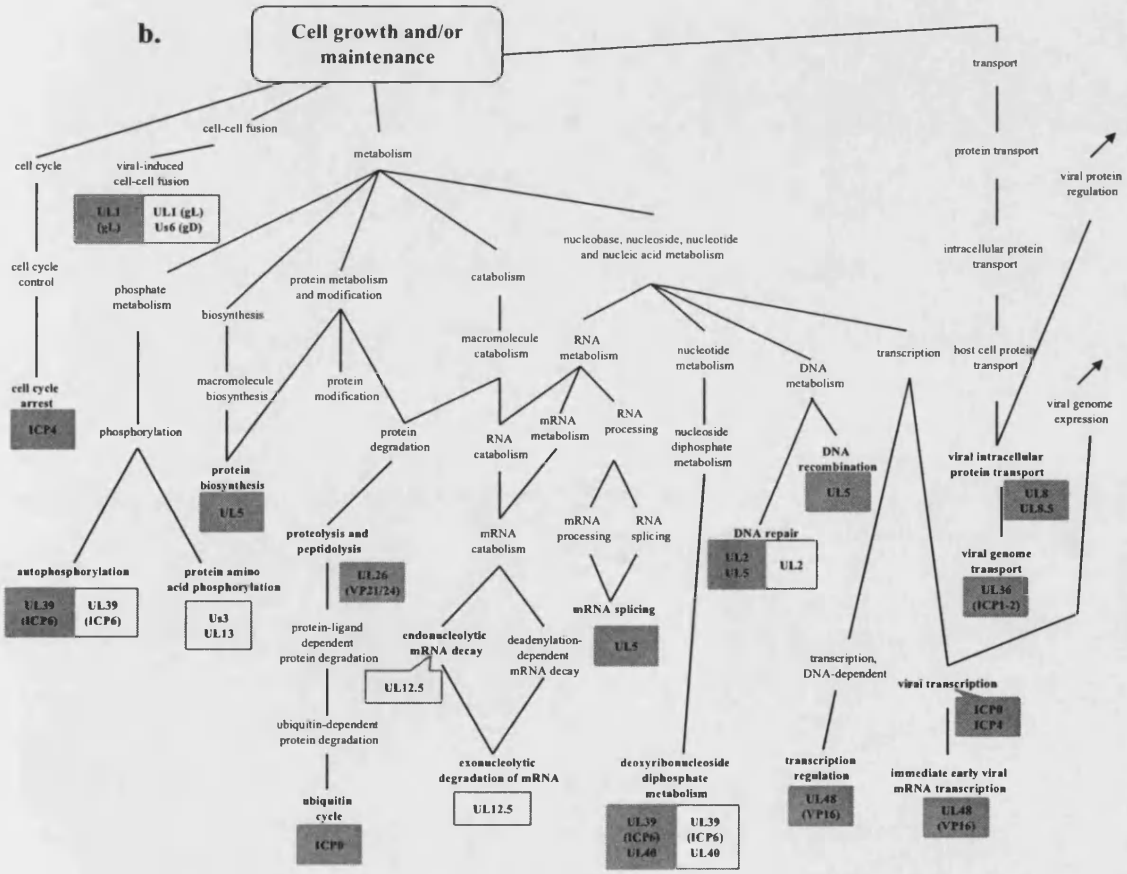
Apparent contradictions in the DAGs can also be explained from knowledge of the biology. The viral proteins involved in attachment and penetration of HSV-1 to uninfected cells, namely UL44 (gC) (Immergluck, Domowicz et al. 1998) and UL1 (gL) (Westra, Glazenburg et al. 1997), maintain apparent functions of attachment and entry of the cell during late infection according to the DAGs (Figure 5.1c). This is clearly untrue and is most likely due to both proteins having time and placement dependent functions. Both of these proteins are found in the virus particle where they are involved in viral entry into cells. However, UL44 (gC) also functions in virulence enhancement (Figure 5.1c) and viral inhibition of host complement neutralisation (Figure 5.1d) (Lubinski, Wang et al. 1998; Lubinski, Wang et al. 1999), and UL1 (gL) possibly functions in viral induced cell-cell fusion (Figure 5.1b) (Browne, Bruun et al. 2001) whilst intracellular. The interpretation of the time dependent gene expression data blurs these distinctions as the GO based annotation encapsulates all information of the proteins time and location independently. For example, UL44 attachment and entry functions may only be manifested in the viral particles, and UL1 cell fusion activity in infected cells *in vitro* most likely reflects its fusion function in virus particles. Therefore, the processes indicated by the DAGs are correct when viewed in complete biological context, but care must be taken in the initial interpretations.

Other technical problems with DNA array data can also lead to misleading interpretation. For example, the HSV-1 microarray used 52 probes to detect 72 transcripts. This means a number of probes detect more than one transcript, as was the case with probe U.1 that detected ORFs UL1 and UL2. UL1 functions early in infection and UL2 functions late in infection but, due to the probe detecting each transcript at different times, their mutual presence at both timepoints is misleading. More extensive microarray with probes for each gene and all splice variants where they exist would circumvent some of these initial shortcomings.

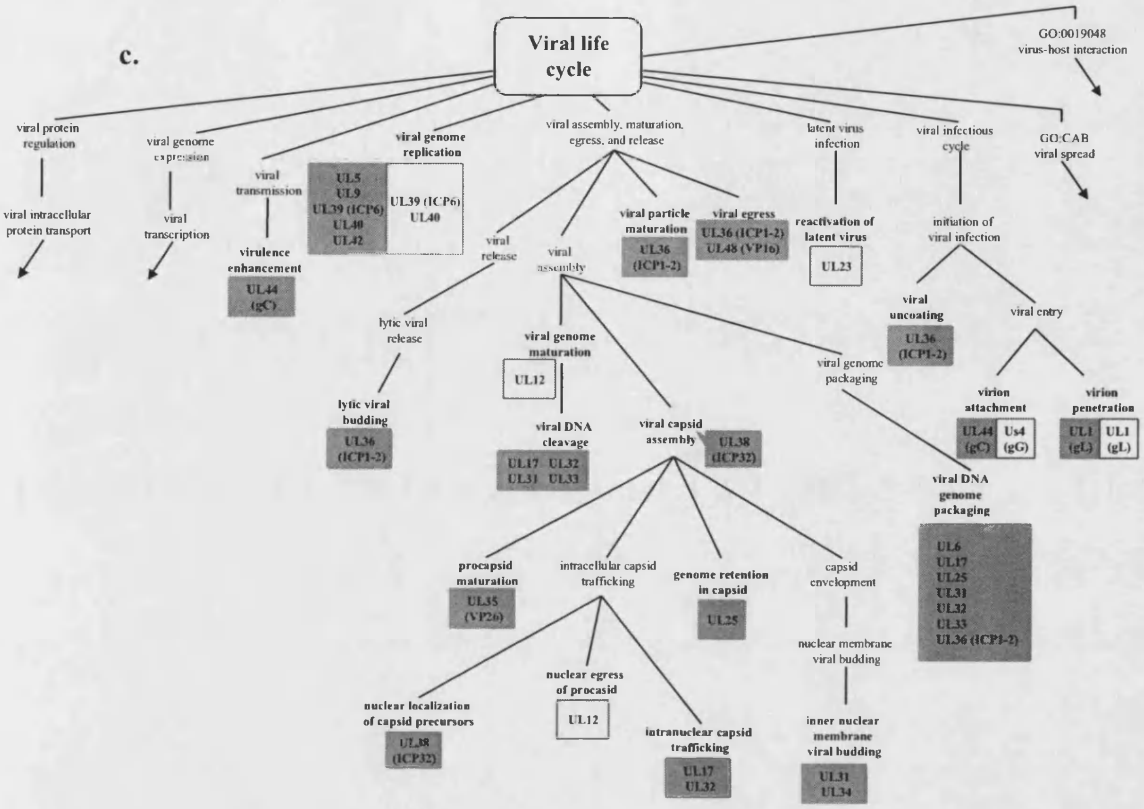
Despite the observations made in this analysis, the microarray and number of time point samples is too limited to allow detailed mechanistic insights. To explore this further at the host gene level, more extensive HCMV datasets were analysed.

Figure 5.1 (opposite). Graphical representation of DAG with Time dependent Gene Product Annotations. (a) An overview of the biological process DAG containing viral terms broken into three sections which are enlarged in (b), (c), and (d). Gene products in white boxes were expressed 2h post-HSV-1 infection; gene products in grey boxes were expressed 8h post-HSV-1 infection by HSV-1 DNA microarrays (Stingley, Ramirez et al. 2000). Terms in bold are accompanied by HSV-1 gene product annotations.

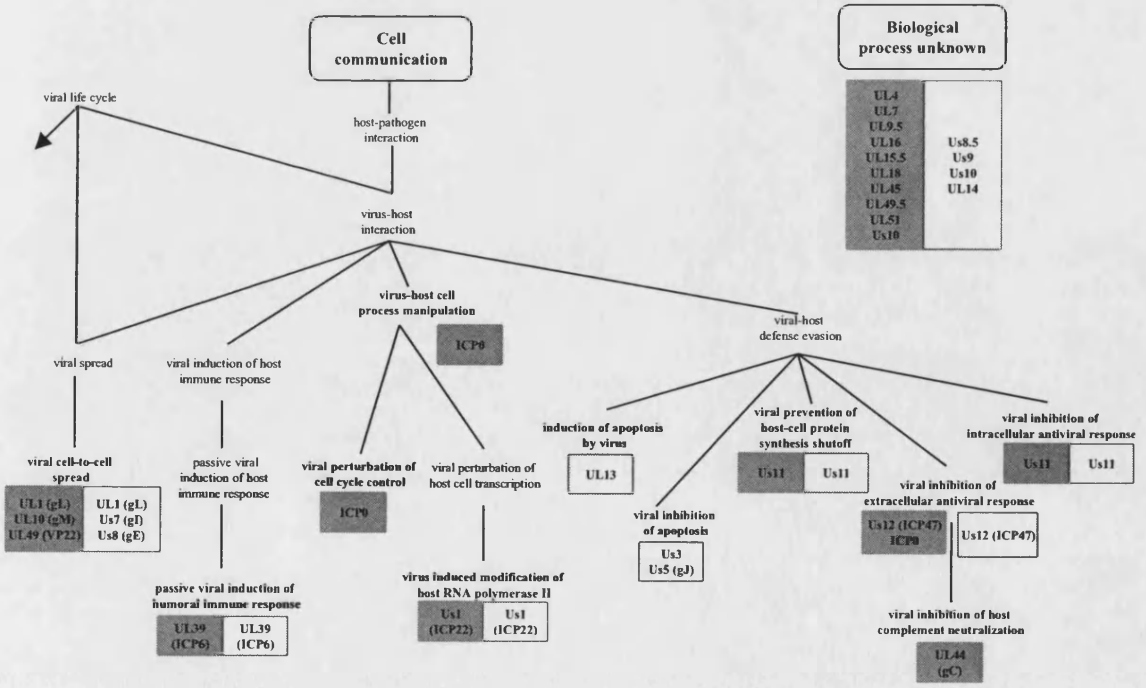

a.
Biological Process
Biological process unknown
Cell growth and/or maintenance
Viral life cycle
Cell communication


b.
Cell growth and/or maintenance
cell-cell fusion
cell cycle
viral-induced cell-cell fusion
UL1 (gL)
UL1 (gL) Us6 (gD)
metabolism
transport
protein transport
viral protein regulation
cell cycle control
phosphate metabolism
biosynthesis
protein metabolism and modification
catabolism
nucleobase, nucleoside, nucleotide and nucleic acid metabolism
intracellular protein transport
cell cycle arrest
ICP4
macromolecule biosynthesis
protein modification
macromolecule catabolism
RNA metabolism
nucleotide metabolism
DNA metabolism
transcription
host cell protein transport
viral genome expression
phosphorylation
protein degradation
RNA catabolism
mRNA metabolism
RNA processing
nucleoside diphosphate metabolism
DNA recombination
UL5
protein biosynthesis
UL5
viral intracellular protein transport
UL8
UL8.5
proteolysis and peptidolysis
UL26 (VP21/24)
mRNA catabolism
mRNA processing
RNA splicing
DNA repair
UL2
UL5
UL2
autophosphorylation
UL39 (ICP6)
UL39 (ICP6)
protein amino acid phosphorylation
Us3
UL13
viral genome transport
UL36 (ICP1-2)
mRNA splicing
UL5
protein-ligand dependent protein degradation
endonucleolytic mRNA decay
deadenylation-dependent mRNA decay
transcription, DNA-dependent
viral transcription
ICP0
ICP4
UL12.5
ubiquitin-dependent protein degradation
exonucleolytic degradation of mRNA
deoxyribonucleoside diphosphate metabolism
transcription regulation
UL48 (VP16)
immediate early viral mRNA transcription
UL48 (VP16)
ubiquitin cycle
ICP0
UL12.5
UL39 (ICP6) UL40
UL39 (ICP6) UL40

c.


Viral life cycle
GO:0019048 virus-host interaction
viral protein regulation
viral intracellular protein transport
viral genome expression
viral transcription
viral transmission
virulence enhancement
UL44 (gC)
viral genome replication
UL5 UL9 UL39 (ICP6) UL40 UL42
UL39 (ICP6) UL40
viral assembly, maturation, egress, and release
viral release
lytic viral release
lytic viral budding
UL36 (ICP1-2)
viral assembly
viral genome maturation
UL12
viral DNA cleavage
UL17 UL32 UL31 UL33
viral capsid assembly
UL38 (ICP32)
procapsid maturation
UL35 (VP26)
intracellular capsid trafficking
nuclear localization of capsid precursors
UL38 (ICP32)
nuclear egress of procasid
UL12
intranuclear capsid trafficking
UL17 UL32
genome retention in capsid
UL25
capsid envelopment
nuclear membrane viral budding
inner nuclear membrane viral budding
UL31 UL34
viral genome packaging
viral DNA genome packaging
UL6 UL17 UL25 UL31 UL32 UL33 UL36 (ICP1-2)
viral particle maturation
UL36 (ICP1-2)
viral egress
UL36 (ICP1-2) UL48 (VP16)
latent virus infection
reactivation of latent virus
UL23
viral infectious cycle
initiation of viral infection
viral uncoating
UL36 (ICP1-2)
viral entry
virion attachment
UL44 (gC) Us4 (gG)
virion penetration
UL1 (gL) UL1 (gL)
GO:CAB viral spread

d.


Cell communication
Biological process unknown
UL4 UL7 UL9.5 UL16 UL15.5 UL18 UL45 UL49.5 UL51 Us10
Us8.5 Us9 Us10 UL14
viral life cycle
host-pathogen interaction
virus-host interaction
viral spread
viral cell-to-cell spread
UL1 (gL) UL10 (gM) UL49 (VP22)
UL1 (gL) Us7 (gI) Us8 (gE)
viral induction of host immune response
passive viral induction of host immune response
passive viral induction of humoral immune response
UL39 (ICP6)
UL39 (ICP6)
virus-host cell process manipulation
ICP0
viral perturbation of cell cycle control
ICP0
viral perturbation of host cell transcription
virus induced modification of host RNA polymerase II
Us1 (ICP22)
Us1 (ICP22)
viral-host defense evasion
induction of apoptosis by virus
UL13
viral inhibition of apoptosis
Us3 Us5 (gJ)
viral prevention of host-cell protein synthesis shutoff
Us11
Us11
viral inhibition of intracellular antiviral response
Us11
Us11
viral inhibition of extracellular antiviral response
Us12 (ICP47) ICP0
Us12 (ICP47)
viral inhibition of host complement neutralization
UL44 (gC)

5.3.2 Re-annotation of Existing Analysed Microarray Data with GO Numbers

Two existing microarray timecourses, *viruses* (infection of fibroblast cells with AD169 and Toledo HCMV strains) and *cells* (infection of fibroblasts and HUVEC cells with the Toledo HCMV strain) had been previously analysed manually using the existing annotation of each probe. This consisted of a probe number, name and, where available, a short description (Figure 5.2). This level of annotation often required further research into the function of the protein before extensive functional analysis could be undertaken, as in the case of Figure 5.2b. In order to cluster the data by a protein characteristic other than expression similarity, such as: site of activation, functional similarity, or process/pathway participation, further data must be collected on each protein. Thus, the probes from the two timecourses *viruses* and *cells*, were re-labelled via GO FINDER using the GO terms annotated to each probe in LocusLink. This allowed the existing analysis to act as control for the new GO term annotations, and for further GO related analysis to be undertaken.

5.3.2.1 Existing Clusters

5.3.2.1.1 Mitochondrial Genes

One of the clusters from the original analysis included a number of genes from the Toledo infected fibroblast and HUVEC *cells* timecourse that are known to increase mitochondrial function in the cell (Figure 5.3). These include members of the respiratory chain, and mitochondrial membrane transporters. It is known that the fibroblast reticular mitochondrial network is perturbed by AD169 infection (McCormick, Smith et al. 2003). Previous analysis of the *cells* timecourse identified an increase in mitochondrial protein expression 6h post Toledo infection in HUVEC cells, while in contrast, fibroblast expression remained unchanged. This is possibly due to the lower metabolic potential of the HUVEC cells, requiring an increase in energy production before viral replication can be completed.

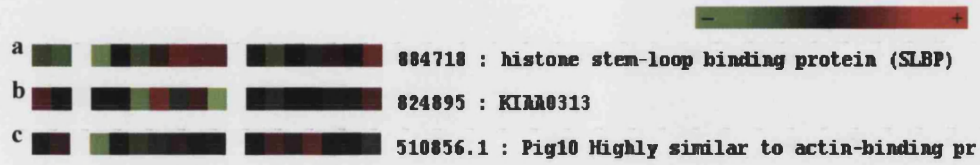


Figure 5.2 Examples of probe annotation. These three proteins are labelled with their probe number, name, and in the case of probe 510856.1, a brief description; pr = protein. Expression levels are depicted as a function of colour: red is positive fold-magnitude expression and green is negative fold-magnitude expression than the median (black) level of expression.

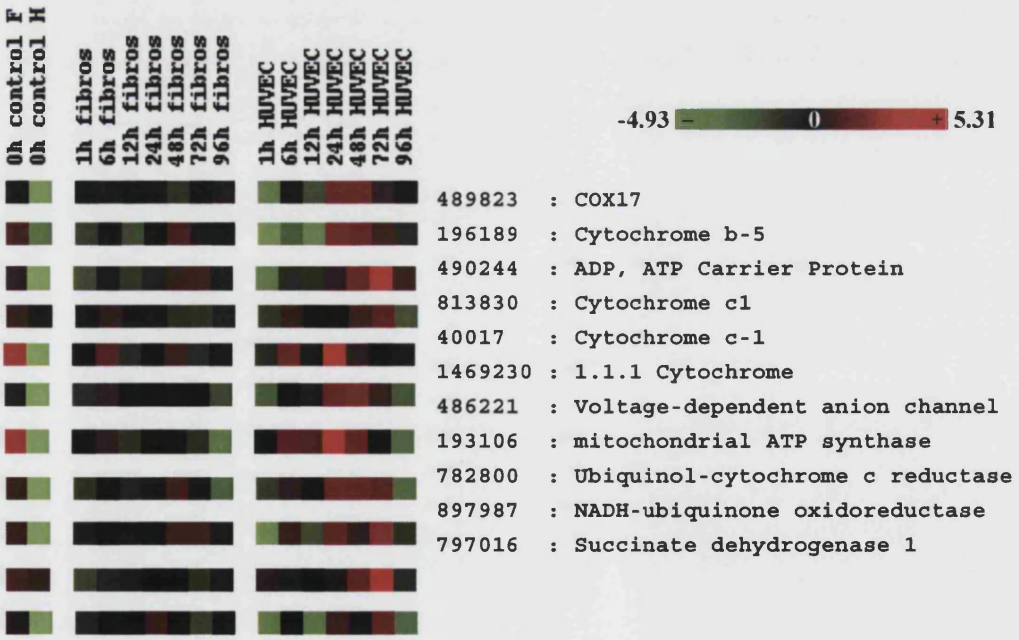


Figure 5.3 Expression of mitochondrial genes increased in Toledo infected HUVEC

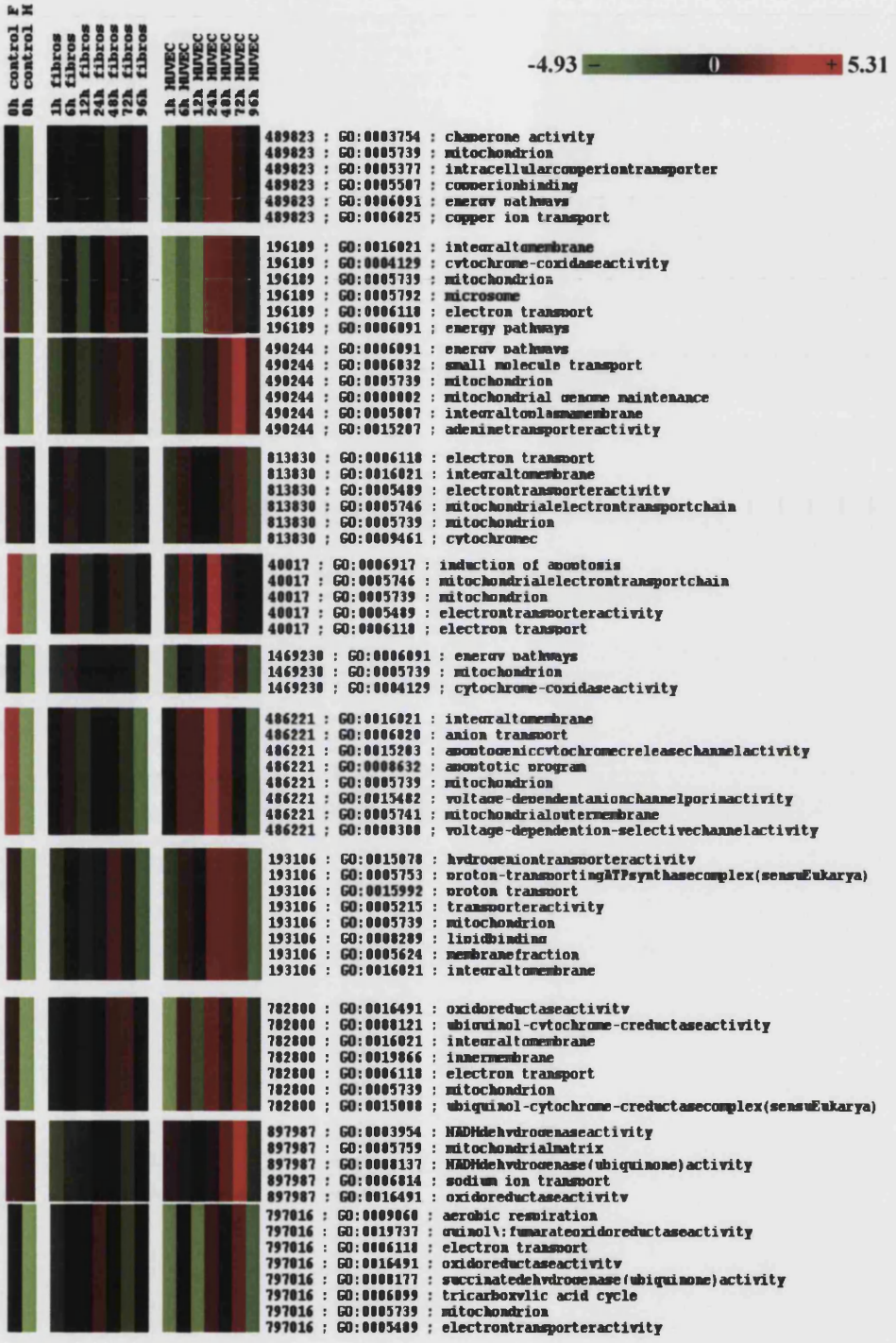


Figure 5.4 Expression of mitochondrial genes increased in Toledo infected HUVEC with GO term annotations

a.

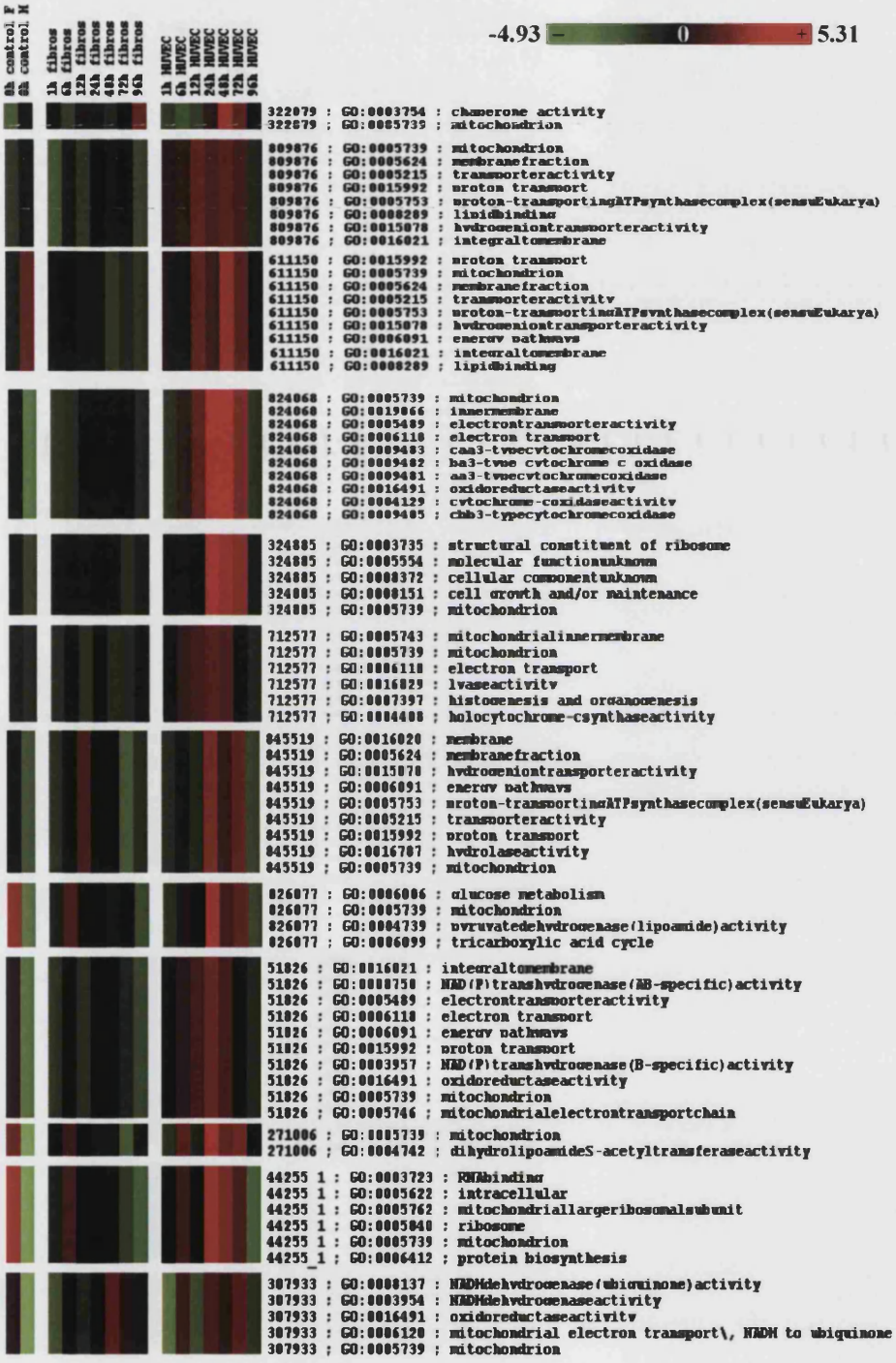
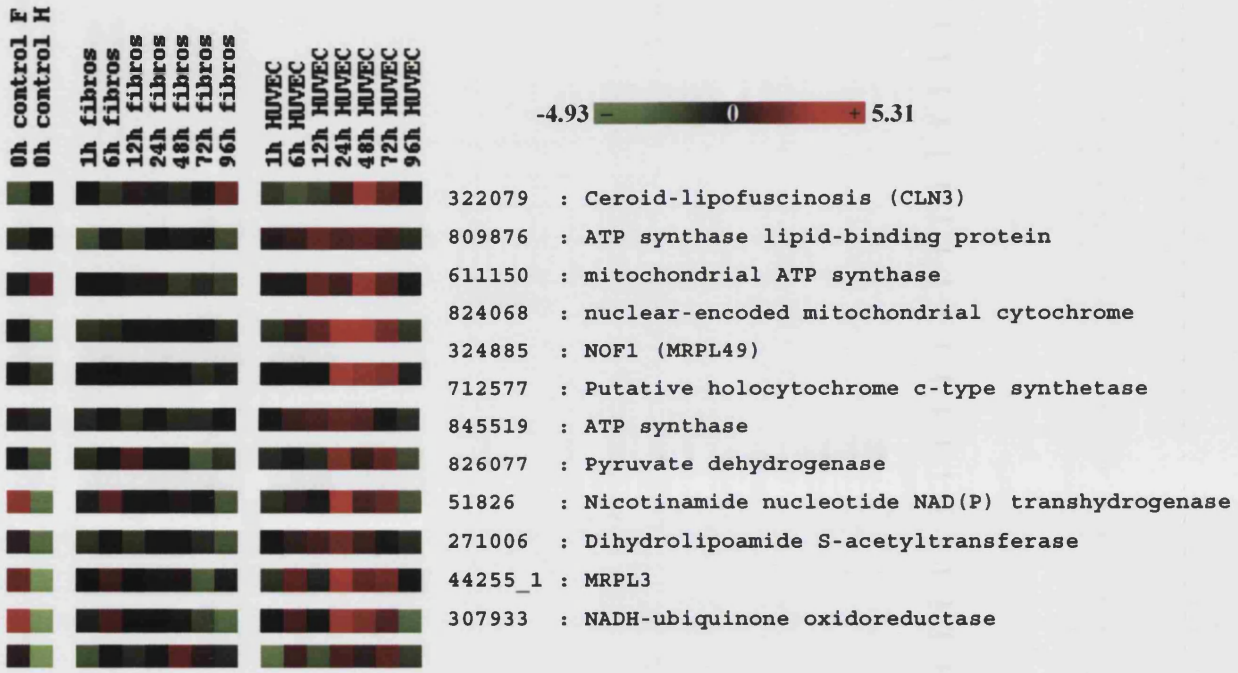


Figure 5.5 (above and opposite) Additional Genes found with increased expression after infection with Toledo in HUVEC. a) annotated with GO terms; b) annotated with names.

b.


0h control F
0h control H
1h fibros
6h fibros
12h fibros
24h fibros
48h fibros
72h fibros
96h fibros
1h HUVEC
6h HUVEC
12h HUVEC
24h HUVEC
48h HUVEC
72h HUVEC
96h HUVEC
-4.93
0
5.31
322079 : Ceroid-lipofuscinosis (CLN3)
809876 : ATP synthase lipid-binding protein
611150 : mitochondrial ATP synthase
824068 : nuclear-encoded mitochondrial cytochrome
324885 : NOF1 (MRPL49)
712577 : Putative holocytochrome c-type synthetase
845519 : ATP synthase
826077 : Pyruvate dehydrogenase
51826 : Nicotinamide nucleotide NAD(P) transhydrogenase
271006 : Dihydrolipoamide S-acetyltransferase
44255_1 : MRPL3
307933 : NADH-ubiquinone oxidoreductase

After reannotating this cluster with GO numbers (Figure 5.4), the cellular compartment term (**GO:0005739:mitochondrion**) is seen annotated to each member; this is expected as all selected proteins function in the mitochondria of the cell. As the cluster was originally manually identified, it is possible that there exist more probes in the *cells* timecourse that share a similar expression pattern to those above. Therefore, a search of the remaining data was conducted looking for additional probes that also function in the mitochondria, by collecting probes that were annotated with the term **GO:0005739:mitochondrion**. A total of 63 additional genes that are functionally related to the mitochondrion were found. Of these, 12 showed similar expression patterns to the original cluster of 11 genes, thereby representing over 100% expansion of genes identified to function in the mitochondrion that are upregulated in Toledo infected HUVEC (Figure 5.5).

Of the twelve new proteins discovered, four (the two ATP synthases: 611150, 845519; nuclear-encoded mitochondrial cytochrome: 824068; and the NADH-ubiquinone oxidoreductase: 307933) were members of the original cluster. A further seven proteins were found to function in the mitochondrion: in the mitochondrial ribosomes (mitochondrial ribosomal protein [MRP] L3 and L49, 324885 [MRPL49], 44255_1 [MRPL3]), the respiratory chain (nicotinamide nucleotide transhydrogenase, 51826; and holocytochrome c-type synthetase [HCCS], 712577), and the Kreb's cycle (pyruvate dehydrogenase, 826077; and dihydroloamide S-acetyltransferase).

This search also revealed an additional protein involved in ATP synthesis, ATP synthase lipid-binding protein (ATPase subunit C) (Yan, Lerner et al. 1994). Located in the mitochondrial membrane, it is a nonenzymatic membrane component of mitochondrial ATPase, and its expression in the cell is regulated by the final member of the cluster, CLN3. CLN3 is a chaperone protein involved in the folding and assembly regulation of a number of proteins in the cell including ATPase subunit C (Janes, Munroe et al. 1996). The CLN3 protein is of additional interest because mutations in this gene lead to juvenile ceroid-liporfuscinosis, or Batten disease, which is a progressive neurological disease caused by accelerated apoptotic cell death. Research into CLN3 has revealed that it has an antiapoptotic effect when over-expressed in NT2 neuronal precursor cells. It is also found to be over-expressed in a number of human cancer cell lines (Rylova, Amalfitano et al. 2002). Blocking of such expression led to inhibited growth and viability of cancer cells and an overall increase in incidence of

apoptosis. Therefore, induction of CLN3 may be involved in the prevention of apoptosis in HCMV Toledo infected HUVEC.

These new insights into the upregulation of mitochondrial genes 6 hours post-Toledo infection in HUVEC support the hypothesis that Toledo has to increase HUVEC viability by increasing mitochondria output before viral replication can occur.

5.3.2.2 DAG Structure Defined Clusters

5.3.2.2.1 Apoptosis Genes

An alternative method of utilising the GO resource is to take advantage of the DAG structure that is inherent to the ontologies. This provides a pre-structured view of gene product hierarchy before further analysis is undertaken. The process 'apoptosis' was chosen as it represents a manageable subsection of the DAGs whilst providing a biologically significant dataset from the array results. Apoptosis features in 14 GO terms categorised broadly into four tiers: apoptosis (**GO:0006915**), regulation of apoptosis (**GO:0042981**), positive (**GO:0043065**) or negative (**GO:0043066**) regulation of apoptosis, and induction of (**GO:0006917**) or anti-apoptosis (**GO:0006916**) (Figure 5.6a). Induction of apoptosis is further subdivided into eight increasingly descriptive terms (Figure 5.6b).

The *viruses* dataset was searched for genes that were annotated with any of the apoptosis terms in the DAG in Figure 5.6. The search yielded 31 different genes that induce, regulate, or are involved in apoptosis with significant expression patterns (Figure 5.7). From the gene expression patterns, the two viruses appear to have different approaches to regulating programmed cell death in the same cell type. Each strain affects different components of the cell cycle to control the fate of the cell.

From examining the genes regulated within the DAG structure, it becomes apparent that there are a number of genes that cannot be labelled to have solely 'positive' or 'negative' control of apoptosis. There are nine genes annotated to terms above the 'positive'-'negative' split in the DAG, i.e., to **GO: 0006915 : apoptosis** and **GO:0042981 : regulation of apoptosis**, indicating their ability to regulate cell death in a positive or negative manner, according to cellular conditions, and gene expression

levels. Likewise, of the 22 genes annotated to 'induction' and 'anti-apoptosis', two genes (594502 : TNFRSF6) and (1939252 : RNF7) are found annotated to both mutually exclusive terms.

No particular viral-strain specific pattern can be deduced from the DAGs, the two strains having a fairly equal representation of gene expression across the DAG. The only noticeable difference is that Toledo upregulates genes that can both activate or inhibit apoptosis (i.e. those genes above the 'positive'-'negative' split), while AD169 upregulates those genes below the split (i.e. that either induce or inhibit apoptosis) The significance of this is unclear.

5.3.2.3 Using Additional Resources in Combination with GO

5.3.2.3.1 LocusLink and KEGG

To further explore the GO-defined apoptosis-related gene expression data, it is useful to visualise their interaction in a diagram. Resources, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa, Goto et al. 2004), that schematically depict cellular processes in pathways and metabolic cycles provide useful additional tools to aid in the study of large datasets such as are produced by microarray experiments (Figure 5.8).

