Applying comparative genomics to poxvirus genome evolution

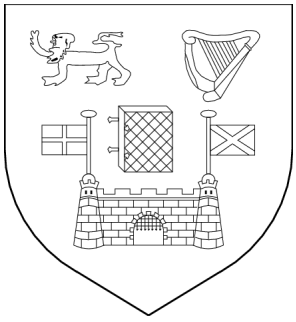
by

Kirsten Bratke

A thesis submitted to
The University of Dublin
for the degree of

Doctor of Philosophy

Department of Genetics
Trinity College
University of Dublin



October, 2009

Declaration

This thesis has not been submitted as an exercise for a degree at any other University. Except where otherwise stated, the work described herein has been carried out by the author alone. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College. The copyright belongs jointly to the University of Dublin and Kirsten Bratke.

Signature of Author .....

Kirsten Bratke

30 October, 2009

Summary

In this thesis I examine the molecular evolution of poxviruses using various comparative genomics approaches. Poxviruses are large DNA viruses that infect vertebrates and insects. Understanding the evolutionary processes that have shaped them can help us explain their infectious cycle and host interactions.

I lay the groundwork for the later analyses presented here by reconstructing a robust phylogenetic tree of the 104 poxvirus genomes whose complete genome sequences were available. This tree is based on concatenated multiple alignments of core families found in all poxviruses, and in order to identify these families, I used homology searches (BLAST) and clustering algorithms (MCL). The resulting tree agreed with previously published phylogenies of smaller groups of poxvirus genomes.

The first study discussed in this thesis aims to identify cases of horizontal gene transfer (HGT) from a host or host-related organism into poxviruses. The main genomic feature used here is conserved synteny, the genomic neighbourhood of the gene in question, which, for most poxviruses, is extremely well conserved and helps to determine whether one or multiple independent HGT events have taken place. The main finding from this study is that taking synteny into account can provide very different conclusions about HGT history than merely looking at sequence homology between host and viral genes. In particular, for thymidine kinase and interleukin-10, multiple HGT events are implied by different, but conserved, syntenic

environments, where phylogenetic relationships favoured single origins.

In the second study, an attempt is made to predict interactions between proteins, based on their phylogenetic profiles, i.e. presence or absence of genes encoding them across all poxviruses. This method assumes that where two proteins are present in a non-monophyletic group of genomes, and therefore have experienced co-gains or co-losses, this implies a biological link. While some predictions can be made, some with plausible functional links, biological relevance is hard to ascertain. When examining all known protein-protein interactions between proteins encoded by vaccinia virus, it becomes very clear that such interactions exist mainly between well-conserved proteins present since the early poxviral lineage, which tend to be located near the genomic centre. Using these characteristics, some tentative predictions of further interactions can be made, although there are too many that are too interconnected to offer a useful shortlist of candidates for further investigation.

The last study presented here takes a close look at each gene in a representative vaccinia virus genome with the aim to identify its evolutionary point and mechanism of origin. Evidence for gene fusion and fission, HGT and de novo origins is sought but the origins of a large proportion of genes (38%) remain undiscovered. A poxviral origin via horizontal transfer from another organism is by far the most common, at 32%.

Acknowledgements

A big thank you is due to Aoife McLysaght for being my supervisor, and to all current and former members of her lab: David González Knowles, Takashi Makino, Åsa Pérez-Bercoff and Daniel Murphy, for helping out, giving ideas and making this a really nice place to work.

Thank you also to other members of the Smurfit Institute of Genetics, past and present, who have helped me at some point during the past four years: Kevin Byrne, David Lynn, Gavin Conant, Meg Woolfit, Lilian Lau, Marie Sémon and Liina Rae.

A collaboration with Stefan Rothenburg and Thomas Dever at the National Institute of Health has been very enjoyable, interesting and productive and I would like to thank them for that.

Financially I was supported both by the Irish Research Council for Science, Engineering and Technology and by a Trinity College Dublin Scholarship and I am most grateful for both.

I also want to thank my parents, for encouraging and supporting me all the way to a doctorate and taking the greatest interest in everything I do.

Most of all, thank you to Daniel Kelleher, for proofreading and asking the best questions. And for making me see this through because he promised me a wedding and a marriage at the end of it.

Contents

List of Tables

List of Figures

Abbreviations

aa	amino acid
ADP	adenosine diphosphate
AmEPV	*Amsacta moorei* entomopox virus
BCE	before the common era
BLAST	Basic Local Alignment Search Tool
bp	base pair(s)
BPSV	bovine papular stomatitis virus
CA	carbonic anhydrase
cDNA	complementary DNA
CE	common era
CEV	cell-associated enveloped virus
CFR	case fatality rate
CGN	conserved gene order
ChK	chemokine(s)
ChK-BP	chemokine binding protein
ChPV	chordopox virus(es)
CK	cytokine(s)
CMLV	camelpox virus(es)
CNPV	canarypox virus
CPXV	cowpox virus
CPXV-BR	cowpox virus strain Brighton Red
CRV	crocodilepox virus
DNA	deoxyribonucleic acid
DP	deoxyribodipyrimidine photolyase
dsDNA	double stranded DNA
dUTP	deoxyuridine triphosphate
EEV	extracellular enveloped virus
EF	enhancing factor
EPV	entomopox virus
FPV-I	fowlpox virus strain Iowa
GAG	glycosaminoglycan
GOI	gene of interest
GP	glutathione peroxidase

GPCR	G-protein coupled receptor
GTPV	goatpox virus
HGT	horizontal gene transfer
HIV	human immunodeficiency virus
HPXV	horsepox virus
HT	horizontally transferred
IEV	intracellular enveloped virion
IL	interleukin(s)
IMV	intracellular mature virion
ITR	inverted terminal repeat
kb	kilobases
KSHV	Kaposi sarcoma-associated herpesvirus
LR	likelihood ratio
LRR	leucine rich repeat
LSDV	lumpy skin disease virus
MAP	mitogen-activated protein
MCL	Markov Cluster Algorithm
MHC	major histocompatibility complex
miRNA	micro RNA
ML	maximum likelihood
MOCV	Molluscum contagiosum virus
MP	maximum parsimony
MPXV	monkeypox virus(es)
mRNA	messenger RNA
MYXV	myxoma virus
NAD	nicotinamide adenine dinucleotide
NCLDV	nucleo-cytoplasmic large DNA virus(es)
NJ	neighbour-joining
nm	nanometres
nt	nucleotide(s)
OPV	orthopox virus(es)
ORF	open reading frame
ORFV	orf virus
PBRC	Poxvirus Bioinformatics Resource Center
PKR	protein kinase R
PPI	protein-protein interaction(s)
PSSM	position-specific scoring matrix
PV	poxvirus(es)
REV	reticuloendotheliosis virus
RING	really interesting new gene
RNA	ribonucleic acid
RPO	RNA polymerase
RPXV	rabbitpox virus

RR	ribonucleotide reductase
RRL	RR, large subunit
RRS	RR, small subunit
SD	standard deviation
Serpin	serine protease inhibitor
SFV	Shope fibroma virus
SPPV	sheeppox virus
TATV	taterapox virus
TK	thymidine kinase
TM	transmembrane
VACV	vaccinia virus(es)
VACV-COP	vaccinia virus strain Copenhagen
VARV	variola virus
VETF	viral early transcription factor
VITF	viral intermediate transcription factor
VLTF	viral late transcription factor
VZV	varicella zoster virus
WHO	World Health Organization
YLDV	yaba-like disease virus

Für Papa

Chapter 1

Introduction

Poxviruses are large double-stranded DNA viruses with genomes of up to 360 kb that infect a wide range of hosts from insects to mammals. The virions are rectangular, approximately 350 × 270 nm in size, surrounded by a 30 nm surface layer. They are classified into entomopox (EPV, insect-infecting) and chordopox viruses (ChPV, vertebrate-infecting), with ChPV sub-divided into at least nine genera. Several members of the *Poxviridae* are associated with human disease, including Molluscum contagiosum virus (MOCV) and monkeypox virus, which has been classified as an emerging virus (Su, 2004) and was responsible for an outbreak of monkeypox in the US in 2003. The most notorious poxvirus is variola, an orthopox virus (OPV), which causes smallpox and was responsible for killing over 500 million people in the 20th century (Mahalingam *et al.*, 2004) until it was declared eradicated by the World Health Organization (WHO) in 1979. Vaccinia is the non-lethal virus used to vaccinate against variola. Cowpox, a closely related OPV, was used in the development of the first vaccine by Edward Jenner in the late 18th century. Vaccinia virus (VACV) swas the first animal virus to be seen under a microscope, grown in tissue culture and chemically analysed (Moss, 2001).

These aspects of poxvirus biology make them interesting organisms to study. The potential for future outbreaks challenges science to learn as

much as possible about how the genomes of these viruses evolve. In this thesis, I present work that uses computational comparative genomic methods to further the understanding of poxvirus evolution. Chapter 2 describes some of these methods, while Chapters 3, 4 and 5 present work done on horizontal gene transfer (HGT) into poxviruses, the use of phyletic patterns for determining function of unknown poxvirus proteins and the history of all genes in vaccinia virus, respectively. Chapter 6 discusses the conclusions drawn from this body of work. In this introduction, I will give an overview of poxvirus biology, genomes and evolution, and discuss in some detail smallpox and its history, pathology and the threat posed by its potential as a bioweapon. Information that is more relevant to my own research will be introduced in the following chapters.

1.1 The History and Biology of Smallpox

1.1.1 History

Smallpox has claimed the lives of "more members of the human population over the span of recorded history than all other infectious diseases combined" (McFadden, 2005). The disease probably originated over 3000 years ago in Egypt but the original host of the virus or manner of transfer into humans is unclear. Among the victims were Queen Mary II of England and Tsar Peter II of Russia, who died on his wedding day (WHO, 2008a). Skin rashes typical of smallpox were discovered on the mummy of Egyptian Pharaoh Ramses V, who died around 1150 BCE, and historical records from China and India from this period also describe a disease which may have been smallpox. As there is no animal reservoir for variola virus (VARV), it probably derives from a closely related OPV present in another animal. Gubser and Smith (2002), who found that the genome sequence of variola virus is most similar to that of camelpox virus, suggested that the two species have evolved from

a closely related ancestral virus that infected rodents. Esposito *et al.* (2006) more specifically suggested gerbils as the origin of variola. A possible scenario for the spread of smallpox therefore lies in a transfer from Egypt to India, where it became an endemic human disease, followed by a spread to China in the first century CE and to Japan 500 years later. Islamic armies spread the smallpox across North Africa and into Southern Europe in the 7th and 8th centuries CE (Fenner *et al.* (1988), Chapter 5; Geddes (2006)). The Crusades in Europe facilitated the spread of variola virus in the West and when Europeans started to settle America in the 16th century, they brought smallpox along with them, killing a large percentage of the native population who had no immunity to the new disease (Diamond, 1997).

Esposito *et al.* (2006), in an analysis of 45 variola isolates, found three main clades that cluster according to their geographic origin. Li *et al.* (2007) used these variola isolates and historical records to show that two primary clades, containing Asian major strains and strains from West Africa, respectively, diverged either 16,000 or 68,000 years ago, depending on records used. Both minor and major forms of the virus were present together in one primary clade, not supporting the traditional view of separation of strains. Figure 1.1 was taken from their paper and shows two hypotheses of VARV origin.

1.1.2 Disease Progression

Smallpox is transmitted mainly through aerosol but also through infected clothing or bedding, and is less transmissible than measles, chickenpox and influenza (Breman and Henderson, 2002). The following outline of disease progression is taken from Fenner *et al.* (1988) and Breman and Henderson (2002) and also holds true for other OPV. The virus enters the respiratory tract and passes into the lymph nodes. There follows a latent incubation period of 4-14 days during which the virus multiplies without

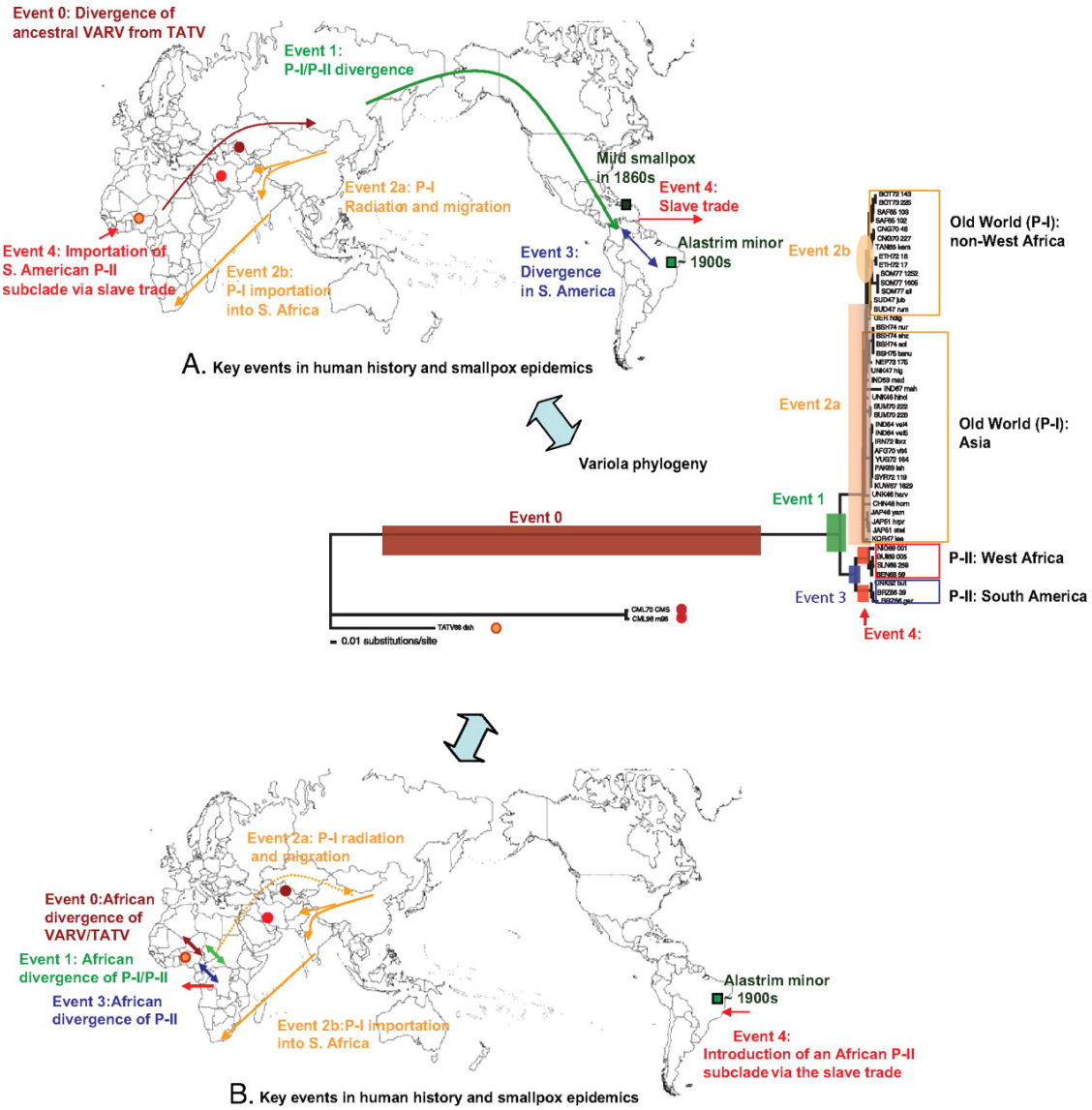


Figure 1.1: Origin of VARV, as suggested by Li *et al.* (2007). Depending on the time points used to calibrate these analyses, different scenarios for VARV evolution and smallpox epidemic routes are inferred. Common to both hypotheses is a divergence of two variola clades into two geographically discrete populations.

causing symptoms. The disease then enters into the prodromal phase, where mucous membranes in the mouth and pharynx are infected. Headache, back pain and fever are common symptoms in this pre-eruptive stage. The virus moves into the capillary epithelium of the skin and causes the characteristic lesions, predominantly on the face and appendages of the patient. The lesions progress through several stages: "herald spots," typically on the forehead, are followed by lesions on other body parts, starting as papules and macules, turn into vesicles and eventually form scabs which leave depigmented scars upon healing. Death from smallpox is usually attributed to toxemia, the presence of toxins in the blood, and to hypotension. Blindness and arthritis can be a side-effect in surviving patients. When infected macrophages migrate to the lymph nodes, cytotoxic T- and B-cells are produced and neutralising antibodies are formed that remain in the system for years, so that smallpox infection leads to lifelong immunity.

90% of smallpox infections are due to variola major, which has a case fatality rate (CFR) of 30%. Variola major can cause different clinical types of disease (Fenner *et al.*, 1988, Chapter 1). The ordinary type just described can be subdivided into confluent, semiconfluent and discrete ordinary-type smallpox, depending on the extent of the rash and the distance of lesions from each other. The discrete subtype was the most common of these and had the lowest CFR. Modified-type smallpox was rare and characterised by a quicker progression of the disease and a lower CFR. This was most often due to vaccination some years previously. Flat-type smallpox was almost always fatal and was identified by the lesions remaining level with the skin and maturing very slowly. Haemorrhagic-type smallpox was similarly rare and fatal. Symptoms presented as above with haemorrhages into the skin and mucous membranes and bleeding from many areas, with patients often dying of heart failure.

Vaccinia virus present in an aerosol is almost completely destroyed within

six hours at 30 - 31°C and 80% humidity, and variola is expected to behave similarly (Henderson *et al.*, 1999).

No treatment other than administration of vaccine up to four days after exposure to variola or the management of symptoms is currently available, but antiviral agents are being investigated. Gleevec, a drug designed to treat chronic myeloid leukemia, has been shown to have some inhibitory effect on the spread of poxviruses, through blocking the release of viral particles from the cell surface (Reeves *et al.*, 2005). Prichard *et al.* (2007) examined the effect of several thymidine analogues on poxviruses. Where thymidine kinase is present and available to phosphorylate these compounds, they showed promising antiviral activity. The drug ST246 also shows promise as an anti-poxviral drug, efficiently reducing the production and release of OPV (Smith *et al.*, 2009). In addition to chemical agents, miRNA-based gene silencing may be another avenue towards gaining control of viral replication, as presented by Barnes *et al.* (2008). Insertion of miRNA homologues into a viral genome may restrict the genome's tissue tropism and prevent pathogenicity, giving an attenuated virus strain.

1.1.3 Eradication

When Jenner made the observation that inoculation with cowpox could protect against smallpox, he provided the first indication that this disease might be controlled. Incidentally, there is now growing recognition of a Dorset farmer, Jesty, who had privately vaccinated his family in the same way some decades before Jenner, as reported by Pead (2003). The vaccines used in the 20th century and still being stocked today contain pulp scraped from vaccinia-infected animal skin. Vaccination provides protection from infection with smallpox for at least ten years and even after that, any symptoms were milder and transmission was less likely than in unvaccinated individuals (WHO, 2008a). The success of vaccinia as a vaccination agent is due to

the high replicative fidelity of the viral DNA polymerase, which limits the ability of smallpox to adapt to its human host to such an extent as to make its antigens unrecognisable by an immune system previously exposed to vaccinia (Bray and Buller, 2004).

Through a global eradication campaign, the last natural case of smallpox was reported in Somalia in 1977, one laboratory infection caused another death in the UK in 1978 and the disease was officially certified eradicated in 1979 (Breman and Henderson, 2002), a momentous event for world health (Figure 1.2). The eradication strategy rested on mass vaccination campaigns and a surveillance system for the detection of cases and the containment of new outbreaks as summarised in Chapter 10 in Fenner *et al.* (1988). One of the reasons eradication was possible was the lack of an animal reservoir host of the virus. Routine vaccinations against smallpox stopped in 1982 because the risk of adverse reactions outweighed the threat of endemic smallpox, leading to a large susceptible population worldwide. The main complications associated with vaccination are severe inflamed eczema, progressive vaccinia, where lesions fail to heal and eventually are fatal and postvaccinial encephalitis, most often seen in children, with a fatality rate of 35% (WHO, 2008a). There has been no record of natural immunity to smallpox.

1.1.4 Bioterrorism

There exists a threat of bioterrorism involving poxviruses. Fear of this biological threat perhaps peaked in the late 1990s when Ken Alibek, the former deputy director of the Soviet Union's civilian bioweapons program defected to the US and revealed that his government had produced large quantities of smallpox for use in intercontinental missiles and was working on the production of more virulent variola strains as a world no longer protected from smallpox presented a military opportunity (Alibek, 1999). With the

Figure 1.2: Postage stamps to commemorate the eradication of smallpox were published by many different countries, here: Guinea and the United Nations.

Soviet Union broken up, and the Cold War over, Alibek felt the threat had diminished but not disappeared as rogue regimes and terrorist groups may be in possession of knowledge gained through his research. In the past, variola has been used as a bioweapon by British forces in North America during the French and Indian Wars (1754 - 1767). By providing native Americans with blankets that had been in contact with smallpox patients they caused epidemics that killed more than 50% of many affected tribes (Crosby, 1976). Whether this was done intentionally or not, in the words of Charles Darwin: "Wherever the European has trod, death seems to pursue the aboriginal" (Darwin, 1839).

Working groups on public health are formulating containment strategies in the event of an outbreak (Henderson *et al.*, 1999) and research laboratories from many different disciplines are concerning themselves with modelling smallpox outbreak, spread and vaccination (Pourbohloul *et al.*, 2005; Burke *et al.*, 2006; Longini *et al.*, 2007). It is generally agreed that surveillance and containment are crucial, with vaccination of hospital workers and other at-risk groups being recommended if threat levels were to rise.

Health authorities have enforced very strict guidelines for working with variola to avoid an outbreak or a genetically engineered variola strain (WHO, 2008b). Only two institutes in the world currently hold variola isolates, the Center for Disease Control and Prevention in Atlanta, United States and the State Research Center of Virology and Biotechnology in Novosibirsk, Russia. Where variola research is carried out by other laboratories, no other OPV strain may be handled at the same time, and no more than 20% of the smallpox genome may be represented by clones in any working laboratory to avoid danger of recombination or intentional reconstruction of an infectious variola virus strain.

1.2 Poxvirus Infectious Cycle

1.2.1 Affected Organisms

Economic loss due to sheep, cattle, camels and even crocodiles infected with poxviruses is a serious problem in some regions of the world and may become more important with global climate change as the spread of insect vectors widens (Babiuk *et al.*, 2008).

Different groups of poxviruses employ different infectious strategies: while OPV cause systemic and potentially lethal infections, MOCV causes acute localised infections, small benign skin tumors with almost no inflammation. These strategies are reflected in different sets of genes modulating virus-host interactions (Senkevich *et al.*, 1996). The effects of poxviruses on their hosts range from causing mild skin lesions to lesions of the respiratory tract to systemic immunosuppression. Ectromelia virus infection in the mouse model, for example, causes a localised growth at the site of inoculation, involvement of regional lymph nodes and then viremia (entering of viruses into the bloodstream), leading to an infection of the liver and spleen and then other organs, which eventually causes death (Turner and Moyer, 1990). The type of disease caused may depend on the route of infection, for example, when inhaled, variola causes a severe systemic illness as described above, but a milder form of the disease manifests when the skin is inoculated with viral particles (Bray and Buller, 2004). Poxvirus lesions often appear in the soft mucosal tissues of the mouth, nose, muzzle, head, neck, teats, tongue and respiratory tract of infected hosts. In deer, poxvirus infection can lead to keratoconjunctivitis.

Poxviruses differ in their ability to infect various host species. ChPV have co-evolved with their hosts to a degree but the high fidelity of the viral DNA polymerase has limited host-specific adaptation and many viruses are pathogenic in several hosts. Chapter 5 will deal with genes responsible for

host range of poxviruses in more detail.

1.2.2 Cell Entry

The virus can be transmitted by vectors like lice (swine) and insects (fowl) as well as by aerosol (sheep) or ingestion (insects). Vaccinia can attach to and enter almost any cell lines, suggesting that receptors allowing this virus entry must be ubiquitous. At least eight proteins are known to form the multiprotein entry-fusion complex, which needs the viral membrane in order to assemble (Senkevich *et al.*, 2005; Bisht *et al.*, 2008). These proteins are conserved in all poxviruses. The components of this complex locate to the viral membrane, even if the formation of the complex is interrupted. Several further proteins are involved in vaccinia cell entry and virus-induced cell-cell fusion, (Turner and Moyer, 2008; Nichols *et al.*, 2008). Entry of a virus particle is facilitated by fusion with the cell's plasma membrane. Cell surface glycosaminoglycans (GAGs) play a role in binding of the virion. Mercer and Helenius (2008) observed that vaccinia virus appears to mimic apoptotic bodies to avoid detection by the immune system and gain cell entry through a process known as macropinocytosis, an invagination of the cell membrane. Once inside the host cell, poxviruses remain in the cytoplasm, never entering the nucleus, unlike many other viruses, and in order to adapt to this environment they provide their own replication machinery to a large extent.

1.2.3 Poxvirus Gene Expression

Poxvirus expression takes place in stages: early, intermediate and late, which are activated in a functional cascade. Proteins expressed early include DNA and RNA polymerase, and other housekeeping proteins, like those involved in DNA replication and nucleotide biosynthesis. This group also contains molecules needed for the next stage of expression, like

intermediate transcription factors. The intermediate phase produces mainly late transcription factors, while during late expression, proteins for virion morphogenesis and packaging are produced. Factors needed for early gene transcription, like mRNA modification enzymes, are in turn packaged inside the virus particles for use in the next infectious cycle. Assarsson *et al.* (2008) used a genome tiling array to simultaneously measure the expression levels of all vaccinia genes during infection. They found 35 genes with immediate-early expression with over half of them having an unknown function and others being involved in modulating the host's immune response and transcribing late genes. Rubins *et al.* (2008) used a pox-human DNA microarray to monitor both viral and host expression in different stages of various infected human cell types. This provided them with a complete transcriptional map of vaccinia and MPV genomes, in which they found no overall relationship between genomic location and transcriptional timing, with the exception of an overrepresentation of early transcripts near the genome termini and late transcripts, often in tandem clusters, in the central region. Genes do, however, appear to be arranged in such a way as to minimise transcriptional interference and make efficient use of the genome space. This is achieved by frequently embedding transcription termination signals within adjacent genes. Three mechanisms allow such a gene to be expressed nonetheless: an early signal may be contained in an adjacent late gene, where it is disregarded during late gene expression, a gene may use the transcription termination signal of an adjacent gene, or a signal may be placed in an adjacent gene transcribed in the opposite direction (Lee-Chen and Niles, 1988; Condit and Niles, 2002). Similarly, termination signals of early genes are not found inside other early genes. Vaccinia early mRNA is detected within 20 minutes after infection and reaches a peak concentration after 1-2 hours, while late mRNA is first detected after 2 hours, reaching a peak around 4 hours post-infection (Moss, 2001). Each vaccinia gene contains its own, simple, transcriptional

promoter, comprising approximately 30 bp upstream of the first nucleotide (Condit and Niles, 2002). The early-stage promoter is AT-rich and has a consensus core sequence: AAAAAATGAAAAAA/TA. Intermediate- and late-stage promoters are also AT-rich but different in sequence. A conserved sequence element is required to achieve a promoter effect, and Eaton *et al.* (2008) showed that the second repeat containing the sequence TAAAT is critical.

There is a termination sequence (TTTTTNT) near the end of most early genes. Several groups have reported occasional instances of read-through transcription, where two or more adjacent genes are transcribed without termination, but this is not a precise mechanism and is not the norm (Bajszar *et al.*, 1983; Golini and Kates, 1984; Cabirac *et al.*, 1986; Macaulay *et al.*, 1987). The mRNA capping enzyme is required to be present for termination, and it acts here as a trans-acting RNA-binding protein that binds to the termination sequence. Early viral transcripts are capped (5' end) and polyadenylated (3' end) as required for ribosome binding and thus resemble the mRNAs of their eukaryotic hosts. Unlike in eukaryotic systems, however, polyadenylation occurs as a consequence of transcription termination, rather than endonucleolytic processing (Shuman and Moss, 1988). Some poxvirus proteins undergo various forms of post-translational modification like glycosylation, phosphorylation, ADP-ribosylation and acylation (VanSlyke and Hruby, 1990). A mRNA decapping enzyme, D10, has also been identified in vaccinia virus, and is conserved in all poxviruses (Parrish *et al.*, 2007).

Intermediate mRNA translation follows DNA replication, as described below, enabling late mRNA translation. These stages are also called the post-replicative stages of gene expression and depend on replication, as they can be inhibited by DNA replication inhibitors as reviewed by Condit and Niles (2002). Late mRNAs have a 5' poly(A) leader sequence, which is perhaps

necessary to resist the deleterious effects of read-through complementary RNA on translation (Moss, 1990).