The KEGG pathway database provides online depictions that are annotated with LocusLink IDs for each gene highlighted in the pathway, allowing for cross-reference between different annotation systems. By cross-referencing the LocusLink IDs of the 31 genes identified in the GO- defined apoptosis gene cluster to the KEGG resource a complex picture of the different apoptosis control mechanisms employed by AD169 and Toledo can be constructed (Figure 5.8).

Complementing GO with such resources allows for the microarray data to be further exploited. It is not possible to derive such complex depictions as Figure 5.8 without the participation of all three resources (GO, LocusLink, KEGG). These compilations of data can also serve to compensate for the lack of annotation that would otherwise compromise studies that rely upon only one resource or another.

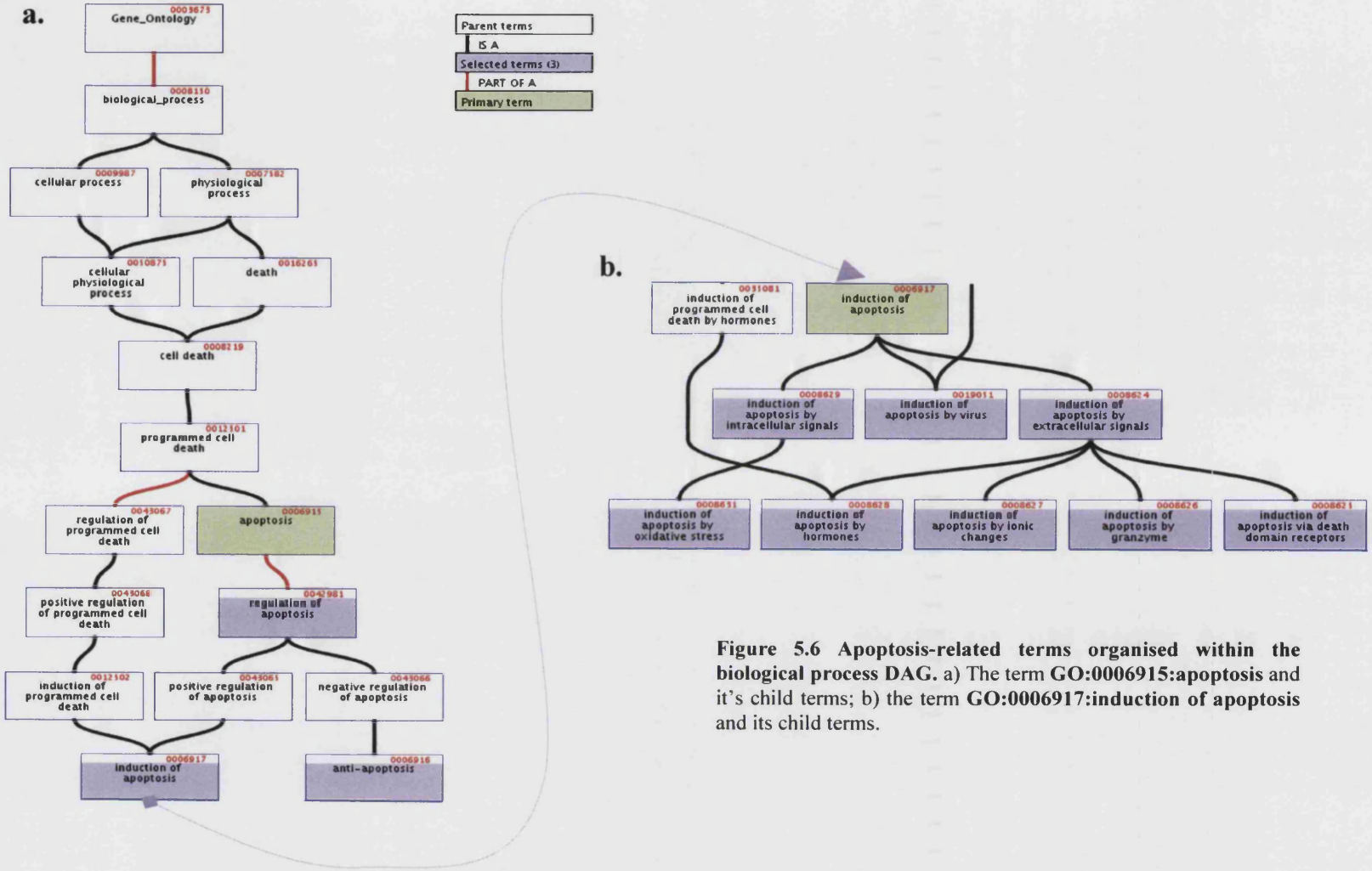
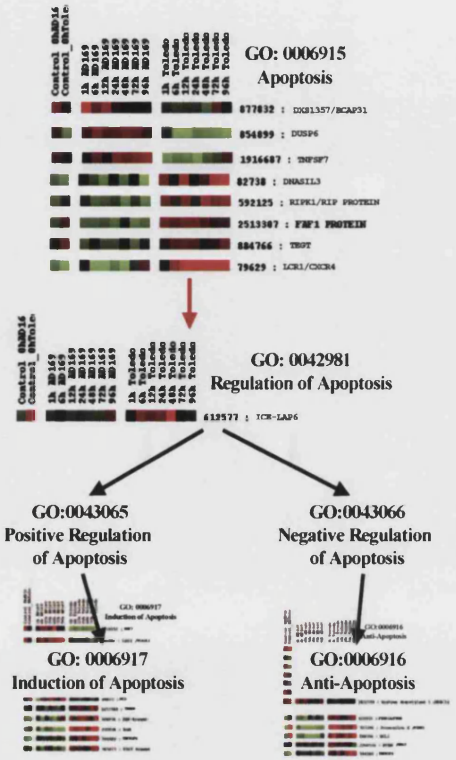


Figure 5.6 Apoptosis-related terms organised within the biological process DAG. a) The term **GO:0006915:apoptosis** and it's child terms; b) the term **GO:0006917:induction of apoptosis** and its child terms.

a



b

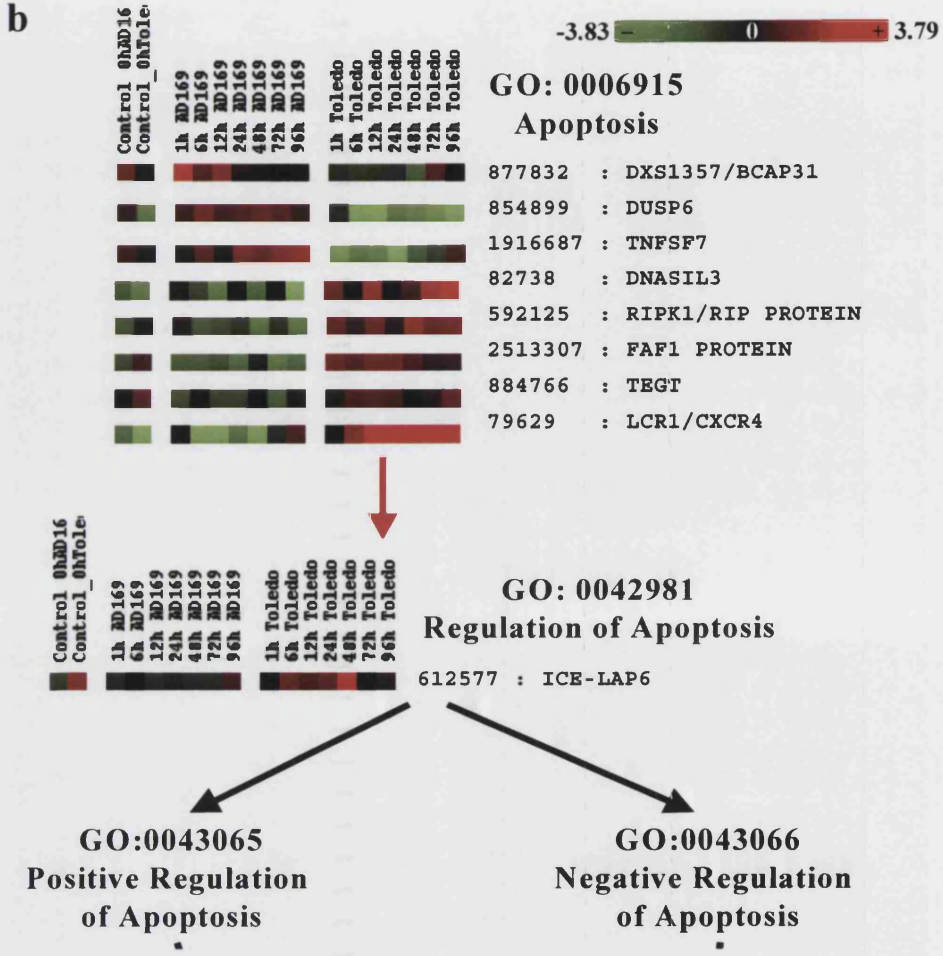
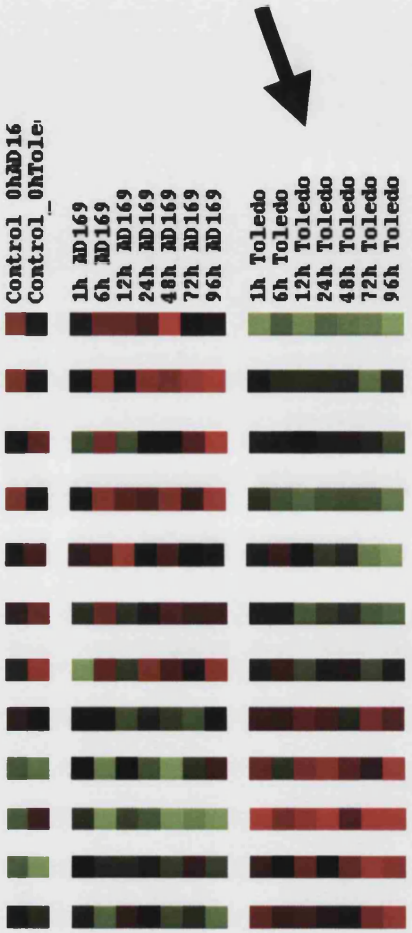
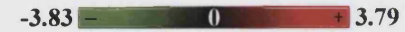


Figure 5.7 (above and opposite) Genes annotated to Apoptosis GO Terms from the *viruses* array data. a) Entire figure; (b) top section of figure enlarged; c) bottom section of figure enlarged (opposite). The genes are organised according to the apoptosis GO term they are annotated to. Those genes annotated with more than one term are displayed beneath the most specific term (i.e. at the lowest level of the DAG). Those that are annotated with two (or more) terms at the same level of the DAG are displayed beneath both terms.

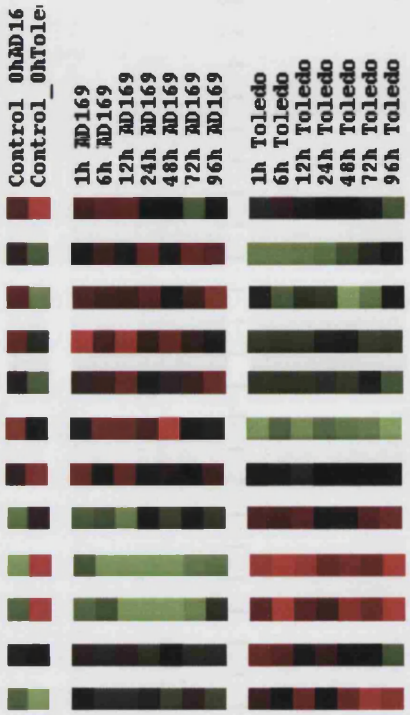
c





GO: 0006917
Induction of Apoptosis

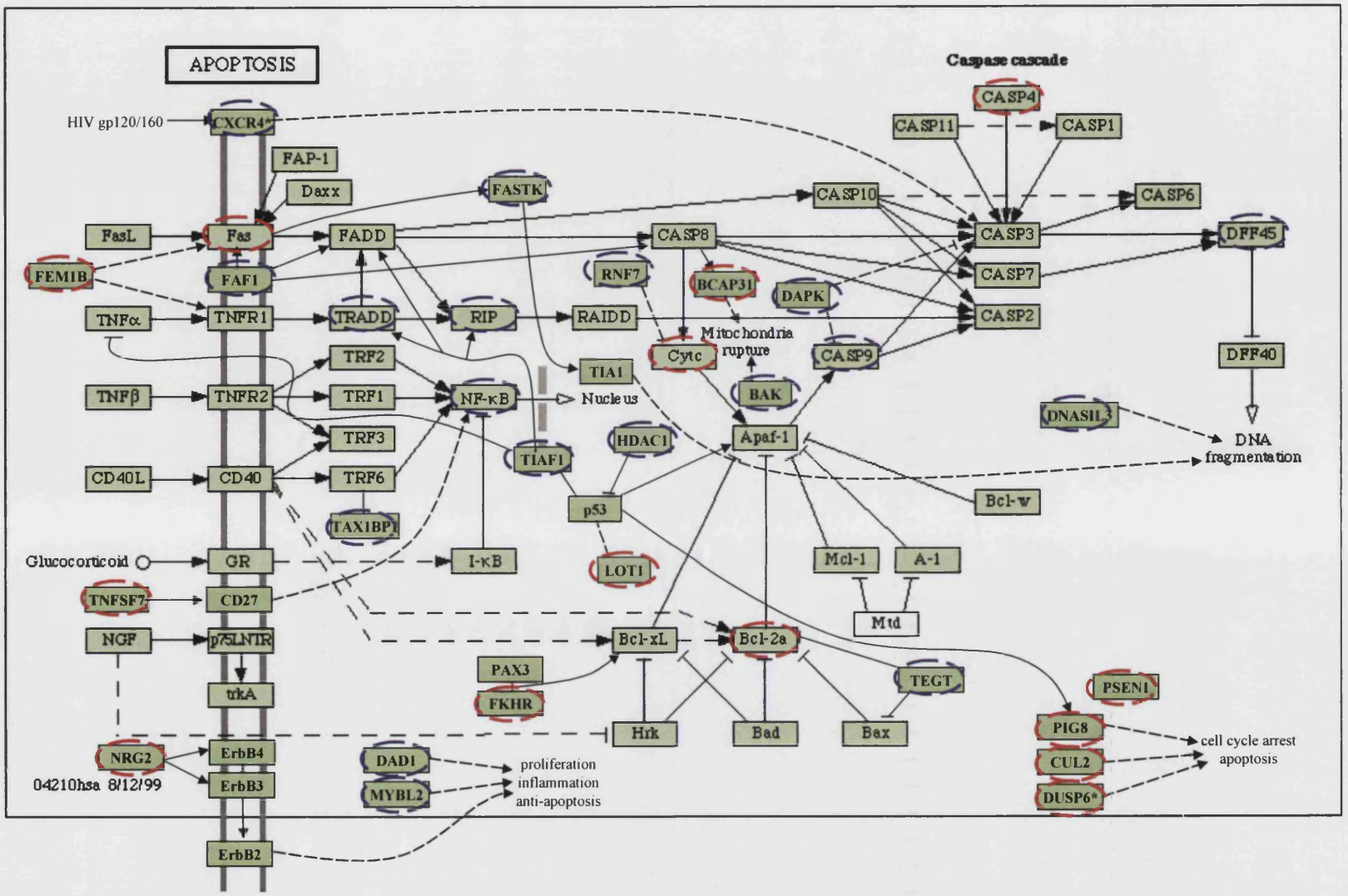
1939252 : RNF7
796904 : LOT1/PLAGL1
788247 : CUL-2
384081 : PIG8/EI24
2508930 : CASP4
248957 : FEM1B
40017 : HCS
1473788 : TRADD
364934 : DAP-KINASE
235938 : BAK
594502 : TNFRSF6
345077 : FAST KINASE



GO:0006916
Anti-Apoptosis

810039 : DAD1
898109 : TAX1BP1
815526 : MYBL2
781222 : TIAF1
789357 : NFkB1
1939252 : RNF7
1012799 : HDAC1
628955 : FOXO1A/FKHR
727390 : PSEN1
288796 : BCL2
1240116 : NTAK/NRG2
594502 : TNFRSF6

Figure 5.8 (opposite) The Apoptosis Cluster genes superimposed upon the KEGG Apoptosis Pathway. Genes circled in blue indicate pro-apoptotic genes downregulated, and anti-apoptotic genes upregulated by AD169 (downregulated by Toledo); genes circled in red indicate pro-apoptotic genes downregulated, and anti-apoptotic genes upregulated by Toledo (downregulated by AD169); CXCR4 and DUSP6 (highlighted by *) can also both be found in figure 5.11.


APOPTOSIS
Caspase cascade
HIV gp120/160
CXCR4*
FAP-1
Daxx
FasL
Fas
FADD
FASTK
CASP10
CASP4
CASP11
CASP1
CASP6
CASP8
CASP3
DFF45
FEM1B
FAF1
RNF7
BCAP31
DAPK
CASP7
TNFα
TNFR1
TRADD
RIP
RAIDD
CASP2
Mitochondria rupture
Cytc
CASP9
DFF40
TRF2
TIA1
TNFβ
TNFR2
TRF1
NF-κB
Nucleus
BAK
DNASIL3
TRF3
HDAC1
Apaf-1
DNA fragmentation
TIAF1
CD40L
CD40
TRF6
Bcl-w
p53
TAX1BP1
Glucocorticoid
GR
I-κB
LOT1
Mcl-1
A-1
TNFSF7
CD27
Mtd
NGF
p75LNTR
Bcl-xL
Bcl-2a
PAX3
TEGT
PSEN1
trkA
FKHR
Hrk
Bad
Bax
PIG8
NRG2
ErbB4
DAD1
proliferation
inflammation
anti-apoptosis
cell cycle arrest
apoptosis
04210hsa 8/12/99
ErbB3
MYBL2
CUL2
DUSP6*
ErbB2

5.3.2.4 GO Term Defined Clusters

5.3.2.4.1 Chemotaxis/MAPK Genes

An advantage of annotating microarray data with GO terms is the added ability to search the data by involvement in a certain pathway, activation in a certain location/protein structure, or gene function. HCMV has been associated with a number of inflammatory and chemotactic immune responses (Streblow, Soderberg-Naucler et al. 1999; Cinatl, Blaheta et al. 2000; Cinatl, Kotchetkov et al. 2000; Streblow, Orloff et al. 2001; Prosch, Priemer et al. 2003; Moutaftsi, Brennan et al. 2004), and modulated chemokine expression (Lecointe, Dugas et al. 2002; Momma, Nagineni et al. 2003; Scholz, Vogel et al. 2004). Therefore, a search of the AD169 and Toledo infection of fibroblasts '*viruses*' data was undertaken for all genes that are involved in the biological process of chemotaxis (**GO:0006935:chemotaxis**). This resulted in six genes (Figure 5.9), four that were clearly upregulated in Toledo infected cells, and two that were upregulated in AD169 infected cells.

This relatively small number of genes may be due to a lack of up-to-date annotation within the Gene Ontology, or to a lack of genes on the array involved in chemotaxis. It is difficult to draw extensive conclusions from such small gene clusters. It is evident, however, that within this cluster there are four genes that are involved in the Mitogen Activated Protein Kinase (MAPK) pathway, which are contrastingly regulated between viral strain infections (Figure 5.9 CXCR4, PIK3CB, and Mp38 for Toledo and p38 for AD169).

A search was thus undertaken for all genes that contained the term 'MAPK'. This revealed ten genes (including the four from the chemotaxis cluster) that are all members of the MAPK pathway (Figure 5.10). Gene 1057458 was later removed from the analysis as it represented a putative gene with mouse characteristics, leaving nine genes with contrasting expression patterns that are involved in the MAPK pathway that are differentially regulated.

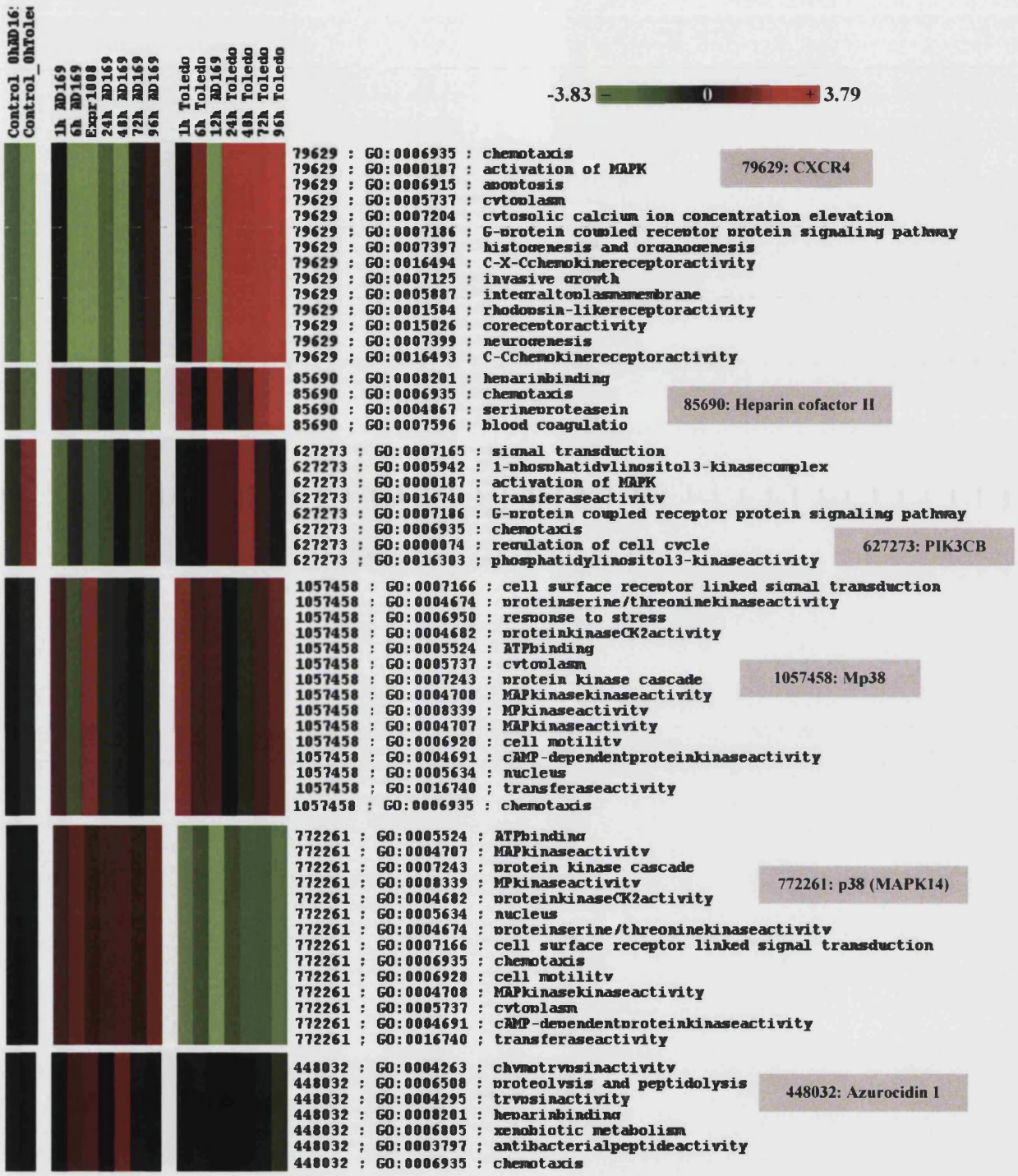


Figure 5.9 Genes Involved in Chemotaxis. A search of the *viruses* dataset revealed six genes annotated with the **GO:0006935:chemotaxis** term.

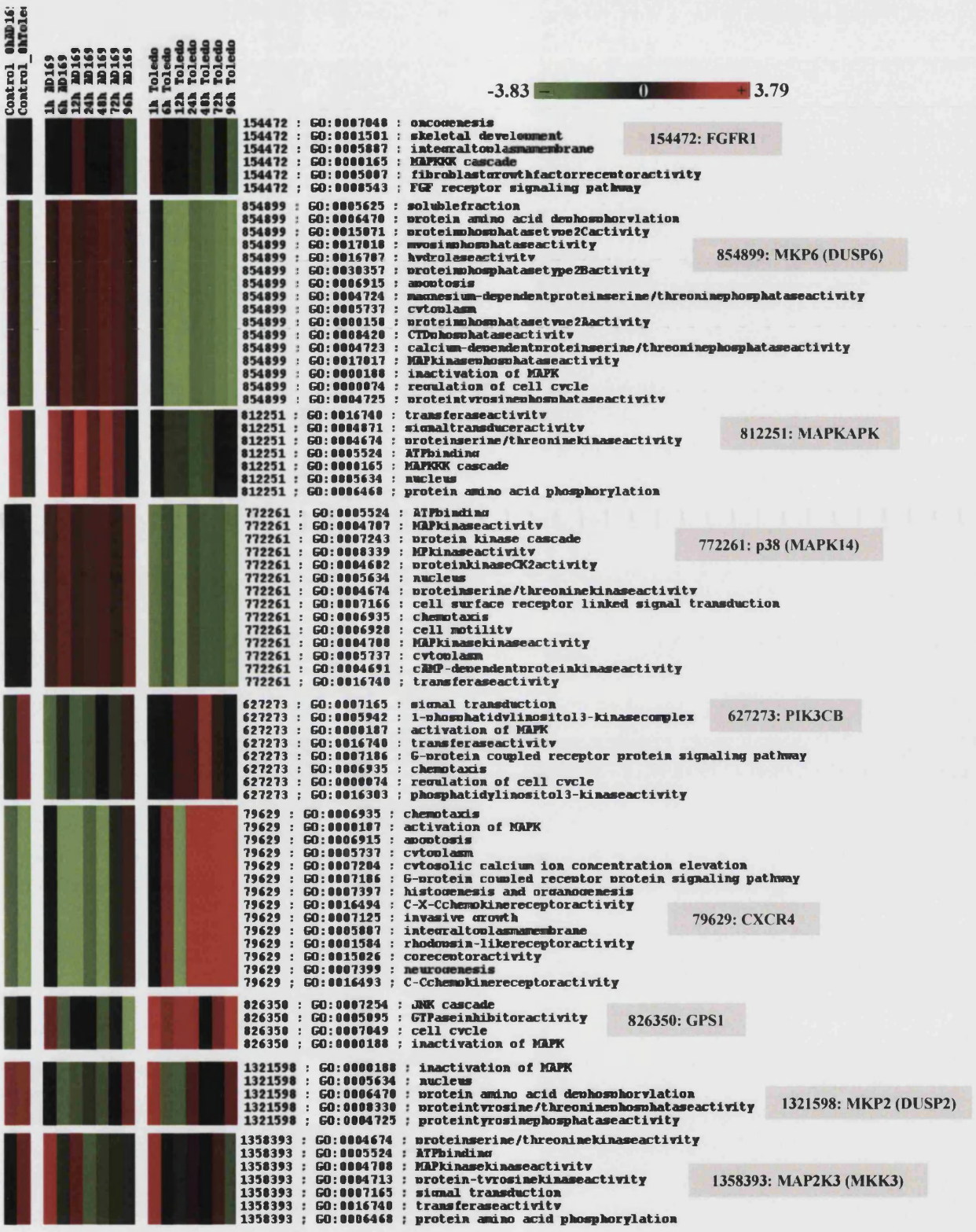


Figure 5.10 Genes Involved in the MAPK Pathway. A search of the *viruses* dataset revealed nine genes annotated with terms relating to the MAPK pathway. Gene names are outlined in grey boxes; synonyms are in brackets.

The MAPK pathway is conserved in all eukaryotes and transduces a wide variety of external cellular signals to affect a number of cellular responses including cell growth, differentiation, inflammation, and apoptosis through the activation of a variety of transcription factors (Schaeffer and Weber 1999; Wilkinson and Millar 2000). Six different MAPK pathways have been described in mammalian systems (Schaeffer and Weber 1999), but the best characterised are the ERK, JNK, and p38 pathways (Figure 5.11). The results of the MAPK search are more readily interpreted in terms of their effects on different components of the three pathways by imposing them upon a simplified map of the process (Figure 5.12). This allows the significance of their varied expression patterns to be examined in context.

Increases in ERK (Rodems and Spector 1998; Johnson, Ma et al. 2001) and p38 (Johnson, Huong et al. 2000) activity in early HCMV infection has been attributed to the virus actively inhibiting cellular phosphatase activity. It has also been shown that stress activated p38 and JNK upregulate the HCMV-IE promoter, leading to the expression of IE1-72 and IE2-86, two viral transcription factors necessary for lytic infection. These data could account for the initial increases in GPS1, MKP2, MKP6 and decreases in p38 and MAPKAPK seen a few hours post infection, reflecting cellular responses to the activation of MAPK pathways in the cell. If MAPK pathways are important for lytic HCMV replication, however, this does not explain why permissive infection occurs in cells that are terminally differentiated and exhibit little or no MAPK activity (Weller 1971), and why latency is established in undifferentiated cells with MAPK activity present.

An alternative hypothesis is that AD169, and in particular Toledo, needs to downregulate the MAPK pathways for full lytic replication. Toledo would do this by upregulation of PIK3CB blocking any SDF-1/CXCR4 signalling, downregulation of FGFR and upregulation of GPS1, which together decrease or block Ras and JNK activity. Additionally upregulation of MKP2 would block ERK activity and downregulation of MKK3 and p38 would reduce p38 activity. Similarly AD169 upregulation of MKP and MKP6/2 would reduce ERK and JNK activity, and downregulation of MKK3 would reduce p38 activity. These observations and hypothesis are supported by evidence that JNK, p38 and ERK inhibit HCMV-IE promoter expression (Sun, Harrowe et al. 2001). If the MAPK pathways inhibit IE promoter activity, this would explain HCMV's mechanism of blocking any early-

infection activated pathways to allow lytic replication. If HCMV is unable to deactivate these pathways this could account for HCMV's inability to lytically replicate in non-differentiated cells where MAPK activity is high. Thus, HCMV's reliance upon differentiation for reactivation from latency could be linked to the deactivation of MAPK activity in differentiated cells (Sissons, Bain et al. 2002).

The HCMV-IE promoter enhancer contains a number of cellular transcription factor binding sites (Figure 5.13) (Thomsen, Stenberg et al. 1984; Boshart, Weber et al. 1985; Meier and Stinski 1996) including six CRE (cAMP responsive element) sites (namely 5 ATF/CREB, 1 AP-1) (Sambucetti, Cherrington et al. 1989). Many of the transcription factors that bind to these sites are upregulated by the MAPK pathways (ELK, SAP1α, c-JUN, c-FOS, CREB, ATF2; Figures 5.11 and 5.12) and are known to increase transcriptional activity of promoters in the cell when activated. Thus, if MAPK pathway activation increases HCMV-IE promoter activity it would be most probably via the CRE sites, which have been previously shown to play a role in HCMV-IE promoter basal activity maintenance (Hunninghake, Monick et al. 1989; Chang, Crawford et al. 1990; Niller and Hennighausen 1990). As demonstrated by Sun, however, increased MAPK activity represses HCMV-IE promoter activity, and the absence of CRE sites does not inhibit MAPK repression of MEKK1 induced HCMV-IE promoter activity (Sun, Harrowe et al. 2001). This indicates that MAPK repression of promoter activity is probably not manifested via the CRE sites; therefore, any MAPK pathway activity related to HCMV infection probably does not involve any CRE site activity. It was also noted by Sun et al that in the absence of MEKK1 induced activity, MAPKs have little or no effect upon HCMV-IE promoter basal transcription. Overall, this suggests that MAPK activation is not necessary for HCMV lytic replication as previously reported (Rodems and Spector 1998; Johnson, Huong et al. 2000; Johnson, Ma et al. 2001). This supports the hypothesis that sustained suppression of ERK, JNK and p38 by HCMV is necessary for lytic replication as identified in the *viruses* expression data.

Figure 5.11 The MAPK signaling pathways

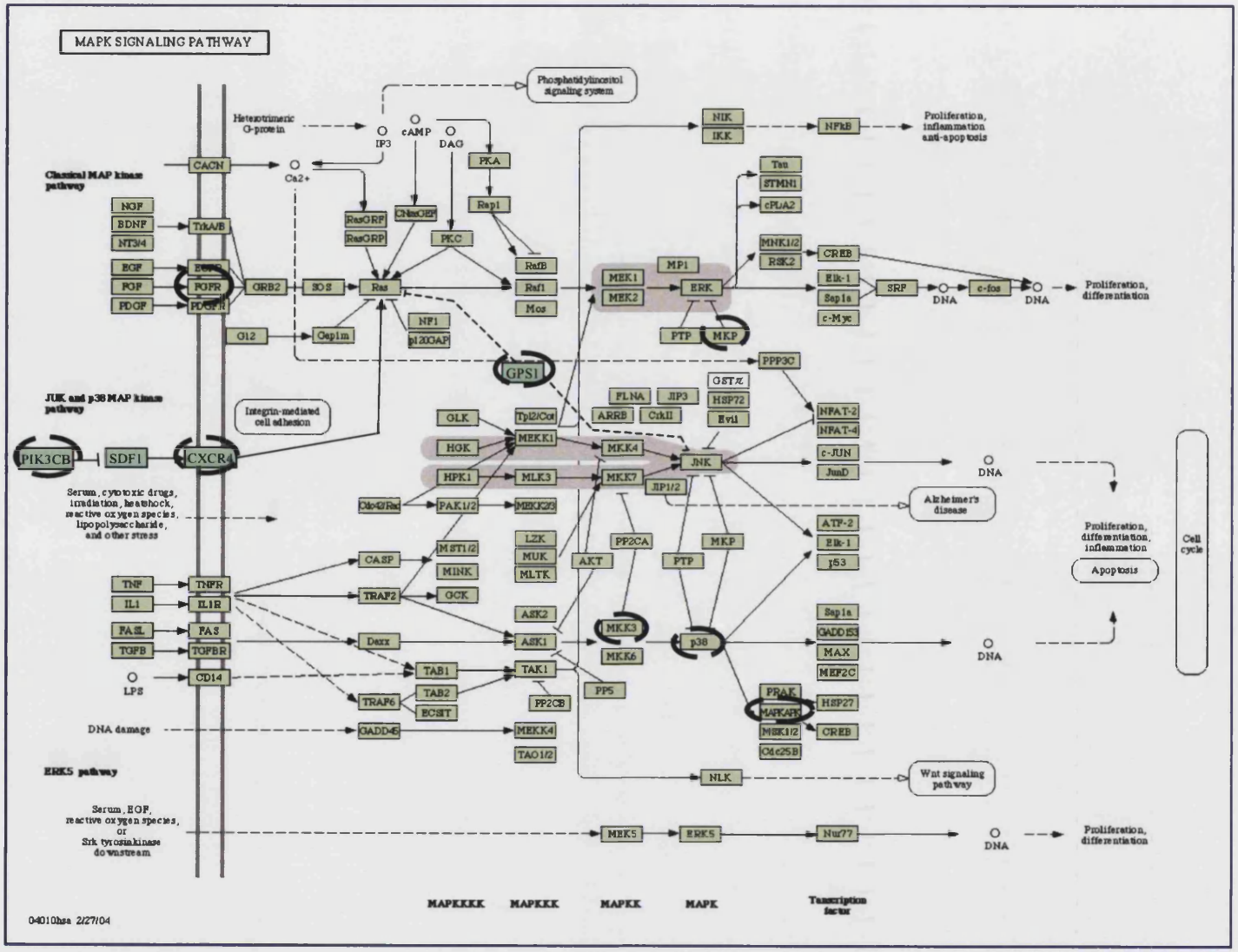
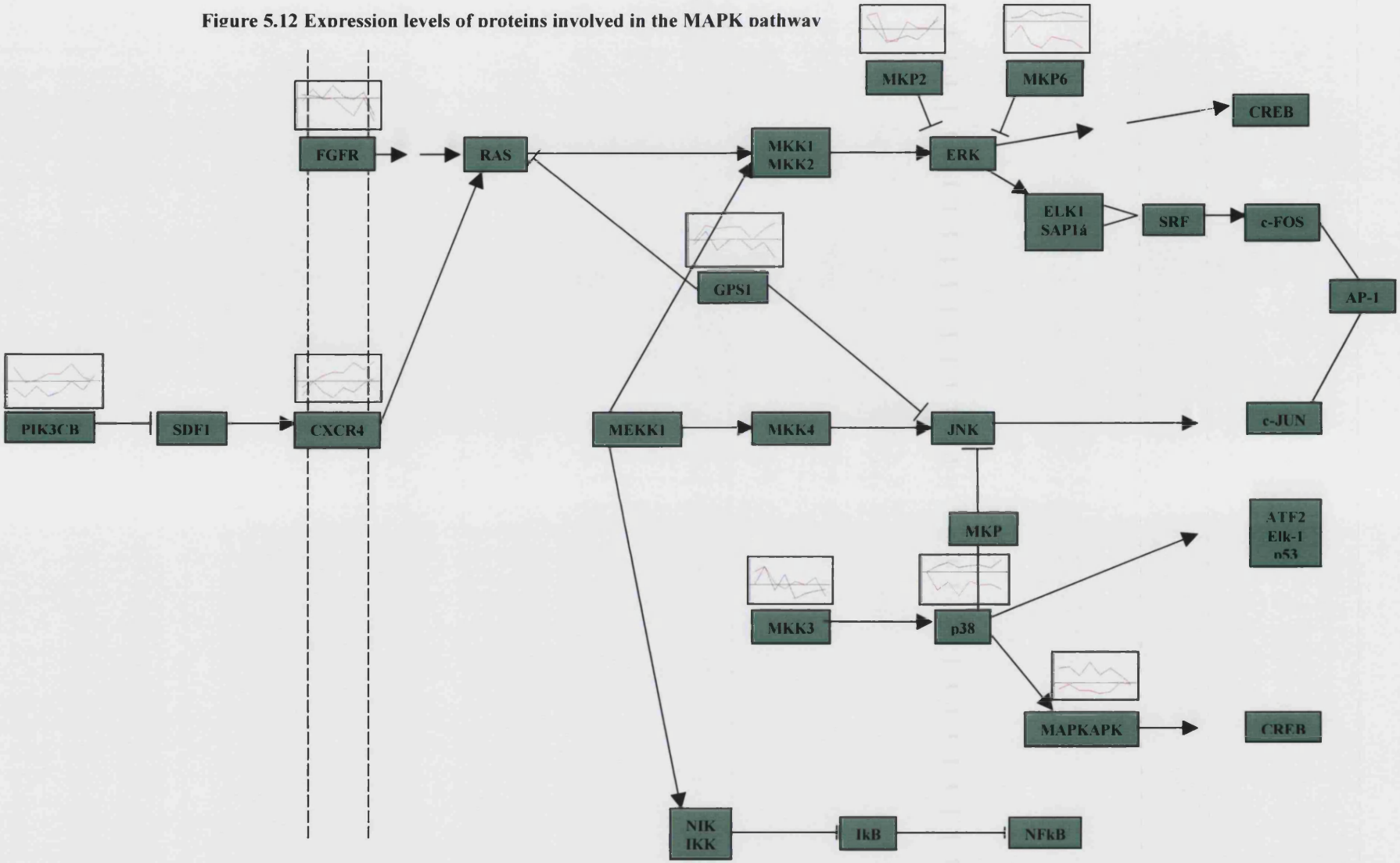


Figure 5.11 (Above) The MAPK signaling pathways. An overview of the three MAPK pathways: ERK, JNK and p38. Proteins present in Figure 5.10 are circled in black. MKP represents both MKP2 (DUSP2) and MKP6 (DUSP6).

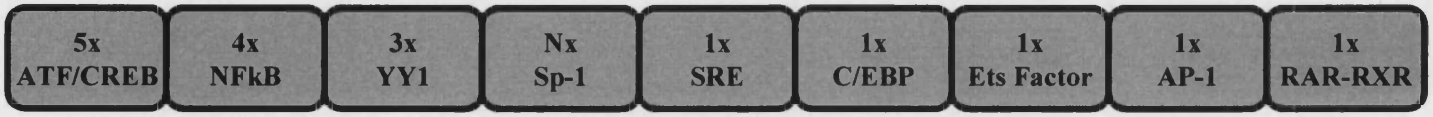
Figure 5.12 (Below) Expression levels of proteins involved in the MAPK pathways. A simplified version of the MAPK pathways adapted from Figure 5.11. Expression levels for AD169 (blue) and Toledo (red) are displayed above their corresponding proteins.

Figure 5.12 Expression levels of proteins involved in the MAPK pathway



MEKK1 activation of the HCMV-IE promoter is not exerted via the MAPK pathway, and Sun et al demonstrate that MEKK1 upregulates NFkB via IKK (Figure 5.12). In the absence of CRE sites, MEKK1 was able to activate HCMV-IE promoter activity through the NFkB binding sites also present in the enhancer region of the promoter (Figure 5.13) (Sun, Harrowe et al. 2001). NFkB is upregulated in cells responding to inflammation and immune reactions; this response also triggers differentiation of peripheral blood monocytes into macrophages. It is at this point that reactivation from latency has been observed in HCMV infected monocytes (Sissons, Bain et al. 2002).

Our data, therefore, indicate that HCMV infection of fibroblasts induces a downregulation of the MAPK pathways ERK, JNK and p38 by Toledo and AD169. This is in contrast to reports that HCMV upregulates MAPK pathways to induce HCMV-IE promoter activity via cellular transcription factors (Rodems and Spector 1998; Johnson, Huong et al. 2000; Johnson, Ma et al. 2001). Our data support the hypothesis by Sun et al that MAPK activity in cells prevents lytic replication of HCMV by suppressing HCMV-IE promoter activation, forcing the virus to establish latency. MAPK activity is higher in undifferentiated cells such as peripheral blood mononuclear cells, known to harbour latent HCMV. Sun also hypothesizes that reactivation is triggered by increases in NFkB expression seen during cellular responses to stress, which coincides with differentiation in macrophages. This is supported by evidence that MAPK pathway activation inhibits HCMV-IE promoter activity induced by NFkB via MEKK1, even in the absence of CRE binding sites.



ATF	**Activating Transcription Factor**
CREB	**cAMP Responsive Element Binding protein**
NFkB	**Nuclear Factor of kappa light polypeptide gene enhancer in B-cells**
YY1	**YingYang-1**
SRE	**Serum Response Element**
C/EBP	**CCAAT/Enhancer Binding Protein**
RAR-RXR	**Retinoic Acid Receptor (Retinoid X Receptor)**

Figure 5.13 Schematic representation of cellular transcription factor binding sites in the HCMV-IE promoter enhancer region. The number of each site present in the region is listed with the name of the transcription factor/factor family that binds to it.

5.4 Conclusion

The quality of bioinformatics results is always dependent upon the quality of the experimental data upon which it is based. The data and subsequent analyses presented here are based upon a multiplicity of infection (MOI) of 1 which only results in infection of approximately 66% of the cells being analysed. Therefore, any results derived or concluded from such experiments must take into account that the data could be representing the responses of uninfected cells, infected cells, or the proximal response of uninfected cells to infected cells. It cannot be assumed that any differences between such experiments are due solely to the viral infection variable that defines each timecourse experiment.

As an alternative system of annotation, however, the Gene Ontology has proven to be a beneficial addition to microarray analysis, providing not only a complementary system of annotation, as in the case of the mitochondrial genes cluster, but also an alternative system of clustering, able to identify unique clusters of functional similarity easily, as in the case of the MAPK genes and Apoptosis genes clusters.

The structure of the Gene Ontology extends the range of analytical possibilities, allowing clustering not only by gene function (apoptosis), but also by gene location (such as the mitochondrion), and involvement in different pathways (MAPK). Its increasing use in a variety of different web sources, such as LocusLink, allows for the collaboration of different data organisation systems – seen in the overlay of microarray data upon KEGG Encyclopedia pathways.

GO does not provide a universal answer to microarray analysis. While the use of GO can aid in the automation of large dataset handling, there are types of analysis that cannot be performed, or enhanced, using GO. In the existing analyses of the datasets there were a number of examples of clusters that could not have been gathered using GO, or a combination of GO and additional resources.

For example, Figure 5.14a is a cluster based upon association with disease of the eye suggested from the initial microarray analysis (Eva Gramoustianou). The Gene Ontology does not encompass disease associations; therefore, although each of the genes in the cluster below was annotated with the term **GO:0007601:vision**, this proved

to be only a partial annotation for the cluster, and a further search of the data using the term proved fruitless. It may be possible, however, to derive such clusters by using the Gene Ontology in collaboration with such additional resources as the Disease Ontology (http://diseaseontology.sourceforge.net/), or the Online Mendelian Inheritance in Man™ (OMIM; Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000. World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/), the latter of which is also closely associated with a number of online resources including LocusLink.

Similarly, the cluster in Figure 5.14b has been gathered by identifying genes that are located at chromosome position 1q21. Chromosome loci is another characteristic not represented in the Gene Ontology; it was, therefore, not unexpected that GO term links could not be found among these 10 genes to enable supplemental analysis. LocusLink is the prime resource for identifying chromosome locus, and is probably the best resource to start with when working with such clusters.

Finally, Figure 5.14c depicts a cluster comprised of genes that regulate chromatin structure. This is an example where the Gene Ontology does utilise the clustering preference (**GO:0006325:establishment and/or maintenance of chromatin architecture**); however, the cluster was not sufficiently annotated to enable detailed investigation. The one variable that will hinder GO reliant analysis of any experimental research is a lack of annotation of participating gene products. Each of the genes in Figure 5.14c was annotated with the term **GO:0000785:chromatin**; however, this is a term in the cellular component ontology and does not give any indication to function. This drawback in GO-oriented microarray analysis is easily overcome by annotating the genes in question as information concerning their function becomes available.

Use of the Gene Ontology in combination with experimental methods is inherently reliant upon quality of annotation. If used alone, poor quality of or incorrect annotation could potentially hinder in-depth analysis of results, as seen in Figure 5.14c. However, use of the Gene Ontology as a method of annotation and analysis has proven beneficial in not only its supplementation of existing analysis, but also in its ability to provide unique new viewpoints of the data, alone and in combination with other resources.

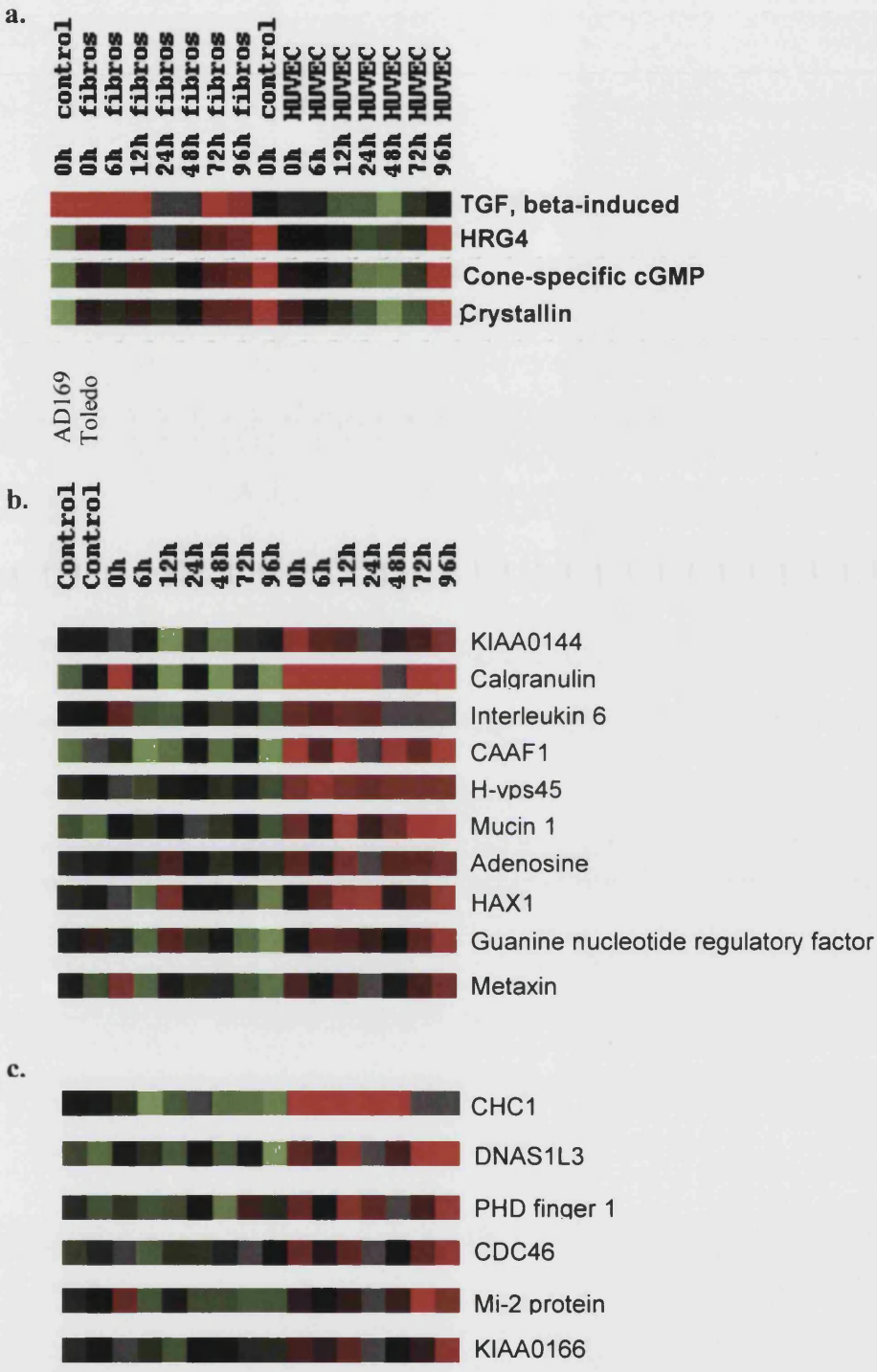


Figure 5.14 Gene clusters determined by characteristics not found in the Gene Ontology. a) genes associated with vision impairment and disease that are essential for transmission of the visual signal b) genes located at chromosomal position 1q21 and c) genes that are involved in the regulation of chromatin structure.

6.0 Discussion

Bioinformatics can be loosely divided into the two areas of static and dynamic informatics. Static informatics involves the organisation and presentation of raw data in a variety of resources. These include, most commonly, primary and compilation databases, ontologies and vocabularies, post-experimental analytical tools, and, when no analysis upon the results has been conducted, secondary databases. The results from static informatics are becoming the backbone of modern biology, both *in vitro* and *in silico,* whether to provide post-laboratory data storage, pre-experimental datasets, or computed data analysis. As such, the importance of static work increases with our reliance upon computational support; however, it is still often overlooked as a valid form of biological research despite its universality in the field.

Dynamic informatics involves more the *in silico* experimentation conducted to analyse and interpret raw data and is usually formulated around a specific question or hypothesis. Dynamic informatics benefits from the speed and accessibility that modern computing provides, allowing data gathered from a wide variety of experiments to be compared simultaneously and in larger quantities. It also encompasses complex data analysis and prediction that was previously impossible due to the computational limitations of the human brain. While increasing in accuracy and acceptance in the scientific community, the results from dynamic studies usually still require laboratory confirmation.

Experimental research is not infallible, and bioinformatics is no exception. All research requires careful analysis to prevent false information being propagated as fact. Mistakes and dubious results are sometimes published due to careless curation or overconfidence in *in silico* methods. This was the case with the discoverers of the first adenylyl cyclase in plants (Ichikawa, Suzuki et al. 1997). Homology was determined using sequence comparison methods; however, while the percentage similarity suggested a high degree of relatedness, the alignment revealed that the plant protein lacked a number of key features common to adenylyl cyclases and the paper was later retracted (Ichikawa, Suzuki et al. 1998). Virology is also subject to this type of error (Rigoutsos, Novotny et al. 2003). Experimental data, therefore, is essential in confirming and controlling bioinformatics research, even when large datasets are considered.

The advantage of many bioinformatics methods, however, is the ability to computationally cross-validate large amounts of data. In Chapter 2, the identification of known herpesvirus-host homologues was used as a sufficient control to allow the probable new homologues to be reported with higher confidence. Likewise, in Chapters 3 and 4 the processes of adding new terms and annotating viral gene products were closely interlinked, with new terms being created and DAGs being rearranged where necessary in order to accurately annotate HHV-1. Chapter 5 utilised previous detailed manual analysis of microarray data as a control for computational annotation. In addition, these chapters not only built upon work previously done, but in many cases added to and enhanced previous work from static resources (Table 6.1) in a feedback loop.

The primary aim of this thesis was to study herpesvirus-host interaction using a range of bioinformatics methods. The work presented is a combination of static and dynamic bioinformatics that demonstrates the equal necessity and validity of both (Table 6.1). Each chapter is underpinned by previous static bioinformatics work, and can be expanded with more dynamic bioinformatics work. There is also a general flow from static (red) to dynamic (green) to experimental (blue) work.

Table 6.1 Previous and future work relating to this thesis.*

Previous Work	Thesis	Future Work
• Creation and maintenance of VIDA • Conceptual gene prediction and translation of human genome	**Chapter 2:** Identification of new herpesvirus gene homologues in the human genome.	• Confirm function of new homologues identified. • Repeat with other virus families. • Repeat with other host genomes. • Repeat with same virus/host with most recent data available. • Update VIDA and Human Genome resources with results.
		feedback
• Creation and Maintenance of the Gene Ontology.	**Chapter 3:** New viral additions to the Gene Ontology.	• Use the new terms to annotate a viral genome. • Continue to Expand the Gene Ontology with new terms.
		feedback
• Chapter 3	**Chapter 4:** Annotation of herpesvirus gene products using the Gene Ontology	• Use annotation to study host-virus interactions. • Annotate more viral genomes.
		feedback
• Chapter 3 and 4 • Microarray hybridisation and initial manual result analysis.	**Chapter 5:** Analysis of host-virus interaction microarray data using the Gene Ontology.	• Confirm viral-host interaction identified. • Use Gene Ontology to analyse other microarray or large experimental dataset results.

*Red text: static informatics work
Green text: dynamic informatics work
Blue text: laboratory experimentation work

7.0 Appendix A: new(*) and existing viral gene ontology terms

term: (delayed) early viral mRNA transcription
goid: GO:0019084
definition: The second round of viral gene transcription; most genes transcribed in this round are necessary for genome replication.
definition_reference: ISBN:0781702534

***term:** active viral induction of cell-mediated immune response
goid: GO:0046737
definition: The intentional, virally-encoded stimulation of a cell-mediated host defense response to viral infection.
definition_reference: ISBN:0781802976

***term:** active viral induction of host immune response
goid: GO:0046732
definition: The intentional, virally-encoded stimulation of a host defense response to viral infection.
definition_reference: ISBN:0781802976

***term:** active viral induction of humoral immune response
goid: GO:0046736
definition: The intentional, virally-encoded stimulation of a host humoral defense response to viral infection.
definition_reference: ISBN:0781802976

***term:** active viral induction of innate immune response
goid: GO:0046738
definition: The intentional, virally-encoded stimulation of an innate host defense response to viral infection.
definition_reference: ISBN:0781802976

term: ambisense viral genome
goid: GO:0019027
definition: A RNA genome that contains coding regions that are either positive sense or negative sense on the same RNA molecule.
definition_reference: ISBN:0121585336

***term:** assemblon
goid: GO:0046808
definition: Antigenically dense structures located at the periphery of nuclei, close to but not abutting nuclear membranes. Assemblons contain the proteins for immature-capsid assembly; they are located at the periphery of a diffuse structure composed of proteins involved in DNA synthesis, which overlaps only minimally with the assemblons. More than one site can be present simultaneously.
definition_reference: PMID:8676489

***term:** autophosphorylation
goid: GO:0046777
definition: The phosphorylation by a protein of one of its own residues.
definition_reference: ISBN:0198506732

term: bipartite viral genome
goid: GO:0019018
definition: A segmented viral genome consisting of two sub-genomic nucleic acids but each nucleic acid is packaged into a different virion.
definition_reference: ISBN:0121585336

***term:** capsomere
goid: GO:0046727
definition: Any of the protein subunits that comprise the closed shell or coat (capsid) of certain viruses.
definition_reference: ISBN:0198506732

***term:** cytoplasmic viral capsid transport
goid: GO:0046743
definition: The directed movement of viral capsid proteins within the cytoplasm of the host cell.
definition_reference: ISBN:0781718325
definition_reference: PMID:11581394

term: DNA viral genome
goid: GO:0019021
definition: A viral genome composed of deoxyribonucleic acid.
definition_reference: ISBN:0121585336

term: dsRNA viral genome
goid: GO:0019023
definition: A viral genome composed of double stranded RNA.
definition_reference: ISBN:0121585336

***term:** enhancement of virulence
goid: GO:0046800
definition: Any process that activates or increases the severity of viral infection and subsequent disease.
definition_reference: PMID:10587354

term: Epstein-Barr Virus-induced receptor activity
goid: GO:001625
definition: none.
definition_reference: none.

***term:** ER membrane viral budding
goid: GO:0046764
definition: The evagination of the nucleocapsid from the host ER membrane system, resulting in envelopment of the virus.
definition_reference: ISBN:0072370319

***term:** ER membrane viral budding during viral capsid envelopment
goid: GO:0046751
definition: The envelopment of a virus, in which the nucleocapsid evaginates from the host ER membrane system, thus acquiring a membrane envelope.
definition_reference: ISBN:0072370319

***term:** ER membrane viral budding during viral capsid re-envelopment
goid: GO:0046748
definition: The re-envelopment of a virus, in which the nucleocapsid evaginates from the host ER membrane system, thus acquiring an additional membrane envelope.
definition_reference: ISBN:0072370319

term: establishment of viral latency
goid: GO:0019043
definition: The process by which a virus reaches a latent state.
definition_reference: ISBN:0781702534

***term:** genome retention in viral capsid
goid: GO:0046815
definition: The processes by which the viral genome is retained within the capsid during genome cleavage and packaging.
definition_reference: PMID:9696839

***term:** Golgi membrane viral budding
goid: GO:0046763
definition: The evagination of the nucleocapsid from the host Golgi membrane system, resulting in envelopment of the virus.
definition_reference: ISBN:0072370319

***term:** Golgi membrane viral budding during viral capsid envelopment
goid: GO:0046750

definition: The envelopment of a virus, in which the nucleocapsid evaginates from the host Golgi membrane system, thus acquiring a membrane envelope.
definition_reference: ISBN:0072370319

***term:** Golgi membrane viral budding during viral capsid re-envelopment
goid: GO:0046747
definition: The re-envelopment of a virus, in which the nucleocapsid evaginates from the host Golgi membrane system, thus acquiring an additional membrane envelope.
definition_reference: ISBN:0072370319

term: helical viral capsid
goid: GO:0019029
definition: The protein coat that surrounds the infective nucleic acid in some virus particles; the subunits are arranged to form a protein helix with the genetic material contained within. Tobacco mosaic virus has such a capsid structure.
definition_reference: ISBN:071673706X

***term:** histone deacetylase inhibitor activity
goid: GO:0046811
definition: Stops, prevents or reduces the activity of histone deacetylase, which catalyzes of the removal of acetyl groups from histones, proteins complexed to DNA in chromatin and chromosomes.
definition_reference: GO:ai
definition_reference: PMID:10482575

***term:** host cell extracellular matrix binding
goid: GO:0046810
definition: Interacting selectively with the extracellular matrix of a host cell.
definition_reference: PMID:7996163

***term:** host cell surface binding
goid: GO:0046812
definition: Interacting selectively with the surface of a host cell.
definition_reference: GO:ai

***term:** host cell surface receptor binding
goid: GO:0046789
definition: Interacting selectively with a receptor on the host cell surface.
definition_reference: GO:ai
definition_reference: PMID:11511370

term: icosahedral viral capsid
goid: GO:0019030
definition: The protein coat that surrounds the infective nucleic acid in some virus particles; the subunits are arranged to form an icosahedron, a solid with 20 faces and 12 vertices. Tobacco satellite necrosis virus has such a capsid structure.
definition_reference: ISBN:0198506732
definition_reference: ISBN:071673706X

term: immediate early viral mRNA transcription
goid: GO:0019085
definition: The transcriptional period of the earliest expressed viral genes, mainly encoding transcriptional regulators.
definition_reference: ISBN:0781702534

term: induction of apoptosis by virus
goid: GO:0019051
definition: Viral processes that result in the induction of apoptosis of infected cells, facilitating release and spread of progeny virions.
definition_reference: ISBN:0781718325

term: initiation of viral infection
goid: GO:0019059
definition: Processes involved in the start of virus infection of cells.

definition_reference: ISBN:0781702534

***term:** inner nuclear membrane viral budding during viral capsid envelopment
goid: GO:0046771
definition: The envelopment of a virus, in which the nucleocapsid evaginates from the host inner nuclear membrane system, thus acquiring a membrane envelope.
definition_reference: ISBN:0072370319

***term:** inner nuclear membrane viral budding during viral capsid re-envelopment
goid: GO:0046769
definition: The re-envelopment of a virus, in which the nucleocapsid evaginates from the host inner nuclear membrane system, thus acquiring an additional membrane envelope.
definition_reference: ISBN:0072370319

***term:** intracellular viral capsid transport
goid: GO:0046801
definition: The directed movement of viral capsid proteins within the host cell.
definition_reference: GO:ai
definition_reference: PMID:9188566

***term:** intracellular virion transport
goid: GO:0046795
definition: The directed movement of a virion within a host cell from one location to another.
definition_reference: GO:ai
definition_reference: PMID:11733033

***term:** intronless viral mRNA-nucleus export
goid: GO:0046784
definition: The export of intronless viral mRNA from the nucleus to the cytoplasm for translation.
definition_reference: PMID:11598019

term: late viral mRNA transcription
goid: GO:0019086
definition: The last group of viral genes to be transcribed during the viral life cycle; genes consist mainly of those encoding structural proteins.
definition_reference: ISBN:0781702534

term: latent virus infection
goid: GO:0019042
definition: A viral process characterized by (a) the lack of efficient expression of all the viral genes that are transcribed during productive infection, and (b) the activation of a unique latent transcriptional progam.
definition_reference: ISBN:0781702534

term: latent virus maintenance
goid: GO:0019044
definition: The processes required for maintaining the latent form of the viral genome within a cell.
definition_reference: ISBN:0781718325

term: latent virus replication
goid: GO:0019045
definition: The processes required for latent viral replication in a cell.
definition_reference: ISBN:0781702534

***term:** lytic ER membrane viral budding
goid: GO:0046757
definition: A form of viral release in which the nucleocapsid evaginates from the host ER membrane system, resulting in envelopment of the virus and cell lysis.
definition_reference: ISBN:0072370319

***term:** lytic Golgi membrane viral budding
goid: GO:0046758

definition: A form of viral release in which the nucleocapsid evaginates from the host Golgi membrane system, resulting in envelopment of the virus and cell lysis.
definition_reference: ISBN:0072370319

***term:** lytic plasma membrane viral budding
goid: GO:0046759
definition: A form of viral release in which the nucleocapsid evaginates from the host nuclear membrane system, resulting in envelopment of the virus and cell lysis.
definition_reference: ISBN:0072370319

term: lytic viral budding
goid: GO:0019078
definition: A form of viral release in which the viral particles bud out through cellular membranes, resulting in cell lysis. It is also a form of viral envelopment.
definition_reference: ISBN:0781702534

***term:** lytic viral exocytosis
goid: GO:0046756
definition: The exit of the virion particle from the host cell by exocytosis, resulting in cell lysis.
definition_reference: ISBN:0072370319

term: lytic viral release
goid: GO:0019077
definition: A viral infection and replication that leads to the destruction (lysis) of the infected cell with the release of virions.
definition_reference: GO:pk

term: multipartite viral genome
goid: GO:0019020
definition: A segmented viral genome consisting of more than three sub-genomic nucleic acids but each nucleic acid is packaged into a different virion.
definition_reference: ISBN:0121585336

***term:** microtubule polymerization
goid: GO:0046785
definition: The addition of tubulin heterodimers to one or both ends of a microtubule.
definition_reference: GO:ai

term: negative regulation of antiviral response
goid: GO:0050687
definition: Any process that stops, prevents or reduces the rate or extent of antiviral mechanisms, thereby facilitating viral replication.
definition_reference: GO:ai

term: negative regulation of retroviral genome replication
goid: GO:0045869
definition: Any process that stops, prevents or reduces the rate of retroviral genome replication.
definition_reference: GO:curators

term: negative regulation of viral genome replication
goid: GO:0045071
definition: Any process that stops, prevents or reduces the rate of viral genome replication.
definition_reference: GO:curators

***term:** negative regulation of viral protein levels
goid: GO:0046725
definition: Any process that reduces the levels of viral proteins in a cell.
definition_reference: GO:ai

term: negative regulation of virion penetration
goid: GO:0046597
definition: Any process that stops, prevents or reduces the rate of the introduction of virus particles into the cell.

definition_reference: GO:ai

term: negative sense viral genome
goid: GO:0019026
definition: A single stranded RNA genome with the opposite nucleotide polarity as mRNA.
definition_reference: ISBN:0121585336

***term:** non-lytic ER membrane viral budding
goid: GO:0046762
definition: A form of viral release in which the nucleocapsid evaginates from the host ER membrane system, resulting in envelopment of the virus without triggering cell lysis.
definition_reference: ISBN:0072370319

***term:** non-lytic Golgi membrane viral budding
goid: GO:0046760
definition: A form of viral release in which the nucleocapsid evaginates from the host Golgi membrane system, resulting in envelopment of the virus without triggering cell lysis.
definition_reference: ISBN:0072370319

***term:** non-lytic plasma membrane viral budding
goid: GO:0046761
definition: A form of viral release in which the nucleocapsid evaginates from the host plasma membrane system, resulting in envelopment of the virus without triggering cell lysis.
definition_reference: ISBN:0072370319

***term:** non-lytic viral budding
goid: GO:0046755
definition: A form of viral release in which the viral particles bud out through cellular membranes without causing cell lysis. It is also a form of viral envelopment.
definition_reference: ISBN:0072370319

***term:** non-lytic viral exocytosis
goid: GO:0046754
definition: The exit of the virion particle from the host cell by exocytosis, without causing cell lysis.
definition_reference: ISBN:0072370319

***term:** non-lytic viral release
goid: GO:0046753
definition: The release of virion particles from the cell that does not result in cell lysis.
definition_reference: ISBN:0072370319

term: non-segmented viral genome
goid: GO:0019016
definition: A viral genome that consists of one continuous nucleic acid molecule.
definition_reference: GO:p.kellum

***term:** nuclear egress of viral procapsid
goid: GO:0046802
definition: The exit of the immature viral procapsid from the nucleus of the host cell.
definition_reference: PMID:9601512
definition_reference: PMID:9765421

***term:** nuclear localization of viral capsid precursors
goid: GO:0046752
definition: The process which accumulates the necessary components for assembly of a capsid in the nucleus.
definition_reference: ISBN:0781718325

***term:** nuclear membrane viral budding
goid: GO:0046765
definition: The evagination of the nucleocapsid from the host nuclear membrane system, resulting in envelopment of the virus.
definition_reference: ISBN:0072370319

***term:** nuclear membrane viral budding during viral capsid envelopment
goid: GO:0046749
definition: The envelopment of a virus, in which the nucleocapsid evaginates from the host nuclear membrane system, thus acquiring a membrane envelope.
definition_reference: ISBN:0072370319

***term:** nuclear membrane viral budding during viral capsid re-envelopment
goid: GO:0046746
definition: The re-envelopment of a virus, in which the nucleocapsid evaginates from the host nuclear membrane system, thus acquiring an additional membrane envelope.
definition_reference: ISBN:0072370319

***term:** nuclear viral capsid transport
goid: GO:0046742
definition: The directed movement of viral capsid proteins within the nucleus of the host cell.
definition_reference: ISBN:0781718325

***term:** outer nuclear membrane viral budding during viral capsid envelopment
goid: GO:0046772
definition: The envelopment of a virus, in which the nucleocapsid evaginates from the host outer nuclear membrane system, thus acquiring a membrane envelope.
definition_reference: ISBN:0072370319

***term:** outer nuclear membrane viral budding during viral capsid re-envelopment
goid: GO:0046770
definition: The re-envelopment of a virus, in which the nucleocapsid evaginates from the host outer nuclear membrane system, thus acquiring an additional membrane envelope.
definition_reference: ISBN:0072370319

***term:** passive viral induction of cell-mediated immune response
goid: GO:0046734
definition: The unintentional stimulation by a virus of a cell-mediated host defense response to viral infection, as part of the viral infectious cycle.
definition_reference: ISBN:0781802976

***term:** passive viral induction of host immune response
goid: GO:0046731
definition: The unintentional stimulation by a virus of a host defense response to viral infection, as part of the viral infectious cycle.
definition_reference: ISBN:0781802976

***term:** passive viral induction of humoral immune response
goid: GO:0046733
definition: The unintentional stimulation by a virus of a host humoral defense response to viral infection, as part of the viral infectious cycle.
definition_reference: ISBN:0781802976

***term:** passive viral induction of innate immune response
goid: GO:0046735
definition: The unintentional stimulation by a virus of an innate host defense response to viral infection, as part of the viral infectious cycle.
definition_reference: ISBN:0781802976
definition_reference: GO:curators

***term:** plasma membrane viral budding
goid: GO:0046766
definition: The evagination of the nucleocapsid from the host plasma membrane system, resulting in envelopment of the virus.
definition_reference: ISBN:0072370319

***term:** plasma membrane viral budding during viral capsid envelopment
goid: GO:0046767

definition: The envelopment of a virus, in which the nucleocapsid evaginates from the host plasma membrane system, thus acquiring a membrane envelope.
definition_reference: ISBN:0072370319

***term:** plasma membrane viral budding during viral capsid re-envelopment
goid: GO:0046768
definition: The re-envelopment of a virus, in which the nucleocapsid evaginates from the host plasma membrane system, thus acquiring an additional membrane envelope.
definition_reference: ISBN:0072370319

term: positive regulation of retroviral genome replication
goid: GO:0045870
definition: Any process that activates or increases the rate of retroviral genome replication.
definition_reference: GO:curators

term: positive regulation of viral genome replication
goid: GO:0045070
definition: Any process that activates or increases the rate of viral genome replication.
definition_reference: GO:ai

***term:** positive regulation of viral protein levels
goid: GO:0046726
definition: Any process that increases the levels of viral proteins in a cell.
definition_reference: GO:ai

term: positive regulation of viral transcription
goid: GO:0050434
definition: Any process that activates or increases the rate of viral transcription.
definition_reference: GO:ai

term: positive regulation of virion penetration
goid: GO:0046598
definition: Any process that activates or increases the rate of the introduction of virus particles into the cell.
definition_reference: GO:ai

term: positive sense viral genome
goid: GO:0019025
definition: A single stranded RNA genome with the same nucleotide polarity as mRNA.
definition_reference: ISBN:0121585336

term: provirus
goid: GO:0019038
definition: The name given to a viral genome after it has been integrated into the host genome; particularly applies to retroviruses and is a required part of the retroviral replication cycle.
definition_reference: ISBN:0121585336

term: provirus integration
goid: GO:0019047
definition: The molecular events that lead to the integration of a viral genome into the host genome.
definition_reference: ISBN:0121585336

term: reactivation of latent virus
goid: GO:0019046
definition: The reactivation of a virus from a latent to a lytic state.
definition_reference: ISBN:0781702534

***term:** recruitment of helicase-primase complex to DNA lesions
goid: GO:0046799
definition: The recruitment of the helicase-primase complex to viral DNA lesions during viral DNA repair.
definition_reference: ISBN:0781718325

***term:** reduction of virulence
goid: GO:0046803
definition: Any process that stops, prevents or reduces the severity of viral infection and subsequent disease.
definition_reference: PMID:10982346

term: regulation of antiviral response
goid: GO:0050688
definition: Any process that modulates the frequency, rate or extent of the antiviral response of a cell or organism.
definition_reference: GO:ai

term: regulation of retroviral genome replication
goid: GO:0045091
definition: Any process that modulates the frequency, rate or extent of retroviral genome replication.
definition_reference: GO:curators

term: regulation of viral genome replication
goid: GO:0045069
definition: Any process that modulates the frequency, rate or extent of viral genome replication.
definition_reference: GO:ai

term: regulation of viral life cycle
goid: GO:0050792
definition: Any process that modulates the rate or extent of the viral life cycle, the set of processes by which a virus reproduces and spreads among hosts.
definition_reference: GO:curators

***term:** regulation of viral protein levels
goid: GO:0046719
definition: Any process that modulates the levels of viral proteins in a cell.
definition_reference: GO:ai

***term:** regulation of viral transcription
goid: GO:0046782
definition: Any process that modulates the frequency, rate or extent of the transcription of the viral genome.
definition_reference: GO:ai

term: regulation of virion penetration
goid: GO:0046596
definition: Any process that modulates the frequency, rate or extent of the introduction of virus particles into the cell.
definition_reference: GO:ai