As an exception to what has just been described, a small number of genes have tandem early and late promoters and are active throughout poxvirus infection, albeit not at the same time (Moss, 1990).

1.2.4 DNA Replication

DNA replication commences after the transcription of early poxvirus genes (about half of the viral genome). Two drugs, etoposide and mitoxantrone, have been shown to inhibit vaccinia replication despite their usual target (type II topoisomerase) not being encoded by poxviruses (Lin *et al.*, 2008). Viral DNA ligase, however, is targeted by these drugs in poxvirus, and as this ligase was shown to bind to a protein homologous to a part of type II topoisomerase, it appears that ligase recruits human topoisomerase to the site of virus biogenesis, thus enabling virus replication. One DNA strand is nicked near the terminus and the complementary strand is formed by DNA polymerase, giving one open and one closed hairpin structure. The open one has an inverted repeat and can fold back on itself and prime replication, which can continue around the hairpin and form concatemers, several DNA molecules joined end to end in alternating orientation (Moyer and Graves, 1981). The junction consists of a duplex copy of the hairpin loop that is present at the end of each poxvirus genome. **Figure 1.3 illustrates this formation of concatemers.**

1.2.5 Poxvirus Assembly and Dispersion

Virion assembly then begins in specific areas of the cytoplasm (viral factories). After late gene expression, the concatemeric DNA is rapidly resolved into unit-length genomes and packaged inside immature virions which then matures and acquires a viral membrane from modified Golgi

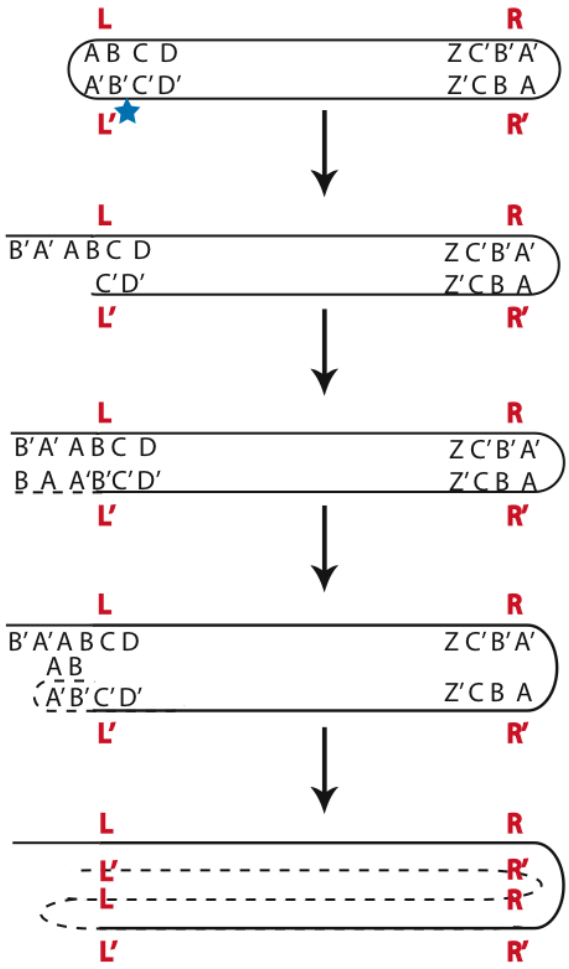


Figure 1.3: Poxvirus replication after Figure 8 in Moyer and Graves (1981). Red letters label the left and right ends of the genome, black letters sequence elements, with a dash indicating complementary sequences. Strand extension after the nick (blue star), and hairpin formation lead to an old genome copy (solid line) attached to a new genome copy (dashed line) in alternating orientation. This is repeated to produce longer concatemeric molecules.

membranes. Resolution requires two copies of the sequence A/T-T_3-A/G-N_{7-9}-A_6 in inverted repeat orientation on the hairpin loop (Merchlinsky, 1990). The intracellular mature virion (IMV) may either remain within the cell and be shed with skin debris or move out to the cell periphery, relying on active transport on microtubules. In the latter scenario, the IMV takes on a second lipid membrane, which is host-derived and thus aids in escaping host detection once the virus is released from the cell. The virus particle can now also be referred to as the intracellular enveloped virion (IEV). The outer membrane fuses with the plasma membrane to release the virion and is labelled cell-associated enveloped virus (CEV) when retained on the cell surface, or extracellular enveloped virus (EEV) if released from the cell. On budding, actin tail formation is triggered, and CEV are located at the tips of these microvilli, which spread the virions to adjacent cells, while EEV mediate more long-range dissemination. In some ChPV, IMV become occluded in a dense protein matrix (occlusion bodies), that are released into the environment after the infected cell degenerates (Roberts and Smith, 2008). Figure 1.4 is from McFadden (2005) and illustrates the different steps of the poxvirus infectious cycle as described in this section.

1.2.6 Virus-Host Interactions

To interfere with the host's immune system, sophisticated viruses like the *Poxviridae* interact with their host cells in a number of different ways as described below. However, there are also remnants of less invasive interactions in the genomes of poxviruses: as the virions do not pass into the host cell's nucleus, they encode the majority of the enzymatic machinery needed for basic housekeeping processes such as DNA replication and transcription. There is evidence that many of these enzymes were initially acquired from the host through horizontal gene transfer (HGT) and have since adapted to more efficiently carry out the virus-specific needs.

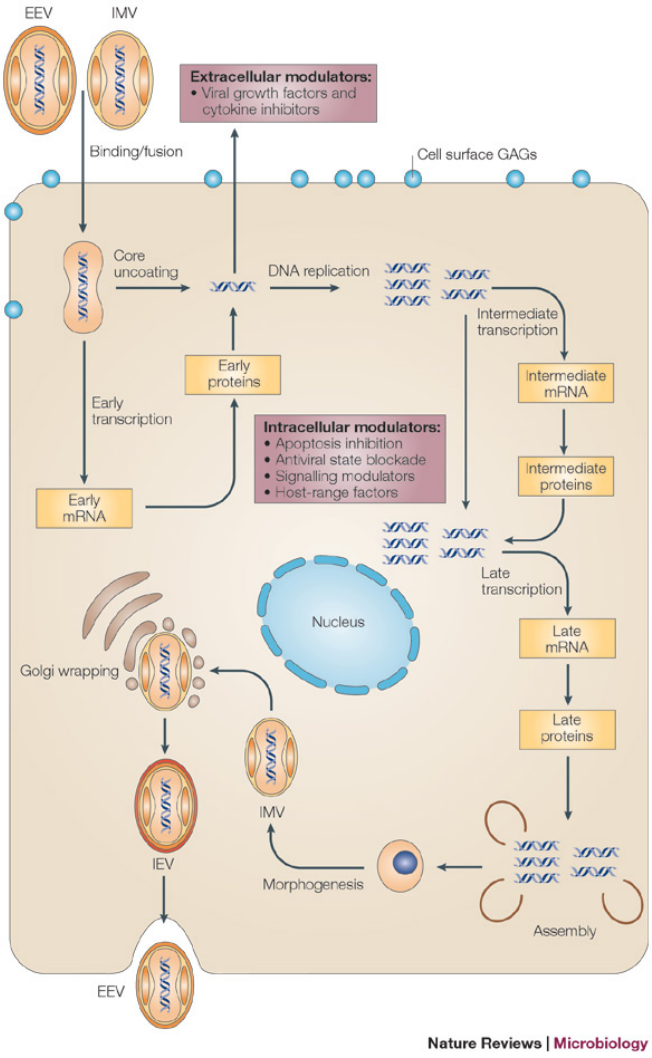


Figure 1.4: Poxvirus Infectious Cycle. Purple boxes list modulators through which poxviruses interact with the host. Figure from McFadden (2005).

Hughes (2002) noted that these proteins, some of which originated early in the history of life, may have been necessary for DNA viruses to evolve. In Chapter 3, a strict and comprehensive search for poxvirus proteins potentially acquired via HGT was carried out and most of the candidates are indeed involved in DNA precursor metabolism and DNA replication.

Many other proteins, not involved in housekeeping functions, may have been captured from the host and because of their homology to the host proteins are now involved in competing with their cellular counterparts, by sequestering the host protein's substrate or binding protein. Reviews of these proteins have been written by Barry and McFadden (1997); Bugert and Darai (2000); Lucas and McFadden (2004). Figure 1.5 was taken from Johnston and McFadden (2003) and illustrates the different ways in which poxviruses interfere with the cellular system.

The expression of viral cytokines (CK) and cytokine receptors and binding proteins is an important strategy for hijacking the host's immune system, as reviewed by Alcami and Smith (1995a). Cytokines like interleukins (IL, see Chapter 3 for more details on IL-10) regulate the inflammatory immune response to pathogens in the cell, via cytokine receptors. Viral homologues of these receptors bind to the cellular molecules, thus inhibiting the cell's immune response pathway. Among these viroreceptors are tumour necrosis factor (TNF) receptor crmB (Hu *et al.*, 1994), IL-receptors, Toll-like receptors (Bowie *et al.*, 2000) and interferon receptors (Alcami and Smith, 1995b; Xu *et al.*, 2008). **Many of these viral immunomodulators have the effect of inhibiting or suppressing the downstream protein complex NF-κB, which acts as a transcription factor (e.g. for cytokines) in response to stress, often via interfering with the adaptor protein MyD88 that is an integral part of the signalling pathway (Bowie *et al.*, 2004).**

Chemokines (ChK) mediate migration of immune cells to sites of

inflammation and poxviruses encode ChK (Senkevich *et al.*, 1996) and ChK receptor/ChK binding protein homologues (Alcami *et al.*, 1998; Seet *et al.*, 2003) to bind targets without inducing downstream signalling. Chemokine binding proteins bind and sequester some ChK in some OPV, parapox and myxoma, but their sequence is not homologous to any known cellular receptors and might be a new development in poxvirus (Bugert and Darai, 2000).

The human complement system is involved in the regulation of humoral immune response and inflammation and can be activated by three independent pathways that lead to the assembly of serine protease complexes. Control factors encoded by OPV inhibit the induction of inflammation and thus diminish the destruction of infected host tissue, allowing further viral infection. Ciulla *et al.* (2005) show that this protection against the host complement system is probably due to the acquisition of an ancestral complement regulator by an ancestral OPV and divergence in the viral lineages since then, to enhance species-specific functions.

Serine protease inhibitors (serpins) are involved in a wide range of processes, for example, the degradation of antigens into peptides for presentation to cytotoxic T-cells; viral serpins thus decrease recognition of pathogens and the following immune response. They were first identified by Pickup *et al.* (1986) in cowpox, where they are responsible for haemorrhage by inhibiting proteases involved in normal coagulation. Serpin homologues are found in many poxviruses (Smith *et al.*, 1989; Afonso *et al.*, 2000; Lee *et al.*, 2001; Afonso *et al.*, 2005). In myxoma, SERP1 inhibits the complement system, decreasing inflammation and lysis of infected tissues (Macen *et al.*, 1993). Serpins appear to be fundamental to the success of the poxvirus as Legrand *et al.* (2005) showed that inactivation along with co-expression of antiviral cytokine IFN-γ leads to a safe virus that does not replicate in mice and produces a good immune response after one immunisation.

Interestingly, Shope Fibroma virus (SFV), another leporipoxvirus, only carries a fragmented copy of SERP1, which might explain its decreased virulence in European rabbits.

Cellular growth factors lead to cell growth and induce angiogenesis, the growth of blood vessels. Encoding homologues allows viruses to compete for receptor binding, thus stimulating growth of infected cells. This contributes to the vascularised nature of lesions formed in bovine papular stomatitis virus (BPSV) (Inder *et al.*, 2007) and many other poxviruses.

Poxviruses are also known to interfere with host cell apoptosis. Where controlled cell death of an infected cell would end a virus particle's infectious cycle, host protein homologues have been shown to inhibit apoptosis and thus prolong infection, e.g. in orf virus (ORFV, Westphal *et al.* (2007)), fowlpox virus (Banadyga *et al.*, 2007) and myxoma virus (Johnston *et al.*, 2005). However, while most poxvirus infections have an anti-apoptotic effect, parapox have been observed by Kruse and Weber (2001) to also be able to induce apoptosis in monocytes, the source of antigen presenting cells, thus reducing antigen presentation. This strategy may be behind the ability of parapox to repeatedly infect their host even after short intervals. The viral protein responsible for this pro-apoptotic behaviour is likely to be a structural component. Poxvirus-encoded apoptosis inhibitors may conceal the presence of infected cells where MHC-I surface proteins would normally present non-self antigens to cytotoxic T-lymphocytes, which would lead to cell death via the activation of the Fas death receptor or the granzyme pathway (reviewed by Everett and McFadden (2002)). Alternatively, viral proteins can sequester immune system proteins involved in the apoptotic response, as described for cytokine receptors above. Inhibition of the pro-apoptotic caspase cascade is achieved by many poxviruses, e.g. in cowpox, where serpin CrmA inhibits caspase-1, which would normally activate a pro-inflammatory cytokine (Ray *et al.*, 1992), and caspase-8 (Zhou *et al.*, 1997).

Semaphorins are large secreted surface molecules, involved in neuronal guidance through remodeling of the cytoskeleton. Vaccinia open reading frame (ORF) A39R is a poxvirus-homologue of this family (Gardner *et al.*, 2001) and by binding to plexin C1, a semaphorin receptor, causes rearrangements of the actin cytoskeleton, which can inhibit phagocytosis, as this strongly depends on cytoskeletal structure, thus favouring spread of the virus. These observations establish semaphorin homologues as yet another route by which poxviruses can evade the host's immune system (Walzer *et al.*, 2005).

Another example of poxvirus proteins mimicking those of their host is vaccinia protein K3L, which is a homologue of the eukaryotic translation initiation factor eIF2. This cellular protein is phosphorylated and thus inactivated by the protein kinase R (PKR) in the presence of a viral infection, which leads to inhibition of protein synthesis and viral replication. K3L acts as a competitive inhibitor, trying to avoid this shutdown of viral activity. This creates a situation, where cellular PKR must be conserved enough to recognise its cellular substrate but flexible enough to avoid equally rapidly evolving substrate mimics like K3L. It has been shown that both PKR and K3L are under positive selection (Elde *et al.*, 2008), illustrating how this "molecular arms race" drives the evolution of both host and pathogen proteins. E3L, another vaccinia protein, inhibits PKR by forming heterodimers, and it too may have contributed to the rapid evolution PKR experiences in vertebrates (Rothenburg *et al.*, 2008). These authors have shown how altering positively selected sites in PKR leads to altered sensitivity to its viral inhibitors.