***term:** replication compartment
goid: GO:0046809
definition: Globular nuclear domains where the transcription and replication of the viral genome occurs. More than one site can be present simultaneously.
definition_reference: PMID:9499108

term: retroviral genome replication
goid: GO:0045090
definition: Any process involved in the replication of a retroviral genome. Retroviruses use RNA as their nucleic acid and reverse transcriptase to copy their genome into the DNA of the host cells chromosomes.
definition_reference: GO:curators
definition_reference: http://cancerweb.ncl.ac.uk/omd/index.html
definition_reference: ISBN:0198506732

term: RNA viral genome
goid: GO:0019022
definition: A viral genome composed of ribonucleic acid. This results in genome replication and expression of genetic information being inextricably linked.

definition_reference: ISBN:0121585336

term: segmented viral genome
goid: GO:0019017
definition: A viral genome that is divided into two or more physically separate molecules of nucleic acid and packaged into a single virion.
definition_reference: ISBN:0121585336

term: ssRNA viral genome
goid: GO:0019024
definition: A viral genome composed of single stranded RNA of either positive or negative sense.
definition_reference: ISBN:0121585336

term: viral antireceptor activity
goid: GO:0019041
definition: none.
definition_reference: GO:curators

term: viral assembly
goid: GO:0019068
definition: A late phase of viral replication during which all the components necessary for the formation of a mature virion collect at a particular site in the cell and the basic structure of the virus particle is formed.
definition_reference: ISBN:0121585336

term: viral assembly intermediate
goid: GO:0019037
definition: Specific locations and structures in the virus infected cell involved in assembling new virions.
definition_reference: ISBN:0781718325

term: viral assembly, maturation, egress, and release
goid: GO:0019067
definition: The processes involved in the assembly, maturation, egress, and release of progeny virions.
definition_reference: ISBN:1555811272

term: viral capsid
goid: GO:0019028
definition: The protein coat that surrounds the infective nucleic acid in some virus particles. It comprises numerous regularly arranged subunits, or capsomeres.
definition_reference: ISBN:0198506732

***term:** viral capsid (sensu Retroviridae)
goid: GO:0046728
definition: The protein coat that surrounds the viral nucleocapsid, which in turn encapsulates the infective nucleic acid in retrovirus particles; the structure is complex, and specific structures and functions are associated with different elements of the capsid.
definition_reference: ISBN:1555811272
definition_reference: ISBN:0122270304

term: viral capsid assembly
goid: GO:0019069
definition: The assembly of a virus capsid from its protein subunits.
definition_reference: ISBN:0781702534

***term:** viral capsid envelopment
goid: GO:0046744
definition: The process by which a capsid acquires a membrane envelope.
definition_reference: ISBN:0781718325

***term:** viral capsid re-envelopment
goid: GO:0046745
definition: The process by which a capsid acquires another membrane envelope, subsequent to acquiring an initial membrane envelope.

definition_reference: ISBN:0781718325

***term:** viral dispersion of host splicing factors
goid: GO:0046781
definition: Viral processes that disperse host splicing factors (snRNPs) to prevent host mRNA splicing, thus reducing host protein production.
definition_reference: ISBN:0781718325

term: viral DNA cleavage
goid: GO:0019071
definition: The cleavage of viral DNA into singular functional units.
definition_reference: ISBN:0121585336

term: viral DNA genome packaging
goid: GO:0019073
definition: The packing of viral DNA into a capsid.
definition_reference: ISBN:0781702534

***term:** viral DNA repair
goid: GO:0046787
definition: The process of restoring viral DNA after damage or errors in replication.
definition_reference: ISBN:0781718325
***term:** viral egress
goid: GO:0046788
definition: The process of moving the (often) incomplete virion to the cell surface in order to be released from the cell. Egress can involve travel through the ER or cytoplasm and will often include final maturation stages of the virion, but it occurs entirely within the cell.
definition_reference: ISBN:0781718325
definition_reference: Ria_Holzerlandt:ria.h@ucl.ac.uk

***term:** viral entry
goid: GO:0046718
definition: The process by which a virion enters a host cell, including virion attachment and penetration.
definition_reference: ISBN:0781718325

term: viral envelope
goid: GO:0019031
definition: The lipid bilayer and associated glycoproteins that surround many types of virus particle.
definition_reference: ISBN:0781718325

term: viral envelope fusion
goid: GO:0019064
definition: A form of viral penetration which involves the fusion of the virion envelope with the cellular membrane.
definition_reference: ISBN:0781702534

term: viral evasion of host immune response
goid: GO:0030683
definition: Avoidance by a virus of the host immune system.
definition_reference: GO:mah

term: viral genome
goid: GO:0019015
definition: The whole of the genetic information of a virus, contained as either DNA or RNA.
definition_reference: ISBN:0198506732

term: viral genome expression
goid: GO:0019080
definition: The achievement of highly specific, quantitative, temporal and spatial control of virus gene expression within the limited genetic resources of the viral genome.
definition_reference: ISBN:0121585336

term: viral genome maturation

goid: GO:0019070
definition: Viral processes that occur on newly synthesized viral genomes.
definition_reference: GO:pk

term: viral genome packaging
goid: GO:0019072
definition: The encapsulation of the viral genome within the capsid.
definition_reference: ISBN:0121585336

term: viral genome replication
goid: GO:0019079
definition: Any process involved directly in viral genome replication, including viral nucleotide metabolism.
definition_reference: ISBN:0781702534

***term:** viral genome transport
goid: GO:0046796
definition: The directed movement of the viral genome(s) within a host cell.
definition_reference: GO:ai
definition_reference: PMID:11090159

term: viral host cell process manipulation
goid: GO:0019054
definition: Alteration of defined cellular processes that viruses target during replication.
definition_reference: GO:pk
definition_reference: GO:mah

term: viral host defense evasion
goid: GO:0019049
definition: The countering of host defenses by active or passive mechanisms.
definition_reference: ISBN:1555811272

term: viral immortalization
goid: GO:0019088
definition: A virus-induced cellular transformation arising in immortalized cells, or cells capable of indefinite replication, due to their ability to produce their own telomerase.
definition_reference: ISBN:0781702534

***term:** viral induction of host immune response
goid: GO:0046730
definition: The induction by a virus of an immune response in the host.
definition_reference: ISBN:0781802976

term: viral infectious cycle
goid: GO:0019058
definition: A set of processes which all viruses follow to ensure survival; includes attachment and entry of the virus particle, decoding of genome information, translation of viral mRNA by host ribosomes, genome replication, and assembly and release of viral particles containing the genome.
definition_reference: ISBN:1555811272

term: viral inhibition of apoptosis
goid: GO:0019050
definition: Viral processes and gene products that result in the inhibition of apoptosis, facilitating prolonged cell survival during viral replication.
definition_reference: ISBN:0781718325

***term:** viral inhibition of cell cycle arrest
goid: GO:0046792
definition: Viral interference in host cell processes that lead cell cycle arrest, allowing cell division to occur.
definition_reference: PMID:9371605

***term:** viral inhibition of expression of host genes with introns

goid: GO:0046779
definition: Viral processes that discriminate against and subsequently inhibit host transcripts containing introns, thus allowing only intronless viral mRNA to be fully processed.
definition_reference: PMID:11598019

term: viral inhibition of extracellular antiviral response
goid: GO:0019053
synonyms: negative regulation of extracellular antiviral response by virus
definition: Viral processes that result in the inhibition of extracellular (adaptive immune response) antiviral mechanisms, thereby facilitating viral replication.
definition_reference: GO:pk

***term:** viral inhibition of host cell protein biosynthesis shutoff
goid: GO:0046773
definition: Viral processes that result in the inhibition of the shutoff of host cell protein biosynthesis that occurs in response to viral infection.
definition_reference: ISBN:0781718325

***term:** viral inhibition of host complement neutralization
goid: GO:0046791
definition: Viral processes that result in the inhibition of complement neutralization of the host cell.
definition_reference: PMID:10587354

***term:** viral inhibition of host cytokine production
goid: GO:0046775
definition: Viral processes that result in the inhibition of host cell cytokine production.
definition_reference: PMID:10859382

***term:** viral inhibition of host mRNA splicing
goid: GO:0046780
definition: Viral processes that inhibit the splicing of host mRNA, thus reducing host protein production.
definition_reference: ISBN:0781718325

term: viral inhibition of intracellular antiviral response
goid: GO:0019052
synonyms: negative regulation of intracellular antiviral response by virus
definition: Viral processes that result in the inhibition of intracellular (innate immune response) antiviral mechanisms, thereby facilitating viral replication.
definition_reference: GO:pk

***term:** viral inhibition of intracellular interferon activity
goid: GO:0046774
definition: Viral processes that result in the inhibition of interferon activity within the host cell.
definition_reference: PMID:10859382

***term:** viral inhibition of MHC class I cell surface presentation
goid: GO:0046776
definition: Viral processes that result in the inhibition of presentation of MHC class I antigen-presenting proteins on the host cell surface.
definition_reference: PMID:10859382

term: viral integration complex
goid: GO:0019035
definition: Virus specific complex of protein required for integrating viral genomes into the host genome.
definition_reference: ISBN:0781718325

term: viral intracellular protein transport
goid: GO:0019060
definition: The directed movement of viral proteins within the host cell.
definition_reference: ISBN:0781702534

term: viral life cycle
goid: GO:0016032

definition: A set of processes by which a virus reproduces. Usually, this is by infection of a host cell, replication of the viral genome, and assembly of progeny virus particles. In some cases the viral genetic material may integrate into the host genome and only subsequently, under particular circumstances, 'complete' its life cycle; see viral infectious cycle (GO:0019058) and its children, and lysogeny (GO:0030069).
definition_reference: GO:mah

term: viral nucleocapsid
goid: GO:0019013
definition: The complete protein-nucleic acid complex that is the packaged form of the genome in a virus particle.
definition_reference: ISBN:0781702534

term: viral particle maturation
goid: GO:0019075
definition: The assembly of the component viral parts into an infectious virion.
definition_reference: ISBN:0781718325

term: viral perturbation of cell cycle regulation
goid: GO:0019055
definition: Viral processes that modulates the rate of the host cell cycle to facilitate virus replication.
definition_reference: ISBN:0781718325

term: viral perturbation of host cell mRNA translation
goid: GO:0019057
definition: The inhibition of transcription of cellular protein-coding genes by host RNA polymerase II.
definition_reference: ISBN:0781718325

term: viral perturbation of host cell transcription
goid: GO:0019056
definition: The inhibition, by viral gene products, of host RNA polymerase II facilitated transcription.
definition_reference: ISBN:0781718325

***term:** viral perturbation of host mRNA processing
goid: GO:0046778
definition: Viral processes that interfere with the processing of mRNA in the host cell.
definition_reference: ISBN:0781718325

***term:** viral perturbation of polysomes
goid: GO:0046783
definition: Viral processes that interfere with and inhibit the assembly and function of polysomes (GO:0005844).
definition_reference: PMID:10438802

***term:** viral portal complex
goid: GO:0046798
definition: A multimeric ring of proteins through which the DNA enters and exits the viral capsid.
definition_reference: PMID:11602732

***term:** viral procapsid
goid: GO:0046729
definition: A stable empty viral capsid produced during the assembly of viruses.
definition_reference: ISBN:1555811272
definition_reference: ISBN:0072370319

***term:** viral procapsid maturation
goid: GO:0046797
definition: The period of virion development during which the capsid components form the immature capsid and encapsulate the viral genome; the capsid often undergoes a number of structural alterations during this period.
definition_reference: PMID:10627558

term: viral protein biosynthesis

goid: GO:0019081
definition: The formation from simpler components of viral proteins.
definition_reference: ISBN:0781702534

term: viral protein processing
goid: GO:0019082
definition: The posttranslational processing of viral proteins.
definition_reference: ISBN:0781702534

term: viral receptor activity
goid: GO:0001618
definition: Combining with a virus component to initiate a change in cell activity.
definition_reference: MGI:dph

term: viral receptor mediated endocytosis
goid: GO:0019065
definition: Endocytosis of the virus particle resulting in the accumulation of virus particles within the cell via cytoplasmic vesicles.
definition_reference: ISBN:0781702534

term: viral regulation of antiviral response
goid: GO:0050690
synonyms: regulation of antiviral response by virus
definition: Any viral process that modulates the frequency, rate or extent of the antiviral response of the host cell or organism.
definition_reference: GO:ai

term: viral release
goid: GO:0019076
definition: The processes by which a virus is released from a cell.
definition_reference: ISBN:0781702534

term: viral replication complex
goid: GO:0019034
definition: Specific locations and structures in the virus infected cell involved in replicating the viral genome.
definition_reference: ISBN:0781718325

***term:** viral replication complex formation and maintenance
goid: GO:0046786
definition: The process of organizing and assembling viral replication proteins in preparation for viral replication.
definition_reference: ISBN:0781718325

term: viral RNA genome packaging
goid: GO:0019074
definition: The packaging of viral RNA into a nucleocapsid.
definition_reference: ISBN:0781718325

***term:** viral scaffold
goid: GO:0046806
definition: A complex of proteins that form a scaffold around which the viral capsid is constructed.
definition_reference: ISBN:0072370319

***term:** viral scaffold assembly and maintenance
goid: GO:0046807
definition: The assembly and maintenance of the viral scaffold (GO:0046806) around which the viral capsid is constructed.
definition_reference: ISBN:0072370319

***term:** viral spread within host
goid: GO:0046739
definition: The dissemination of infectious virion particles within an infected host.

definition_reference: ISBN:0781718325

***term:** viral spread within host, cell to cell
goid: GO:0046740
definition: The process of viral dissemination within an infected host organism where infectious virion particles are passed from infected to uninfected host cells.
definition_reference: ISBN:0781718325
***term:** viral spread within host, tissue to tissue
goid: GO:0046741
definition: The process of viral dissemination within an infected host organism where infectious virion particles are passed from infected to uninfected host tissue.
definition_reference: ISBN:0781718325

term: viral tegument
goid: GO:0019033
definition: A structure lying between the capsid and envelope of a virus, varying in thickness and often distributed asymmetrically.
definition_reference: ISBN:0721662544

term: viral transcription
goid: GO:0019083
definition: The mechanisms involved in viral gene transcription, especially referring to those with temporal properties unique to viral transcription.
definition_reference: ISBN:0781702534

term: viral transcriptional complex
goid: GO:0019036
definition: Specific locations and structures in the virus infected cell involved in transcribing the viral genome.
definition_reference: ISBN:0781718325

term: viral transformation
goid: GO:0019087
definition: Any virus-induced change in the morphological, biochemical, or growth parameters of a cell.
definition_reference: ISBN:0781702534

term: viral translocation
goid: GO:0019066
definition: The translocation of an entire virus particle across the host cell plasma membrane.
definition_reference: ISBN:0781702534

term: viral transmission
goid: GO:0019089
definition: The transfer of virions in order to create new infection.
definition_reference: ISBN:0781702534

term: viral uncoating
goid: GO:0019061
definition: A general term applied to the events that occur after penetration; refers to the 'uncoating' of the viral genome from the nucleocapsid core.
definition_reference: ISBN:0781702534

term: viral-cell fusion molecule activity
goid: GO:0019039
definition: none.
definition_reference: Pfam PF00523
definition_reference: GO:curators

term: viral-induced cell-cell fusion
goid: GO:0006948
definition: The process of syncytia-forming cell-cell fusion, caused by a virus.
definition_reference: ISBN:0781718325

term: virion
goid: GO:0019012
definition: The complete fully infectious extracellular virus particle.
definition_reference: ISBN:0781718325

term: virion attachment
goid: GO:0019062
definition: The processes involved in the specific binding of a viral antireceptor to a cell surface receptor.
definition_reference: ISBN:0781702534

***term:** virion attachment, binding of host cell surface coreceptor
goid: GO:0046814
definition: The process during virion attachment where a virion binds to a host cell surface receptor after an initial binding event has occurred, resulting in the fusion of the virion and host cell membranes and the initiation of viral entry.
definition_reference: ISBN:0879694971

***term:** virion attachment, binding of host cell surface receptor
goid: GO:0046813
definition: The process during virion attachment where a virion binds to a host cell receptor, resulting in a conformational change of the virus protein.
definition_reference: ISBN:0879694971

***term:** virion binding
goid: GO:0046790
definition: Interacting selectively with a virion, either by binding to components of the capsid or the viral envelope.
definition_reference: GO:ai

term: virion penetration
goid: GO:0019063
definition: The processes required for the introduction of virus particles into the cell.
definition_reference: ISBN:0781702534

***term:** virion transport
goid: GO:0046794
definition: The directed movement of a virion into, out of, or within a host cell.
definition_reference: GO:ai

***term:** virion transport vesicle
goid: GO:0046816
definition: A vesicle used to transport the partial or complete virion between cellular compartments.
definition_reference: PMID:7933124

term: virus induced gene silencing
goid: GO:0009616
definition: Specific posttranscriptional gene inactivation ('silencing') both of viral gene(s), and host gene(s) homologous to the viral genes. This silencing is triggered by viral infection, and occurs through a specific decrease in the level of mRNA of both host and viral genes.
definition_reference: GO:jl

term: virus-host interaction
goid: GO:0019048
definition: Interactions, directly with the host cell macromolecular machinery, to allow virus replication.
definition_reference: ISBN:0781718325

***term:** virus-induced modification of host RNA polymerase II
goid: GO:0046793
definition: The viral induction of modification to the host RNA polymerase II.
definition_reference: PMID:7637000

8.0 Appendix B: evidence codes and references for HHV-1 gene product annotations

GI Number	Gene	Gene Product	Go number	Reference(s)	Code
1944537	RL1	γ134.5, ICP34.5	GO:0046773	(Roizman and Knipe 2001)	TAS
1944537	RL1	γ134.5, ICP34.5	GO:0046792	(Brown, MacLean et al. 1997)	TAS
1944537	RL1	γ134.5, ICP34.5	GO:0019208	(Roizman and Knipe 2001)	TAS
1944537	RL1	γ134.5, ICP34.5	GO:0005515	(Roizman and Knipe 2001)	TAS
1944537	RL1	γ134.5, ICP34.5	GO:0008372	ND	ND
59500	RL2	α0, ICP0	GO:0019055	(Hobbs and DeLuca 1999)	TAS
59500	RL2	α0, ICP0	GO:0019053	(Eidson, Hobbs et al. 2002)	TAS
59500	RL2	α0, ICP0	GO:0006512	(Roizman and Knipe 2001)	TAS
59500	RL2	α0, ICP0	GO:0019054	(Hobbs and DeLuca 1999)	TAS
59500	RL2	α0, ICP0	GO:0019083	(Roizman and Knipe 2001)	TAS
59500	RL2	α0, ICP0	GO:0046811	(Hobbs and DeLuca 1999)	TAS
59500	RL2	α0, ICP0	GO:0005515	(Roizman and Knipe 2001)	TAS
59500	RL2	α0, ICP0	GO:0046818	(Roizman and Knipe 2001)	TAS
59502	UL1	gL	GO:0019063	(Roizman and Knipe 2001)	TAS
59502	UL1	gL	GO:0019064	(Roizman and Knipe 2001)	TAS
59502	UL1	gL	GO:0005554	ND	ND
59502	UL1	gL	GO:0019031	(Roizman and Knipe 2001)	TAS
59503	UL2		GO:0006281	(Sekino, Bruner et al. 2000)	NAS
59503	UL2		GO:0004844	(Roizman and Knipe 2001)	TAS
59503	UL2		GO:0005634	(Roizman and Knipe 2001)	TAS
59504	UL3		GO:0000004	ND	ND
59504	UL3		GO:0005554	ND	ND
59504	UL3		GO:0046818	(Roizman and Knipe 2001)	TAS
59505	UL4		GO:0000004	ND	ND
59505	UL4		GO:0005554	ND	ND
59505	UL4		GO:0019012	(Roizman and Knipe 2001)	TAS
59505	UL4		GO:0046818	(Roizman and Knipe 2001)	TAS
59507	UL5		GO:0019079	(Roizman and Knipe 2001)	TAS
59507	UL5		GO:0003678	(Roizman and Knipe 2001)	TAS
59507	UL5		GO:0016887	(Roizman and Knipe 2001)	TAS
59507	UL5		GO:0003677	(Biswas and Weller 2001)	TAS
59507	UL5		GO:0005524	(Roizman and Knipe 2001)	TAS
59507	UL5		GO:0046809	(Roizman and Knipe 2001)	TAS
59507	UL5		GO:0019034	(Roizman and Knipe 2001)	TAS
59506	UL6		GO:0019073	(Roizman and Knipe 2001)	TAS
59506	UL6		GO:0005554	ND	ND
59506	UL6		GO:0005634	(Roizman and Knipe 2001)	TAS
59506	UL6		GO:0046798	(Newcomb, Juhas et al. 2001)	TAS
59508	UL7		GO:0000004	ND	ND
59508	UL7		GO:0005554	ND	ND
59508	UL7		GO:0008372	ND	ND
59509	UL8		GO:0019060	(Roizman and Knipe 2001)	TAS
59509	UL8		GO:0045740	(Roizman and Knipe 2001)	TAS
59509	UL8		GO:0005515	(Roizman and Knipe 2001)	TAS
59509	UL8		GO:0019034	(Roizman and Knipe 2001)	TAS
VIDAUL8.5	UL8.5		GO:0000004	ND	ND
VIDAUL8.5	UL8.5		GO:0005554	ND	ND
VIDAUL8.5	UL8.5		GO:0008372	ND	ND
59511	UL9		GO:0019079	(Roizman and Knipe 2001)	TAS
59511	UL9		GO:0016887	(Roizman and Knipe 2001)	TAS
59511	UL9		GO:0004003	(Isler and Schaffer 2001)	NAS
59511	UL9		GO:0003688	(Roizman and Knipe 2001)	TAS
59511	UL9		GO:0005524	(Roizman and Knipe 2001)	TAS
59511	UL9		GO:0003677	(Roizman and Knipe 2001)	TAS
59511	UL9		GO:0046809	(Roizman and Knipe 2001)	TAS
59511	UL9		GO:0005634	(Roizman and Knipe 2001)	TAS
VIDAUL9.5	UL9.5		GO:0000004	ND	ND
VIDAUL9.5	UL9.5		GO:0005554	ND	ND
VIDAUL9.5	UL9.5		GO:0008372	ND	ND
59510	UL10	gM	GO:0046740	(MacLean, Robertson et al. 1993)	TAS
59510	UL10	gM	GO:0005554	ND	ND
59510	UL10	gM	GO:0005886	(Roizman and Knipe 2001)	TAS
59510	UL10	gM	GO:0019031	(Roizman and Knipe 2001)	TAS
VIDAUL10.5	UL10.5		GO:0000004	ND	ND
VIDAUL10.5	UL10.5		GO:0005554	ND	ND
VIDAUL10.5	UL10.5		GO:0008372	ND	ND

59512	UL11		GO:0046744	(Roizman and Knipe 2001)	TAS
59512	UL11		GO:0046788	(Baines and Roizman 1992)	TAS
59512	UL11		GO:0046745	(Baines and Roizman 1992)	TAS
59512	UL11		GO:0019060	(Baines and Roizman 1992)	TAS
59512	UL11		GO:0005554	ND	ND
59512	UL11		GO:0005634	(Baines, Jacob et al. 1995)	TAS
59512	UL11		GO:0046818	(Baines, Jacob et al. 1995)	TAS
59512	UL11		GO:0005637	(Baines, Jacob et al. 1995)	TAS
59512	UL11		GO:0012505	(Baines, Jacob et al. 1995)	TAS
59512	UL11		GO:0019033	(Loomis, Bowzard et al. 2001)	TAS
59513	UL12		GO:0019070	(Roizman and Knipe 2001)	TAS
59513	UL12		GO:0046802	(Shao, Rapp et al. 1993)	TAS
59513	UL12		GO:0004527	(Roizman and Knipe 2001)	TAS
59513	UL12		GO:0003677	(Roizman and Knipe 2001)	TAS
59513	UL12		GO:0004519	(Roizman and Knipe 2001)	TAS
59513	UL12		GO:0005634	(Roizman and Knipe 2001)	TAS
VIDAUL12.5	UL12.5		GO:0000291	(Bronstein, Weller et al. 1997)	TAS
VIDAUL12.5	UL12.5		GO:0000294	(Bronstein, Weller et al. 1997)	TAS
VIDAUL12.5	UL12.5		GO:0004519	(Bronstein, Weller et al. 1997)	TAS
VIDAUL12.5	UL12.5		GO:0004527	(Bronstein, Weller et al. 1997)	TAS
VIDAUL12.5	UL12.5		GO:0019030	(Bronstein, Weller et al. 1997)	TAS
59514	UL13		GO:0019051	(Hagglund, Munger et al. 2002)	TAS
59514	UL13		GO:0006468	(Roizman and Knipe 2001)	TAS
59514	UL13		GO:0004672	(Roizman and Knipe 2001)	TAS
59514	UL13		GO:0019012	(Roizman and Knipe 2001)	TAS
59846	UL14		GO:0046740	(Roizman and Knipe 2001)	TAS
59846	UL14		GO:0005554	ND	ND
59846	UL14		GO:0005737	(Cunningham, Davison et al. 2000)	TAS
59846	UL14		GO:0019033	(Roizman and Knipe 2001)	TAS
59846	UL14		GO:0046818	(Cunningham, Davison et al. 2000)	TAS
59501	UL15		GO:0019073	(Roizman and Knipe 2001)	TAS
59501	UL15		GO:0019071	(Baines, Poon et al. 1994)	TAS
59501	UL15		GO:0005524	(Yu and Weller 1998)	TAS
59501	UL15		GO:0046729	(Sheaffer, Newcomb et al. 2001)	TAS
59501	UL15		GO:0005634	(Yu and Weller 1998)	TAS
59501	UL15		GO:0046809	(Yu and Weller 1998)	TAS
59516	UL16		GO:0019073	(Roizman and Knipe 2001)	TAS
59516	UL16		GO:0019071	(Roizman and Knipe 2001)	TAS
59516	UL16		GO:0005554	ND	ND
59516	UL16		GO:0046808	(Nalwanga, Rempel et al. 1996)	TAS
59516	UL16		GO:0005737	(Nalwanga, Rempel et al. 1996)	TAS
59516	UL16		GO:0005634	(Nalwanga, Rempel et al. 1996)	TAS
59516	UL16		GO:0019012	(Nalwanga, Rempel et al. 1996)	TAS
59516	UL16		GO:0046809	(Nalwanga, Rempel et al. 1996)	TAS
59517	UL17		GO:0019071	(Roizman and Knipe 2001)	TAS
59517	UL17		GO:0019073	(Roizman and Knipe 2001)	TAS
59517	UL17		GO:0046742	(Roizman and Knipe 2001)	TAS
59517	UL17		GO:0005554	ND	ND
59517	UL17		GO:0005634	(Roizman and Knipe 2001)	TAS
59517	UL17		GO:0019033	(Roizman and Knipe 2001)	TAS
VIDAUL15.5	UL15.5		GO:0000004	ND	ND
VIDAUL15.5	UL15.5		GO:0005554	ND	ND
VIDAUL15.5	UL15.5		GO:0008372	ND	ND
59518	UL18	VP23	GO:0019073	(Roizman and Knipe 2001)	TAS
59518	UL18	VP23	GO:0019071	(Roizman and Knipe 2001)	TAS
59518	UL18	VP23	GO:0005554	ND	ND
59518	UL18	VP23	GO:0005634	(Roizman and Knipe 2001)	TAS
59518	UL18	VP23	GO:0019030	(Roizman and Knipe 2001)	TAS
59519	UL19	VP5, ICP5	GO:0000004	ND	ND
59519	UL19	VP5, ICP5	GO:0005554	ND	ND
59519	UL19	VP5, ICP5	GO:0005634	(Roizman and Knipe 2001)	TAS
59519	UL19	VP5, ICP5	GO:0019030	(Roizman and Knipe 2001)	TAS
59519	UL19	VP5, ICP5	GO:0046727	(Roizman and Knipe 2001)	TAS
59520	UL20		GO:0019060	(Baines, Ward et al. 1991)	TAS
59520	UL20		GO:0046788	(Roizman and Knipe 2001)	TAS
59520	UL20		GO:0005554	ND	ND
59520	UL20		GO:0005795	(Roizman and Knipe 2001)	TAS
59520	UL20		GO:0019031	(Roizman and Knipe 2001)	TAS
59520	UL20		GO:0005635	(Roizman and Knipe 2001)	TAS

59520	UL20		GO:0046816	(Ward, Campadelli-Fiume et al. 1994)	TAS
VIDAUL20.5	UL20.5		GO:0000004	ND	ND
VIDAUL20.5	UL20.5		GO:0005554	ND	ND
VIDAUL20.5	UL20.5		GO:0046818	(Roizman and Knipe 2001)	TAS
59521	UL21		GO:0000226	(Takakuwa, Goshima et al. 2001)	TAS
59521	UL21		GO:0046795	(Takakuwa, Goshima et al. 2001)	TAS
59521	UL21		GO:0046801	(Takakuwa, Goshima et al. 2001)	TAS
59521	UL21		GO:0046785	(Takakuwa, Goshima et al. 2001)	TAS
59521	UL21		GO:0005875	(Takakuwa, Goshima et al. 2001)	TAS
59521	UL21		GO:0008017	(Takakuwa, Goshima et al. 2001)	TAS
59521	UL21		GO:0005737	(Takakuwa, Goshima et al. 2001)	TAS
59521	UL21		GO:0019033	(Takakuwa, Goshima et al. 2001)	NAS
59522	UL22	gH	GO:0046740	(Roizman and Knipe 2001)	TAS
59522	UL22	gH	GO:0019063	(Roizman and Knipe 2001)	TAS
59522	UL22	gH	GO:0046788	(Roizman and Knipe 2001)	TAS
59522	UL22	gH	GO:0006948	(Roizman and Knipe 2001)	TAS
59522	UL22	gH	GO:0005554	ND	ND
59522	UL22	gH	GO:0019031	(Roizman and Knipe 2001)	TAS
59524	UL23	ICP36	GO:0019046	(Tenser 1991)	IDA
59524	UL23	ICP36	GO:0019206	(Roizman and Knipe 2001)	TAS
59524	UL23	ICP36	GO:0008372	ND	ND
59523	UL24		GO:0000004	ND	ND
59523	UL24		GO:0005554	ND	ND
59523	UL24		GO:0016020	(Roizman and Knipe 2001)	TAS
59525	UL25		GO:0019073	(Roizman and Knipe 2001)	TAS
59525	UL25		GO:0019063	(Roizman and Knipe 2001)	TAS
59525	UL25		GO:0046815	(Ogasawara, Suzutani et al. 2001)	TAS
59525	UL25		GO:0003677	(Ogasawara, Suzutani et al. 2001)	TAS
59525	UL25		GO:0005634	(Ogasawara, Suzutani et al. 2001)	TAS
59525	UL25		GO:0019030	(Roizman and Knipe 2001)	TAS
59525	UL25		GO:0005737	(Ogasawara, Suzutani et al. 2001)	TAS
59525	UL25		GO:0046729	(Ogasawara, Suzutani et al. 2001)	TAS
59526	UL26	VP24 & VP21	GO:0006508	(Roizman and Knipe 2001)	TAS
59526	UL26	VP24 & VP21	GO:0046807	(Roizman and Knipe 2001)	TAS
59526	UL26	VP24 & VP21	GO:0004252	(Roizman and Knipe 2001)	TAS
59526	UL26	VP24 & VP21	GO:0005634	(Roizman and Knipe 2001)	TAS
59526	UL26	VP24 & VP21	GO:0046806	(Roizman and Knipe 2001)	TAS
1944539	UL26.5	ICP35, VP22a	GO:0046752	(Nicholson, Addison et al. 1994)	TAS
1944539	UL26.5	ICP35, VP22a	GO:0046807	(Roizman and Knipe 2001)	TAS
1944539	UL26.5	ICP35, VP22a	GO:0005554	ND	ND
1944539	UL26.5	ICP35, VP22a	GO:0005634	(Roizman and Knipe 2001)	TAS
1944539	UL26.5	ICP35, VP22a	GO:0005737	(Nicholson, Addison et al. 1994)	TAS
1944539	UL26.5	ICP35, VP22a	GO:0046806	(Roizman and Knipe 2001)	TAS
59527	UL27	gB	GO:0019064	(Roizman and Knipe 2001)	TAS
59527	UL27	gB	GO:0046810	(Spear, Shieh et al. 1992)	TAS
59527	UL27	gB	GO:0019039	(Cai, Person et al. 1988; Turner, Bruun et al. 1998) (Davis-Poynter, Bell et al. 1994)	TAS
59527	UL27	gB	GO:0019031	(Roizman and Knipe 2001)	TAS
VIDAUL27.5	UL27.5		GO:0000004	ND	ND
VIDAUL27.5	UL27.5		GO:0005554	ND	ND
VIDAUL27.5	UL27.5		GO:0005737	(Chang, Menotti et al. 1998)	TAS
59528	UL28	ICP18.5	GO:0019073	(Roizman and Knipe 2001)	TAS
59528	UL28	ICP18.5	GO:0019071	(Roizman and Knipe 2001)	TAS
59528	UL28	ICP18.5	GO:0005554	ND	ND
59528	UL28	ICP18.5	GO:0005737	(Koslowski, Shaver et al. 1997)	TAS
59528	UL28	ICP18.5	GO:0005634	(Roizman and Knipe 2001)	TAS
59529	UL29	ICP8	GO:0019079	(Roizman and Knipe 2001)	TAS
59529	UL29	ICP8	GO:0046799	(Boehmer 1998)	TAS
59529	UL29	ICP8	GO:0046786	(Roizman and Knipe 2001)	TAS

59529	UL29	ICP8	GO:0003697	(Roizman and Knipe 2001)	TAS
59529	UL29	ICP8	GO:0019034	(Roizman and Knipe 2001)	TAS
59529	UL29	ICP8	GO:0046809	(Roizman and Knipe 2001)	TAS
59530	UL30		GO:0019079	(Roizman and Knipe 2001)	TAS
59530	UL30		GO:0003887	(Roizman and Knipe 2001)	TAS
59530	UL30		GO:0008408	(Roizman and Knipe 2001)	TAS
59530	UL30		GO:0046809	(Roizman and Knipe 2001)	TAS
59531	UL31		GO:0019071	(Ye, Vaughan et al. 2000)	NAS
59531	UL31		GO:0019073	(Ye, Vaughan et al. 2000)	NAS
59531	UL31		GO:0046771	(Reynolds, Ryckman et al. 2001)	TAS
59531	UL31		GO:0005554	ND	ND
59531	UL31		GO:0005634	(Roizman and Knipe 2001)	TAS
59533	UL32		GO:0019071	(Lamberti and Weller 1998)	TAS
59533	UL32		GO:0019073	(Roizman and Knipe 2001)	TAS
59533	UL32		GO:0046742	(Lamberti and Weller 1998)	TAS
59533	UL32		GO:0005554	ND	ND
59533	UL32		GO:0005737	(Lamberti and Weller 1998)	TAS
59533	UL32		GO:0005634	(Roizman and Knipe 2001)	TAS
59533	UL32		GO:0046809	(Lamberti and Weller 1998)	TAS
59532	UL33		GO:0019071	(al-Kobaisi, Rixon et al. 1991)	TAS
59532	UL33		GO:0019073	(Roizman and Knipe 2001)	TAS
59532	UL33		GO:0005554	ND	ND
59532	UL33		GO:0005737	(Reynolds, Fan et al. 2000)	TAS
59532	UL33		GO:0046809	(Reynolds, Fan et al. 2000)	TAS
59532	UL33		GO:0005634	(Roizman and Knipe 2001)	TAS
59534	UL34		GO:0046771	(Reynolds, Ryckman et al. 2001)	TAS
59534	UL34		GO:0005554	ND	ND
59534	UL34		GO:0019012	(Roizman and Knipe 2001)	TAS
59534	UL34		GO:0005635	(Ye, Vaughan et al. 2000)	TAS
59534	UL34		GO:0005641	(Ye, Vaughan et al. 2000)	TAS
59535	UL35	VP26	GO:0046797	(Chi and Wilson 2000)	TAS
59535	UL35	VP26	GO:0005554	ND	ND
59535	UL35	VP26	GO:0019030	(Roizman and Knipe 2001)	TAS
59535	UL35	VP26	GO:0005634	(Roizman and Knipe 2001)	TAS
59536	UL36	ICP1-2	GO:0046788	(Desai 2000)	TAS
59536	UL36	ICP1-2	GO:0019061	(Desai 2000)	TAS
59536	UL36	ICP1-2	GO:0019075	(Desai 2000)	TAS
59536	UL36	ICP1-2	GO:0019078	(Desai 2000)	TAS
59536	UL36	ICP1-2	GO:0005554	ND	ND
59536	UL36	ICP1-2	GO:0005634	(Desai 2000)	TAS
59536	UL36	ICP1-2	GO:0019033	(Roizman and Knipe 2001)	TAS
59536	UL36	ICP1-2	GO:0005737	(Desai 2000)	TAS
59537	UL37		GO:0046749	(Desai, Sexton et al. 2001)	TAS
59537	UL37		GO:0019075	(Desai, Sexton et al. 2001)	TAS
59537	UL37		GO:0046743	(Desai, Sexton et al. 2001)	TAS
59537	UL37		GO:0046788	(Desai, Sexton et al. 2001)	TAS
59537	UL37		GO:0046745	(Desai, Sexton et al. 2001)	TAS
59537	UL37		GO:0005554	ND	ND
59537	UL37		GO:0019033	(Roizman and Knipe 2001)	TAS
59537	UL37		GO:0005737	(Roizman and Knipe 2001)	TAS
59537	UL37		GO:0005634	(Roizman and Knipe 2001)	TAS
59538	UL38	ICP32, VP19C	GO:0046752	(Rixon, Addison et al. 1996)	TAS
59538	UL38	ICP32, VP19C	GO:0019069	(Roizman and Knipe 2001)	TAS
59538	UL38	ICP32, VP19C	GO:0003677	(Roizman and Knipe 2001)	TAS
59538	UL38	ICP32, VP19C	GO:0019030	(Newcomb, Trus et al. 1993)	TAS
59538	UL38	ICP32, VP19C	GO:0005634	(Rixon, Addison et al. 1996)	TAS
59539	UL39	ICP6	GO:0009186	(Roizman and Knipe 2001)	TAS
59539	UL39	ICP6	GO:0019079	(Roizman and Knipe 2001)	TAS
59539	UL39	ICP6	GO:0046733	(Salvucci, Bonneau et al. 1995)	IDA
59539	UL39	ICP6	GO:0046777	(Roizman and Knipe 2001)	TAS
59539	UL39	ICP6	GO:0004672	(Roizman and Knipe 2001)	TAS
59539	UL39	ICP6	GO:0004748	(Roizman and Knipe 2001)	TAS
59539	UL39	ICP6	GO:0005971	(Roizman and Knipe 2001)	TAS
59540	UL40		GO:0009186	(Roizman and Knipe 2001)	TAS
59540	UL40		GO:0019079	(Roizman and Knipe 2001)	TAS
59540	UL40		GO:0004748	(Roizman and Knipe 2001)	TAS
59540	UL40		GO:0005971	(Roizman and Knipe 2001)	TAS
59541	UL41	vhs	GO:0006355	(Fenwick and Clark 1982)	TAS
59541	UL41	vhs	GO:0046776	(Hill, Barnett et al. 1994)	TAS
59541	UL41	vhs	GO:0046775	(Suzutani, Nagamine et al. 2000)	TAS
59541	UL41	vhs	GO:0046774	(Suzutani, Nagamine et al. 2000)	TAS
59541	UL41	vhs	GO:0046783	(Sydiskis and Roizman 1968)	TAS
59541	UL41	vhs	GO:0019049	(Suzutani, Nagamine et al.	TAS

				2000)	
59541	UL41	vhs	GO:0019054	(Roizman and Knipe 2001)	TAS
59541	UL41	vhs	GO:0000294	(Elgadi and Smiley 1999)	TAS
59541	UL41	vhs	GO:0019033	(Roizman and Knipe 2001)	TAS
59541	UL41	vhs	GO:0005737	(Elgadi, Hayes et al. 1999)	TAS
59542	UL42		GO:0019079	(Roizman and Knipe 2001)	TAS
59542	UL42		GO:0030337	(Roizman and Knipe 2001)	TAS
59542	UL42		GO:0003677	(Roizman and Knipe 2001)	TAS
59542	UL42		GO:0046809	(Roizman and Knipe 2001)	TAS
59543	UL43		GO:0000004	ND	ND
59543	UL43		GO:0005554	ND	ND
59543	UL43		GO:0008372	ND	ND
VIDAUL43.5	UL43.5		GO:0000004	ND	ND
VIDAUL43.5	UL43.5		GO:0005554	ND	ND
VIDAUL43.5	UL43.5		GO:0046808	(Roizman and Knipe 2001)	TAS
59544	UL44	gC, VP7.5	GO:0046800	(Lubinski, Wang et al. 1999)	TAS
59544	UL44	gC, VP7.5	GO:0046791	(Fries, Friedman et al. 1986) (Kostavasili, Sahu et al. 1997)	TAS
59544	UL44	gC, VP7.5	GO:0019062	(Herold, WuDunn et al. 1991) (WuDunn and Spear 1989)	TAS
59544	UL44	gC, VP7.5	GO:0005554	ND	ND
59544	UL44	gC, VP7.5	GO:0019031	(Roizman and Knipe 2001)	TAS
59545	UL45		GO:0000004	ND	ND
59545	UL45		GO:0005554	ND	ND
59545	UL45		GO:0019031	(Visalli and Brandt 2002)	NAS
1944540	UL46	VP11/12	GO:0000004	ND	ND
1944540	UL46	VP11/12	GO:0005554	ND	ND
1944540	UL46	VP11/12	GO:0005641	(Willard 2002)	TAS
1944540	UL46	VP11/12	GO:0005737	(Willard 2002)	TAS
1944540	UL46	VP11/12	GO:0019033	(Roizman and Knipe 2001)	TAS
1944540	UL46	VP11/12	GO:0005886	(Willard 2002)	TAS
59547	UL47	VP13/14	GO:0000004	ND	ND
59547	UL47	VP13/14	GO:0005554	ND	ND
59547	UL47	VP13/14	GO:0019033	(Roizman and Knipe 2001)	TAS
59547	UL47	VP13/14	GO:0005634	(Donnelly and Elliott 2001)	TAS
59547	UL47	VP13/14	GO:0005737	(Donnelly and Elliott 2001)	TAS
59548	UL48	VP16, ICP25, αTIF	GO:0019085	(Roizman and Knipe 2001)	TAS
59548	UL48	VP16, ICP25, αTIF	GO:0046782	(Roizman and Knipe 2001)	TAS
59548	UL48	VP16, ICP25, αTIF	GO:0046788	(Mossman, Sherburne et al. 2000)	TAS
59548	UL48	VP16, ICP25, αTIF	GO:0005515	(Roizman and Knipe 2001)	TAS
59548	UL48	VP16, ICP25, αTIF	GO:0016563	(Roizman and Knipe 2001)	TAS
59548	UL48	VP16, ICP25, αTIF	GO:0005634	(Roizman and Knipe 2001)	TAS
59548	UL48	VP16, ICP25, αTIF	GO:0005737	(Roizman and Knipe 2001)	TAS
59548	UL48	VP16, ICP25, αTIF	GO:0019033	(Roizman and Knipe 2001)	TAS
59549	UL49	VP22	GO:0046740	(Elliott and O'Hare 1997)	TAS
59549	UL49	VP22	GO:0003682	(Pomeranz and Blaho 1999)	TAS
59549	UL49	VP22	GO:0019033	(Roizman and Knipe 2001)	TAS
59549	UL49	VP22	GO:0005634	(Pomeranz and Blaho 1999) (Elliott and O'Hare 2000)	TAS
59549	UL49	VP22	GO:0005737	(Pomeranz and Blaho 1999; Elliott and O'Hare 2000)	TAS
1944541	UL49.5		GO:0000004	ND	ND
1944541	UL49.5		GO:0005554	ND	ND
1944541	UL49.5		GO:0019031	(McGeoch, Dolan et al. 1986; Dolan, McKie et al. 1992)	TAS
59550	UL50	dUTPase	GO:0006399	(Roizman and Knipe 2001)	TAS
59550	UL50	dUTPase	GO:0009117	(Roizman and Knipe 2001)	TAS
59550	UL50	dUTPase	GO:0004170	(Roizman and Knipe 2001)	TAS
59550	UL50	dUTPase	GO:0008372	ND	ND
59551	UL51		GO:0000004	ND	ND
59551	UL51		GO:0005554	ND	ND
59551	UL51		GO:0005641	(Daikoku, Ikenoya et al. 1998)	TAS
59551	UL51		GO:0005634	(Daikoku, Ikenoya et al. 1998)	TAS
59551	UL51		GO:0019012	(Daikoku, Ikenoya et al. 1998)	TAS
59552	UL52		GO:0019079	(Roizman and Knipe 2001)	TAS
59552	UL52		GO:0003677	(Biswas and Weller 2001)	TAS
59552	UL52		GO:0003678	(Roizman and Knipe 2001)	TAS
59552	UL52		GO:0003896	(Roizman and Knipe 2001)	TAS
59552	UL52		GO:0019034	(Roizman and Knipe 2001)	TAS
59552	UL52		GO:0046809	(Roizman and Knipe 2001)	TAS

59553	UL53	gK	GO:0046788	(Roizman and Knipe 2001)	TAS
59553	UL53	gK	GO:0046801	(Jayachandra, Baghian et al. 1997)	TAS
59553	UL53	gK	GO:0005554	ND	ND
59553	UL53	gK	GO:0005641	(Jayachandra, Baghian et al. 1997)	TAS
59553	UL53	gK	GO:0005634	(Jayachandra, Baghian et al. 1997)	TAS
59553	UL53	gK	GO:0005794	(Foster, Rybachuk et al. 2001)	TAS
59553	UL53	gK	GO:0005789	(Rajcani and Kudelova 1999)	TAS
59553	UL53	gK	GO:0019031	(Roizman and Knipe 2001)	TAS
59553	UL53	gK	GO:0005635	(Rajcani and Kudelova 1999)	TAS
59554	UL54	α27, ICP27	GO:0019056	(Roizman and Knipe 2001)	TAS
59554	UL54	α27, ICP27	GO:0045071	(Rice, Su et al. 1989)	TAS
59554	UL54	α27, ICP27	GO:0046779	(Hardwicke and Sandri-Goldin 1994; Hardy and Sandri-Goldin 1994)	TAS
59554	UL54	α27, ICP27	GO:0045070	(Roizman and Knipe 2001)	TAS
59554	UL54	α27, ICP27	GO:0046781	(Phelan, Carmo-Fonseca et al. 1993)	TAS
59554	UL54	α27, ICP27	GO:0046780	(Roizman and Knipe 2001)	TAS
59554	UL54	α27, ICP27	GO:0046784	(Sandri-Goldin 1998)	TAS
59554	UL54	α27, ICP27	GO:0046782	(Roizman and Knipe 2001)	TAS
59554	UL54	α27, ICP27	GO:0016564	(Roizman and Knipe 2001)	TAS
59554	UL54	α27, ICP27	GO:0003723	(Roizman and Knipe 2001)	TAS
59554	UL54	α27, ICP27	GO:0005737	(Roizman and Knipe 2001)	TAS
59554	UL54	α27, ICP27	GO:0005634	(Roizman and Knipe 2001)	TAS
59554	UL54	α27, ICP27	GO:0046809	(Roizman and Knipe 2001)	TAS
59555	UL55		GO:0000004	ND	ND
59555	UL55		GO:0005554	ND	ND
59555	UL55		GO:0046808	(Roizman and Knipe 2001)	TAS
59555	UL55		GO:0005634	(Roizman and Knipe 2001)	TAS
1944542	UL56		GO:0046803	(Roizman and Knipe 2001)	TAS
1944542	UL56		GO:0005554	ND	ND
1944542	UL56		GO:0019012	(Roizman and Knipe 2001)	TAS
1944542	UL56		GO:0005634	(Roizman and Knipe 2001)	TAS
59558	RS1	α4, ICP4	GO:0019083	(Roizman and Knipe 2001)	TAS
59558	RS1	α4, ICP4	GO:0007050	(Song, Yeh et al. 2001)	TAS
59558	RS1	α4, ICP4	GO:0019055	(Song, Yeh et al. 2001)	TAS
59558	RS1	α4, ICP4	GO:0046782	(Roizman and Knipe 2001)	TAS
59558	RS1	α4, ICP4	GO:0003677	(Roizman and Knipe 2001)	TAS
59558	RS1	α4, ICP4	GO:0030528	(Roizman and Knipe 2001)	TAS
59558	RS1	α4, ICP4	GO:0046809	(Knipe and Smith 1986)	TAS
59558	RS1	α4, ICP4	GO:0005634	(Roizman and Knipe 2001)	TAS
59559	Us1	α22, ICP22	GO:0046793	(Ogle and Roizman 1999)	TAS
59559	Us1	α22, ICP22	GO:0005554	ND	ND
59559	Us1	α22, ICP22	GO:0046818	(Ogle and Roizman 1999)	TAS
59559	Us1	α22, ICP22	GO:0005634	(Ogle and Roizman 1999)	TAS
59559	Us1	α22, ICP22	GO:0005737	(Ogle and Roizman 1999)	TAS
VIDAUs1.5	Us1.5		GO:0019051	(Hagglund, Munger et al. 2002)	TAS
VIDAUs1.5	Us1.5		GO:0005554	ND	ND
VIDAUs1.5	Us1.5		GO:0008372	ND	ND
59560	Us2		GO:0000004	ND	ND
59560	Us2		GO:0005554	ND	ND
59560	Us2		GO:0005634	(Roizman and Knipe 2001)	TAS
59560	Us2		GO:0019033	(Roizman and Knipe 2001)	TAS
59561	Us3		GO:0006468	(Roizman and Knipe 2001)	TAS
59561	Us3		GO:0019050	(Roizman and Knipe 2001)	TAS
59561	Us3		GO:0004672	(Roizman and Knipe 2001)	TAS
59561	Us3		GO:0008372	ND	ND
59562	Us4	gC	GO:0019062	(Tran, Kissner et al. 2000)	TAS
59562	Us4	gC	GO:0005554	ND	ND
59562	Us4	gC	GO:0019031	(Roizman and Knipe 2001)	TAS
59563	Us5	gJ	GO:0019050	(Roizman and Knipe 2001)	TAS
59563	Us5	gJ	GO:0005554	ND	ND
59563	Us5	gJ	GO:0008372	ND	ND
59564	Us6	gD, VP17/18	GO:0019064	(Cai, Person et al. 1988; Davis-Poynter, Bell et al. 1994; Turner, Bruun et al. 1998)	TAS
59564	Us6	gD, VP17/18	GO:0043066	(Zhou, Galvan et al. 2000)	TAS
59564	Us6	gD, VP17/18	GO:0046789	(Roizman and Knipe 2001)	TAS
59564	Us6	gD, VP17/18	GO:0046814	(Roizman and Knipe 2001)	TAS
59564	Us6	gD, VP17/18	GO:0019031	(Roizman and Knipe 2001)	TAS
59565	Us7	gI	GO:0046740	(Roizman and Knipe 2001)	TAS
59565	Us7	gI	GO:0005554	ND	ND
59565	Us7	gI	GO:0005794	(McMillan and Johnson 2001)	TAS

59565	Us7	gI	GO:0019031	(Roizman and Knipe 2001)	TAS
59566	Us8	gE	GO:0046740	(Roizman and Knipe 2001)	TAS
59566	Us8	gE	GO:0005554	ND	ND
59566	Us8	gE	GO:0005794	(McMillan and Johnson 2001)	TAS
59566	Us8	gE	GO:0019031	(Roizman and Knipe 2001)	TAS
1944544	Us8.5		GO:0000004	ND	ND
1944544	Us8.5		GO:0005554	ND	ND
1944544	Us8.5		GO:0005730	(Georgopoulou, Kakkanas et al. 1995)	TAS
59567	Us9		GO:0000004	ND	ND
59567	Us9		GO:0005554	ND	ND
59567	Us9		GO:0019033	(Roizman and Knipe 2001)	TAS
59567	Us9		GO:0019031	(Roizman and Knipe 2001)	TAS
59568	Us10		GO:0000004	ND	ND
59568	Us10		GO:0005554	ND	ND
59568	Us10		GO:0005634	(Yamada, Daikoku et al. 1997)	TAS
59568	Us10		GO:0019030	(Yamada, Daikoku et al. 1997)	TAS
59568	Us10		GO:0019033	(Roizman and Knipe 2001)	TAS
59569	Us11		GO:0019052	(Cassady and Gross 2002)	TAS
59569	Us11		GO:0046773	(Roizman and Knipe 2001)	TAS
59569	Us11		GO:0003723	(Roizman and Knipe 2001)	TAS
59569	Us11		GO:0015935	(Roizman and Knipe 2001)	TAS
59569	Us11		GO:0019033	(Roizman and Knipe 2001)	TAS
59569	Us11		GO:0005730	(Roizman and Knipe 2001)	TAS
59570	Us12	α47, ICP47	GO:0019053	(Roizman and Knipe 2001)	TAS
59570	Us12	α47, ICP47	GO:0048019	(Ahn, Meyer et al. 1996; Tomazin, Hill et al. 1996)	TAS
59570	Us12	α47, ICP47	GO:0005783	(Neumann, Kraas et al. 1997)	TAS

Appendix B References:

Ahn, K., T. H. Meyer, et al. (1996). "Molecular mechanism and species specificity of TAP inhibition by herpes simplex virus ICP47." Embo J **15**(13): 3247-55.

al-Kobaisi, M. F., F. J. Rixon, et al. (1991). "The herpes simplex virus UL33 gene product is required for the assembly of full capsids." Virology **180**(1): 380-8.

Baines, J. D., R. J. Jacob, et al. (1995). "The herpes simplex virus 1 UL11 proteins are associated with cytoplasmic and nuclear membranes and with nuclear bodies of infected cells." J Virol **69**(2): 825-33.

Baines, J. D., A. P. Poon, et al. (1994). "The herpes simplex virus 1 UL15 gene encodes two proteins and is required for cleavage of genomic viral DNA." J Virol **68**(12): 8118-24.

Baines, J. D. and B. Roizman (1992). "The UL11 gene of herpes simplex virus 1 encodes a function that facilitates nucleocapsid envelopment and egress from cells." J Virol **66**(8): 5168-74.

Baines, J. D., P. L. Ward, et al. (1991). "The UL20 gene of herpes simplex virus 1 encodes a function necessary for viral egress." J Virol **65**(12): 6414-24.

Biswas, N. and S. K. Weller (2001). "The UL5 and UL52 subunits of the herpes simplex virus type 1 helicase-primase subcomplex exhibit a complex interdependence for DNA binding." J Biol Chem **276**(20): 17610-9.

Boehmer, P. E. (1998). "The herpes simplex virus type-1 single-strand DNA-binding protein, ICP8, increases the processivity of the UL9 protein DNA helicase." J Biol Chem **273**(5): 2676-83.

Bronstein, J. C., S. K. Weller, et al. (1997). "The product of the UL12.5 gene of herpes simplex virus type 1 is a capsid-associated nuclease." J Virol **71**(4): 3039-47.

Brown, S. M., A. R. MacLean, et al. (1997). "The herpes simplex virus virulence factor ICP34.5 and the cellular protein MyD116 complex with proliferating cell nuclear antigen through the 63-amino-acid domain conserved in ICP34.5, MyD116, and GADD34." J Virol **71**(12): 9442-9.

Cai, W. Z., S. Person, et al. (1988). "Functional regions and structural features of the gB glycoprotein of herpes simplex virus type 1. An analysis of linker insertion mutants." J Mol Biol **201**(3): 575-88.

Cassady, K. A. and M. Gross (2002). "The herpes simplex virus type 1 U(S)11 protein interacts with protein kinase R in infected cells and requires a 30-amino-acid sequence adjacent to a kinase substrate domain." J Virol **76**(5): 2029-35.

Chang, Y. E., L. Menotti, et al. (1998). "UL27.5 is a novel gamma2 gene antisense to the herpes simplex virus 1 gene encoding glycoprotein B." J Virol **72**(7): 6056-64.

Chi, J. H. and D. W. Wilson (2000). "ATP-Dependent localization of the herpes simplex virus capsid protein VP26 to sites of procapsid maturation." J Virol **74**(3): 1468-76.

Cunningham, C., A. J. Davison, et al. (2000). "Herpes simplex virus type 1 gene UL14: phenotype of a null mutant and identification of the encoded protein." J Virol **74**(1): 33-41.

Daikoku, T., K. Ikenoya, et al. (1998). "Identification and characterization of the herpes simplex virus type 1 UL51 gene product." J Gen Virol **79 (Pt 12)**: 3027-31.

Davis-Poynter, N., S. Bell, et al. (1994). "Analysis of the contributions of herpes simplex virus type 1 membrane proteins to the induction of cell-cell fusion." J Virol **68**(11): 7586-90.

Desai, P., G. L. Sexton, et al. (2001). "A null mutation in the gene encoding the herpes simplex virus type 1 UL37 polypeptide abrogates virus maturation." J Virol **75**(21): 10259-71.

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." J Virol **74**(24): 11608-18.

Dolan, A., E. McKie, et al. (1992). "Status of the ICP34.5 gene in herpes simplex virus type 1 strain 17." J Gen Virol **73 (Pt 4)**: 971-3.

Donnelly, M. and G. Elliott (2001). "Nuclear localization and shuttling of herpes simplex virus tegument protein VP13/14." J Virol **75**(6): 2566-74.

Eidson, K. M., W. E. Hobbs, et al. (2002). "Expression of herpes simplex virus ICP0 inhibits the induction of interferon-stimulated genes by viral infection." J Virol **76**(5): 2180-91.

Elgadi, M. M., C. E. Hayes, et al. (1999). "The herpes simplex virus vhs protein induces endoribonucleolytic cleavage of target RNAs in cell extracts." J Virol **73**(9): 7153-64.

Elgadi, M. M. and J. R. Smiley (1999). "Picornavirus internal ribosome entry site elements target RNA cleavage events induced by the herpes simplex virus virion host shutoff protein." J Virol **73**(11): 9222-31.

Elliott, G. and P. O'Hare (1997). "Intercellular trafficking and protein delivery by a herpesvirus structural protein." Cell **88**(2): 223-33.

Elliott, G. and P. O'Hare (2000). "Cytoplasm-to-nucleus translocation of a herpesvirus tegument protein during cell division." J Virol **74**(5): 2131-41.

Fenwick, M. L. and J. Clark (1982). "Early and delayed shut-off of host protein synthesis in cells infected with herpes simplex virus." J Gen Virol **61 (Pt 1)**: 121-5.

Foster, T. P., G. V. Rybachuk, et al. (2001). "Glycoprotein K specified by herpes simplex virus type 1 is expressed on virions as a Golgi complex-dependent glycosylated species and functions in virion entry." J Virol **75**(24): 12431-8.

Fries, L. F., H. M. Friedman, et al. (1986). "Glycoprotein C of herpes simplex virus 1 is an inhibitor of the complement cascade." J Immunol **137**(5): 1636-41.

Georgopoulou, U., A. Kakkanas, et al. (1995). "Characterization of the US8.5 protein of herpes simplex virus." Arch Virol **140**(12): 2227-41.

Hagglund, R., J. Munger, et al. (2002). "U(S)3 protein kinase of herpes simplex virus 1 blocks caspase 3 activation induced by the products of U(S)1.5 and U(L)13 genes and modulates expression of transduced U(S)1.5 open reading frame in a cell type-specific manner." J Virol **76**(2): 743-54.

Hardwicke, M. A. and R. M. Sandri-Goldin (1994). "The herpes simplex virus regulatory protein ICP27 contributes to the decrease in cellular mRNA levels during infection." J Virol **68**(8): 4797-810.

Hardy, W. R. and R. M. Sandri-Goldin (1994). "Herpes simplex virus inhibits host cell splicing, and regulatory protein ICP27 is required for this effect." J Virol **68**(12): 7790-9.

Herold, B. C., D. WuDunn, et al. (1991). "Glycoprotein C of herpes simplex virus type 1 plays a principal role in the adsorption of virus to cells and in infectivity." J Virol **65**(3): 1090-8.

Hill, A. B., B. C. Barnett, et al. (1994). "HLA class I molecules are not transported to the cell surface in cells infected with herpes simplex virus types 1 and 2." J Immunol **152**(6): 2736-41.

Hobbs, W. E., 2nd and N. A. DeLuca (1999). "Perturbation of cell cycle progression and cellular gene expression as a function of herpes simplex virus ICP0." J Virol **73**(10): 8245-55.

Isler, J. A. and P. A. Schaffer (2001). "Origin binding protein-containing protein-DNA complex formation at herpes simplex virus type 1 oriS: role in oriS-dependent DNA replication." J Virol **75**(15): 6808-16.

Jayachandra, S., A. Baghian, et al. (1997). "Herpes simplex virus type 1 glycoprotein K is not essential for infectious virus production in actively replicating cells but is required for efficient envelopment and translocation of infectious virions from the cytoplasm to the extracellular space." J Virol **71**(7): 5012-24.

Knipe, D. M. and J. L. Smith (1986). "A mutant herpesvirus protein leads to a block in nuclear localization of other viral proteins." Mol Cell Biol **6**(7): 2371-81.

Koslowski, K. M., P. R. Shaver, et al. (1997). "The pseudorabies virus UL28 protein enters the nucleus after coexpression with the herpes simplex virus UL15 protein." J Virol **71**(12): 9118-23.

Kostavasili, I., A. Sahu, et al. (1997). "Mechanism of complement inactivation by glycoprotein C of herpes simplex virus." J Immunol **158**(4): 1763-71.

Lamberti, C. and S. K. Weller (1998). "The herpes simplex virus type 1 cleavage/packaging protein, UL32, is involved in efficient localization of capsids to replication compartments." J Virol **72**(3): 2463-73.

Loomis, J. S., J. B. Bowzard, et al. (2001). "Intracellular trafficking of the UL11 tegument protein of herpes simplex virus type 1." J Virol **75**(24): 12209-19.

Lubinski, J., L. Wang, et al. (1999). "In vivo role of complement-interacting domains of herpes simplex virus type 1 glycoprotein gC." J Exp Med **190**(11): 1637-46.

MacLean, C. A., L. M. Robertson, et al. (1993). "Characterization of the UL10 gene product of herpes simplex virus type 1 and investigation of its role in vivo." J Gen Virol **74 (Pt 6)**: 975-83.

McGeoch, D. J., A. Dolan, et al. (1986). "Complete DNA sequence of the short repeat region in the genome of herpes simplex virus type 1." Nucleic Acids Res **14**(4): 1727-45.

McMillan, T. N. and D. C. Johnson (2001). "Cytoplasmic domain of herpes simplex virus gE causes accumulation in the trans-Golgi network, a site of virus envelopment and sorting of virions to cell junctions." J Virol **75**(4): 1928-40.

Mossman, K. L., R. Sherburne, et al. (2000). "Evidence that herpes simplex virus VP16 is required for viral egress downstream of the initial envelopment event." J Virol **74**(14): 6287-99.

Nalwanga, D., S. Rempel, et al. (1996). "The UL 16 gene product of herpes simplex virus 1 is a virion protein that colocalizes with intranuclear capsid proteins." Virology **226**(2): 236-42.

Neumann, L., W. Kraas, et al. (1997). "The active domain of the herpes simplex virus protein ICP47: a potent inhibitor of the transporter associated with antigen processing." J Mol Biol **272**(4): 484-92.

Newcomb, W. W., R. M. Juhas, et al. (2001). "The UL6 gene product forms the portal for entry of DNA into the herpes simplex virus capsid." J Virol **75**(22): 10923-32.

Newcomb, W. W., B. L. Trus, et al. (1993). "Structure of the herpes simplex virus capsid. Molecular composition of the pentons and the triplexes." J Mol Biol **232**(2): 499-511.

Nicholson, P., C. Addison, et al. (1994). "Localization of the herpes simplex virus type 1 major capsid protein VP5 to the cell nucleus requires the abundant scaffolding protein VP22a." J Gen Virol **75 (Pt 5)**: 1091-9.

Ogasawara, M., T. Suzutani, et al. (2001). "Role of the UL25 gene product in packaging DNA into the herpes simplex virus capsid: location of UL25 product in the capsid and demonstration that it binds DNA." J Virol **75**(3): 1427-36.

Ogle, W. O. and B. Roizman (1999). "Functional anatomy of herpes simplex virus 1 overlapping genes encoding infected-cell protein 22 and US1.5 protein." J Virol **73**(5): 4305-15.

Phelan, A., M. Carmo-Fonseca, et al. (1993). "A herpes simplex virus type 1 immediate-early gene product, IE63, regulates small nuclear ribonucleoprotein distribution." Proc Natl Acad Sci U S A **90**(19): 9056-60.

Pomeranz, L. E. and J. A. Blaho (1999). "Modified VP22 localizes to the cell nucleus during synchronized herpes simplex virus type 1 infection." J Virol **73**(8): 6769-81.

Rajcani, J. and M. Kudelova (1999). "Glycoprotein K of herpes simplex virus: a transmembrane protein encoded by the UL53 gene which regulates membrane fusion." Virus Genes **18**(1): 81-90.

Reynolds, A. E., Y. Fan, et al. (2000). "Characterization of the U(L)33 gene product of herpes simplex virus 1." Virology **266**(2): 310-8.

Reynolds, A. E., B. J. Ryckman, et al. (2001). "U(L)31 and U(L)34 proteins of herpes simplex virus type 1 form a complex that accumulates at the nuclear rim and is required for envelopment of nucleocapsids." J Virol **75**(18): 8803-17.

Rice, S. A., L. S. Su, et al. (1989). "Herpes simplex virus alpha protein ICP27 possesses separable positive and negative regulatory activities." J Virol **63**(8): 3399-407.

Rixon, F. J., C. Addison, et al. (1996). "Multiple interactions control the intracellular localization of the herpes simplex virus type 1 capsid proteins." J Gen Virol **77 (Pt 9)**: 2251-60.

Roizman, B. and D. M. Knipe (2001). Chapter 72: Herpes Simplex Viruses and Their Replication. Fields Virology. D. M. Knipe and P. M. Howley. Philadelphia, Lippincott Williams & Wilkins. **2:** 2399-2460.

Salvucci, L. A., R. H. Bonneau, et al. (1995). "Polymorphism within the herpes simplex virus (HSV) ribonucleotide reductase large subunit (ICP6) confers type specificity for recognition by HSV type 1-specific cytotoxic T lymphocytes." J Virol **69**(2): 1122-31.

Sandri-Goldin, R. M. (1998). "ICP27 mediates HSV RNA export by shuttling through a leucine-rich nuclear export signal and binding viral intronless RNAs through an RGG motif." Genes Dev **12**(6): 868-79.

Sekino, Y., S. D. Bruner, et al. (2000). "Selective inhibition of herpes simplex virus type-1 uracil-DNA glycosylase by designed substrate analogs." J Biol Chem **275**(47): 36506-8.

Shao, L., L. M. Rapp, et al. (1993). "Herpes simplex virus 1 alkaline nuclease is required for efficient egress of capsids from the nucleus." Virology **196**(1): 146-62.

Sheaffer, A. K., W. W. Newcomb, et al. (2001). "Herpes simplex virus DNA cleavage and packaging proteins associate with the procapsid prior to its maturation." J Virol **75**(2): 687-98.

Song, B., K. C. Yeh, et al. (2001). "Herpes simplex virus gene products required for viral inhibition of expression of G1-phase functions." Virology **290**(2): 320-8.

Spear, P. G., M. T. Shieh, et al. (1992). "Heparan sulfate glycosaminoglycans as primary cell surface receptors for herpes simplex virus." Adv Exp Med Biol **313**: 341-53.

Suzutani, T., M. Nagamine, et al. (2000). "The role of the UL41 gene of herpes simplex virus type 1 in evasion of non-specific host defence mechanisms during primary infection." J Gen Virol **81**(Pt 7): 1763-71.

Sydiskis, R. J. and B. Roizman (1968). "The sedimentation profiles of cytoplasmic polyribosomes in mammalian cells productively and abortively infected with herpes simplex virus." Virology **34**(3): 562-5.

Takakuwa, H., F. Goshima, et al. (2001). "Herpes simplex virus encodes a virion-associated protein which promotes long cellular processes in over-expressing cells." Genes Cells **6**(11): 955-66.

Tenser, R. B. (1991). "Role of herpes simplex virus thymidine kinase expression in viral pathogenesis and latency." Intervirology **32**(2): 76-92.

Tomazin, R., A. B. Hill, et al. (1996). "Stable binding of the herpes simplex virus ICP47 protein to the peptide binding site of TAP." Embo J **15**(13): 3256-66.

Tran, L. C., J. M. Kissner, et al. (2000). "A herpes simplex virus 1 recombinant lacking the glycoprotein G coding sequences is defective in entry through apical surfaces of polarized epithelial cells in culture and in vivo." Proc Natl Acad Sci U S A **97**(4): 1818-22.

Turner, A., B. Bruun, et al. (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**(1): 873-5.

Visalli, R. J. and C. R. Brandt (2002). "Mutation of the herpes simplex virus 1 KOS UL45 gene reveals dose dependent effects on central nervous system growth." Arch Virol **147**(3): 519-32.

Ward, P. L., G. Campadelli-Fiume, et al. (1994). "Localization and putative function of the UL20 membrane protein in cells infected with herpes simplex virus 1." J Virol **68**(11): 7406-17.

Willard, M. (2002). "Rapid directional translocations in virus replication." J Virol **76**(10): 5220-32.

WuDunn, D. and P. G. Spear (1989). "Initial interaction of herpes simplex virus with cells is binding to heparan sulfate." J Virol **63**(1): 52-8.

Yamada, H., T. Daikoku, et al. (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)**: 2923-31.

Ye, G. J., K. T. Vaughan, et al. (2000). "The herpes simplex virus 1 U(L)34 protein interacts with a cytoplasmic dynein intermediate chain and targets nuclear membrane." J Virol **74**(3): 1355-63.

Yu, D. and S. K. Weller (1998). "Genetic analysis of the UL 15 gene locus for the putative terminase of herpes simplex virus type 1." Virology **243**(1): 32-44.

Zhou, G., V. Galvan, et al. (2000). "Glycoprotein D or J delivered in trans blocks apoptosis in SK-N-SH cells induced by a herpes simplex virus 1 mutant lacking intact genes expressing both glycoproteins." J Virol **74**(24): 11782-91.

9.0 Bibliography

Ahn, K., T. H. Meyer, et al. (1996). "Molecular mechanism and species specificity of TAP inhibition by herpes simplex virus ICP47." Embo J **15**(13): 3247-55.

Alba, M. M., R. Das, et al. (2001a). "Genomewide function conservation and phylogeny in the Herpesviridae." Genome Res **11**(1): 43-54.

Alba, M. M., D. Lee, et al. (2001). "VIDA: a virus database system for the organization of animal virus genome open reading frames." Nucleic Acids Res **29**(1): 133-6.

Alcami, A. and U. H. Koszinowski (2000). "Viral mechanisms of immune evasion." Trends Microbiol **8**(9): 410-8.

al-Kobaisi, M. F., F. J. Rixon, et al. (1991). "The herpes simplex virus UL33 gene product is required for the assembly of full capsids." Virology **180**(1): 380-8.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Altschul, S. F. and E. V. Koonin (1998). "Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases." Trends in Biochemical Sciences **23**(11): 444-447.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Armstrong, J. A., H. G. Pereira, et al. (1961). "Observations on the virus of infectious bovine rhinotracheitis, and its affinity with the Herpesvirus group." Virology **14**: 276-85.

Arrand, J. R., L. Rymo, et al. (1981). "Molecular cloning of the complete Epstein-Barr virus genome as a set of overlapping restriction endonuclease fragments." Nucleic Acids Res **9**(13): 2999-3014.

Arvanitakis, L., E. Geras-Raaka, et al. (1997). "Human herpesvirus KSHV encodes a constitutively active G-protein-coupled receptor linked to cell proliferation." Nature **385**(6614): 347-50.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Asher, Y., M. Heller, et al. (1969). "Incorporation of lipids into herpes simplex virus particles." J Gen Virol **4**(1): 65-76.

Attwood, T. K., P. Bradley, et al. (2003). "PRINTS and its automatic supplement, prePRINTS." Nucleic Acids Res **31**(1): 400-2.

Baines, J. D., R. J. Jacob, et al. (1995). "The herpes simplex virus 1 UL11 proteins are associated with cytoplasmic and nuclear membranes and with nuclear bodies of infected cells." J Virol **69**(2): 825-33.

Baines, J. D., A. P. Poon, et al. (1994). "The herpes simplex virus 1 UL15 gene encodes two proteins and is required for cleavage of genomic viral DNA." J Virol **68**(12): 8118-24.

Baines, J. D. and B. Roizman (1992). "The UL11 gene of herpes simplex virus 1 encodes a function that facilitates nucleocapsid envelopment and egress from cells." J Virol **66**(8): 5168-74.