1.3 Poxvirus Genomes

Poxviruses are brick-shaped and measure approximately $300 \times 250 \times 200$ nm. Their genomes are composed of a single linear double-stranded DNA molecule

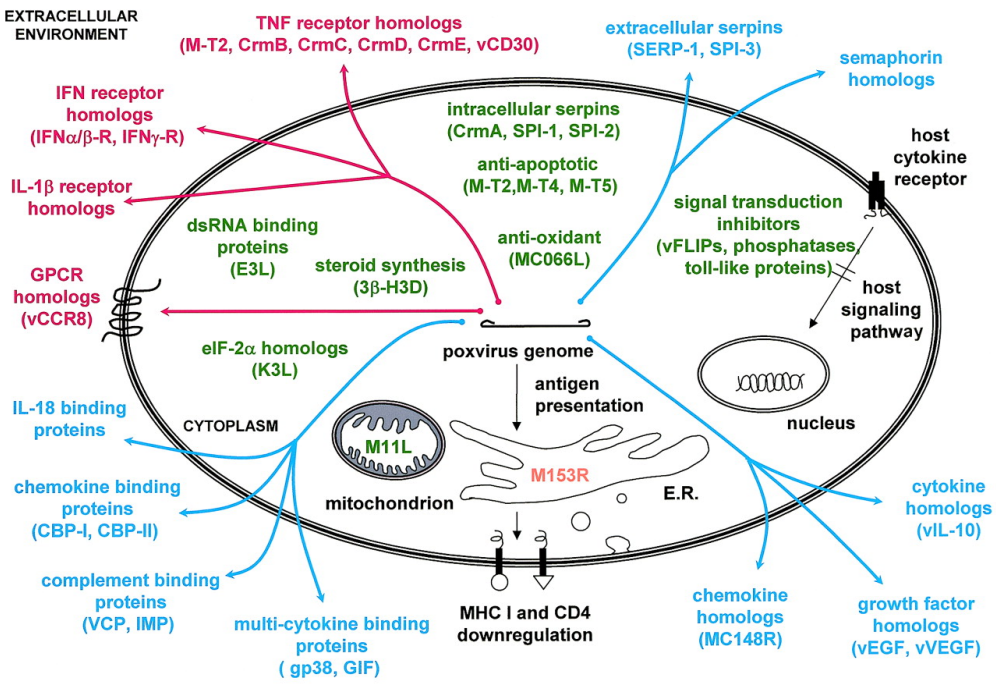


Figure 1.5: Immunomodulatory Proteins. Different colours indicate different various modes of interfering with the host immune system: virostealth (orange), virotransduction (green), viromimicry through viroreceptors (red) and viromimicry through virokines (blue). Abbreviations: GPCR - G-protein coupled receptor; GIF - granulocyte-macrophage colony-stimulating factor/IL-2 factor; IMP - inflammation modulatory protein; vEGF - viral epidermal growth factor; vVEGF, viral vascular endothelial growth factor. Figure from Johnston and McFadden (2003).

ranging in size from 134 kb to 360 kb. Poxvirus genes contain no introns and little overlap, so the genome space is used very efficiently, as dictated by limited virus particle size. The proportion of the genome that is taken up by intergenic space ranges from under 5% in most clade II poxviruses to just under 20% in some OPV (data not shown). Those virus genera that diverged after OPV are collectively referred to as clade II poxviruses (Hughes and Friedman, 2005): yatapox, leporipox, deerpox, capripox and suipox (Figure 2.2 on page 51).

The genome ends in inverted terminal repeat (ITR) regions, which are initially identical sequences with opposite orientation that include an AT-rich crosslinked hairpin-shaped terminus connecting the two DNA strands. Immediately adjacent to the hairpin loops are series of tandem repeats, measuring 3.5 kb in vaccinia. ITR regions are of variable length in poxviruses, ranging from 0 to 12,000 bp, as noted in Table 1.1. Notably, in most variola virus strains, this region is smaller than 1 kb, consistent with the early observation by Mackett and Archard (1979), that the terminal repeats of variola did not cross-hybridise, evidently because they are not homologous.

Deletion of sequences from the near-terminal regions can often cause attenuation of the virus, for example in rabbitpox virus and cowpox virus strain Brighton Red (CPXV-BR), where many mutants with restricted host range exhibited deletions in the terminal regions (Pickup *et al.*, 1984). The GC-content of poxviruses varies dramatically, from 18% in EPV, to almost 65% in parapox, as noted in Table 1.1.

Over 100 genomes have been fully sequenced and many attempts have been made to reconstruct a robust phylogeny of poxviruses (McLysaght *et al.*, 2003; Gubser *et al.*, 2004; Hughes and Friedman, 2005; Xing *et al.*, 2006). As part of the work described in this thesis, I have reconstructed a comprehensive more recent phylogenetic tree (Section 2.4). Table 1.1 lists the viruses that all analyses in this thesis are based on. The main poxvirus dichotomy is

that between entomopox (EPV) and chordopox (ChPV), infecting insects and vertebrates, respectively. Within ChPV, orthopox viruses (OPV) form the largest group with currently 78 fully sequenced genomes. The reference poxvirus strain is vaccinia virus Copenhagen (VACV-COP), sequenced by Goebel *et al.* (1990), the non-pathogenic variola vaccine, which contains 198 protein-coding regions in 192 kb. The genes were named according to their position in a *Hin*DIII restriction fragment followed by the letter L or R depending on the direction of the ORF. It was recognised early on that there is a well-conserved genome structure within ChPV (Mackett and Archard, 1979) and that genes in the syntenic core of the genomes carry out the same functions in different viruses, while genes in the ITR region fulfill species-specific functions (Mercer *et al.*, 1995). Representative OPV were cleaved with different restriction enzymes, thus creating a map of cleavage sites to compare across genomes, which suggested that "a central conserved region, terminal, covalent crosslinks and terminal repetitions containing subsets of common sequences are general features of OPV genome structure" (Mackett and Archard, 1979). The syntenic core of poxviruses will be discussed in more detail in Chapter 3.

Table 1.1: Complete Poxvirus Genomes

Genus	Name	Abbreviation	Accession	Host	Size (kb)	Genes	GC (%)	ITR (kb)	Data Set
Entomopox B	Melanoplus sanguinipes EPV	MsEPV	AF063866	grasshopper, locust	236.1	267	18.27	7.2	06/08
Entomopox B	Amsacta moorei EPV	AmEPV	AF250284	moth	232.4	292	17.78	9.5	06/08
Crocodilepox	Crocodilepox	CRV	DQ356948	crocodile	190.1	173	61.92	1.8	06/08
Avipox	Canarypox	AY318871	CNPV	birds	359.9	328	30.37	6.5	06/08
Avipox	Fowlpox isolate Iowa	FWPV-I	AF198100	chicken, turkey	288.5	260	30.89	9.5	06/08
Avipox	Fowlpox isolate Munich	FWPV-M	AJ581527	chicken, turkey	266.1	241	30.76	10.2	06/08
Molluscipox	Molluscum contagiosum virus	MOCV	U60315	human	190.3	163	63.36	4.7	06/08
Parapox	Bovine papular stomatitis virus	BPSV	AY386265	cattle, human	134.4	131	64.50	1.2	06/08
Parapox	Orf virus strain IA82	ORFV-I	AY386263	sheep, goat, human	137.2	130	64.34	3.1	06/08
Parapox	Orf virus strain SA00	ORFV-S	AY386264	sheep, goat, human	140.0	130	63.44	3.9	06/08
Parapox	Orf virus strain NZ2	ORFV-N	DQ184476	sheep, goat, human	137.8	132	64.34	3.4	06/08
Orthopox	Ectromelia	ECTV	AF012825	mouse	209.8	173	33.18	9.4	06/08
Orthopox	Cowpox strain Germany	CPXV-GER	DQ437593	cattle, rodents, human	228.3	219	33.49	7.4	08
Orthopox	Cowpox strain Brighton Red	CPXV-BR	AF482758	cattle, rodents, human	224.5	233	33.40	9.7	06/08
Orthopox	Monkeypox strain USA2003_44	MPXV-USA44	DQ011153	primates	198.8	198	33.08	8.8	06/08
Orthopox	Monkeypox strain USA2003_39	MPXV-USA39	DQ011157	primates	198.8	198	33.08	8.8	06/08
Orthopox	Monkeypox isolate Sierra Leone	MPXV-SIE	AY741551	primates	198.8	177	33.09	8.7	06/08
Orthopox	Monkeypox strain Liberia	MPXV-LIB	DQ011156	primates	200.3	198	33.09	9.4	06/08
Orthopox	Monkeypox strain COP	MPXV-COP	AY753185	primates	199.5	177	33.06	8.8	06/08
Orthopox	Monkeypox strain WRAIR	MPXV-WR	AY603973	primates	199.2	177	33.06	8.7	06/08
Orthopox	Monkeypox strain Zaire1979	MPXV-Z79	DQ011155	primates	197.0	202	33.09	6.4	06/08
Orthopox	Monkeypox strain Congo	MPXV-CON	DQ011154	primates	197.2	200	33.09	6.5	06/08
Orthopox	Monkeypox strain Zaire96	MPXV-Z96	AF380138	primates	196.9	191	33.09	6.4	06/08
Orthopox	Taterapox	TATV	DQ437594	rodents	198.1	225	33.26	4.8	08
Orthopox	Camelpox virus CMS	CMLV-CMS	AY009089	camel	202.2	265	33.20	6.1	06/08
Orthopox	Camelpox virus M96	CMLV-M96	AF438165	camel	205.7	211	33.17	7.7	06/08
Orthopox	Variola strain UK-1952 Butler	VARV-UK52	DQ441447	human	188.3	211	32.69	0.5	08
Orthopox	Variola strain Brazil-1966	VARV-BRZ66	DQ441419	human	188.1	211	32.69	0.5	08
Orthopox	Variola minor	VARV-MIN	Y16780	human	187.0	206	32.71	0.09	06/08
Orthopox	Variola strain Niger-1969	VARV-NIG69	DQ441434	human	186.9	209	32.69	0.1	08
Orthopox	Variola strain Benin-1968	VARV-BEN68	DQ441416	human	187.1	209	32.70	0.2	08
Orthopox	Variola strain Sierra Leone-1969	VARV-SIE69	DQ441437	human	187.0	208	32.69	0.2	08
Orthopox	Variola strain Guinea-1969	VARV-GUI69	DQ441426	human	186.9	207	32.70	0.2	08

Continued on next page ...

Genus	Name	Abbreviation	Accession	Host	Size (kb)	Genes	GC (%)	ITR (kb)	Data Set
Orthopox	Variola strain Korea-1947	VARV-KOR47	DQ441432	human	186.4	202	32.72	0.7	08
Orthopox	Variola strain Japan-1946	VARV-JAP46	DQ441429	human	186.7	202	32.71	0.5	08
Orthopox	Variola strain India-1953-ndel	VARV-IND53-nd	DQ441428	human	186.7	202	32.71	0.5	08
Orthopox	Variola strain Japan-1951-stwl	VARV-JAP51-stwl	DQ441431	human	186.1	202	32.71	0.5	08
Orthopox	Variola strain Japan-1951-hrpr	VARV-JAP51-hrpr	DQ441430	human	186.2	201	32.71	0.5	08
Orthopox	Variola strain China-1948	VARV-CHN48	DQ437582	human	186.7	205	32.71	0.7	08
Orthopox	Variola strain UK-1946-hrvy	VARV-UK46-hrvy	DQ441444	human	185.8	202	32.70	0.5	08
Orthopox	Variola strain UK-1946-hndn	VARV-UK46-hndn	DQ441445	human	186.1	199	32.70	0.5	08
Orthopox	Variola strain Sumatra-1970-222	VARV-SUM70-222	DQ437591	human	185.4	202	32.70	0.2	08
Orthopox	Variola strain Sumatra-1970-228	VARV-SUM70-228	DQ441442	human	185.4	200	32.70	0.2	08
Orthopox	Variola strain India-1964-vel-7124	VARV-IND64-vel4	DQ437585	human	186.7	205	32.71	0.7	08
Orthopox	Variola strain India-1964-vel-7125	VARV-IND64-vel5	DQ437586	human	186.1	203	32.71	0.6	08
Orthopox	Variola strain Afghanistan-1970	VARV-AFG70	DQ437580	human	185.9	203	32.71	0.5	08
Orthopox	Variola strain Kuwait-1967	VARV-KUW67	DQ441433	human	185.9	200	32.70	0.5	08
Orthopox	Variola strain Yugoslavia-1972	VARV-YUG72	DQ441448	human	185.9	201	32.70	0.5	08
Orthopox	Variola strain Syria-1972	VARV-SYR72	DQ437592	human	185.9	203	32.71	0.5	08
Orthopox	Variola strain Iran-1972	VARV-IRN72	DQ437587	human	185.8	203	32.71	0.5	08
Orthopox	Variola strain Pakistan-1969	VARV-PAK69	DQ437589	human	185.9	203	32.71	0.5	08
Orthopox	Variola strain India-1967	VARV-IND67	X69198	human	185.6	197	32.73	0.06	06/08
Orthopox	Variola strain India-1953-mad	VARV-IND53-mad	DQ441427	human	186.1	201	32.70	0.5	08
Orthopox	Variola strain Germany-1958	VARV-GER58	DQ437584	human	184.9	201	32.71	0	08
Orthopox	Variola strain UK-1947	VARV-UK47	DQ441446	human	185.0	201	32.71	0	08
Orthopox	Variola strain Nepal-1973	VARV-NEP73	DQ437588	human	185.7	203	32.71	0.5	08
Orthopox	Variola strain Bangladesh-1974-nur	VARV-BAN74-nur	DQ441420	human	186.3	198	32.70	0.7	08
Orthopox	Variola strain Bangladesh-1975	VARV-BAN75	L22579	human	186.1	189	32.69	0.7	06/08
Orthopox	Variola strain Bangladesh-1975-banu	VARV-BAN75-banu	DQ437581	human	186.0	201	32.70	0.7	08
Orthopox	Variola strain Bangladesh-1974-shz	VARV-BAN74-shz	DQ441421	human	186.3	198	32.70	0.7	08
Orthopox	Variola strain Bangladesh-1974-sol	VARV-BAN74-sol	DQ441422	human	186.3	198	32.70	0.7	08
Orthopox	Variola strain Tanzania-1965	VARV-Tan65	DQ441443	human	185.8	201	32.70	0.3	08
Orthopox	Variola strain Congo-1970-46	VARV-CON70-46	DQ437583	human	186.6	202	32.68	0.7	08
Orthopox	Variola strain Congo-1970-227	VARV-CON70-227	DQ441423	human	186.7	201	32.69	0.7	08
Orthopox	Variola strain South Africa-1965-102	VARV-SA65-102	DQ441435	human	186.1	201	32.69	0.5	08
Orthopox	Variola strain South Africa-1965-103	VARV-SA65-103	DQ441436	human	185.9	201	32.69	0.5	08
Orthopox	Variola strain Botswana-1972	VARV-BOT72	DQ441417	human	185.9	202	32.68	0.5	08
Orthopox	Variola strain Botswana-1973	VARV-BOT73	DQ441418	human	185.9	202	32.69	0.5	08

Continued on next page ...