Baines, J. D., P. L. Ward, et al. (1991). "The UL20 gene of herpes simplex virus 1 encodes a function necessary for viral egress." J Virol **65**(12): 6414-24.

Bairoch, A. (2000). "The ENZYME database in 2000." Nucleic Acids Res **28**(1): 304-5.

Ball, C. A., K. Dolinski, et al. (2000). "Integrating functional genomic information into the Saccharomyces genome database." Nucleic Acids Res **28**(1): 77-80.

Bateman, A., E. Birney, et al. (2000). "The Pfam protein families database." Nucleic Acids Res **28**(1): 263-6.

Bateman, A., L. Coin, et al. (2004). "The Pfam protein families database." Nucleic Acids Res **32 Database issue**: D138-41.

Belanger, C., A. Gravel, et al. (2001). "Human herpesvirus 8 viral FLICE-inhibitory protein inhibits Fas-mediated apoptosis through binding and prevention of procaspase-8 maturation." J Hum Virol **4**(2): 62-73.

Berriz, G. F., J. V. White, et al. (2003). "GoFish finds genes with combinations of Gene Ontology attributes." Bioinformatics **19**(6): 788-9.

Biassoni, R., C. Cantoni, et al. (2001). "Human natural killer cell receptors and co-receptors." Immunol Rev **181**: 203-14.

Biswas, M., J. F. O'Rourke, et al. (2002). "Applications of InterPro in protein annotation and genome analysis." Brief Bioinform **3**(3): 285-95.

Biswas, N. and S. K. Weller (2001). "The UL5 and UL52 subunits of the herpes simplex virus type 1 helicase-primase subcomplex exhibit a complex interdependence for DNA binding." J Biol Chem **276**(20): 17610-9.

Blaschke, C. and A. Valencia (2002). "Automatic ontology construction from the literature." Genome Inform Ser Workshop Genome Inform **13**: 201-13.

Boeckmann, B., A. Bairoch, et al. (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." Nucleic Acids Res **31**(1): 365-70.

Boehmer, P. E. (1998). "The herpes simplex virus type-1 single-strand DNA-binding protein, ICP8, increases the processivity of the UL9 protein DNA helicase." J Biol Chem **273**(5): 2676-83.

Boname, J. M. and P. G. Stevenson (2001). "MHC class I ubiquitination by a viral PHD/LAP finger protein." Immunity **15**(4): 627-36.

Bono, H., I. Nikaido, et al. (2003). "Comprehensive analysis of the mouse metabolome based on the transcriptome." Genome Res **13**(6B): 1345-9.

Boppana, S. B., R. F. Pass, et al. (1992). "Symptomatic congenital cytomegalovirus infection: neonatal morbidity and mortality." Pediatr Infect Dis J **11**(2): 93-9.

Boshart, M., F. Weber, et al. (1985). "A very strong enhancer is located upstream of an immediate early gene of human cytomegalovirus." Cell **41**(2): 521-30.

Bronstein, J. C., S. K. Weller, et al. (1997). "The product of the UL12.5 gene of herpes simplex virus type 1 is a capsid-associated nuclease." J Virol **71**(4): 3039-47.

Brown, S. M., A. R. MacLean, et al. (1997). "The herpes simplex virus virulence factor ICP34.5 and the cellular protein MyD116 complex with proliferating cell nuclear antigen through the 63-amino-acid domain conserved in ICP34.5, MyD116, and GADD34." J Virol **71**(12): 9442-9.

Browne, H., B. Bruun, et al. (2001). "Plasma membrane requirements for cell fusion induced by herpes simplex virus type 1 glycoproteins gB, gD, gH and gL." J Gen Virol **82**(Pt 6): 1419-22.

Brunovskis, P. and H. J. Kung (1995). "Retrotransposition and herpesvirus evolution." Virus Genes **11**(2-3): 259-70.

Bucher, P., K. Karplus, et al. (1996). "A flexible motif search technique based on generalized profiles." Comput Chem **20**(1): 3-23.

Bugert, J. J. and G. Darai (2000). "Poxvirus homologues of cellular genes." Virus Genes **21**(1-2): 111-33.

Bult, C. J., J. A. Blake, et al. (2004). "The Mouse Genome Database (MGD): integrating biology with the genome." Nucleic Acids Res **32 Database issue**: D476-81.

Cai, W. Z., S. Person, et al. (1988). "Functional regions and structural features of the gB glycoprotein of herpes simplex virus type 1. An analysis of linker insertion mutants." J Mol Biol **201**(3): 575-88.

Camon, E., M. Magrane, et al. (2003). "The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro." Genome Res **13**(4): 662-72.

Cantor, M. N., I. N. Sarkar, et al. (2003). "An evaluation of hybrid methods for matching biomedical terminologies: mapping the gene ontology to the UMLS." Stud Health Technol Inform **95**: 62-7.

Cartier, A., T. Komai, et al. (2003). "The Us3 protein kinase of herpes simplex virus 1 blocks apoptosis and induces phosporylation of the Bcl-2 family member Bad." Exp Cell Res **291**(1): 242-50.

Cassady, K. A. and M. Gross (2002). "The herpes simplex virus type 1 U(S)11 protein interacts with protein kinase R in infected cells and requires a 30-amino-acid sequence adjacent to a kinase substrate domain." J Virol **76**(5): 2029-35.

Cebo, C., G. Vergoten, et al. (2002). "Lectin activities of cytokines: functions and putative carbohydrate-recognition domains." Biochim Biophys Acta **1572**(2-3): 422-34.

Chambers, J., A. Angulo, et al. (1999). "DNA microarrays of the complex human cytomegalovirus genome: profiling kinetic class with drug sensitivity of viral gene expression." J Virol **73**(7): 5757-66.

Chang, Y. E., L. Menotti, et al. (1998). "UL27.5 is a novel gamma2 gene antisense to the herpes simplex virus 1 gene encoding glycoprotein B." J Virol **72**(7): 6056-64.

Chang, Y. N., S. Crawford, et al. (1990). "The palindromic series I repeats in the simian cytomegalovirus major immediate-early promoter behave as both strong basal enhancers and cyclic AMP response elements." J Virol **64**(1): 264-77.

Chaston, T. B. and B. A. Lidbury (2001). "Genetic 'budget' of viruses and the cost to the infected host: a theory on the relationship between the genetic capacity of viruses, immune evasion, persistence and disease." Immunol Cell Biol **79**(1): 62-6.

Chee, M. S., A. T. Bankier, et al. (1990). "Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169." Curr Top Microbiol Immunol **154**: 125-69.

Chee, M. S., S. C. Satchwell, et al. (1990a). "Human cytomegalovirus encodes three G protein-coupled receptor homologues." Nature **344**(6268): 774-7.

Chi, J. H. and D. W. Wilson (2000). "ATP-Dependent localization of the herpes simplex virus capsid protein VP26 to sites of procapsid maturation." J Virol **74**(3): 1468-76.

Cinatl, J., Jr., R. Blaheta, et al. (2000). "Decreased neutrophil adhesion to human cytomegalovirus-infected retinal pigment epithelial cells is mediated by virus-

induced up-regulation of Fas ligand independent of neutrophil apoptosis." J Immunol **165**(8): 4405-13.

Cinatl, J., Jr., R. Kotchetkov, et al. (2000). "The antisense oligonucleotide ISIS 2922 prevents cytomegalovirus-induced upregulation of IL-8 and ICAM-1 in cultured human fibroblasts." J Med Virol **60**(3): 313-23.

Clark, J., S. Edwards, et al. (2002). "Identification of amplified and expressed genes in breast cancer by comparative hybridization onto microarrays of randomly selected cDNA clones." Genes Chromosomes Cancer **34**(1): 104-14.

Corpet, F. (1988). "Multiple sequence alignment with hierarchical clustering." Nucleic Acids Res **16**(22): 10881-90.

Coscoy, L., D. J. Sanchez, et al. (2001). "A novel class of herpesvirus-encoded membrane-bound E3 ubiquitin ligases regulates endocytosis of proteins involved in immune recognition." J Cell Biol **155**(7): 1265-73.

Cunningham, C., A. J. Davison, et al. (2000). "Herpes simplex virus type 1 gene UL14: phenotype of a null mutant and identification of the encoded protein." J Virol **74**(1): 33-41.

Daikoku, T., K. Ikenoya, et al. (1998). "Identification and characterization of the herpes simplex virus type 1 UL51 gene product." J Gen Virol **79 (Pt 12)**: 3027-31.

Dairaghi, D. J., D. R. Greaves, et al. (1998). "Abduction of Chemokine Elements by Herpesviruses." Seminars in Virology **8**(5): 377-385.

Damania, B. and R. C. Desrosiers (2001). "Simian homologues of human herpesvirus 8." Philos Trans R Soc Lond B Biol Sci **356**(1408): 535-43.

Dargan, D. J., F. E. Jamieson, et al. (1997). "The published DNA sequence of human cytomegalovirus strain AD169 lacks 929 base pairs affecting genes UL42 and UL43." J Virol **71**(12): 9833-6.

Davison, A. J., A. Dolan, et al. (2003). "The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome." J Gen Virol **84**(Pt 1): 17-28.

Davison, A. J. and J. E. Scott (1986). "The complete DNA sequence of varicella-zoster virus." J Gen Virol **67 (Pt 9)**: 1759-816.

Davis-Poynter, N., S. Bell, et al. (1994). "Analysis of the contributions of herpes simplex virus type 1 membrane proteins to the induction of cell-cell fusion." J Virol **68**(11): 7586-90.

Davis-Poynter, N. J., M. Degli-Esposti, et al. (1999). "Murine cytomegalovirus homologues of cellular immunomodulatory genes." Intervirology **42**(5-6): 331-41.

Davis-Poynter, N. J. and H. E. Farrell (1996). "Masters of deception: a review of herpesvirus immune evasion strategies." Immunol Cell Biol **74**(6): 513-22.

Day, A. J. (1994). "The C-type carbohydrate recognition domain (CRD) superfamily." Biochem Soc Trans **22**(1): 83-8.

Dayhoff, M. O., R. M. Schwartz, et al. (1978). Atlas of Protein Sequence and Structure. M. O. Dayhoff. Washington, DC, NBRF. **5 (3):** 345.

Dennis, G., Jr., B. T. Sherman, et al. (2003). "DAVID: Database for Annotation, Visualization, and Integrated Discovery." Genome Biol **4**(5): P3.

Desai, P., G. L. Sexton, et al. (2001). "A null mutation in the gene encoding the herpes simplex virus type 1 UL37 polypeptide abrogates virus maturation." J Virol **75**(21): 10259-71.

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." J Virol **74**(24): 11608-18.

Dolan, A., F. E. Jamieson, et al. (1998). "The genome sequence of herpes simplex virus type 2." J Virol **72**(3): 2010-21.

Dolan, A., E. McKie, et al. (1992). "Status of the ICP34.5 gene in herpes simplex virus type 1 strain 17." J Gen Virol **73 (Pt 4)**: 971-3.

Doniger, S. W., N. Salomonis, et al. (2003). "MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data." Genome Biol **4**(1): R7.

Donnelly, M. and G. Elliott (2001). "Nuclear localization and shuttling of herpes simplex virus tegument protein VP13/14." J Virol **75**(6): 2566-74.

Dunn, W., C. Chou, et al. (2003). "Functional profiling of a human cytomegalovirus genome." Proc Natl Acad Sci U S A **100**(24): 14223-8.

Eidson, K. M., W. E. Hobbs, et al. (2002). "Expression of herpes simplex virus ICP0 inhibits the induction of interferon-stimulated genes by viral infection." J Virol **76**(5): 2180-91.

Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci U S A **95**(25): 14863-8.

Elgadi, M. M., C. E. Hayes, et al. (1999). "The herpes simplex virus vhs protein induces endoribonucleolytic cleavage of target RNAs in cell extracts." J Virol **73**(9): 7153-64.

Elgadi, M. M. and J. R. Smiley (1999). "Picornavirus internal ribosome entry site elements target RNA cleavage events induced by the herpes simplex virus virion host shutoff protein." J Virol **73**(11): 9222-31.

Elliott, G. and P. O'Hare (1997). "Intercellular trafficking and protein delivery by a herpesvirus structural protein." Cell **88**(2): 223-33.

Elliott, G. and P. O'Hare (2000). "Cytoplasm-to-nucleus translocation of a herpesvirus tegument protein during cell division." J Virol **74**(5): 2131-41.

Epstein, M. A. (1962). "Observations on the mode of release of herpes virus from infected HeLa cells." J Cell Biol **12**: 589-97.

Fenwick, M. L. and J. Clark (1982). "Early and delayed shut-off of host protein synthesis in cells infected with herpes simplex virus." J Gen Virol **61 (Pt l)**: 121-5.

Fiers, W., R. Contreras, et al. (1976). "Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene." Nature **260**(5551): 500-7.

Fiers, W., R. Contreras, et al. (1978). "Complete nucleotide sequence of SV40 DNA." Nature **273**(5658): 113-20.

Fisher, S., O. Genbacev, et al. (2000). "Human cytomegalovirus infection of placental cytotrophoblasts in vitro and in utero: implications for transmission and pathogenesis." J Virol **74**(15): 6808-20.

Flybase Consortium. (2003). "The FlyBase database of the Drosophila genome projects and community literature." Nucleic Acids Res **31**(1): 172-5.

Foster, T. P., G. V. Rybachuk, et al. (2001). "Glycoprotein K specified by herpes simplex virus type 1 is expressed on virions as a Golgi complex-dependent glycosylated species and functions in virion entry." J Virol **75**(24): 12431-8.

Fries, L. F., H. M. Friedman, et al. (1986). "Glycoprotein C of herpes simplex virus 1 is an inhibitor of the complement cascade." J Immunol **137**(5): 1636-41.

Fruehling, S. and R. Longnecker (1997). "The immunoreceptor tyrosine-based activation motif of Epstein-Barr virus LMP2A is essential for blocking BCR-mediated signal transduction." Virology **235**(2): 241-51.

Garavelli, J. S. (2003). "The RESID Database of Protein Modifications: 2003 developments." Nucleic Acids Res **31**(1): 499-501.

Garcia-Hernandez, M., T. Z. Berardini, et al. (2002). "TAIR: a resource for integrated Arabidopsis data." Funct Integr Genomics **2**(6): 239-53.

Gene Ontology Consortium. (2001). "Creating the gene ontology resource: design and implementation." Genome Res **11**(8): 1425-33.

Georgopoulou, U., A. Kakkanas, et al. (1995). "Characterization of the US8.5 protein of herpes simplex virus." Arch Virol **140**(12): 2227-41.

Glenn, M., L. Rainbow, et al. (1999). "Identification of a spliced gene from Kaposi's sarcoma-associated herpesvirus encoding a protein with similarities to latent membrane proteins 1 and 2A of Epstein-Barr virus." J Virol **73**(8): 6953-63.

Gompels, U. A., J. Nicholas, et al. (1995). "The DNA sequence of human herpesvirus-6: structure, coding content, and genome evolution." Virology **209**(1): 29-51.

Gotoh, O. (1982). "An improved algorithm for matching biological sequences." J Mol Biol **162**(3): 705-8.

Gough, J., K. Karplus, et al. (2001). "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure." J Mol Biol **313**(4): 903-19.

Gouzy, J., F. Corpet, et al. (1999). "Whole genome protein domain analysis using a new method for domain clustering." Comput Chem **23**(3-4): 333-40.

Gouzy, J., P. Eugene, et al. (1997). "XDOM, a graphical tool to analyse domain arrangements in any set of protein sequences." Comput Appl Biosci **13**(6): 601-8.

Gribaldo, S., D. Casane, et al. (2003). "Functional divergence prediction from evolutionary analysis: a case study of vertebrate hemoglobin." Mol Biol Evol **20**(11): 1754-9.

Guettler, S., E. N. Jackson, et al. (2003). "ESTs from the basidiomycete Schizophyllum commune grown on nitrogen-replete and nitrogen-limited media." Fungal Genet Biol **39**(2): 191-8.

Haft, D. H., J. D. Selengut, et al. (2003). "The TIGRFAMs database of protein families." Nucleic Acids Res **31**(1): 371-3.

Hagglund, R., J. Munger, et al. (2002). "U(S)3 protein kinase of herpes simplex virus 1 blocks caspase 3 activation induced by the products of U(S)1.5 and U(L)13 genes and modulates expression of transduced U(S)1.5 open reading frame in a cell type-specific manner." J Virol **76**(2): 743-54.

Hahn, G., R. Jores, et al. (1998). "Cytomegalovirus remains latent in a common precursor of dendritic and myeloid cells." Proc Natl Acad Sci U S A **95**(7): 3937-42.

Hammarstrom, S. (1999). "The carcinoembryonic antigen (CEA) family: structures, suggested functions and expression in normal and malignant tissues." Semin Cancer Biol **9**(2): 67-81.

Hardwicke, M. A. and R. M. Sandri-Goldin (1994). "The herpes simplex virus regulatory protein ICP27 contributes to the decrease in cellular mRNA levels during infection." J Virol **68**(8): 4797-810.

Hardy, W. R. and R. M. Sandri-Goldin (1994). "Herpes simplex virus inhibits host cell splicing, and regulatory protein ICP27 is required for this effect." J Virol **68**(12): 7790-9.

Harhay, G. P. and J. W. Keele (2003). "Positional candidate gene selection from livestock EST databases using Gene Ontology." Bioinformatics **19**(2): 249-55.

Harris, M. A., J. Clark, et al. (2004). "The Gene Ontology (GO) database and informatics resource." Nucleic Acids Res **32 Database issue**: D258-61.

Harris, T. W., N. Chen, et al. (2004). "WormBase: a multi-species resource for nematode biology and genomics." Nucleic Acids Res **32 Database issue**: D411-7.

Henikoff, J. G., E. A. Greene, et al. (2000). "Increased coverage of protein families with the blocks database servers." Nucleic Acids Res **28**(1): 228-30.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proc Natl Acad Sci U S A **89**(22): 10915-9.

Henikoff, S., J. G. Henikoff, et al. (1999). "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations." Bioinformatics **15**(6): 471-9.

Hennig, S., D. Groth, et al. (2003). "Automated Gene Ontology annotation for anonymous sequence data." Nucleic Acids Res **31**(13): 3712-5.

Herold, B. C., D. WuDunn, et al. (1991). "Glycoprotein C of herpes simplex virus type 1 plays a principal role in the adsorption of virus to cells and in infectivity." J Virol **65**(3): 1090-8.

Higgins, D. G. and W. R. Taylor (2000). "Multiple sequence alignment." Methods Mol Biol **143**: 1-18.

Hill, A. B., B. C. Barnett, et al. (1994). "HLA class I molecules are not transported to the cell surface in cells infected with herpes simplex virus types 1 and 2." J Immunol **152**(6): 2736-41.

Hill, D. P., D. A. Begley, et al. (2004). "The mouse Gene Expression Database (GXD): updates and enhancements." Nucleic Acids Res **32 Database issue**: D568-71.

Hill, D. P., J. A. Blake, et al. (2002). "Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies." Genome Res **12**(12): 1982-91.

Hill, J. M., W. J. Lukiw, et al. (2001). "Gene expression analyzed by microarrays in HSV-1 latent mouse trigeminal ganglion following heat stress." Virus Genes **23**(3): 273-80.

Hiscock, D. and C. Upton (2000). "Viral Genome DataBase: storing and analyzing genes and proteins from complete viral genomes." Bioinformatics **16**(5): 484-5.

Hobbs, W. E., 2nd and N. A. DeLuca (1999). "Perturbation of cell cycle progression and cellular gene expression as a function of herpes simplex virus ICP0." J Virol **73**(10): 8245-55.

Hodges, P. E., P. M. Carrico, et al. (2002). "Annotating the human proteome: the Human Proteome Survey Database (HumanPSD) and an in-depth target database for G protein-coupled receptors (GPCR-PD) from Incyte Genomics." Nucleic Acids Res **30**(1): 137-41.

Holzerlandt, R., C. Orengo, et al. (2002). "Identification of new herpesvirus gene homologs in the human genome." Genome Res **12**(11): 1739-48.

Huang, H., W. C. Barker, et al. (2003). "iProClass: an integrated database of protein family, function and structure information." Nucleic Acids Res **31**(1): 390-2.

Hughes, A. L. (2002). "Origin and evolution of viral interleukin-10 and other DNA virus genes with vertebrate homologues." J Mol Evol **54**(1): 90-101.

Hulo, N., C. J. Sigrist, et al. (2004). "Recent improvements to the PROSITE database." Nucleic Acids Res **32 Database issue**: D134-7.

Hunninghake, G. W., M. M. Monick, et al. (1989). "The promoter-regulatory region of the major immediate-early gene of human cytomegalovirus responds to T-lymphocyte stimulation and contains functional cyclic AMP-response elements." J Virol **63**(7): 3026-33.

Ichikawa, T., Y. Suzuki, et al. (1997). "Identification and role of adenylyl cyclase in auxin signalling in higher plants." Nature **390**(6661): 698-701.

Ichikawa, T., Y. Suzuki, et al. (1998). "Identification and role of adenylyl cyclase in auxin signalling in higher plants." Nature **396**(6709): 390.

IHGSC, I. H. G. S. C. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Immergluck, L. C., M. S. Domowicz, et al. (1998). "Viral and cellular requirements for entry of herpes simplex virus type 1 into primary neuronal cells." J Gen Virol **79 (Pt 3)**: 549-59.

Irmler, M., M. Thome, et al. (1997). "Inhibition of death receptor signals by cellular FLIP." Nature **388**(6638): 190-5.

Ishido, S., C. Wang, et al. (2000). "Downregulation of major histocompatibility complex class I molecules by Kaposi's sarcoma-associated herpesvirus K3 and K5 proteins." J Virol **74**(11): 5300-9.

Isler, J. A. and P. A. Schaffer (2001). "Origin binding protein-containing protein-DNA complex formation at herpes simplex virus type 1 oriS: role in oriS-dependent DNA replication." J Virol **75**(15): 6808-16.

Janes, R. W., P. B. Munroe, et al. (1996). "A model for Batten disease protein CLN3: functional implications from homology and mutations." FEBS Lett **399**(1-2): 75-7.

Jayachandra, S., A. Baghian, et al. (1997). "Herpes simplex virus type 1 glycoprotein K is not essential for infectious virus production in actively replicating cells but is required for efficient envelopment and translocation of infectious virions from the cytoplasm to the extracellular space." J Virol **71**(7): 5012-24.

Jenner, R. G., M. M. Alba, et al. (2001). "Kaposi's sarcoma-associated herpesvirus latent and lytic gene expression as revealed by DNA arrays." J Virol **75**(2): 891-902.

Jenner, R. G. and C. Boshoff (2002). "The molecular pathology of Kaposi's sarcoma-associated herpesvirus." BBA - Reviews on Cancer **1602**(1): 1-22.

Jensen, L. J., R. Gupta, et al. (2003). "Prediction of human protein function according to Gene Ontology categories." Bioinformatics **19**(5): 635-42.

Jenssen, T. K., A. Laegreid, et al. (2001). "A literature network of human genes for high-throughput analysis of gene expression." Nat Genet **28**(1): 21-8.

Johnson, R. A., S. M. Huong, et al. (2000). "Activation of the mitogen-activated protein kinase p38 by human cytomegalovirus infection through two distinct pathways: a novel mechanism for activation of p38." J Virol **74**(3): 1158-67.

Johnson, R. A., X. L. Ma, et al. (2001). "The role of MKK1/2 kinase activity in human cytomegalovirus infection." J Gen Virol **82**(Pt 3): 493-7.

Jones, J. O. and A. M. Arvin (2003). "Microarray analysis of host cell gene transcription in response to varicella-zoster virus infection of human T cells and fibroblasts in vitro and SCIDhu skin xenografts in vivo." J Virol **77**(2): 1268-80.

Kanehisa, M., S. Goto, et al. (2004). "The KEGG resource for deciphering the genome." Nucleic Acids Res **32 Database issue**: D277-80.

Kantor, R., R. Machekano, et al. (2001). "Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database: an expanded data model integrating natural language text and sequence analysis programs." Nucleic Acids Res **29**(1): 296-9.

Karsch-Mizrachi, I. and B. F. Ouellette (2001). "The GenBank sequence database." Methods Biochem Anal **43**: 45-63.

Kellam, P. and M. M. Alba (2002). "Virus bioinformatics: databases and recent applications." Appl Bioinformatics **1**(1): 37-42.

Kellam, P. and X. Liu (2003). Chapter 14: Experimental Use of DNA Arrays. Bioinformatics: genes, proteins and computers. C. Orengo, D. Jones and J. Thornton. Oxford, BIOS Scientific Publishers Ltd. **1:** 217-228.

Khan, S., G. Situ, et al. (2003). "GoFigure: automated Gene Ontology annotation." Bioinformatics **19**(18): 2484-5.

King, O. D., R. E. Foulger, et al. (2003). "Predicting gene function from patterns of annotation." Genome Res **13**(5): 896-904.

King, O. D., J. C. Lee, et al. (2003). "Predicting phenotype from patterns of annotation." Bioinformatics **19 Suppl 1**: i183-9.

Knipe, D. M., P. M. Howley, et al. (2001). Field's Virology. Philadelphia, Lippincott Williams & Wilkins.

Knipe, D. M. and J. L. Smith (1986). "A mutant herpesvirus protein leads to a block in nuclear localization of other viral proteins." Mol Cell Biol **6**(7): 2371-81.

Kohonen, T. (1995). Self-Organizing Maps. Heidelberg, Springer.

Koslowski, K. M., P. R. Shaver, et al. (1997). "The pseudorabies virus UL28 protein enters the nucleus after coexpression with the herpes simplex virus UL15 protein." J Virol **71**(12): 9118-23.

Kostavasili, I., A. Sahu, et al. (1997). "Mechanism of complement inactivation by glycoprotein C of herpes simplex virus." J Immunol **158**(4): 1763-71.

Kotenko, S. V., S. Saccani, et al. (2000). "Human cytomegalovirus harbors its own unique IL-10 homolog (cmvIL-10)." Proc Natl Acad Sci U S A **97**(4): 1695-700.

Kulikova, T., P. Aldebert, et al. (2004). "The EMBL Nucleotide Sequence Database." Nucleic Acids Res **32 Database issue**: D27-30.

Lafontaine, D., S. Mercure, et al. (1997). "Update of the viroid and viroid-like sequence database: addition of a hepatitis delta virus RNA section." Nucleic Acids Res **25**(1): 123-5.

Lagreid, A., T. R. Hvidsten, et al. (2003). "Predicting gene ontology biological process from temporal gene expression patterns." Genome Res **13**(5): 965-79.

Lalani, A. S., J. W. Barrett, et al. (2000). "Modulating chemokines: more lessons from viruses." Immunol Today **21**(2): 100-6.

Lamberti, C. and S. K. Weller (1998). "The herpes simplex virus type 1 cleavage/packaging protein, UL32, is involved in efficient localization of capsids to replication compartments." J Virol **72**(3): 2463-73.

Lecointe, D., N. Dugas, et al. (2002). "Human cytomegalovirus infection reduces surface CCR5 expression in human microglial cells, astrocytes and monocyte-derived macrophages." Microbes Infect **4**(14): 1401-8.

Letunic, I., R. R. Copley, et al. (2004). "SMART 4.0: towards genomic data integration." Nucleic Acids Res **32 Database issue**: D142-4.

Ljungman, P. (1996). "Cytomegalovirus infections in transplant patients." Scand J Infect Dis Suppl **100**: 59-63.

Loomis, J. S., J. B. Bowzard, et al. (2001). "Intracellular trafficking of the UL11 tegument protein of herpes simplex virus type 1." J Virol **75**(24): 12209-19.

Lord, P. W., R. D. Stevens, et al. (2003). "Semantic similarity measures as tools for exploring the gene ontology." Pac Symp Biocomput: 601-12.

Lorenzo, M. E., J. U. Jung, et al. (2002). "Kaposi's sarcoma-associated herpesvirus K3 utilizes the ubiquitin-proteasome system in routing class major histocompatibility complexes to late endocytic compartments." J Virol **76**(11): 5522-31.

Lubinski, J., L. Wang, et al. (1999). "In vivo role of complement-interacting domains of herpes simplex virus type 1 glycoprotein gC." J Exp Med **190**(11): 1637-46.

Lubinski, J. M., L. Wang, et al. (1998). "Herpes simplex virus type 1 glycoprotein gC mediates immune evasion in vivo." J Virol **72**(10): 8257-63.

Macken, C., H. Lu, et al. (2001). The value of a database in surveillance and vaccine selection. Options for the Control of Influenza IV. A. D. M. E. Osterhaus, N. Cox and A. W. Hampson. Amsterdam, Elsevier Science: 103-106.

MacLean, C. A., L. M. Robertson, et al. (1993). "Characterization of the UL10 gene product of herpes simplex virus type 1 and investigation of its role in vivo." J Gen Virol **74 (Pt 6)**: 975-83.

Maxwell, P. (1999). "Carcinoembryonic antigen: cell adhesion molecule and useful diagnostic marker." Br J Biomed Sci **56**(3): 209-14.

McCarter, J. P., M. D. Mitreva, et al. (2003). "Analysis and functional classification of transcripts from the nematode Meloidogyne incognita." Genome Biol **4**(4): R26.

McCormick, A. L., V. L. Smith, et al. (2003). "Disruption of mitochondrial networks by the human cytomegalovirus UL37 gene product viral mitochondrion-localized inhibitor of apoptosis." J Virol **77**(1): 631-41.

McCray, A. T., A. C. Browne, et al. (2002). "The lexical properties of the gene ontology." Proc AMIA Symp: 504-8.

McFadden, G. and P. M. Murphy (2000). "Host-related immunomodulators encoded by poxviruses and herpesviruses." Curr Opin Microbiol **3**(4): 371-8.

McGeoch, D. J. (2001). "Molecular evolution of the gamma-Herpesvirinae." Philos. Trans. R. Soc. Lond. B. Biol. Sci. **356**(1408): 421-35.

McGeoch, D. J. and S. Cook (1994). "Molecular phylogeny of the alphaherpesvirinae subfamily and a proposed evolutionary timescale." J Mol Biol **238**(1): 9-22.

McGeoch, D. J., S. Cook, et al. (1995). "Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses." J Mol Biol **247**(3): 443-58.

McGeoch, D. J., M. A. Dalrymple, et al. (1988). "The complete DNA sequence of the long unique region in the genome of herpes simplex virus type 1." J Gen Virol **69 (Pt 7)**: 1531-74.

McGeoch, D. J. and A. J. Davison (1999). "The descent of human herpesvirus 8." Semin Cancer Biol **9**(3): 201-9.

McGeoch, D. J., A. Dolan, et al. (1986). "Complete DNA sequence of the short repeat region in the genome of herpes simplex virus type 1." Nucleic Acids Res **14**(4): 1727-45.

McGeoch, D. J., A. Dolan, et al. (1985). "Sequence determination and genetic content of the short unique region in the genome of herpes simplex virus type 1." J Mol Biol **181**(1): 1-13.

McMillan, T. N. and D. C. Johnson (2001). "Cytoplasmic domain of herpes simplex virus gE causes accumulation in the trans-Golgi network, a site of virus envelopment and sorting of virions to cell junctions." J Virol **75**(4): 1928-40.

Means, R. E., S. Ishido, et al. (2002). "Multiple endocytic trafficking pathways of MHC class I molecules induced by a Herpesvirus protein." Embo J **21**(7): 1638-49.

Meier, J. L. and M. F. Stinski (1996). "Regulation of human cytomegalovirus immediate-early gene expression." Intervirology **39**(5-6): 331-42.

Miyazaki, S., H. Sugawara, et al. (2004). "DDBJ in the stream of various biological data." Nucleic Acids Res **32 Database issue**: D31-4.

Mocarski, E. S. and C. T. Courcelle (2001). Chapter 76: Cytomegaloviruses and Their Replication. Fields Virology. D. M. Knipe and P. M. Howley. Philadelphia, Lippincott Williams & Wilkins. **2:** 2629-2673.

Momma, Y., C. N. Nagineni, et al. (2003). "Differential expression of chemokines by human retinal pigment epithelial cells infected with cytomegalovirus." Invest Ophthalmol Vis Sci **44**(5): 2026-33.

Montague, M. G. and C. A. Hutchison, 3rd (2000). "Gene content phylogeny of herpesviruses." Proc Natl Acad Sci U S A **97**(10): 5334-9.

Moore, P. S., C. Boshoff, et al. (1996). "Molecular mimicry of human cytokine and cytokine response pathway genes by KSHV." Science **274**(5293): 1739-44.

Moore, P. S. and Y. Chang (2001). Chapter 82: Kaposi's Sarcoma-Associated Herpesvirus. Fields Virology. D. M. Knipe and P. M. Howley. Philadelphia, Lippincott Williams & Wilkins. **2:** 2803-2833.

Moses, A. V., M. A. Jarvis, et al. (2002). "A functional genomics approach to Kaposi's sarcoma." Ann N Y Acad Sci **975**: 180-91.

Mossman, K. L., R. Sherburne, et al. (2000). "Evidence that herpes simplex virus VP16 is required for viral egress downstream of the initial envelopment event." J Virol **74**(14): 6287-99.

Moutaftsi, M., P. Brennan, et al. (2004). "Impaired lymphoid chemokine-mediated migration due to a block on the chemokine receptor switch in human cytomegalovirus-infected dendritic cells." J Virol **78**(6): 3046-54.

Mulder, N. J., R. Apweiler, et al. (2003). "The InterPro Database, 2003 brings increased coverage and new features." Nucleic Acids Res **31**(1): 315-8.

Nalwanga, D., S. Rempel, et al. (1996). "The UL 16 gene product of herpes simplex virus 1 is a virion protein that colocalizes with intranuclear capsid proteins." Virology **226**(2): 236-42.

Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol **48**(3): 443-53.

Neumann, L., W. Kraas, et al. (1997). "The active domain of the herpes simplex virus protein ICP47: a potent inhibitor of the transporter associated with antigen processing." J Mol Biol **272**(4): 484-92.

New, D. C. and Y. H. Wong (2003). "CC chemokine receptor-coupled signalling pathways." Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai) **35**(9): 779-88.

Newcomb, W. W., R. M. Juhas, et al. (2001). "The UL6 gene product forms the portal for entry of DNA into the herpes simplex virus capsid." J Virol **75**(22): 10923-32.

Newcomb, W. W., B. L. Trus, et al. (1993). "Structure of the herpes simplex virus capsid. Molecular composition of the pentons and the triplexes." J Mol Biol **232**(2): 499-511.

Nicholas, J., V. Ruvolo, et al. (1997). "A single 13-kilobase divergent locus in the Kaposi sarcoma-associated herpesvirus (human herpesvirus 8) genome contains nine open reading frames that are homologous to or related to cellular proteins." J Virol **71**(3): 1963-74.

Nicholson, P., C. Addison, et al. (1994). "Localization of the herpes simplex virus type 1 major capsid protein VP5 to the cell nucleus requires the abundant scaffolding protein VP22a." J Gen Virol **75 (Pt 5)**: 1091-9.

NIH-Newsroom (2003). International Consortium Completes Human Genome Project, National Human Genome Research Institute.

Niller, H. H. and L. Hennighausen (1990). "Phytohemagglutinin-induced activity of cyclic AMP (cAMP) response elements from cytomegalovirus is reduced by cyclosporine and synergistically enhanced by cAMP." J Virol **64**(5): 2388-91.

Novotny, J., I. Rigoutsos, et al. (2001). "In silico structural and functional analysis of the human cytomegalovirus (HHV5) genome." J Mol Biol **310**(5): 1151-66.

Ogasawara, M., T. Suzutani, et al. (2001). "Role of the UL25 gene product in packaging DNA into the herpes simplex virus capsid: location of UL25 product in the capsid and demonstration that it binds DNA." J Virol **75**(3): 1427-36.

Ogle, W. O. and B. Roizman (1999). "Functional anatomy of herpes simplex virus 1 overlapping genes encoding infected-cell protein 22 and US1.5 protein." J Virol **73**(5): 4305-15.

Orengo, C. (2003). Chapter 3: Sequence Comparison Methods. Bioinformatics: genes, proteins and computers. C. Orengo, D. Jones and J. Thornton. Oxford, BIOS Scientific Publishers Ltd. **1:** 29-48.

Orengo, C. A., A. D. Michie, et al. (1997). "CATH--a hierarchic classification of protein domain structures." Structure **5**(8): 1093-108.

Page, R. D. M. and E. C. Holmes (1998). Chapter 5: Measuring Genetic Change. Molecular Evolution, A Phylogenetic Approach. R. D. M. Page and E. C. Holmes. Oxford, Blackwell Science Ltd.: 135-146.

Palmer, L. E., A. L. O'Shaughnessy, et al. (2003). "A survey of canine expressed sequence tags and a display of their annotations through a flexible web-based interface." J Hered **94**(1): 15-22.

Pass, R. F. (2001). Chapter 77: Cytomegalovirus. Fields Virology. D. M. Knipe and P. M. Howley. Philadelphia, Lippincott Williams & Wilkins. **2:** 2675-2705.

Patterson, D., J. Bleskan, et al. (1999). "Human phosphoribosylformylglycineamide amidotransferase (FGARAT): regional mapping, complete coding sequence, isolation of a functional genomic clone, and DNA sequence analysis." Gene **239**(2): 381-91.

Pearson, H. (2001). "Biology's name game." Nature **411**(6838): 631-2.

Pennisi, E. (2003). HUMAN GENOME: A Low Number Wins the GeneSweep Pool. Science. **300:** 1484.

Perry, L. J. and D. J. McGeoch (1988). "The DNA sequences of the long repeat region and adjoining parts of the long unique region in the genome of herpes simplex virus type 1." J Gen Virol **69 (Pt 11)**: 2831-46.

Phelan, A., M. Carmo-Fonseca, et al. (1993). "A herpes simplex virus type 1 immediate-early gene product, IE63, regulates small nuclear ribonucleoprotein distribution." Proc Natl Acad Sci U S A **90**(19): 9056-60.

Ploegh, H. L. (1998). "Viral strategies of immune evasion." Science **280**(5361): 248-53.

Polson, A. G., D. Wang, et al. (2002). "Modulation of host gene expression by the constitutively active G protein-coupled receptor of Kaposi's sarcoma-associated herpesvirus." Cancer Res **62**(15): 4525-30.

Pomeranz, L. E. and J. A. Blaho (1999). "Modified VP22 localizes to the cell nucleus during synchronized herpes simplex virus type 1 infection." J Virol **73**(8): 6769-81.

Ponting, C. P. and R. R. Russell (2002). "The natural history of protein domains." Annu Rev Biophys Biomol Struct **31**: 45-71.

Prosch, S., C. Priemer, et al. (2003). "Proteasome inhibitors: a novel tool to suppress human cytomegalovirus replication and virus-induced immune modulation." Antivir Ther **8**(6): 555-67.

Quackenbush, J. (2001). "Computational analysis of microarray data." Nat Rev Genet **2**(6): 418-27.

Raftery, M., A. Muller, et al. (2000). "Herpesvirus homologues of cellular genes." Virus Genes **21**(1-2): 65-75.

Rajcani, J. and M. Kudelova (1999). "Glycoprotein K of herpes simplex virus: a transmembrane protein encoded by the UL53 gene which regulates membrane fusion." Virus Genes **18**(1): 81-90.

Raychaudhuri, S., J. T. Chang, et al. (2002). "Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature." Genome Res **12**(1): 203-14.

Reynolds, A. E., Y. Fan, et al. (2000). "Characterization of the U(L)33 gene product of herpes simplex virus 1." Virology **266**(2): 310-8.

Reynolds, A. E., B. J. Ryckman, et al. (2001). "U(L)31 and U(L)34 proteins of herpes simplex virus type 1 form a complex that accumulates at the nuclear rim and is required for envelopment of nucleocapsids." J Virol **75**(18): 8803-17.

Rhee, S. Y., W. Beavis, et al. (2003). "The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community." Nucleic Acids Res **31**(1): 224-8.

Rice, S. A., L. S. Su, et al. (1989). "Herpes simplex virus alpha protein ICP27 possesses separable positive and negative regulatory activities." J Virol **63**(8): 3399-407.

Rigoutsos, I., J. Novotny, et al. (2003). "In silico pattern-based analysis of the human cytomegalovirus genome." J Virol **77**(7): 4326-44.

Rixon, F. J., C. Addison, et al. (1996). "Multiple interactions control the intracellular localization of the herpes simplex virus type 1 capsid proteins." J Gen Virol **77 (Pt 9)**: 2251-60.

Rodems, S. M. and D. H. Spector (1998). "Extracellular signal-regulated kinase activity is sustained early during human cytomegalovirus infection." J Virol **72**(11): 9173-80.

Roizman, B. and D. M. Knipe (2001). Chapter 72: Herpes Simplex Viruses and Their Replication. Fields Virology. P. M. Howley. Philadelphia, Lippincott Williams & Wilkins. **2:** 2399-2460.

Roizman, B. and P. E. Pellet (2001). Chapter 71: The Family *Herpesviridae*: A Brief Introduction. Fields Virology. D. M. Knipe and P. M. Howley. Philadelphia, Lippincott Williams & Wilkins. **2:** 2381-2397.

Russo, J. J., R. A. Bohenzky, et al. (1996). "Nucleotide sequence of the Kaposi sarcoma-associated herpesvirus (HHV8)." Proc Natl Acad Sci U S A **93**(25): 14862-7.

Rylova, S. N., A. Amalfitano, et al. (2002). "The CLN3 gene is a novel molecular target for cancer drug discovery." Cancer Res **62**(3): 801-8.

Salvucci, L. A., R. H. Bonneau, et al. (1995). "Polymorphism within the herpes simplex virus (HSV) ribonucleotide reductase large subunit (ICP6) confers type specificity for recognition by HSV type 1-specific cytotoxic T lymphocytes." J Virol **69**(2): 1122-31.

Sambucetti, L. C., J. M. Cherrington, et al. (1989). "NF-kappa B activation of the cytomegalovirus enhancer is mediated by a viral transactivator and by T cell stimulation." Embo J **8**(13): 4251-8.

Sandri-Goldin, R. M. (1998). "ICP27 mediates HSV RNA export by shuttling through a leucine-rich nuclear export signal and binding viral intronless RNAs through an RGG motif." Genes Dev **12**(6): 868-79.

Saurin, A. J., K. L. Borden, et al. (1996). "Does this have a familiar RING?" Trends Biochem Sci **21**(6): 208-14.

Schaeffer, H. J. and M. J. Weber (1999). "Mitogen-activated protein kinases: specific messages from ubiquitous messengers." Mol Cell Biol **19**(4): 2435-44.

Schaffer, A. A., Y. I. Wolf, et al. (1999). "IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices." Bioinformatics **15**(12): 1000-11.

Scholz, M., J. U. Vogel, et al. (2004). "Thrombin stimulates IL-6 and IL-8 expression in cytomegalovirus-infected human retinal pigment epithelial cells." Int J Mol Med **13**(2): 327-31.

Schug, J., S. Diskin, et al. (2002). "Predicting gene ontology functions from ProDom and CDD protein domains." Genome Res **12**(4): 648-55.

Schweitzer, B., V. Taylor, et al. (1998). "Neural membrane protein 35 (NMP35): a novel member of a gene family which is highly expressed in the adult nervous system." Mol Cell Neurosci **11**(5-6): 260-73.

Sekino, Y., S. D. Bruner, et al. (2000). "Selective inhibition of herpes simplex virus type-1 uracil-DNA glycosylase by designed substrate analogs." J Biol Chem **275**(47): 36506-8.

Servant, F., C. Bru, et al. (2002). "ProDom: automated clustering of homologous domains." Brief Bioinform **3**(3): 246-51.

Shao, L., L. M. Rapp, et al. (1993). "Herpes simplex virus 1 alkaline nuclease is required for efficient egress of capsids from the nucleus." Virology **196**(1): 146-62.

Shchelkunov, S. N., P. F. Safronov, et al. (1998). "The genomic sequence analysis of the left and right species-specific terminal region of a cowpox virus strain reveals unique sequences and a cluster of intact ORFs for immunomodulatory and host range proteins." Virology **243**(2): 432-60.

Sheaffer, A. K., W. W. Newcomb, et al. (2001). "Herpes simplex virus DNA cleavage and packaging proteins associate with the procapsid prior to its maturation." J Virol **75**(2): 687-98.

Sinclair, J. and P. Sissons (1996). "Latent and persistent infections of monocytes and macrophages." Intervirology **39**(5-6): 293-301.

Sissons, J. G., M. Bain, et al. (2002). "Latency and reactivation of human cytomegalovirus." J Infect **44**(2): 73-7.

Skaletskaya, A., L. M. Bartle, et al. (2001). "A cytomegalovirus-encoded inhibitor of apoptosis that suppresses caspase-8 activation." Proc Natl Acad Sci U S A **98**(14): 7829-34.

Skepper, J. N., A. Whiteley, et al. (2001). "Herpes simplex virus nucleocapsids mature to progeny virions by an envelopment --> deenvelopment --> reenvelopment pathway." J Virol **75**(12): 5697-702.

Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-7.

Soderberg-Naucler, C. and J. Y. Nelson (1999). "Human cytomegalovirus latency and reactivation - a delicate balance between the virus and its host's immune system." Intervirology **42**(5-6): 314-21.

Somia, N. V., M. J. Schmitt, et al. (1999). "LFG: an anti-apoptotic gene that provides protection from Fas-mediated cell death." Proc Natl Acad Sci U S A **96**(22): 12667-72.

Song, B., K. C. Yeh, et al. (2001). "Herpes simplex virus gene products required for viral inhibition of expression of G1-phase functions." Virology **290**(2): 320-8.

Spear, P. G., M. T. Shieh, et al. (1992). "Heparan sulfate glycosaminoglycans as primary cell surface receptors for herpes simplex virus." Adv Exp Med Biol **313**: 341-53.

Sprague, J., D. Clements, et al. (2003). "The Zebrafish Information Network (ZFIN): the zebrafish model organism database." Nucleic Acids Res **31**(1): 241-3.

Spriggs, M. K. (1996). "One step ahead of the game: viral immunomodulatory molecules." Annu Rev Immunol **14**: 101-30.

Stingley, S. W., J. J. Ramirez, et al. (2000). "Global analysis of herpes simplex virus type 1 transcription using an oligonucleotide-based DNA microarray." J Virol **74**(21): 9916-27.

Streblow, D. N., S. L. Orloff, et al. (2001). "Do pathogens accelerate atherosclerosis?" J Nutr **131**(10): 2798S-2804S.

Streblow, D. N., C. Soderberg-Naucler, et al. (1999). "The human cytomegalovirus chemokine receptor US28 mediates vascular smooth muscle cell migration." Cell **99**(5): 511-20.

Sun, B., G. Harrowe, et al. (2001). "Modulation of human cytomegalovirus immediate-early gene enhancer by mitogen-activated protein kinase kinase kinase-1." J Cell Biochem **83**(4): 563-73.

Suzutani, T., M. Nagamine, et al. (2000). "The role of the UL41 gene of herpes simplex virus type 1 in evasion of non-specific host defence mechanisms during primary infection." J Gen Virol **81**(Pt 7): 1763-71.

Swanton, C., D. J. Mann, et al. (1997). "Herpes viral cyclin/Cdk6 complexes evade inhibition by CDK inhibitor proteins." Nature **390**(6656): 184-7.

Sydiskis, R. J. and B. Roizman (1968). "The sedimentation profiles of cytoplasmic polyribosomes in mammalian cells productively and abortively infected with herpes simplex virus." Virology **34**(3): 562-5.

Takakuwa, H., F. Goshima, et al. (2001). "Herpes simplex virus encodes a virion-associated protein which promotes long cellular processes in over-expressing cells." Genes Cells **6**(11): 955-66.

Tanoue, J., M. Yoshikawa, et al. (2002). "The GeneAround GO viewer." Bioinformatics **18**(12): 1705-6.

Teglund, S., A. Olsen, et al. (1994). "The pregnancy-specific glycoprotein (PSG) gene cluster on human chromosome 19: fine structure of the 11 PSG genes and identification of 6 new genes forming a third subgroup within the carcinoembryonic antigen (CEA) family." Genomics **23**(3): 669-84.

Teichmann, S. A. (2002). "The constraints protein-protein interactions place on sequence divergence." J Mol Biol **324**(3): 399-407.

Tenser, R. B. (1991). "Role of herpes simplex virus thymidine kinase expression in viral pathogenesis and latency." Intervirology **32**(2): 76-92.

Thome, M., P. Schneider, et al. (1997). "Viral FLICE-inhibitory proteins (FLIPs) prevent apoptosis induced by death receptors." Nature **386**(6624): 517-21.

Thomsen, A. R., A. Nansen, et al. (2003). "Regulation of T cell migration during viral infection: role of adhesion molecules and chemokines." Immunol Lett **85**(2): 119-27.

Thomsen, D. R., R. M. Stenberg, et al. (1984). "Promoter-regulatory region of the major immediate early gene of human cytomegalovirus." Proc Natl Acad Sci U S A **81**(3): 659-63.

Todd, A. E., C. A. Orengo, et al. (2001). "Evolution of function in protein superfamilies, from a structural perspective." J Mol Biol **307**(4): 1113-43.

Tomazin, R., A. B. Hill, et al. (1996). "Stable binding of the herpes simplex virus ICP47 protein to the peptide binding site of TAP." Embo J **15**(13): 3256-66.

Tortorella, D., B. E. Gewurz, et al. (2000). "Viral subversion of the immune system." Annu Rev Immunol **18**: 861-926.

Tran, L. C., J. M. Kissner, et al. (2000). "A herpes simplex virus 1 recombinant lacking the glycoprotein G coding sequences is defective in entry through apical surfaces of polarized epithelial cells in culture and in vivo." Proc Natl Acad Sci U S A **97**(4): 1818-22.

Troyanskaya, O., M. Cantor, et al. (2001). "Missing value estimation methods for DNA microarrays." Bioinformatics **17**(6): 520-5.

Tulipano, P. K., W. S. Millar, et al. (2003). "Linking molecular imaging terminology to the gene ontology (GO)." Pac Symp Biocomput: 613-23.

Turner, A., B. Bruun, et al. (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**(1): 873-5.

Valdar, W. S. J. and D. T. Jones (2003). Chapter 4: Amino Acid Residue Conservation. Bioinformatics: genes, proteins and computers. C. Orengo, D. Jones and J. Thornton. Oxford, BIOS Scientific Publishers Ltd. **1:** 49-64.

Vieira, J., P. O'Hearn, et al. (2001). "Activation of Kaposi's sarcoma-associated herpesvirus (human herpesvirus 8) lytic replication by human cytomegalovirus." J Virol **75**(3): 1378-86.

Visalli, R. J. and C. R. Brandt (2002). "Mutation of the herpes simplex virus 1 KOS UL45 gene reveals dose dependent effects on central nervous system growth." Arch Virol **147**(3): 519-32.

Vlazny, D. A., A. Kwong, et al. (1982). "Site-specific cleavage/packaging of herpes simplex virus DNA and the selective maturation of nucleocapsids containing full-length viral DNA." Proc Natl Acad Sci U S A **79**(5): 1423-7.

Vogel, C., C. Berzuini, et al. (2004). "Supra-domains: evolutionary units larger than single protein domains." J Mol Biol **336**(3): 809-23.

Voigt, S., G. R. Sandford, et al. (2001). "Identification and characterization of a spliced C-type lectin-like gene encoded by rat cytomegalovirus." J Virol **75**(2): 603-11.

Wagener, C. and S. Ergun (2000). "Angiogenic properties of the carcinoembryonic antigen-related cell adhesion molecule 1." Exp Cell Res **261**(1): 19-24.

Wagner, E. K., J. J. Ramirez, et al. (2002). "Practical approaches to long oligonucleotide-based DNA microarray: lessons from herpesviruses." Prog Nucleic Acid Res Mol Biol **71**: 445-91.

Wang, H. W., T. V. Sharp, et al. (2002). "Characterization of an anti-apoptotic glycoprotein encoded by Kaposi's sarcoma-associated herpesvirus which resembles a spliced variant of human survivin." Embo J **21**(11): 2602-15.

Ward, P. L., G. Campadelli-Fiume, et al. (1994). "Localization and putative function of the UL20 membrane protein in cells infected with herpes simplex virus 1." J Virol **68**(11): 7406-17.

Waterman, M. S. and M. Eggert (1987). "A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons." J Mol Biol **197**(4): 723-8.

Weller, T. H. (1971). "The cytomegaloviruses: ubiquitous agents with protean clinical manifestations. I." N Engl J Med **285**(4): 203-14.

Westra, D. F., K. L. Glazenburg, et al. (1997). "Glycoprotein H of herpes simplex virus type 1 requires glycoprotein L for transport to the surfaces of insect cells." J Virol **71**(3): 2285-91.

Wheeler, D. L., D. M. Church, et al. (2004). "Database resources of the National Center for Biotechnology Information: update." Nucleic Acids Res **32 Database issue**: D35-40.

Wilkinson, M. G. and J. B. Millar (2000). "Control of the eukaryotic cell cycle by MAP kinase signaling pathways." Faseb J **14**(14): 2147-57.

Willard, M. (2002). "Rapid directional translocations in virus replication." J Virol **76**(10): 5220-32.

Wootton, J. C. and S. Federehen (1993). "Statistics of local complexity in amino acid sequences and sequence databases." Computational Chemistry **17**: 179.

Wren, J. D. and H. R. Garner (2004). "Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network." Bioinformatics **20**(2): 191-8.

WuDunn, D. and P. G. Spear (1989). "Initial interaction of herpes simplex virus with cells is binding to heparan sulfate." J Virol **63**(1): 52-8.

Xie, H., A. Wasserman, et al. (2002). "Large-scale protein annotation through gene ontology." Genome Res **12**(5): 785-94.

Yamada, H., T. Daikoku, et al. (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)**: 2923-31.

Yamashita, A., Y. Watanabe, et al. (1997). "Microtubule-associated coiled-coil protein Ssm4 is involved in the meiotic development in fission yeast." Genes Cells **2**(2): 155-66.

Yan, W. L., T. J. Lerner, et al. (1994). "Sequence analysis and mapping of a novel human mitochondrial ATP synthase subunit 9 cDNA (ATP5G3)." Genomics **24**(2): 375-7.

Ye, G. J., K. T. Vaughan, et al. (2000). "The herpes simplex virus 1 U(L)34 protein interacts with a cytoplasmic dynein intermediate chain and targets nuclear membrane." J Virol **74**(3): 1355-63.

Yeh, I., P. D. Karp, et al. (2003). "Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO)." Bioinformatics **19**(2): 241-8.

Yu, D., M. C. Silva, et al. (2003). "Functional map of human cytomegalovirus AD169 defined by global mutational analysis." Proc Natl Acad Sci U S A **100**(21): 12396-401.

Yu, D. and S. K. Weller (1998). "Genetic analysis of the UL 15 gene locus for the putative terminase of herpes simplex virus type 1." Virology **243**(1): 32-44.

Zeeberg, B. R., W. Feng, et al. (2003). "GoMiner: a resource for biological interpretation of genomic and proteomic data." Genome Biol **4**(4): R28.

Zhou, G., V. Galvan, et al. (2000). "Glycoprotein D or J delivered in trans blocks apoptosis in SK-N-SH cells induced by a herpes simplex virus 1 mutant lacking intact genes expressing both glycoproteins." J Virol **74**(24): 11782-91.

Zhou, G. Q., V. Baranov, et al. (1997). "Highly specific monoclonal antibody demonstrates that pregnancy-specific glycoprotein (PSG) is limited to syncytiotrophoblast in human early and term placenta." Placenta **18**(7): 491-501.

Letter

Identification of New Herpesvirus Gene Homologs in the Human Genome

Ria Holzerlandt,[1] Christine Orengo,[2] Paul Kellam,[1,4] and M. Mar Albà[1,3]

[1]*Wohl Virion Centre, Department of Immunology and Molecular Pathology, and* [2]*Biomolecular Structure and Modelling Unit, Department of Biochemistry, University College London, London W1T 4JF, United Kingdom*

Viruses are intracellular parasites that use many cellular pathways during their replication. Large DNA viruses, such as herpesviruses, have captured a repertoire of cellular genes to block or mimic host immune responses, apoptosis regulation, and cell-cycle control mechanisms. We have conducted a systematic search for all homologs of herpesvirus proteins in the human genome using position-specific scoring matrices representing herpesvirus protein sequence domains, and pair-wise sequence comparisons. The analysis shows that ~13% of the herpesvirus proteins have clear sequence similarity to products of the human genome. Different human herpesviruses vary in their numbers of human homologs, indicating distinct rates of gene acquisition in different lineages. Our analysis has identified new families of herpesvirus/human homologs from viruses including human herpesvirus 5 (human cytomegalovirus; HCMV) and human herpesvirus 8 (Kaposi's sarcoma–associated herpesvirus; KSHV), which may play important roles in host-virus interactions.

Viruses are obligate intracellular parasites and, as such, use many normal cellular pathways and components during their replication cycle. Large DNA viruses may contain up to a few hundred open reading frames (ORFs). Among the proteins they encode, we can distinguish between those that have essential viral functions, such as genome replication and capsid assembly, and those that are involved in direct interaction with the host, effecting immune evasion, cell proliferation, and apoptosis control (Ploegh 1998; Tschopp et al. 1998). Many of the latter genes are likely to have been acquired from the host to mimic or block normal cellular functions (Moore et al. 1996; Alcami and Koszinowski 2000; McFadden and Murphy 2000). Identifying and understanding the functions of such "acquired" viral proteins may lead to the development of therapeutic strategies to combat persistent viral infection.

An approach to the identification of virus proteins that interfere with the host system is to search for homologs in the host genome. Until recently, the fraction of host genome sequence data available for analysis, and the quality of annotation of such data, has limited the identification of such homologs. The publication of the draft of the human genome and conceptual translated products (Lander et al. 2001) enables us to conduct, for the first time, a comprehensive assessment of homologous proteins between a vertebrate genome and viral ORFs. There are two methods particularly applicable to mass analysis of sequence databases. The first involves searching of individual protein sequences against a database using pair-wise sequence comparison algorithms, and has previously been used to identify individual virus/host homologs. Viral proteins, however, are subject to high mutation rates, and that may cloud or mask true homology. A second, more sensitive approach is to search databases with amino acid sequence motifs that are conserved between related proteins. Motifs can be defined as regions of amino acid sequence that are more highly conserved than the rest of the protein owing to functional constraints. An accurate representation of such motifs can be obtained by constructing position-specific scoring matrices (PSSMs) that store the frequency of occurrence of different amino acids along the motif.

In the present study, we focus on the analysis of herpesviruses, one of the best-characterized large DNA virus families. Typically, each herpesvirus genome contains between 70 and 120 ORFs, with the exception of human cytomegalovirus (HCMV), which codes for up to 220 ORFs. The herpesviruses infect a wide range of animal hosts and—on the basis of differences in genome content, organization, and cellular tropism—have been divided into three subfamilies: the alphaherpesviruses, betaherpesviruses, and gammaherpesviruses. There are a number of herpesviruses that have yet to be categorized in a herpesvirus subfamily, including channel catfish herpesvirus, and these are classified as "other" in this study (see Table 1; ICTV 2000). Eight different herpesviruses, encompassing all three subfamilies, are known to infect humans. Herpesviruses persist and replicate their genomes in the nucleus and acquire host genes by an ill-defined process (Brunovskis and Kung 1995; Chaston and Lidbury 2001). Most of these acquired genes are located in regions outside the five gene blocks common to all herpesvirus genomes. Previous work by others and ourselves has identified a set of 26 ORFs that are conserved across all herpesviruses (McGeoch and Davison 1999; Albà et al. 2001a). The remaining herpesvirus genes are present in all members of a virus subfamily, present in a subset of viruses in a subfamily, or unique to a particular virus. Many of these potentially important proteins, however, remain uncharacterized.

We have recently developed a virus database, VIDA (Albà et al. 2001b), in which all herpesvirus ORFs are grouped together into homologous protein families (HPFs), each defined by one or more conserved amino acid regions (motifs). To identify human proteins that are related to the herpesvirus protein families, we have constructed PSSMs for all HPF-defining motifs and used them to perform sensitive searches

[3]Present address: Grup de Recerca en Informàtica Biomèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, 08003 Barcelona, Spain.
[4]Corresponding author.
E-MAIL p.kellam@ucl.ac.uk; FAX 44-020-7679-9555.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.334302. Article published online before print in October 2002.

Table 1. Herpesvirus-Human Homologs

Function class	Viral function (VIDA)	HPF[1]	Virus[2]	GenBank[3]	Human function
DNA replication	DNA polymerase	1	a,b,g	8393995	polymerase (DNA-directed), α
		293	o	15303524	polymerase (DNA directed), δ 1
	helicase/primase	16	a,b,g	5523990	DNA helicase
Nucleotide repair/ metabolism	uracil–DNA glycosylase	8	a,b,g	6224979	uracil–DNA glycosylase
	ribonucleotide reduct. large sub.	24	a,b,g	4506749	ribonucleotide reductase M1 polypeptide
	ribonucleotide reduct. small sub.	33	a,g	4557845	ribonucleotide reductase M2 polypeptide
	thymidylate synthase	92	a-,g-	15297069	thymidylate synthetase
	dihydrofolate reductase	141	g-,b-	15297069	dihydrofolate reductase
	dUTP pyrophosphatase	S	CCHV ORF49	4503423	dUTP pyrophosphatase
		S	SaHV-1 ORF49	14756895	dUTP pyrophosphatase
	thymidine kinase	S	CCHV ORF5	11430716	thymidine kinsae 2, mitochondrial
	DNA methyltransferase	S	RaHV-1 54_21	4503351	DNA (cytosine-5-)–methyltransferase 1
Enzyme	protein kinase	29	a,b,g-	14746991	serine/threonine-protein kinase PRP4
		40	a,o	4505649	protein kinase cdc2-related PCTAIRE-2
		214	o	9994197	G protein-coupled receptor kinase 7
		S	RaHV-1 54_2	14741902	CamKI–like protein kinase
	phospholipase-like protein	328	a-	5174497	endothelial cell–derived lipase precursor
	b-1,6-N-acetylglucosaminyltransf.	S	BoHV-4 ORF3-4	11431963	glucosaminyl (*N*–acetyl) transferase 3
	serine protease	S	CCHV ORF47	4505577	paired basic amino acid cleaving system 4
Gene expression regulation	transcriptional activator	74	a	5174653	ring finger protein (C3H2C3 type) 6
	bZIP domain	174	a-	4504809	jun B proto–oncogene
Glycoprotein	glycoprotein OX-2-like	194	b-	730246	OX-2 membrane glycoprotein precursor
	glycoprotein OX-2-like	242	g-	730246	OX-2 membrane glycoprotein precursor
Host-virus interaction	TNFR receptor	13	HHV-5 UL144	4507571	tumor necrosis factor receptor, member 14
	virion–assoc. host shutoff factor	48	a	14738228	flap structure–specific endonuclease 1
	viral interferon regulatory factor	89	g-	4504723	interferon regulatory factor 2
		243	g-	13629153	interferon consensus seq. binding prot. 1
		S	HHV-8 vIRF-3	4505287	interferon regulatory factor 4
	G protein-coupled receptor	27	b,g-	13643500	chemokine (C–C motif) receptor 2
		248	b-	4758468	G protein–coupled receptor 50
		S	EHV-2, ORF 74	4502639	chemokine (C–C motif) receptor 5
	complement binding protein	10	g-	10835143	decay accelerating factor for complement
	viral cyclin	102	g-	14767736	cyclin D1
	viral interleukin 10	140	g-	10835141	interleukin 10
	viral interleukin 6	273	g-	10834984	interleukin 6 (interferon, β 2)
	viral interleukin 17	S	HVS-2 ORF13	4504651	interleukin 17
	vBcl-2	161	g-	4502363	BCL2–antagonist-killer 1
		259	g-	4557355	B–cell lymphoma protein 2 α
		850	MeHV-1 ORF1	11433559	BCL2-like 10 (apoptosis facilitator)
	MHC I downregulation	150	g-	8923613	hypothetical protein FLJ20668
	viral FLICE–inhibitory protein	256	g-	14731507	CASP8 and FADD–like apoptosis regulator
		S	EHV-2 E8	4505229	Fas (TNFRSF6)–associated via death domain
	CxC chemokine vIL8	531	a-	10834978	interleukin 8
	vMIP-I	225	g-	5174671	small inducible cytokine subf. A, member 26
	α chemokine	321	b-	4885589	small inducible cytokine subf. B, member 9B
	β chemokine	387	b-	5174671	small inducible cytokine subf. A, member 26
	vMIP-III	S	HHV-8 K4.1	4506829	small inducible cytokine subf. A, member 17
	signal transduction protein	316	RRV, R1	12056967	Fc fragment of IgG, receptor for (CD16)
	CARD–like apoptotic protein	355	EHV-2, E10	4502379	CARD–like apoptotic protein
	U-PAR antigen CD59	352	HVS-2, ORF15	13639271	CD59 antigen p18-20

Table 1. (*Continued*)

Function class	Viral function (VIDA)	HPF[1]	Virus[2]	GenBank[3]	Human function
	natural killer (NK) cell decoy pr.	S	HHV-5 UL18	5031745	major histocompatibility complex, class I, E
	colony-stimulating factor I	S	HHV-4 BARF1	4885123	CD80 antigen
	C-type lectin-like protein	S	RCMV lectin	4504883	killer cell lectin–like receptor subf. C, member 2
	semaphorin homolog	S	AlHV-1 A3	4504237	sema domain, Ig domain, GPI memb. anchor
	MHC1 heavy chain	S	RCMV R144	9665232	major histocompatibility complex, class I
Unknown	unknown	258	a-	4504883	killer cell lectin–like receptor subf. C, member 2
	Unknown	S	GaHV-1 UL45	4504883	killer cell lectin–like receptor subf. C, member 2
	Unknown	S	HHV-5 UL1	14764567	pregnancy specific beta-1-glycoprotein 5
	Unknown	S	HHV-5 US21	6912468	lifeguard

[1]HPF: homologous protein family no. S indicates singleton. HPF details can be visualised by searching VIDA by HPF number in http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA.html (Herpesviridae link).
[2]a indicates alphaherpesvirus; b, betaherpesvirus; g, gammaherpesvirus; o, other; — only a subset of subfamily members are represented. For singletons, virus abbreviation and gene name are given: CCHV, channel catfish herpesvirus; SaHV-1, salmonid herpesvirus 1; RaHV-1, ranid herpesvirus 1; BoHV-4, bovine herpesvirus 4; HHV-8, human herpesvirus 8; EHV-2, equine herpesvirus 2; HVS-2, saimiriine herpesvirus 2; MeHV-1, meleagrid herpesvirus 1; HHV-5, human herpesvirus 5; HHV-4, human herpesvirus 4; RCMV, rat cytomegalovirus; AHIV-1, alcelaphine herpesvirus 1; and GaHV-1, gallid herpesvirus 1.
[3]GenBank protein accession no. (GI number). Only the human protein that hit with the lowest E-value is shown.

of the translated human genome products. Mapping of homologs in the human genome has been complemented by BLAST-based pair-wise sequence comparison searches (Altschul et al. 1990, 1997). Our analysis has resulted in the identification of protein families or singleton proteins that show clear homology with gene products in the human genome, including new host-virus homologs in human herpesvirus (HHV) 5 (HCMV) and HHV-8 (Kaposi's sarcoma–associated herpesvirus; KSHV).

RESULTS

Herpesvirus Proteins With Human Homologs

The identification of herpesvirus/human homologs was undertaken by searching the set of conceptual and known protein sequences derived from the public Human Genome Project (Lander et al. 2001) against herpesvirus protein sequences in the virus database VIDA (Albà et al. 2001b) using two different sequence-similarity search methods. The first method was based on PSSMs derived from predefined viral protein motifs in VIDA. The second used BLAST-based pair-wise sequence comparisons with the collection of singleton viral proteins and a representative set of viral proteins that share <95% sequence identity (N95-rep, see Methods).

Careful examination of putative homologs showed that 39 herpesvirus HPFs and 20 singleton proteins had significant sequence similarity to human gene products (Table 1). This represented 13% of all herpesvirus ORFs in GenBank. Sequence similarity between herpesvirus and human proteins was clearly related to functional similarity, based on previous experimental data. However, functional similarity is defined here in a broad sense, meaning the viral proteins participate in the given functional network. This is because viral proteins can change from the precise mechanistic function of the host homolog in subtle ways after acquisition by the virus while still maintaining the broader function. For example, the HHV-8 viral cyclin participates in the cell cycle as a cyclin D homolog but, unlike the host cyclin D, is not negatively regulated (Swanton et al. 1997). The use of PSSMs to perform database searches was more sensitive than using N95-reps with BLASTP, as six of the 39 HPF homologs could only be detected by the first method. One homolog, however, complement binding protein, could only be identified using BLASTP.

Approximately 54% of the combined HPF and singleton hits corresponded to proteins classified in VIDA as being involved in host-virus interaction, primarily effecting immune and/or apoptosis controls. Of the remaining homologs, 32% have functions that can be generally termed metabolic (being "enzymes," involved in "DNA replication," or involved in "nucleotide repair/metabolism"). Homologs to capsid constituents or capsid assembly proteins were not detected. Approximately 42% of the HPFs and singletons that showed homology with human proteins did not contain any HHV ORF members. This method can therefore be used to annotate gene products from non-HHVs for which complete host genome sequence information is still unavailable.

Identification of New Virus-Human Homologs

Of special interest was the identification of human homologs for herpesvirus protein families and singletons of unknown function. The new homologs may provide putative functional annotations for several herpesvirus and/or human proteins. New herpesvirus/human protein families were found for the US12 (unique short) HCMV protein family, the UL1 (unique long) HCMV protein, the gallid/meleagrid herpesvirus UL45 protein family, and the K3/K5 HHV-8 family (Fig. 1).

HCMV US21 is a distant member of a larger HCMV protein family, the US12 protein family, encompassing gene products US12 to US21 (Chee et al. 1990). The US21 showed significant overall sequence similarity to three human proteins: lifeguard, CGI-119, and PP1201. Other members of the

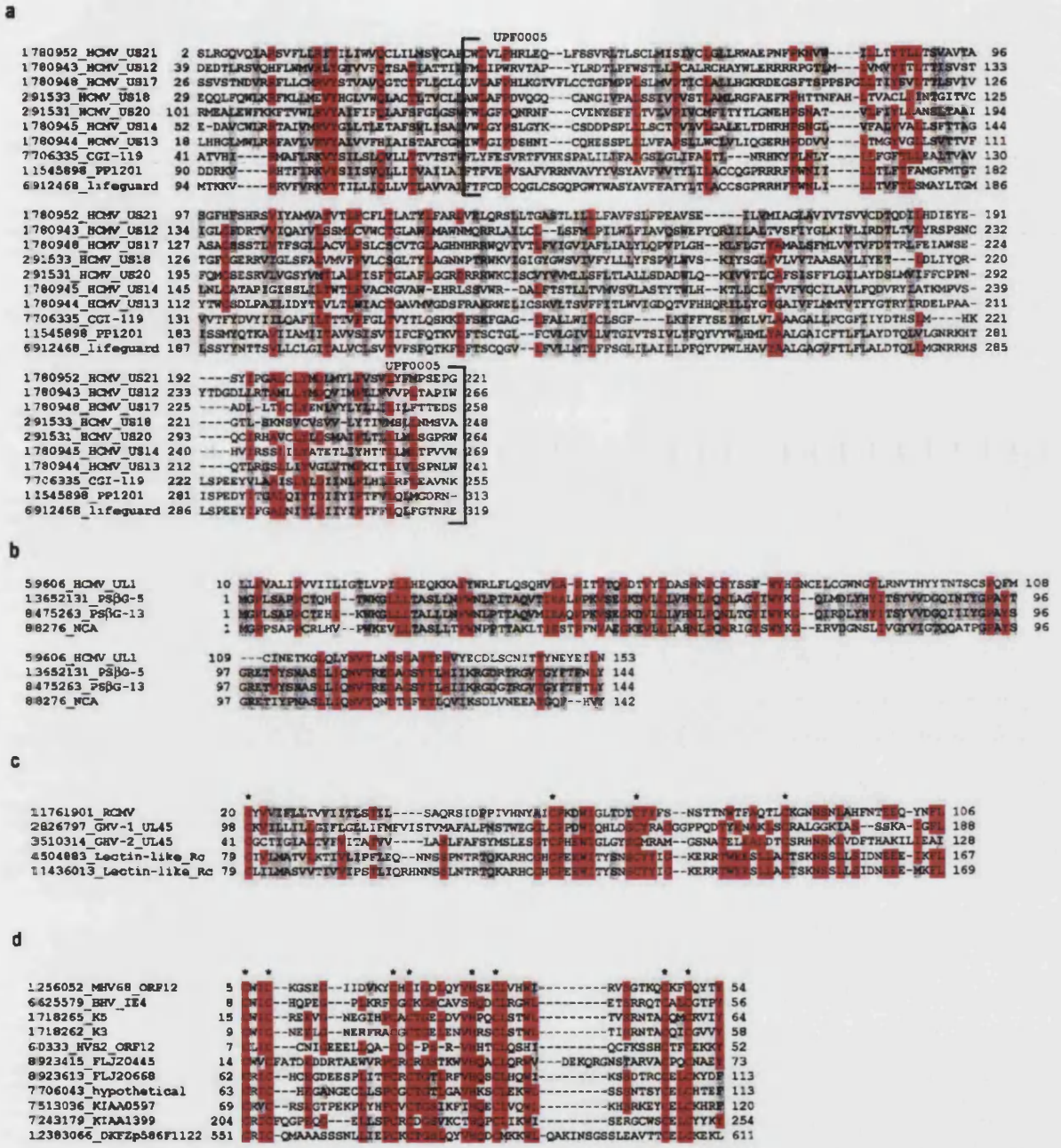


Figure 1 Alignment of new herpesvirus/human homologs. Proteins are labeled with GenBank identification number (GI) and a short description. Amino acids that are shaded red share identity across ≥50% of the alignment; amino acids shaded grey share similarity across ≥50% of the alignment. (*a*) Herpesvirus US12 protein family members, human lifeguard protein, and two additional human proteins. The Pfam UPF0005 domain is indicated. (*b*) HCMV UL1, two PSG proteins (PSβG 5 and 13), and one member of the carcinoembryonic antigen subfamily (NCA, nonreacting antigen). (*c*) A representative from each of the herpesvirus protein families found to contain C-type lectin domains and two natural killer receptors (NKG2-A). The four conserved cysteines, important for disulphide bond formation in the carbohydrate recognition domain, are indicated. (*d*) K3/K5 herpesvirus protein family with six human homologs. Cysteine/histidine conserved residues in the BKS (BHV-4 [bovine herpesvirus 4], KSHV, and swinepox) motif are indicated.