Genus	Name	Abbreviation	Accession	Host	Size (kb)	Genes	GC (%)	ITR (kb)	Data Set
Orthopox	Variola strain Sudan-1947-juba	VARV-SUD47-j	DQ441440	human	186.3	200	32.69	1.0	08
Orthopox	Variola strain Sudan-1947-rumbec	VARV-SUD47-r	DQ441441	human	186.4	201	32.70	1.1	08
Orthopox	Variola strain Somalia-1977-1252	VARV-SOM77-1252	DQ441438	human	184.2	200	32.69	0	08
Orthopox	Variola strain Somalia-1977-1605	VARV-SOM77-1605	DQ441439	human	184.2	200	32.69	0	08
Orthopox	Variola strain Somalia-1977-2479	VARV-SOM77-2479	DQ437590	human	186.2	202	32.67	1.0	08
Orthopox	Variola strain Ethiopia-1972-16	VARV-ETH72-16	DQ441424	human	186.6	203	32.68	1.2	08
Orthopox	Variola strain Ethiopia-1972-17	VARV-ETH72-17	DQ441425	human	186.6	201	32.68	1.2	08
Orthopox	Cowpox strain GRI-90	CPXV-GRI	X94355	cattle, rodents, human	223.7	212	33.65	8.3	06/08
Orthopox	Horsepox	HPXV	DQ792504	horse	212.6	236	33.09	7.5	08
Orthopox	Rabbitpox	RPXV	AY484669	rabbit	197.7	184	33.48	10.0	06/08
Orthopox	Vaccinia MVA	VACV-MVA	AY603355	human, cattle	166.7	193	33.48	4.0	06/08
Orthopox	Vaccinia strain DUKE	VACV-DUKE	DQ439815	human, cattle	200.0	225	33.40	12.5	08
Orthopox	Vaccinia strain 3737	VACV-3737	DQ377945	human, cattle	199.4	216	33.31	14.2	06/08
Orthopox	Vaccinia strain Acambis clone 3	VACV-AC3	AY313848	human, cattle	197.4	229	33.47	10.2	06/08
Orthopox	Vaccinia strain Acambis clone 2000	VARCV-AC2000	AY313847	human, cattle	199.2	235	33.43	16.4	06/08
Orthopox	Vaccinia WR	VACV-WR	AY243312	human, cattle	194.7	218	33.34	10.2	06/08
Orthopox	Vaccinia Tian Tan	VACV-TAN	AF095689	human, cattle	189.3	243	33.21	9.6	06/08
Orthopox	Vaccinia Copenhagen	VACV-COP	M35027	human, cattle	191.7	262	33.40	12.1	06/08
Orthopox	Vaccinia strain Lister clone	VACV-LIS	DQ121394	human, cattle	189.4	196	33.72	6.4	08
Orthopox	Vaccinia strain LC16mO	VACV-LCO	AY678277	human, cattle	189.2	282	33.71	6.3	06/08
Orthopox	Vaccinia strain LC16m8	VACV-LC8	AY678275	human, cattle	189.2	282	33.71	6.3	06/08
Orthopox	Vaccinia strain Lister	VACV-L	AY678276	human, cattle	189.1	786 *	33.76	6.3	06
Yatapox	Yaba monkey tumor virus	YMTV	AY386371	primates	134.7	140	29.83	2.0	06/08
Yatapox	Yaba-like disease virus	YLDV	AJ293568	primates	144.6	151	27.00	1.9	06/08
Yatapox	Tanapox virus isolate Kenya	TPV-KEN	EF420156	primates	144.6	156	26.99	1.9	08
Yatapox	Tanapox virus isolate RoC	TPV-Roc	EF420157	primates	144.6	155	26.99	1.9	08
Leporipox	Rabbit fibroma virus	RFV	AF170722	rabbit	159.9	165	39.53	12.4	06/08
Leporipox	Myxoma virus strain Lausanne	MYXV-LAU	AF170726	rabbit	161.8	170	43.56	11.6	06/08
Leporipox	Myxoma virus strain 6918	MYXV-6918	EU552530	rabbit	161.8	171	43.53	11.6	08
Deerpox	Deerpox isolate W-848-83	DPV-W83	AY689436	deer	166.3	169	26.16	5.0	06/08
Deerpox	Deerpox isolate W-1170-84	DPV-W84	AY689437	deer	170.6	170	27.03	7.1	06/08
Capripox	Sheeppox virus	SPPV	NC_004002	sheep, goat	150.0	148	25.01	2.2	06/08
Capripox	Sheeppox virus A	SPPV-A	AY077833	sheep, goat	150.1	444 *	25.00	2.3	06
Capripox	Sheeppox virus NISKHI	SPPV-N	AY077834	sheep, goat	149.7	444 *	25.00	2.1	06
Capripox	Goatpox virus Pellor	GTPV-PEL	NC_004003	sheep, goat	149.6	150	25.31	2.1	06/08

Continued on next page ...

Genus	Name	Abbreviation	Accession	Host	Size (kb)	Genes	GC (%)	ITR (kb)	Data Set
Capripox	Goatpox virus G20-LKV	GTPV-LKV	AY077836	sheep, goat	149.7	456 *	25.31	2.2	06
Capripox	† LSDV isolate LW 1959	LSDV-59	AF409138	cattle	150.5	159	25.92	2.4	06/08
Capripox	† LSDV isolate NW-LW	LSDV-LW	AF409137	cattle	150.8	158	25.91	2.3	06/08
Capripox	† LSDV isolate NI-2490	LSDV-2490	AF325528	cattle	150.8	156	25.91	2.4	06/08
Suipox	Swinepox	SWPV	AF410153	swine	146.5	150	27.40	3.7	06/08

* ORFs annotated using *getorf* program
† Lumpy skin disease virus

1.4 Poxvirus Evolution

1.4.1 The Origin of Viruses

Most research done on the origin of viruses can be categorised into one of three conflicting theories, which state that they were either precursors to cellular systems (the virus first hypothesis) or represent breakaway elements from cellular genomes (the escape hypothesis) or are progressively simplified unicellular organisms (the reduction hypothesis). The original virus first hypothesis, suggesting that viruses were directly ancestral to cells (d'Herelle, 1924), is no longer acceptable given that viruses require an intracellular development stage for their reproduction and that it is unlikely for free molecules to have evolved to produce an ancestral ribosome for the production of proteins.

Bamford *et al.* (2002) proposed to use similarities of structural features, what they called the "self" of the virus, to imply relatedness between virus groups, and used this approach to suggest that the ancestral viruses were present before the division of cellular life into its current three domains. As viral organisms outnumber host organisms by at least one order of magnitude, these authors depicted viruses as being the main source of DNA in higher eukaryotes, crucial to the origin and maintenance of all forms of life. Advances in imaging techniques now allow researchers a more sophisticated search for structural similarities. Khayat *et al.* (2005), for example, investigated the crystal structure of the major capsid protein of the virus infecting the archaeal *Sulfolobus* species, and found that it is very similar to that of *Paramecium* virus and phage PRD1. This connection between viruses infecting all domains of life suggests that at least some lineages originated before the separation of archaea, bacteria and eukarya. Given the extreme niches that archaea occupy, and the incompatibility between the archaeal and bacterial transcription systems, however, a huge

number of adaptive changes need to have taken place since the hypothetical common ancestor of these viruses (Prangishvili *et al.*, 2001). Some authors suggest that a new viral classification should incorporate these shared structural characteristics between viruses infecting all three domains to a greater extent (Krupovic and Bamford, 2008).

Koonin *et al.* (2006) pointed out some gaps in the work by Bamford *et al.* (2002), in the connection between viral and cellular evolution. Instead, they proposed a scenario of viral evolution based on the fact that many key genes for viral replication and morphogenesis are shared by all viruses and no cellular life forms, despite many virus groups not sharing a conventional common origin. They postulated an Ancient Virus World, a contingent of viral genes that retains its identity throughout evolution. The principal virus lineages are thus thought to have emerged from the primordial pool: RNA viruses, followed by retroid elements and DNA viruses. The viruses developed prior to typical cells and evolved into intracellular parasites later. The emergence of eukaryotic cells provided the second melting pot of virus evolution where major groups of eukaryotic viruses emerged after recombination of genes from bacteriophages, archaeal viruses, plasmids and eukaryotes. A general trend towards increased complexity is noticable in these elements.

The discovery and sequencing of the giant amoeba-infecting mimivirus, with 1.2 megabases and 911 genes (Raoult *et al.*, 2004), gave new information to the arguments around viral origins, as its genome is larger than those of some cellular organisms, making the reduction hypothesis less intuitively likely: Moreira and Brochier-Armanet (2008), working on mimivirus, and Filee *et al.* (2008) on several large DNA viruses, showed that these viruses gain a large proportion of their genetic content from hosts or pathogens co-infecting their hosts, indicating viral gene accretion rather than reduction.

The different strategies of extant viral lineages may indicate a poly-

phyletic origin and further complicate the issue; DNA and RNA viruses represent complexity versus simplicity; relative independence from versus heavy reliance on the host; slow rates of evolution (10^{-7} to 10^{-9} nt subs/site/year) and a tendency to codiverge with their hosts versus rapid evolution (10^{-3} to 10^{-4} nt subs/site/year) and frequent jumping of species boundaries (Holmes, 2008). But all viruses do share some overall similarities like the structure of a protein coat enclosing some nucleoprotein filament (Forterre, 2006).

It appears that the origin and evolution of viruses is unlikely to be described by one simple sequence of events but is intertwined with that of cellular life forms, with which viruses interchange genomic material to the present day. Perhaps, viruses even provided a shuttle system between early cellular organisms, transporting genes between early cells in divisible lipid vessels (Burnett, 2006).

1.4.2 Poxvirus-Specific Evolution

Regardless of how viruses may have originated, less controversial observations can be made about the relationship of poxviruses to other, similar virus groups. Poxviruses are part of a monophyletic group of Nucleo-Cytoplasmic Large DNA viruses (NCLDV, Iyer *et al.* (2001)), which also include asfarviruses, iridoviruses, phycodnavirus and mimivirus. Using sensitive homology searches, the authors found that conserved families in all NCLDV point to the presence of several features in theses viruses, for example, an elaborate system for replication and expression, capsid and membrane virion proteins, and an electron transfer system.

Horizontal gene transfer (HGT) is common between viruses and their hosts, and genes acquired from the host often confound virus phylogenies. As poxviruses operate exclusively in the cytoplasm of the host cell and thus provide their own replication machinery, it is likely that they have acquired

parts of this from the host (Iyer *et al.*, 2006). Many genes involved in the poxvirus' manipulation of the host's immune system are thought to be derived from the host, too (Bugert and Darai, 2000; Hughes, 2002; Lucas and McFadden, 2004), also described in Section 1.2.6.

Filee *et al.* (2008) also presented evidence, based on homology searches, for extensive lateral acquisition of cellular genes by NCLDV. Previously, NCLDV were thought to have evolved through simplification from cellular organisms along with accretion of genes of bacterial origin. Some core genes thought to be ancestral now appear to have been acquired by HGT, suggesting that the ancestral genome may be a lot smaller than anticipated (30-40 genes). The larger the genome, the more horizontally transferred (HT) genes it contains, which is incompatible with genome reduction from a cellular ancestor. Instead, the authors suggested that NCLDV evolved by growth from a simple DNA virus through HGT. Interestingly, it was shown that most HT genes are located near the tips of the genome. The relevance of genomic location in the search for HT genes is revisited in Chapter 3.

Poxviruses are thought to co-evolve with their host to some extent (Iyer *et al.*, 2006). A poignant example of poxvirus-host co-evolution is provided by the release of myxoma virus (MYXV) in Australia, as recently retold by Morales *et al.* (2009). South American MYXV and Californian MYXV each cause benign cutaneous fibroma at the site of inoculation in their native hosts *Sylvilagus brasiliensis* and *S.bachmani*, respectively. Both cause the lethal form of infection, myxomatosis, in European rabbit *Oryctolagus cuniculus*, characterised by skin lesions, systemic immunosuppression and high mortality. When strains of MYXV were released in Australia and France in the early 1950s, they quickly became endemic in Europe and led to the co-evolution of attenuated viral strains and MYXV-resistant rabbits. Strains isolated in Europe 43 years apart are still 99.95% identical, indicating the extent of evolutionary constraints imposed by the co-evolution.

Unlike in RNA viruses, evolutionary change in DNA viruses cannot be observed in real time, due to much lower rates of nucleotide substitution.

1.4.3 Recombination in Poxviruses

Apart from recombination playing a part in priming DNA replication and repairing DNA breaks, there is also evidence for inter-molecular recombination leading to sequence rearrangements between poxviruses: Ball (1987) noted high-frequency homologous recombination in vaccinia, when thymidine kinase (TK) is inserted between direct repeats (1.5 kb). Under non-selective TK conditions, the inserted gene was often excised as the repeats were targets for both inter- and intra-molecular recombination.

Recombination requires poxviral DNA polymerase and probably occurs via a single-strand annealing mechanism. Yao and Evans (2001) find that recombination of linear molecules in vaccinia virus-infected cells requires only a very small region of sequence homology. In a study of recombination between plasmids containing the VACV TK gene and a TK-deficient virus, Spyropoulos *et al.* (1988) found recombination products derived from one or more rounds of recombination. While these experiments involved transfected DNA it is assumed that the results are applicable to viral genomes.

As reviewed by Bennink and Yewdell (1990), vaccinia is often used as a vector for introducing extrinsic gene products from a variety of organisms into a model, to determine the specificity of T-lymphocytes and studying the processing and presentation of T-lymphocytes. This indicates how readily vaccinia virus recombinants are created in the lab. Similarly, many clinical gene therapy trials are now underway that use poxvirus vectors, as reviewed by Moroziewicz and Kaufman (2005), most of which deal with different forms of cancer. Another reason why poxviruses are a favoured gene therapy vector is their location in the cytoplasm where they will not disrupt normal gene regulation or risk long-term genetic instability like other viruses.

Recombination as a means to drive genome evolution has been detected in the extremities of the coding regions of Variola strains (Esposito *et al.*, 2006), and some strains of avipox virus in the Galapagos islands are thought to have derived through recombination of other strains (Thiel *et al.*, 2005). Malignant rabbit fibroma virus has been shown to be a recombinant between 10 kb SFV sequence and 90% MYXV sequence (Upton *et al.*, 1988). Physical restriction maps of four isolates of capripoxviruses indicate that type 4 isolates (represented by Yemen goat poxvirus) are the result of recombination between type 2 (Iraq goat poxvirus) and type 3 isolates (Kenya cattle poxvirus, Gershon *et al.* (1989)).

The low level of homology required for recombination also explains how replicating poxviruses might acquire host DNA via horizontal transfer, as host cDNAs often contain a poly(dA)-poly(dT) end homologous to the sequences of poxviral promoters and could thus provide the homology needed for recombination with a broken virus genome (Yao and Evans, 2001).

1.4.4 Role of Poxviruses in Nucleus Formation

Takemura (2001) speculated as to the role of poxviruses in the origin of the eukaryotic nucleus. He looked at the spread of different types of eukaryotic polymerases. δ-polymerase is important in archaea and present in some viruses, like adeno- and herpesviruses, while α-polymerase is expressed by vaccinia and variola and is similar to eukaryotic molecules. The author suggested that eukaryotic α-pol is derived from pox-like organisms which had some eukaryote-like characteristics. This is supported by poxviruses being independent from the host nucleus, which means they could have infected non-nucleated cells like ancestral eukaryotes. This hypothesis for the origin of the eukaryotic nucleus states that a symbiotic contact of an OPV ancestor with an archaebacterium that already contained a δ-like polymerase led to the first kind of eukaryote. These claims are supported by the observation that δ-

pol can initiate chain elongation and can substitute for α-pol in vitro, showing that the ancestor of δ-pol could participate in DNA replication initiation without any α-like polymerase. Mitochondria use nuclear γ-polymerase, so their own polymerase, used by the free living ancestor to the symbiont, must have at some point migrated to the nucleus, and a similar situation might explain the origin of the α-pol. An alternative explanation has adeno- and herpesviruses acquiring the eukaryotic δ-pol through HGT.

Independently, Bell (2001) proposed a similar idea the same year, also advocating a complex DNA virus as the ancestor of the nucleus. Several nuclear characteristics are thought to be derived from the viral ancestor, including mRNA capping, linear chromosomes and the separation of transcription from translation. A viral origin of the nucleus would also explain the existence of nuclear pores, as viruses on their own need a way of transporting nucleic acid through their membranes. The nuclear pores would have had to develop independently, if the alternative theory of an archaeal ancestor to the nucleus is true (e.g. proposed by Martin and Muller (1998)).

1.4.5 Smallpox Evolution and HIV Resistance

The smallpox virus was implicated in an evolutionary detective story (Galvani and Slatkin, 2003) when it emerged that variola, instead of the previously championed bubonic plague, may have provided the selective pressure necessary for the observed high frequency of the CCR5-Δ32 allele, which coincidentally provides cells with HIV resistance. This allele prevents the expression of a chemokine receptor on the cell surface, which is exploited in viral entry into the cell, thus conferring resistance to HIV-1. Recent origin and European distribution imply it has been selected for, but HIV has not been endemic long enough to provide this kind of driving force. Rather, variola, with its weak but continuous and direct transmission and tendency to kill hosts before reproductive age, appears to have been the predominant

selective pressure. This implies that the CCR5-Δ32 deletion confers full protection against smallpox mortality, which is biologically plausible due to the similar nature of cell entry by HIV and poxviruses (Galvani and Slatkin, 2003).

Subsequently, however, Hummel *et al.* (2005) found the CCR5-Δ32 allele in 2900 year old Bronze Age skeletons and also note that it was not found less frequently among plague victims. This means that the mutation is actually older than previously thought. Sabeti *et al.* (2005) confirm a prehistoric origin of the deletion and, in addition, find that the pattern of variation at this locus is in fact consistent with neutral evolution. These results imply that variola may, in fact, not have provided the selective pressure leading to present-day HIV resistance.

1.5 Poxvirus Bioinformatics

The classical approach to poxvirus genetics saw researchers isolating and characterising genes based on their phenotype, where their presence or absence produced conditionally lethal or drug-resistant mutants, as reviewed by Condit and Niles (1990). Today, bioinformatics methods are increasingly used to discover a gene's functions and evolution. Molecular evolution is the study of how genes and other genetic elements have changed at the sequence level. Because of the vast amount of data now at the researcher's disposal, computational work is useful where traditional lab work has become unfeasible. Virus genomes even provided the foundation of comparative genomics, as their small size and rapid generation turnover provided an attractive model for early genomics studies (Koonin and Dolja, 2006). Comparative genomics studies of different viral families continue to yield important insights into evolutionary mechanisms, clarify the evolution of virus families and shed light on complex host-virus adaptations (McGeoch and Davison, 1999; Hughes and Friedman, 2000; Herniou *et al.*, 2001; Wang

et al., 2007; Charrel *et al.*, 2008; Liu *et al.*, 2009). Work presented in this thesis aims to elucidate different aspects of poxvirus genome evolution through computational analysis.

Multiple sequence comparisons, for example, allow for inferring phylogenetic trees, and although poxvirus trees have been reconstructed before (McLysaght *et al.*, 2003; Gubser *et al.*, 2004; Hughes and Friedman, 2005; Xing *et al.*, 2006), those presented here are based on newer data and include more genomes than any others. A protein family classification of poxviruses, which is stricter and therefore more reliable than those presented elsewhere (e.g.,Upton *et al.* (2003)) is attempted, also. Searches for horizontal gene transfer (HGT) have been sporadically carried out for various groups of viruses, including poxviruses (Hughes, 2002), but in Chapter 3, a method of distinguishing between single and multiple HGT events from host to poxvirus is introduced, based on stricter requirements and a new method that takes into account gene order. The analysis of phyletic patterns is made possible through large numbers of genome sequences and the one described in Chapter 4 is the first comprehensive study carried out on all poxvirus genomes.

Poxviruses are interesting organisms to study not just because of the potential benefit researchers can gain from understanding virus evolutionary history, but also because, due to their cytoplasmic location and similarity to eukaryotic systems as discussed above, they can serve as models for eukaryotic transcription, mRNA processing, DNA replication and recombination.

Chapter 2

Materials and Methods

This chapter introduces the software used for the research presented in this thesis. It also covers the analyses that led to a poxvirus family classification and inferred phylogenetic tree, as these results are used in all subsequent chapters and can therefore be considered materials.

2.1 Software

All results and analyses carried out in this thesis were organised in a MySQL database system. All programs written for this thesis were written in Perl 5.

Homology Searches BLAST, the Basic Local Alignment Search Tool (Altschul *et al.*, 1990) carries out rapid sequence comparisons. One of the output parameters is the E-value (for EXPECT), which gives the expected number of chance matches in a database of a given size. The default value, ten, means that ten matches are expected to be found by chance, and therefore do not necessarily imply evolutionarily significant homology. As E-values approach zero they are equivalent to p-values of the significance of the sequence similarity. An E-value threshold is set by the user when searching for orthologues of a sequence in a database, with a lower threshold defining a more stringent search. If the E-value ascribed to a match is greater than

the *E*-value threshold, the match will not be reported. Reported matches will be referred to as BLAST hits below.

In PSI-BLAST, a newer database search algorithm by the same group (Altschul *et al.*, 1997), iterative searches are run on the same sequence. Residues that are conserved in sequences deemed to be genuine, or meaningful, homologues by the user have a higher impact on subsequent iterations, as each search is based on the profile constructed from the previous search. This profile is called a position-specific scoring matrix (PSSM). This type of homology search is useful for detecting distant but potentially biologically significant similarities, for example in the search for new members of a breast cancer susceptibility gene family (Altschul *et al.*, 1997) or when identifying host homologues to poxvirus genes as described in Chapter 3, where homologues from a donor genome may not always be detectable by regular BLAST.

Sequence Alignments Wherever sequences were aligned in this thesis, the MUSCLE algorithm was used (Edgar, 2004). Its statistical accuracy is the same as that of more established program T_Coffee, written by Notredame *et al.* (2000), but it is faster and more reliable for large numbers of sequences and large sequences, which suits the viral sequences used here.

Phylogenetic Packages Different packages for inferring phylogenetic trees are used to cover the three popular approaches: neighbour-joining (NJ), maximum parsimony (MP) and maximum likelihood (ML). They are used here as described in the relevant sections below.

Clustal W (Thompson *et al.*, 1994) is a widely-used program for dealing with multiple sequence alignments that can construct NJ trees.

Phylip (Felsenstein, 2005) is a package of programs for inferring phylogenies using many different methods. In this thesis, the main program used is *protpars*, which construct phylogenetic trees from protein sequences based

on MP methods.

PhyML was introduced by Guindon and Gascuel (2003) as a newer, fast and reliable algorithm using the maximum likelihood principle. Its core is based on a hill-climbing algorithm that adjusts tree topology and branch lengths simultaneously. An initial tree is built from the given data using distance-based methods and each iteration of the algorithm then improves this estimate.

Other Software The Markov Cluster Algorithm (MCL), written by van Dongen (2000, http://www.micans.org/mcl/), was used to define protein families as described in Section 2.3 below. It finds a cluster structure in graphs by a mathematical bootstrapping procedure, details of which are beyond the scope of this methods section. The graph analysed here is a representation of all poxvirus proteins used and all the BLAST hits they find in the database.

2.2 Data

All available poxvirus proteins were downloaded from GenBank in June 2006 and proteins were predicted for unannotated complete genomes using the EMBOSS *getorf* application (Rice *et al.*, 2000), yielding a total of 17,185 protein sequences from 54 completely sequenced genomes and many other partial genome sequences. In April 2008 it became apparent that many newly added poxvirus sequences would provide a lot more information for these analyses and thus Chapters 4 and 5 are based on this later set of 24,514 virus sequences from 104 completely sequenced genomes and many partial genome sequences, which do not include any annotated by *getorf*. Four complete genomes were removed from the 2008 set as the annotation of their ORFs by *getorf* was found to be too inclusive and therefore unsuitable. The family classification discussed in Section 2.3 below was repeated with

the 2008 set. For a summary of which genomes were included in these studies and how many proteins are encoded by each, see Table 1.1 on page 25, which also shows a column marking the genomes belonging to the 2006/2008 data sets.

2.3 Family Classification

The proteins were then compared in a local all-against-all BLASTp search with an *E*-value threshold of 1, and options -b 1500, -v 1500. The latter two options determine the number of hits to show alignments for, and the number of hits to show one-line descriptions for, respectively. Using values larger than the defaults of 250 and 500 was necessary here to avoid missing hits for sequences with many orthologues in the database, which was the case for over 4000 poxviral proteins.

16,148 of the 2006 data set and 24,326 of the 2008 set form at least one BLAST hit pair; 15,167 and 24,054, respectively, display homology to more than one other poxvirus protein with an *E*-value of 1 or less. This result set is too large and coarse to be used for family classification without further refinement: the 2006 set consists of 677 single-linkage groups made up of 1.2 million BLAST hits, the 2008 set consists of 147 groups of hits made up of 4.0 million BLAST hits. In order to make these homology search results more meaningful, instead of just applying a stricter *E*-value threshold, a clustering algorithm, MCL, was used to determine the most robust homology connections in this large result set. The BLAST results are represented as a graph, with proteins (nodes) being connected by BLAST hits (edges). A natural cluster within this graph is characterised by the presence of many edges between its members. It is our aim to find these natural clusters as they are likely to correspond to real protein families, having many BLAST hits connecting their protein members. MCL simulates stochastic flow along the edges, by computing random walks through the graph between two arbitrary

nodes. It is expected that these random walks are unlikely to go from one natural cluster to another. Two parameters called inflation and centering are used to control the granularity of the clusters that are reported by the algorithm, i.e., how inclusive we want our protein families to be. A matrix of scores between each two homologous poxvirus proteins that have reciprocal BLAST hits was created, for use in the *mcxload* application of MCL. The score used was the product of the percent hit of the two proteins and the normalised log_{10} of the E-value with which they hit each other. As the reciprocal hits often do not have the same E-value and percent hit, an average between the two is used for the score. The application uses this matrix to build a graph from the proteins and their homology scores for input into the main *mcl* algorithm. Sixty-four combinations of the inflation and centering parameters were examined, each taking the values 1.2, 1.4, 1.6, 1.8, 2.0, 2.2, 2.4 and 2.6. These parameters control cluster granularity, usually in a range of 1.2 to 5.0, where higher numbers result in more fine-grained clusters. Their settings were chosen to maximise certain family counts as described below. For an illustration of how different parameter settings influence resulting family sets, see Figure 2.1. Three sets were chosen for analyses of the 2008 set:

Set 1 This setting includes the fewest number of families overall but, at 28, the largest number of core families. Core families are defined here as those that include one (and only one) protein member from each completely sequenced genome and they are listed in Table 2.1. The inflation and centering parameters are both at 1.2 in this setting. This is also the set in which most poxvirus proteins are classified into families, i.e., fewest singletons (583). While this set with its many core families is most useful for determining the poxvirus phylogeny and assigning most proteins to a family, it is expected that some large families could reasonably be split into two or more in a less inclusive setting. This is exemplified in one (designated

Family 1 in Set 1) with 1,132 members, which are distributed among 44 and 54 families in Sets 2 and 3, respectively.

There are 437 families in this set. In the 2006 set, this same setting was used for all analyses and grouped proteins into 33 core families, also listed in Table 2.1.

Set 2 This set of families includes the largest number of conserved single copy families at 59. A single copy family is defined as one that does not include any paralogues from any genome. It is considered conserved if at least one family member is an EPV protein and at least one other a ChPV protein, thus avoiding very small genus-specific families. Inflation is 2.2 and centering 1.2 in this setting. There are 650 families in this set.