US12 protein family, including an HPF that groups six of them in VIDA, did not initially hit any human proteins, but multiple sequence alignments revealed the true extent of amino acid similarity between all these proteins (Fig. 1a). The herpesvirus and human proteins also matched the protein family domain UPF0005 in the Pfam database (Bateman et al.

2000), a putative seven-transmembrane region domain. Lifeguard is the human homolog of the rat protein neuromembrane protein 35, proposed to protect against Fas-mediated apoptosis (Somia et al. 1999).

HCMV UL1 showed sequence similarity to the pregnancy-specific glycoprotein 5 (PSG-5) and other members of the human carcinoembryonic antigen (CEA) protein family. The PSGs, a subgroup of the CEA family, are mainly expressed in the placenta and are secreted into the maternal circulation, possibly regulating immune system responses. The region of sequence similarity covered about two thirds of the UL1 protein and the N-terminal region of PSG and CEA subgroup proteins (Fig. 1b).

The protein family represented by UL45 in gallid (includes Marek's disease herpesvirus) and meleagrid herpesviruses shows homology with human C-type (calcium-dependent) lectin domain containing natural killer (NK)–cell receptor proteins. Two other herpesvirus proteins, from rat cytomegalovirus (RCMV) and from a different gallid herpesvirus strain (GenBank accession no. Y14300), also show significant sequence similarity to C-type lectin domain containing NK-cell receptors. The presence of C-type lectin domain in the RCMV protein was recently reported (Voigt et al. 2001) which now clearly extends to homologs in some avian herpesviruses. NK-cell receptors interact with HLA (human leukocyte antigen) class I antigens and facilitate triggering or inhibition of NK cell–mediated cytotoxicity (Biassoni et al. 2001). C-type lectins contain a carbohydrate recognition domain, which includes four conserved cysteine residues forming two disulphide bonds. These conserved cysteines are also present in the herpesvirus C-type lectin-like homologs (Fig. 1c).

The K3/K5 protein family in VIDA contains a highly conserved zinc finger motif identified in the proteins K3 and K5 from HHV-8, IE1 in bovine herpesvirus 4 (BHV-4), and ORF12 in murine herpesvirus 68 (MHV-68). An additional gene, ORF 12 in saimiriine herpesvirus 2 (HVS-2), a singleton in VIDA, did not initially hit any human gene product. However, it also contains the same conserved motif and should therefore be considered a member of the family (Nicholas et al. 1997). The motif is known as the BKS (BHV-4, KSHV, and swinepox) motif, a member of the PHD/LAP zinc finger class (C4HC3), but clearly differing from PHD/LAP zinc fingers owing to its distinct spacing of the cysteine/histidine residues. K3 and K5 from HHV-8 have been recently discovered to down-regulate MHC class 1 molecules in infected cells (Coscoy and Ganem 2000). We identified six unannotated human proteins, including three identified by pair-wise searches (Jenner and Boshoff 2002), that contain this highly conserved BKS finger motif (Fig. 1d). In the herpesvirus proteins, the motif is always found in the N terminus, but in one human protein, it appeared in the central part of the peptide, whereas in another, the counterpart of murine axotrophin, at the C terminus.

Human Homologs in HHVs

Our analysis provides an estimate of the number of homologs between the eight different HHVs and the translated products from their host genome. A total of 34 different HHV proteins, including HPFs and singletons, showed significant homology with human proteins (Fig. 2). This represents a minimum estimate, as some proteins may still be functionally homolo-

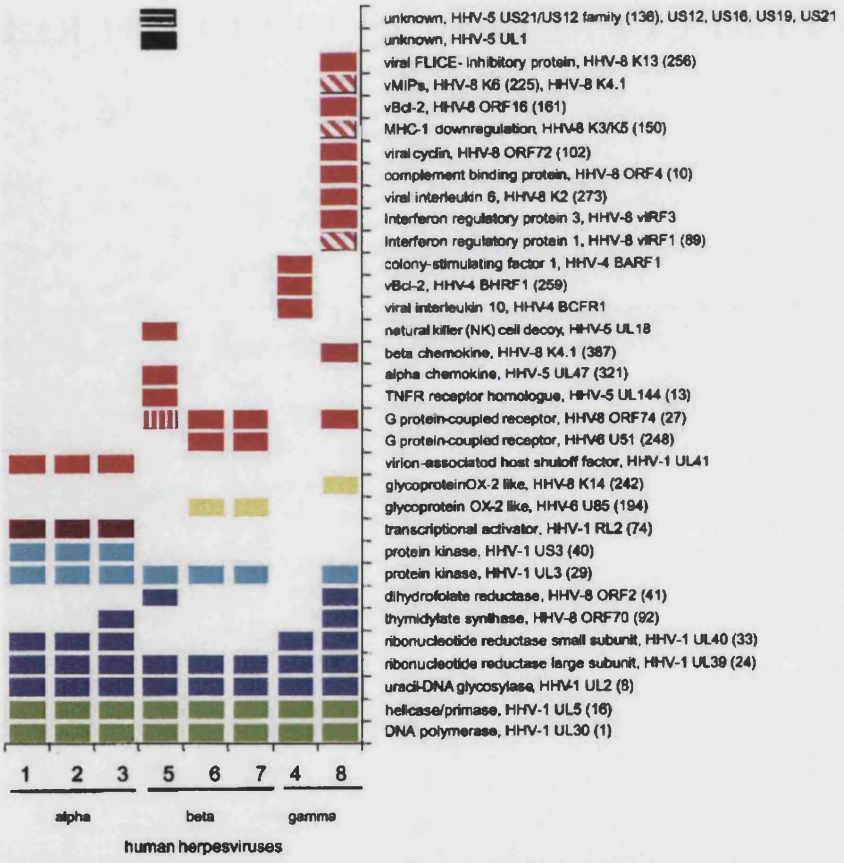


Figure 2 Human herpesvirus (HHV) proteins with human homologs. Alternative names for the HHVs are HHV-1, human simplex virus 1; HHV-2, human simplex virus 2; HHV-3, varicella zoster virus; HHV-4, Epstein-Barr virus; HHV-5, human cytomegalovirus; and HHV-8, Kaposi's sarcoma–associated herpesvirus. Labels show the virus protein function, the name of a member of the HPF (homologous protein family) or singleton, and, for HPFs, the corresponding number in brackets. All the annotations and HPF numbers are taken from VIDA. Note that in some cases more than one HPF/singleton, shown as separate rows in Table 1, are shown together here. This corresponds to highly divergent families. The graph is color coded according to functional class: light green, DNA replication; dark blue, nucleotide repair/metabolism; light blue, enzyme; purple, gene expression regulation; yellow, glycoprotein; red, host-virus interaction; and black, unknown. Diagonal lines within a box indicate two gene copies (per viral genome); vertical lines, three copies; and horizontal lines, 10 copies.

gous but not show significant sequence similarity, and the total number of genes in the human genome is still uncertain (Lander et al. 2001).

Four human homologs are known to be present in all HHVs (i.e., DNA-dependent DNA polymerase, helicase/primase, uracil-DNA glycosylase, and ribonucleotide reductase large subunit), and these were all correctly identified by our methods. An additional protein family, protein kinase HHV-1 UL13, is present in all HHVs except in HHV-4. It is known that the gammaherpesviruses share a common evolutionary branch with the betaherpesvirus, and that the alphaherpesvirus forms a separate lineage (McGeoch and Davison 1999; Albà et al. 2001a). One of the human homologs, ribonucleotide reductase small subunit, is found in the alpha- and gammaherpesviruses, but not in the betaherpesviruses, indicating that it has been lost in the latter lineage. There are three human homologs that appear to be alphaherpesvirus-specific: protein kinase HHV-1 US3, transcriptional activator HHV-1 ICP0 (infected cell protein), and host shutoff factor HHV-1 UL41. This compares to seven homologs that are betaherpesvirus specific and 14 that are gammaherpesvirus specific. Of particular interest are two human homologs that appear in disparate positions in the herpesvirus evolutionary tree: thymidylate synthase in HHV-3 (varicella zoster virus) and in HHV-8 (Kaposi's sarcoma–associated herpesvirus); dihydrofolate reductase in HHV-5 (HCMV) and HHV-8. Independent acquisition of these genes from the host genome, multiple gene loss events in different herpesvirus lineages, or gene transfer between virus genomes could explain their distribution.

The total proportion of human homologs in the different HHVs varies. Using the number of gene products in the corresponding herpesvirus genome GenBank entries (Table 1 in Albà et al. 2001a), this percentage is 11% to 16% of the genes in human alphaherpesviruses, 9% to 11% in the human betaherpesviruses, 10% of the genes in HHV-4, and 30% in the HHV-8 genome. HHV-8 contains a markedly higher proportion of human homologous genes, indicating a higher degree of recent gene transfer from the host genome.

Dynamics of Host Gene Acquisition in the Gammaherpesviruses

Human homologs that are present in all or a large proportion of the herpesvirus genomes, such as DNA polymerase or uracil-DNA glycosylase, are likely to have been acquired from a distant host by an ancestral herpesvirus. Other genes appear to have been acquired more recently, appearing only in a subset of viruses. From the 59 HPFs and singletons that showed homology with human proteins, only 16 were present in alphaherpesviruses, 17 in betaherpesviruses, and 32 in gammaherpesviruses. More than half (54%) of these homologs have host-virus interaction functions. Gammaherpesvirus genomes are particularly rich in genes that have a human counterpart. Therefore, a more detailed analysis of the distribution of gammaherpesvirus-specific human homologs in complete gammaherpesvirus genomes was undertaken (Fig. 3).

Phylogenetic reconstruction of the fully sequenced gammaherpesvirus subfamily members (McGeoch et al. 2000; Montague and Hutchison 2000; Albà et al. 2001a) has established that HHV-4 forms a separate lineage, the lymphocrytpo or gamma-1-herpesviruses 1. The remaining fully sequenced gammaherpesviruses, which include HHV-8, form the rhadino or gamma-2-herpesviruses lineage. The relative positions of alcelaphine herpesvirus 1 (AlHV-1), equine herpesvirus 2 (EHV-2), and MHV-68 within the gammaherpesvirus 2 are still ill-defined, although recent work shows that MHV-68 is probably more closely related to the primate herpesvirus (Fig. 3; McGeoch et al. 2000; Albà et al. 2001a). The presence of human homologs in the different genomes is consistent

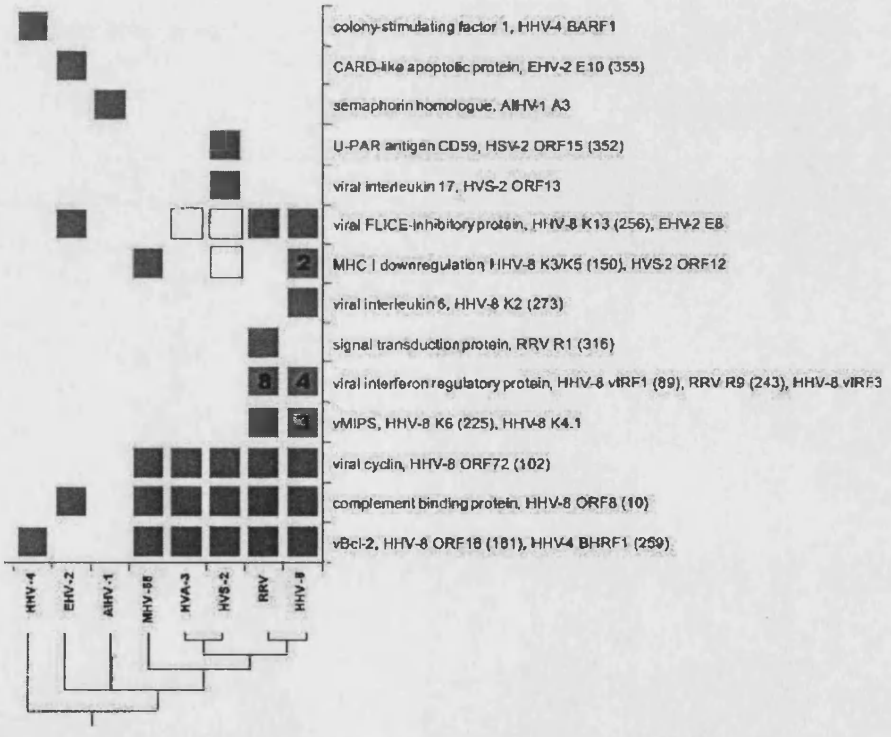


Figure 3 Gammaherpesvirus-specific proteins involved in host-virus interactions that have human homologs. Boxes indicate the presence of a particular gene(s) in a virus genome. Numbers in boxes represent copies within a genome. Labels show the virus protein function, the name of a member of the HPF (homologous protein family) or singleton, and, for HPFs, the corresponding number in brackets. All the annotations and HPF numbers are taken from VIDA. Note that in some cases more than one HPF/singleton, shown as separate rows in Table 1, is shown together. This corresponds to highly divergent families. The HPF/singletons that are not present in Table 1 are represented as unfilled boxes. These are herpesvirus proteins for which we did not identify human homologs in the database searches but that, nevertheless, can be grouped together, by function and residue conservation, with other herpesvirus HPF/singletons for which we could detect human homologs. A consensus phylogenetic tree of the gammaherpesvirus is shown at the bottom. This was generated as described for all HPFs from complete herpesvirus genomes (Albà et al. 2001a).

within the different gammaherpesvirus groups defined by gene-content phylogenetics (Fig. 3); however, some of the homologs show a complex distribution. For example, ORF12, a homolog of the K3/K5 HHV-8 genes, is also present in MHV-68 and HVS-2 but not in the HHV-8 closely related primate herpesviruses ateline herpesvirus 3 (AtHV-3) and *Macaca mulatta* rhadinovirus (RRV). Therefore, the gene may have been lost on several occasions. Another explanation would be independent acquisition from the host genome in HHV-8, MHV-68, and HVS-2, although the fact that the gene is in equivalent positions in these genomes would favor the former. In other homolog cases, a single event of gene acquisition is easier to delineate; for example, the interferon regulatory factor and the macrophage inflammatory protein families are only found in RRV and HHV-8; they are at the same loci in both genomes and hence were presumably captured before host speciation by an ancestor of these two viruses.

DISCUSSION

The publication of the human genome has provided the opportunity to analyze host-parasite interactions in a new light. Herpesviruses capture genes from their host and use them to their own advantage. In the present study, we have analyzed virus-host protein homology using consistent cross-comparative methods for herpesviruses proteins and gene products of the human genome. The study has allowed us to derive a global picture of cellular functions for which herpesviruses have captured and evolved their own counterparts.

Sequence similarity alone revealed a minimum estimate of human homologs in different HHV genomes to be ~9% to 16% of virus genes, with the exception of HHV-8, which is ~30% of viral genes. The reason for a higher percentage of homologs in this virus, and in gammaherpesviruses in general, is unclear but may relate to properties of the cell types infected by this subfamily of herpesviruses. Most of the herpesvirus/human homologs identified correspond to proteins involved in immune modulation and apoptotic control. These proteins are normally specific to one or a few viruses, and they often show a complex distribution across the herpesvirus phylogeny tree (Fig. 3). They are, therefore, likely to contribute to the adaptation of the virus to different hosts or different cellular tropisms. This is in contrast to a more stable group of homologs, composed of proteins involved in DNA replication and nucleotide metabolism, components of the well-conserved virus (and host) DNA genome replication machinery.

In our analysis, we have used PSSMs representing herpesvirus protein motifs to increase sensitivity over pair-wise sequence comparison-based searches. The method has allowed us to identify a number of new herpesvirus/human homologs. The new putative functions require experimental testing but are of interest. The HCMV US12 protein family, composed of 10 members, has homology with lifeguard and related human proteins (CGI-119). Lifeguard is known to inhibit the apoptosis signal mediated by the Fas receptor, and therefore, the related HCMV proteins may also have an antiapoptotic role. Viral proteins that interfere with Fas-mediated apoptosis have already been described in gammaherpesviruses (Belanger et al. 2001) but not in betaherpesviruses. This is surprising as HCMV also replicates in cells of the haematopoietic system, namely, monocytes/macrophages. From our analysis, HCMV potentially encodes a repertoire of anti-Fas apoptosis homologs distinct form the gammaherpesvirus FLIP homologs. Interestingly, in the cowpox virus, a member of the Poxviridae family, a gene termed SR1, of unknown function but similar to the CGI-119 protein, was also identified (Shchelkunov et al. 1998).

Homology was found between the HCMV UL1 gene product and the CEA/PSG human protein family. Known functions for the CEA family include involvement in cell adhesion, signal transduction, and possibly innate immunity (Hammarstrom 1999). The PSGs, a subgroup of the CEA family, are mainly expressed in the placenta and are secreted into the maternal circulation, possibly regulating immune system responses. HCMV infection, which is usually benign in immunocompetent individuals, can have catastrophic consequences during pregnancy (Fisher et al. 2000). Infection of the placenta has a 30% to 40% risk of intrauterine virus transmission to the foetus. Similarity of UL1 to PSGs could subsequently be related to the pathology of HCMV during pregnancy or to general immune modulation in the host.

In the present study, we have also detected human gene products that contain the virus BKS ring finger domain, characteristic of K3 and K5 HHV-8 proteins, indicating a possible common origin and shared function for proteins containing this domain. The BKS domain has not previously been reported in mammals. K3 and K5 from HHV-8 have been recently discovered to down-regulate MHC class 1 molecules in infected cells (Coscoy and Ganem 2000; Coscoy et al. 2001); therefore, the BKS domain may be common to virus and host proteins involved in regulating cellular membrane proteins.

We have detected sequence homology with human proteins for ~13% of all known herpesvirus proteins. The question remains whether the remaining 87% can be considered exclusively viral. It is likely that a fraction may still be functional homologs with global sequence similarity too limited to be detectable by the methods used here. In addition, our methods will not detect very small sequence motifs such as phosphorylation and protein binding sites. Therefore, viral proteins such as HHV8 K15, which contains a tumour necrosis factor receptor–associated factor binding domain (Glenn et al. 1999), or EBV LMP-2A, which contains immunoreceptor tyrosine-based activation motif sequences (Fruehling and Longnecker 1997), are not detected here.

A further confounding factor for detection of viral homologs is the rapid evolution of some viral sequences. It has been estimated that herpesvirus proteins typically evolve one or two orders of magnitude more rapidly than host proteins (McGeoch and Cook 1994), and this may quickly mask any common sequence identifiable ancestry of two proteins. For example, one known human/herpesvirus homolog, thymidine kinase, is present in all known herpesviruses. Because of very limited sequence similarity, however, it could not be identified using our methods; although a human thymidine kinase mitochondrial homolog of the channel catfish herpesvirus thymidine kinase protein was detected. Human homologs of the MHV-68 serpin (serine protease inhibitor) M1 were similarly not identified using sequence similarity searches.

For proteins with viral structural functions, such as capsid constituents and capsid assembly proteins, which make a large proportion of herpesvirus genome coding capacity (20% of the genes of HHV-1), no resemblance to any human protein could be found. This is perhaps not surprising, as these have "viral-only" functions. Recently, however, another method of formulating functional hypotheses of viral proteins, in silico protein structure prediction using threading

techniques, has been applied to herpesvirus proteins. This was performed for all proteins of HCMV, yielding complete structural identifications for 36 viral proteins, only eight of which were previously known. These included some HCMV structural proteins (Novotny et al. 2001).

The relative number of homologs between herpesviruses and the human genome may also increase as the prediction methods and number of human gene products from the human genome becomes more accurate. This is highlighted by failure to detect the sequence-based homology between human and herpesvirus α-*N*-formylglycineamide ribonucleotide aminotransferase (FGARAT), or between human dUTPase and the dUTPase protein family found in all alpha- and gammaherpesviruses (HPF 43). Neither of the human predicted protein data sets contains FGARAT, even though a human FGARAT gene was recently reported (Patterson et al. 1999), and until recently neither contained the human homolog dUTP pyrophosphatase (GenBank accession no. 18583771), which shares homology with its human herpesvirus counterparts. Additional homologs for non-HHV may be identified when their host genome sequence becomes available. The reverse of this argument applies equally to herpesvirus proteins. Many of the ORFs in the herpesvirus genomes are only conceptual translations from the virus genome sequence and are, therefore, predicted hypothetical proteins. Most of the hypothetical proteins are singletons, of which only 4% showed homology with human proteins, in contrast to 10% of the herpesvirus protein families. The analysis of the expression of all ORFs using methods such as DNA array-based profiling (Chambers et al. 1999; Stingley et al. 2000; Jenner et al. 2001) will establish if these potential products are expressed during the virus cycle. Overall, the continued, virus-focused searching of constantly growing protein databases using cross-comparable methods is likely to increase our understanding of the relationship between virus and host.

METHODS

Initial Data Sets

All complete herpesvirus ORFs are available in the viral database VIDA (Albà et al. 2001b). In VIDA, the ORFs are organized into HPFs according to amino acid sequence motifs shared between the proteins, as determined by the XDOM algorithm (Gouzy et al. 1997). In some instances, HPFs contain several proteins from the same virus species. This is owing to the existence of proteins from different strains or to the presence of more than one copy of the gene in the virus genome. Each HPF is annotated with a functional description and functional class, and can contain proteins from any or all of the three herpesvirus subfamilies. The functional descriptions in VIDA include a representative gene name (e.g., "protein kinase, HHV-1 UL13" is a protein kinase family that includes gene UL13 product from HHV-1), and they are used throughout this paper to designate HPFs. When no homology with other herpesvirus proteins can be found, ORFs are represented as singleton proteins in VIDA. A total of 393 homologous multiprotein families (HPFs) and 494 singleton proteins were used in the analysis. This comprises all herpesvirus ORFs from VIDA (4054 nonredundant proteins), including all eight HHVs. VIDA can be accessed at http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA.html.

The conceptual protein translations of two human genome databases were searched in this study: The collection of human genome gene products at the National Centre for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/genome/guide/human/) and the Ensembl Project at the European Bioinformatics Institute (http://www.ensembl.org/). Both databases were downloaded by anonymous FTP and stored locally. The two databases were concatenated into a single library, and low-complexity protein segments were masked using the SEG program with default parameters (Wootton and Federehen 1993).

Construction of Motif PSSMs

Herpesvirus HPFs containing two or more proteins are defined by one or more amino acid motifs conserved across all members of the family. The large majority of HPFs are identified by a single motif (371 out of 393). However, there are 11 HPFs that contain two conserved motifs, eight HPFs that contain three conserved motifs, and three HPFs that share four motifs. The motifs, in the form of multiple alignments, were used to construct PSSMs using the program PSI-BLAST (Altschul et al. 1997). Taking into account that some families contain more than one motif, the total number of PSSMs we constructed was 429.

Construction of a Herpesvirus Protein Data Set at the 95% Identity Level

A data set of all individual herpesvirus proteins with <95% sequence identity was constructed. The representative proteins were selected by computing the global amino acid identity of each protein in each of the HPFs and grouping the proteins into subsets that shared ≥95% sequence identity using the programs HOMOL and SEQCLUSTER, respectively (Orengo et al. 1997). An ORF was then selected at random from each 95% subset (an N95-rep) and used to perform pairwise sequence similarity searches of the human protein databases. For example, nine proteins from HPF 13 (protein kinase, HHV-1 UL13) were selected to represent the 33 proteins it comprised.

Database Searches and Sequence Analysis

The IMPALA program (Schaffer et al. 1999) was used to run searches against the 429 PSSMs derived from the motifs in VIDA. An E-value cutoff of 0.01 and default parameters were used. The collection of singleton protein sequences was searched with both BLASTP (Altschul et al. 1990) and PSI-BLAST (Altschul et al. 1997), with default parameters and an E-value cutoff of 0.01. PSI-BLAST uses iterative profile construction and is more computationally expensive but generally more sensitive. As PSI-BLAST did not reveal any additional singleton homologs, N95-reps were then searched against the human protein library using BLASTP with the same parameters as above.

All database hits were examined and curated manually based on sequence alignments, conserved domain regions, functional annotation, and reference to the literature. The manual inspection of putative homologs led to the removal of some of the initial hits, which appeared to be caused by compositional bias rather than true homology. When appropriate, additional proteins from different organisms were retrieved from GenBank for sequence alignment construction. The alignments were produced by the program MULTALIN (Corpet 1988) and, when necessary, manually edited using JALVIEW (http://www2.ebi.ac.uk/~michele/jalview/contents.html/) and further visualized using BOXSHADE (http://bioweb.pasteur.fr/seqanal/interfaces/boxshade.html/). Analysis of homologous families also included searching the domain database at the NCBI, which is linked to the Pfam (Bateman et al. 2000) and SMART (Schultz et al. 2000) domain databases, using reverse position-specific BLAST (RPS-BLAST; Altschul et al. 1997).

Phylogenetic Tree Construction

Herpesvirus phylogenetic trees based on the gene content of 19 complete herpesvirus genomes were previously constructed (Albà et al. 2001a). For this type of reconstruction, phylogenetic profiles were obtained by considering the protein families as molecular function characters for which different viruses were positive (1) or negative (0). Maximum parsimony and distance methods (neighbor-joining) were applied to the phylogenetic profiles to construct phylogenetic trees. The tree shown in Figure 3 represents a consensus tree from such methods (Albà et al. 2001a).

ACKNOWLEDGMENTS

We thank Robin Weiss for support and critical reading of the manuscript. This work was funded by the Biotechnology and Biological Sciences Research Council (BBSRC; R.H. and M.M.A) and the Medical Research Council (MRC; C.O. and P.K.).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Albà, M.M., Das, R., Orengo, C.A., and Kellam, P. 2001a. Genomewide function conservation and phylogeny in the Herpesviridae. *Genome Res.* **11:** 43–54.

Albà, M.M., Lee, D., Pearl, F.M., Shepherd, A.J., Martin, N., Orengo, C.A., and Kellam, P. 2001b. VIDA: A virus database system for the organization of animal virus genome open reading frames. *Nucleic Acids Res.* **29:** 133–136.

Alcami, A. and Koszinowski, U.H. 2000. Viral mechanisms of immune evasion. *Trends Microbiol.* **8:** 410–418.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L. 2000. The Pfam protein families database. *Nucleic Acids Res.* **28:** 263–266.

Belanger, C., Gravel, A., Tomoiu, A., Janelle, M.E., Gosselin, J., Tremblay, M.J., and Flamand, L. 2001. Human herpesvirus 8 viral FLICE-inhibitory protein inhibits Fas-mediated apoptosis through binding and prevention of procaspase-8 maturation. *J. Hum. Virol.* **4:** 62–73.

Biassoni, R., Cantoni, C., Pende, D., Sivori, S., Parolini, S., Vitale, M., Bottino, C., and Moretta, A. 2001. Human natural killer cell receptors and co-receptors. *Immunol. Rev.* **181:** 203–214.

Brunovskis, P. and Kung, H. J. 1995. Retrotransposition and herpesvirus evolution. *Virus Genes* **11:** 259–270.

Chambers, J., Angulo, A., Amaratunga, D., Guo, H., Jiang, Y., Wan, J.S., Bittner, A., Frueh, K., Jackson, M.R., Peterson, P.A., et al. 1999. DNA microarrays of the complex human cytomegalovirus genome: Profiling kinetic class with drug sensitivity of viral gene expression. *J. Virol.* **73:** 5757–5766.

Chaston, T.B. and Lidbury, B.A. 2001. Genetic "budget" of viruses and the cost to the infected host: A theory on the relationship between the genetic capacity of viruses, immune evasion, persistence and disease. *Immunol. Cell Biol.* **79:** 62–66.

Chee, M.S., Satchwell, S.C., Preddie, E., Weston, K.M., and Barrell, B.G. 1990. Human cytomegalovirus encodes three G protein–coupled receptor homologs. *Nature* **344:** 774–777.

Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16:** 10881–10890.

Coscoy, L. and Ganem, D. 2000. Kaposi's sarcoma–associated herpesvirus encodes two proteins that block cell surface display of MHC class I chains by enhancing their endocytosis. *Proc. Natl. Acad. Sci.* **97:** 8051–8056.

Coscoy, L., Sanchez, D.J., and Ganem, D. 2001. A novel class of herpesvirus-encoded membrane-bound E3 ubiquitin ligases regulates endocytosis of proteins involved in immune recognition. *J. Cell. Biol.* **155:** 1265–1273.

Fisher, S., Genbacev, O., Maidji, E., and Pereira, L. 2000. Human cytomegalovirus infection of placental cytotrophoblasts in vitro and in utero: Implications for transmission and pathogenesis. *J. Virol.* **74:** 6808–6820.

Fruehling, S. and Longnecker, R. 1997. The immunoreceptor tyrosine-based activation motif of Epstein-Barr virus LMP2A is essential for blocking BCR-mediated signal transduction. *Virology* **235:** 241–251.

Glenn, M., Rainbow, L., Aurad, F., Davison, A., and Schulz, T.F. 1999. Identification of a spliced gene from Kaposi's sarcoma–associated herpesvirus encoding a protein with similarities to latent membrane proteins 1 and 2A of Epstein-Barr virus. *J. Virol.* **73:** 6953–6963.

Gouzy, J., Eugene, P., Greene, E.A., Kahn, D., and Corpet, F. 1997. XDOM: A graphical tool to analyze domain arrangements in any set of protein sequences. *Comput. Appl. Biosci.* **13:** 601–608.

Hammarstrom, S. 1999. The carcinoembryonic antigen (CEA) family: Structures, suggested functions and expression in normal and malignant tissues. *Semin. Cancer Biol.* **9:** 67–81.

International Committee on Taxonomy of Viruses (ICTV). 2000. *Virus taxonomy: The classification and nomenclature of viruses. The seventh report of the International Committee on Taxonomy of Viruses.* Academic Press, San Diego, CA.

Jenner, R.G. and Boshoff, C. 2002. The molecular pathology of Kaposi's sarcoma–associated herpesvirus. *Biochim. Biophys. Acta* **1602:** 1–22.

Jenner, R.G., Albà, M.M., Boshoff, C., and Kellam, P. 2001. Kaposi's sarcoma–associated herpesvirus latent and lytic gene expression as revealed by DNA arrays. *J. Virol.* **75:** 891–902.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

McFadden, G. and Murphy, P.M. 2000. Host-related immunomodulators encoded by poxviruses and herpesviruses. *Curr. Opin. Microbiol.* **3:** 371–378.

McGeoch, D.J. and Cook, S. 1994. Molecular phylogeny of the *Alphaherpesvirinae* subfamily and a proposed evolutionary timescale. *J. Mol. Biol.* **238:** 9–22.

McGeoch, D.J. and Davison, A.J. 1999. The descent of human herpesvirus 8. *Semin. Cancer Biol.* **9:** 201–209.

McGeoch, D.J., Dolan, A., and Ralph, A.C. 2000. Toward a comprehensive phylogeny for mammalian and avian herpesviruses. *J. Virol.* **74:** 10401–10406.

Montague, M.G. and Hutchison III, C.A. 2000. Gene content phylogeny of herpesviruses. *Proc. Natl. Acad. Sci.* **97:** 5334–5339.

Moore, P.S., Boshoff, C., Weiss, R.A., and Chang, Y. 1996. Molecular mimicry of human cytokine and cytokine response pathway genes by KSHV. *Science* **274:** 1739–1744.

Nicholas, J., Ruvolo, V., Zong, J., Ciufo, D., Guo, H.G., Reitz, M.S., and Hayward, G.S. 1997. A single 13-kilobase divergent locus in the Kaposi sarcoma–associated herpesvirus (human herpesvirus 8) genome contains nine open reading frames that are homologous to or related to cellular proteins. *J. Virol.* **71:** 1963–1974.

Novotny, J., Rigoutsos, I., Coleman, D., and Shenk, T. 2001. In silico structural and functional analysis of the human cytomegalovirus (HHV5) genome. *J. Mol. Biol.* **310:** 1151–1166.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH: A hierarchic classification of protein domain structures. *Structure* **5:** 1093–1108.

Patterson, D., Bleskan, J., Gardiner, K., and Bowersox, J. 1999. Human phosphoribosylformylglycineamide amidotransferase (FGARAT): Regional mapping, complete coding sequence, isolation of a functional genomic clone, and DNA sequence analysis. *Gene* **239:** 381–391.

Ploegh, H.L. 1998. Viral strategies of immune evasion. *Science* **280:** 248–253.

Schaffer, A.A., Wolf, Y.I., Ponting, C.P., Koonin, E.V., Aravind, L., and Altschul, S.F. 1999. IMPALA: Matching a protein sequence against a collection of PSI-BLAST–constructed position-specific score matrices. *Bioinformatics* **15:** 1000–1111.

Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., and Bork, P. 2000. SMART: A web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28:** 231–234.

Shchelkunov, S.N., Safronov, P.F., Totmenin, A.V., Petrov, N.A., Ryazankina, O.I., Gutorov, V.V., and Kotwal, G.J. 1998. The genomic sequence analysis of the left and right species-specific terminal region of a cowpox virus strain reveals unique sequences and a cluster of intact ORFs for immunomodulatory and host range proteins. *Virology* **243:** 432–460.

Somia, N.V., Schmitt, M.J., Vetter, D.E., Van Antwerp, D., Heinemann, S.F., and Verma, I.M. 1999. LFG: An anti-apoptotic gene that provides protection from Fas-mediated cell death. *Proc. Natl. Acad. Sci.* **96:** 12667–12672.
Stingley, S.W., Ramirez, J.J., Aguilar, S.A., Simmen, K., Sandri-Goldin, R.M., Ghazal, P., and Wagner, E.K. 2000. Global analysis of herpes simplex virus type 1 transcription using an oligonucleotide-based DNA microarray. *J. Virol.* **74:** 9916–9927.
Swanton, C., Mann, D.J., Fleckenstein, B., Neipel, F., Peters, G., and Jones, N. 1997. Herpes viral cyclin/Cdk6 complexes evade inhibition by CDK inhibitor proteins. *Nature* **390:** 184–187.
Tschopp, J., Thome, M., Hofmann, K., and Meinl, E. 1998. The fight of viruses against apoptosis. *Curr. Opin. Genet. Dev.* **8:** 82–87.
Voigt, S., Sandford, G.R., Ding, L., and Burns, W.H. 2001. Identification and characterization of a spliced C-type lectin-like gene encoded by rat cytomegalovirus. *J. Virol.* **75:** 603–611.
Wootton, J.C. and Federehen, S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Computational Chem.* **17:** 179.

WEB SITE REFERENCES

http://bioweb.pasteur.fr/seqanal/interfaces/boxshade.html; BOXSHADE.
http://www.ensembl.org; Ensembl Project at the European Bioinformatics Institute.
http://www2.ebi.ac.uk/~michele/jalview/contents.html; JALVIEW
http://www.ncbi.nlm.nih.gov/genome/guide/human; National Centre for Biotechnology Information.
http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA.html; VIDA.

Received October 26, 2001; accepted in revised form August 13, 2002.