Set 3 This set of families is the one which, at 574, includes the largest number of families with more than three members. This minimises the number of uninformative mini-families. Inflation is 2.6 and centering 2.4 in this setting. There are 705 families in this set.

2.3.1 Comparison with Another Set

A set of poxvirus orthologous clusters that is frequently referred to was first defined by Upton *et al.* (2003). They find a set of 49 gene families conserved in all 21 genomes sequenced at the time, with 41 additional families conserved in all ChPV. These were compared to the 2006 data set described above, which included 33 core and a further 34 ChPV families. Upton's set allowed for the presence of paralogues in some families, which our set does not, but this only accounts for four families excluded from our set.

The only genus not represented in Upton's genome set is parapox and the families missing in our set are always also missing a member from another genome included in Upton's analysis, so that the added 33 complete genomes do not explain the discrepancy. Instead, our family definitions are simply

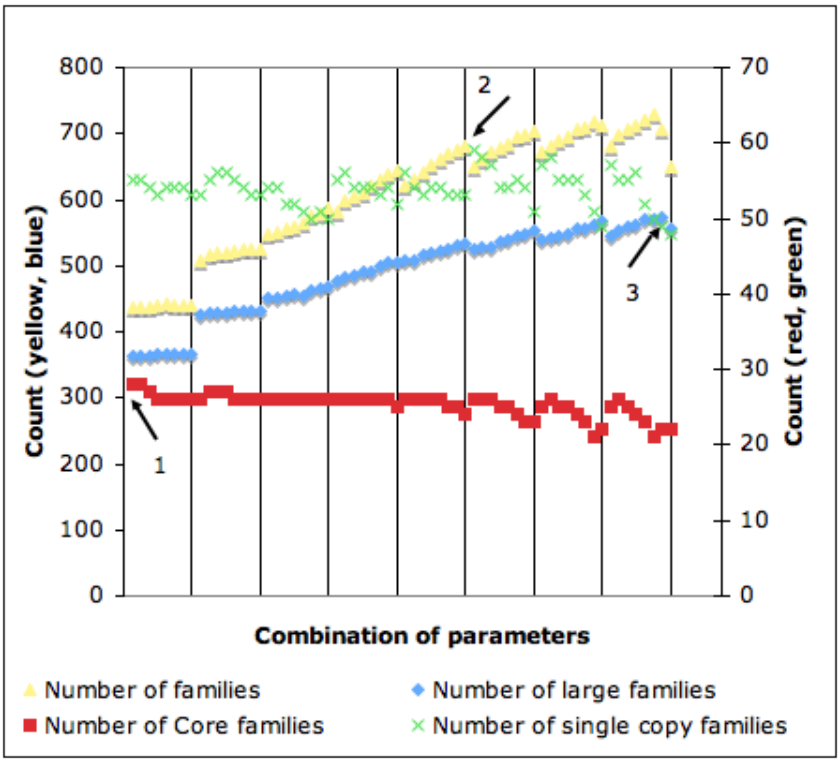


Figure 2.1: MCL Family Output. The numbers of different types of protein family clustered by MCL at various combinations of the inflation and centering parameters are shown in four series. The vertical lines separate sets of increasing inflation parameters from 1.2 to 2.6, in steps of 0.2. Within each vertical segment, the centering parameter increases, by 0.2, from 1.2 to 2.6. The overall number of families formed by one setting (yellow series) and the number of families formed with more than three members (blue series) are scored on the primary y-axis on the left, while the number of core families formed (red series) and the number of single-copy families formed (green series) are scored against the secondary y-axis on the right. The arrows indicate which parameter combination was chosen for Sets 1, 2 and 3, respectively, as described in the main text.

stricter than those of Upton *et al.* (2003), who manually edit their families by adding orthologues with an *E*-value above their threshold on the basis of domain conservation. As our family definition process is automated, no further members are manually included.

2.4 Phylogeny Reconstruction

2.4.1 The 2006 Set

The MUSCLE algorithm (Edgar, 2004) was used to generate protein sequence alignments of the 33 core families from the 2006 data set, and the alignments were concatenated.

A concatenated alignment was used to construct a tree, rather than building a consensus tree, which is another approach to inferring species phylogenies from a set of genes. Gadagkar *et al.* (2005) compared these methods for a set of 66 mammalian species and measured the accuracy of the inferred NJ trees by the percentage of clades that were predicted correctly. They did not correct for gene-specific differences in substitution patterns for the concatenated sets and find that despite this lack of specificity, every gene family that is added to the set improves the accuracy of the inferred tree. While adding more genes also improved consensus tree accuracy, concatenation performed consistently better. The authors recommend calculating a gene support frequency along with the bootstrap support of concatenated gene sets, as concatenation can lead to amplification of weak phylogenetic signals. In the phylogenetic tree constructed here and shown in Figure 2.2 this advice has been followed and both gene support frequency and bootstrap support is given. Degnan and Rosenberg (2006) showed that for any species tree with more than five species, gene tree discordance is so common that using a consensus tree as the species tree is a statistically inconsistent way of finding the true species phylogeny. A concatenation

Vaccinia Gene	Putative Function	Data Set
A2L	Transcription factor VLTF-3	2006/2008
A3L	Viral core protein P4b	2006/2008
A5R	DNA-dependent RNA polymerase subunit RPO19	2008
A7L	Subunit of transcription factor VETF (82 kDa)	2006/2008
A10L	Viral core protein precursor P4a	2006/2008
A11R	Protein involved in morphogenesis	2006/2008
A16L	Myristylated membrane protein	2006/2008
A18R	DNA helicase	2006/2008
A22R	Holliday junction endonuclease / palmitylprotein	2006
A23R	Subunit of transcription factor VITF-3 (45 kDa)	2006
A24R	DNA-dependent RNA polymerase subunit RPO132	2006/2008
A29L	DNA-dependent RNA polymerase subunit RPO35	2008
A32L	ATPase / DNA packaging protein	2006/2008
D5R	NTPase	2006/2008
D6R	Subunit of transcription factor VETF (70 kDa)	2006/2008
D11L	ATPase / transcription termination factor	2006
D12L	Subunit of mRNA capping enzyme	2006/2008
E1L	Subunit of poly-A polymerase	2006/2008
E6R	Unknown IMV protein	2006/2008
E9L	DNA polymerase	2006/2008
E10R	Sulfhydryl oxidase	2006/2008
F9L	Membrane protein / S-S bond formation pathway protein	2006
F10L	Serine/threonine kinase	2006/2008
G5R	Protein involved in membrane formation	2006/2008
G7L	Structural virion core protein	2006
G9R	Myristylated membrane protein	2006/2008
H2R	Unknown IMV protein	2006/2008
H6R	DNA topoisomerase type IB	2006/2008
I7L	Viral core cysteine proteinase	2006/2008
I8R	RNA helicase	2006/2008
J3R	Subunit of poly-A polymerase / cap methyltransferase	2006
J5L	IMV membrane protein	2006/2008
J6R	DNA-dependent RNA polymerase subunit RPO147	2006/2008
L1R	IMV envelope protein	2006
L3L	Unknown IMV protein	2006/2008

Table 2.1: Core Families from Set 1

approach to infer species phylogeny from multiple gene loci has been used in poxviruses (McLysaght *et al.*, 2003; Gubser *et al.*, 2004; Hughes and Friedman, 2005; Xing *et al.*, 2006), other viruses (Wang *et al.*, 2007; Herniou *et al.*, 2001), bacteria (Susko *et al.*, 2006) and eukaryotes (Phillips *et al.*, 2004).

Based on this concatenated sequence alignment, phylogenetic trees of 54 completely sequenced poxviruses (Table 1.1) were reconstructed by different methods: NJ (*protdist* from Phylip package and Clustal W), MP (*protpars* from Phylip package) and ML (PhyML package). The best models for ML analysis were determined using PhyML by experimenting with a range of settings for several parameters (proportion of invariable sites, gamma parameter, substitution model, number of substitution rate categories) and choosing the ones that maximised the likelihood value LnL. This value indicates the probability of observing the data at hand, given a certain model. Preliminary tests had shown that the JTT and WAG models of nucleotide substitution were the best for this analysis and it was then tested how LnL changed for several test runs of the tree when these models and a range of values for the other parameters are used (data not shown).

The JTT and TN93 models were chosen for protein and nucleotide alignments, respectively. The JTT model was described by Jones, Taylor and Thornton (Jones *et al.*, 1992) and has an underlying mutation probability matrix, declaring the likelihood that any given amino acid with change to a specific other one, calculated from a large data set. The TN93 model, introduced by Tamura and Nei (1993), uses a matrix to model the rate of change between any two nucleotides. It does not assume equal nucleotide (nt) frequencies or transition/transversion frequencies. The proportion of invariable sites was set to zero, and the gamma parameter, measuring rate heterogeneity, was left to be estimated by the algorithm, while the number of substitution rate categories was fixed at 6, giving a trade-off between

computational time needed and flexibility of the model.

To make bootstrapping possible, a less computationally intensive, representative phylogeny including at most two members of each genus was constructed. The ML tree shown in Figure 2.2 agrees with previously reported phylogenies in terms of the relative positions of genera (McLysaght *et al.*, 2003; Gubser *et al.*, 2004; Hughes and Friedman, 2005) but includes many recently sequenced genomes. The ML and NJ trees inferred here were identical except for NJ grouping parapox and molluscipox as sister clades with high bootstrap support. The MP tree agreed with this grouping of parapox and molluscipox and also supports capripox as the outgroup of suipox and deerpox, rather than suipox and capripox being monophyletic as suggested by the ML topology. Single gene trees of the 33 core families were also constructed using PhyML, and their congruency with the concatenated gene tree was noted (Figure 2.2). At most nodes, the gene trees were in perfect agreement with the inferred species tree, lending further support to the topology. To infer more confident within-genus relationships for genera with many members (avipox, orthopox, parapox, capripox), orthologue families shared by all members of a genus were aligned at the protein level and the gaps were copied into the nucleotide sequence to generate nt alignments. The nt alignments were concatenated and the tree of the genus was inferred using PhyML with the TN93 model. For comparison the trees were also constructed using NJ (with both the Phylip package and Clustal W). NJ and ML results were always congruent and the bootstrapped ML results are shown in Figure 2.2. MP trees were not relied on heavily as they are most useful when using well-conserved and closely related sequences, which is not often the case for viruses (McCormack and Clewley, 2002). MP therefore perform poorly for branches of differing lengths and are known to frequently suffer from long-branch attraction.

2.4.2 Adding the 2008 Set

Clustal W (Thompson *et al.*, 1994) was used to infer an NJ phylogenetic tree based on an alignment of 28 core families (see Table 2.1) in 104 genomes. This tree can be seen in Figure 2.3, and the blue bootstrap values (as percentage) in Figure 2.4 refer to this tree. While it provides the backbone for the topology presented in Figure 2.4, the large number of genomes and small number of families used to determine their relationship leave some unsatisfactory branches where the bootstrap values are low or where the inferred topology is in conflict with the accepted poxvirus phylogeny. We have used concatenated alignments of larger numbers of protein families present in single copy in subsets of the 104 genomes to resolve these branches. The tree shown in Figure 2.4 is therefore a composite tree. Differently coloured bootstrap values are derived from different subset trees as follows.

The inferred monophyly of MOCV and crocodilepox (CRV) seen in Figure 2.3 is an artefact, according to the author of the crocodilepox genome paper (Afonso *et al.*, 2006), and when using 36 single copy families present in CRV, MOCV, BPSV, fowlpox virus strain Iowa (FPV-I) and *Amsacta moorei* EPV (AmEPV) to draw a NJ tree, the topology observed is that in Figure 2.4, represented by orange bootstrap values.

The relationship of genera in clade II poxviruses was clarified using a ML tree based on 95 families from all clade II genomes plus OPV outgroup, and this corresponds to the green bootstraps in Figure 2.4.

Black bootstraps are derived from an NJ tree of OPV representatives plus outgroup based on 83 shared single copy gene families.

Monkeypox viruses (MPXV) form two clades, in agreement with previous analyses containing fewer MPXV genomes (Likos *et al.*, 2005), which we termed clade I (includes MPXV strains Z96, CON, Z79) and clade II (includes strains SIE, LIB, WRA, COP, USA39 and USA44). As there was very weak bootstrap support for a branch within clade II MPXV, the within-genus

topology was resolved with a NJ tree of MPXV plus outgroup based on 128 single copy families, and the bootstrap values are noted in grey in Figure 2.4.

The 49 variola isolates that are fully sequenced and were analysed by Esposito *et al.* (2006) are difficult to separate in the poxvirus tree due to their high sequence similarity. An ML tree based on 114 single copy families present in all variola genomes, camelpox viruses (CMLV), taterapox virus (TATV) and the ECTV outgroup agrees to a large extent with the topology favoured by Esposito *et al.* (2006) and an NJ tree of the same genomes, and is presented by purple bootstraps in Figure 2.4.

The poorly supported topology of eleven vaccinia isolates and closely related genomes horsepox and rabbitpox is resolved by a NJ tree of 100 families present in single copy in these genomes and closest outgroup cowpox virus strain GRI-90. This tree is the source of the red bootstrap values. Horsepox (HPXV) is so similar to vaccinia that it is placed within the vaccinia cluster in both this and the poxvirus core tree. The author of the horsepox genome paper considers HPXV an outgroup to rabbitpox virus (RPXV) and vaccinia (Tulman *et al.*, 2006) and this seems to be supported by some gene analyses presented in Chapter 5 but the analyses that are based on concatenated sequence alignments point to a closer relationship of vaccinia and HPXV, which is represented in Figure 2.4.

All ML trees were inferred with PhyML, using the JTT model of nucleotide substitution, as described above.

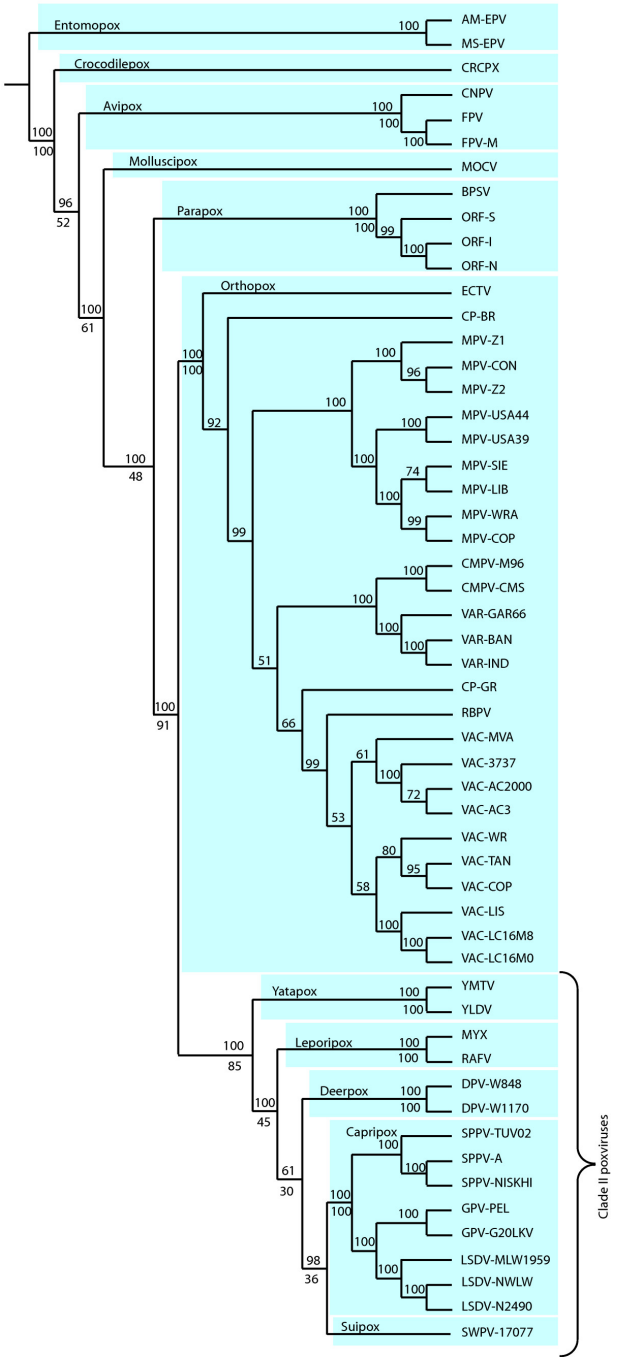


Figure 2.2: ML Poxvirus Tree 2006, a schematic representation of the tree topology. Branch lengths are not to scale. Numbers above the branches are percent bootstrap values, numbers below the branches indicate the percentage of gene trees that supported the branch. Taxonomic groups are labelled and shaded for clarity. The term "clade II" poxviruses is used for simplicity as per the literature (Hughes and Friedman, 2005) and refers to the group of yatapox, deerpox, capripox and suipox, as per convention.

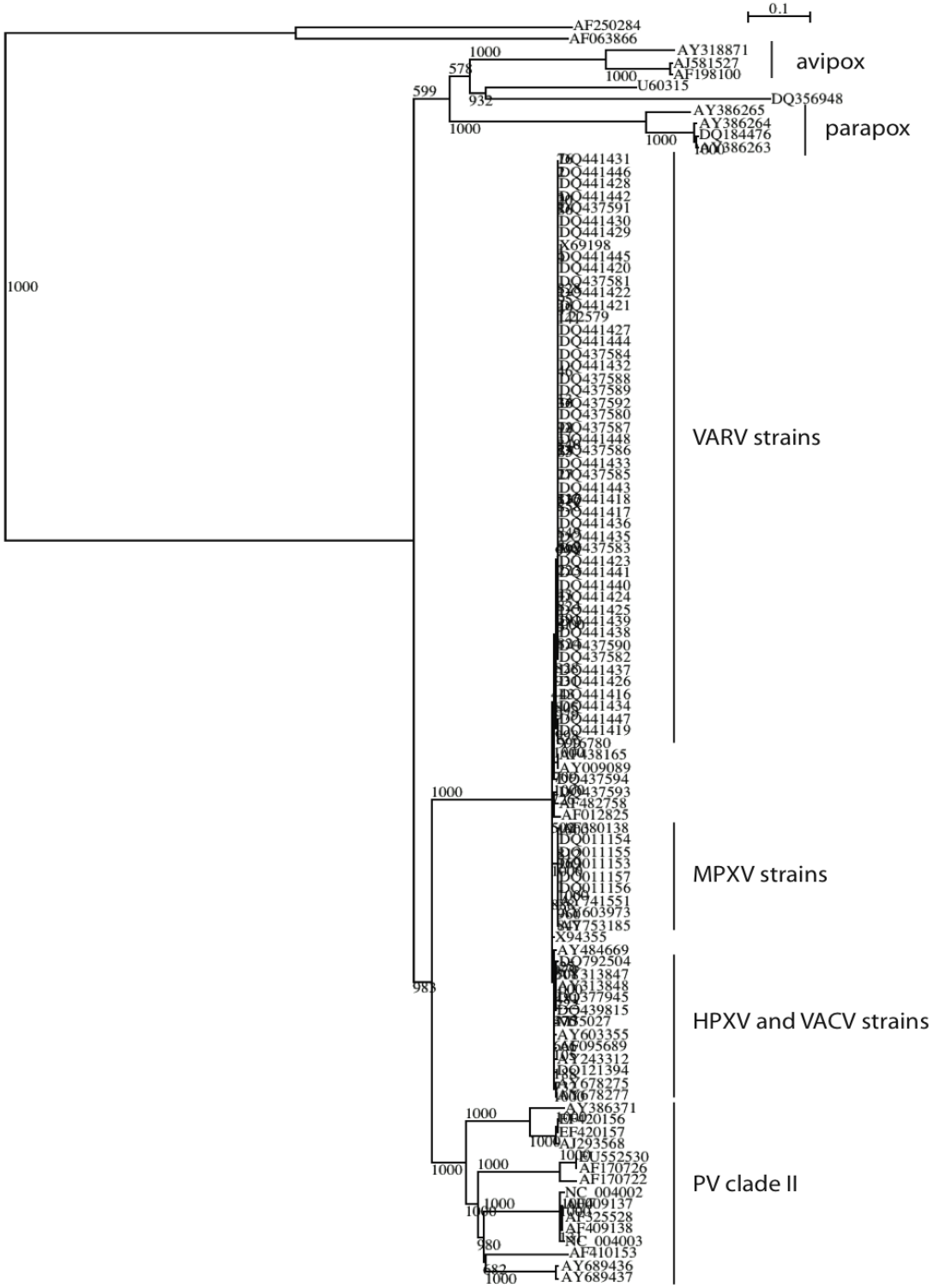


Figure 2.3: NJ Poxvirus Tree 2008, based on concatenated alignments of 28 core families. This tree was visualised with NJplot (Perriere and Gouy, 1996) and labelled with some clade names for clarification.

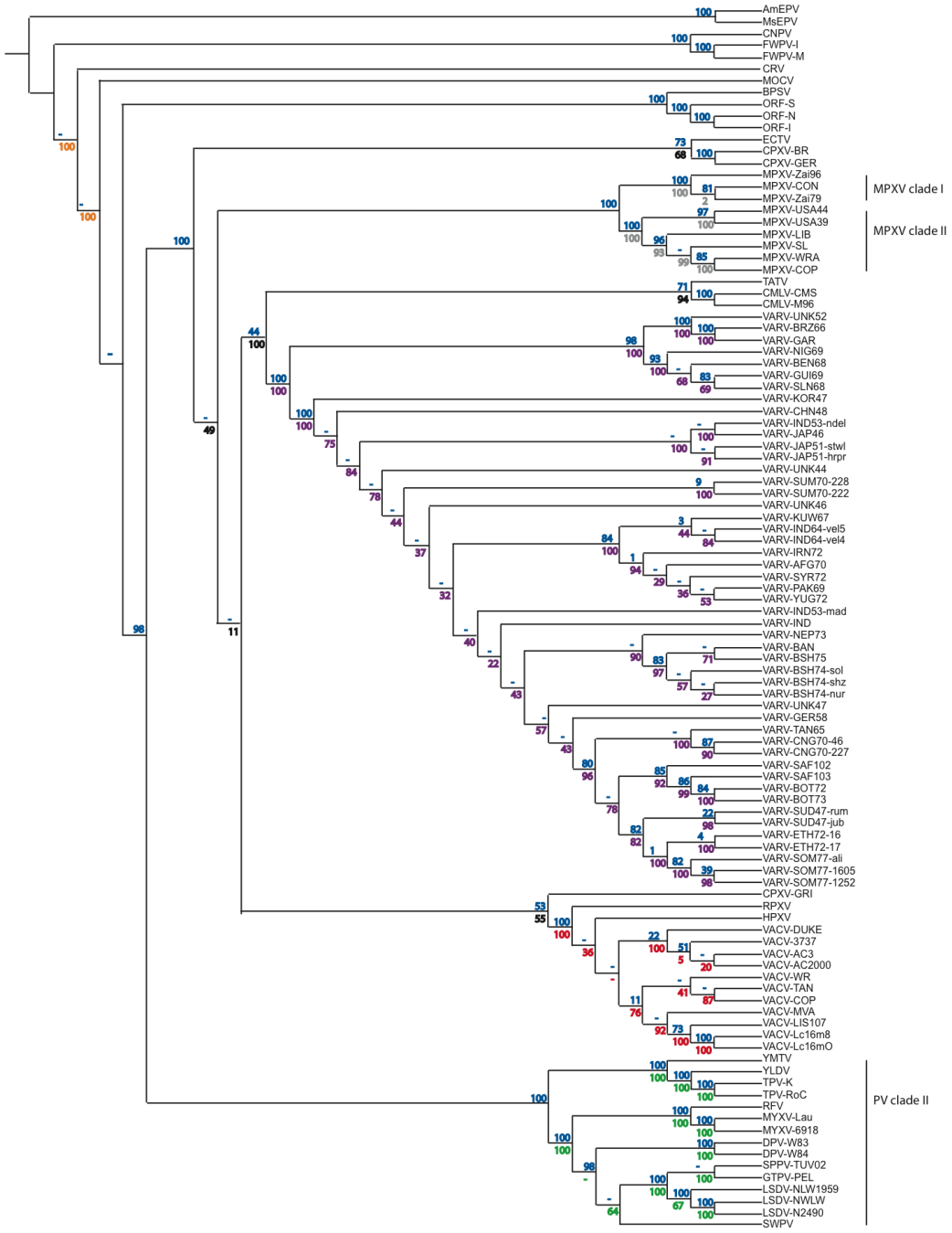


Figure 2.4: Composite Poxvirus Tree 2008, including all 104 completely sequenced genomes. Concatenated alignments of larger numbers of protein families present in single copy in subsets of the 104 genomes are used to draw sub-trees and thus resolve weakly supported branches. Bootstrap values in different colours correspond to these sub-trees. Orange: 36 families in CRV, MOCV, BPSV, FPV-I, AmEPV. Green: 95 families in clade II. Black: 83 families in OPV. Grey: 128 families in MPXV. Purple: 114 families in variola. Red: 100 families in VACV, HPXV, RPXV. Further details are found in the main text.

Chapter 3

Identification of Multiple Independent Horizontal Gene Transfers into Poxviruses using a Comparative Genomics Approach

3.1 Introduction

As described in Section 1.2.6 above, poxviruses contain many genes with sequence and functional similarity to host genes which were probably acquired by horizontal gene transfer (HGT). Here we carry out a comprehensive but strict search for poxvirus genes that show potential host gene capture (HGT from host to virus) in their evolutionary history, and try to determine whether one or several transfer events took place, using a novel method that incorporates comparative genomic information into the investigation of HGT.

The work presented in this chapter has been published under the same title in *BMC Evolutionary Biology* (Bratke and McLysaght, 2008).

3.1.1 Horizontal Gene Transfer

Horizontal gene transfer (also: lateral gene transfer) describes an event that differs from the regular "vertical" transfer of genes from parent to offspring: for example, host gene capture in viruses, where pieces of host DNA are incorporated into the newly formed virion. HGT is thought to play a major role in evolution, as there has apparently been frequent transfer of protein domains or entire genes within and between all divisions of life with important implications for molecular phylogeny studies (Doolittle, 1999; Brown, 2003; Choi and Kim, 2007).

Especially in pathogens, horizontally transferred genes (HT genes) are of enormous interest as they often increase virulence by encoding homologues to the host's immune system and interfering with its processes (Barry and McFadden, 1997; Lawrence and Ochman, 1998; Ochman *et al.*, 2000).

Multiple independent horizontal gene transfer events are thus particularly significant; they highlight the advantage that is conferred on the virus genome by a gene that has been gained and maintained several times. The fact that poxviruses do not enter the nucleus poses a practical problem when it comes to the mechanism of host gene capture: how HGT from host to virus takes place is not yet known, but as poxvirus genes lack introns, retroposition through a processed mRNA intermediate of host genes is probably involved. Vaccinia ligase has the ability to seal nicked substrates containing one or more RNA strands, and Sekiguchi and Shuman (1997) suggested that, in poxvirus HGT, cellular mRNA is joined to vaccinia by ligase, which requires only limited complementarity that may be found, for example in the poly-A-tail of the mRNA aligning with the AT-rich vaccinia genome. In this model, the mRNA is then copied into cDNA by host reverse transcriptase or vaccinia DNA polymerase and has thus found its way into the virus genome.

3.1.1.1 Detecting HGT

There are several approaches to detecting HGT, all of which are best used in conjunction with each other (Lawrence and Ochman, 2002). The first evidence usually comes from sequence similarity to a gene in a distantly related organism or in the viral host. Phylogenetic gene trees, which include sequences from both viruses and cellular organisms, can also provide an indication of host gene capture where virus genes group closely with homologues from the host or organisms related to the host, i.e., where gene history differs from species history (Shackelton and Holmes, 2004). Uncertainties in topology and rooting of the tree may hamper their use, however. Atypical base composition of a particular gene, for example, a GC-content that differs signicantly from that of most other genes in the genome, provides another piece of evidence that HGT may have recently taken place and this approach has been used to detect host-derived regions in poxviruses (Da Silva and Upton, 2005) and other small genomes (Tsirigos and Rigoutsos, 2005). This signal, however, is expected to fade over time and therefore this approach is only useful for detecting recent HGT.

These difficulties are compounded when one wishes to test whether there have been multiple independent HGT events of the same gene into related viruses. This may be indicated by a polyphyletic relationship in the phylogenetic tree (see Figure 3.1), but such a tree would require strong support (e.g., high bootstrap values) for several branches in order to be reliable - frequently, this support is not there. A mixed pattern of presence and absence of a particular gene in a group of closely related organisms is another indication for horizontal transfers of the gene in question, but an ancestral origin with loss in some genomes can also explain this observation. As the conclusions from these different methods may disagree, additional information is important.

We use comparative gene order information to distinguish between a

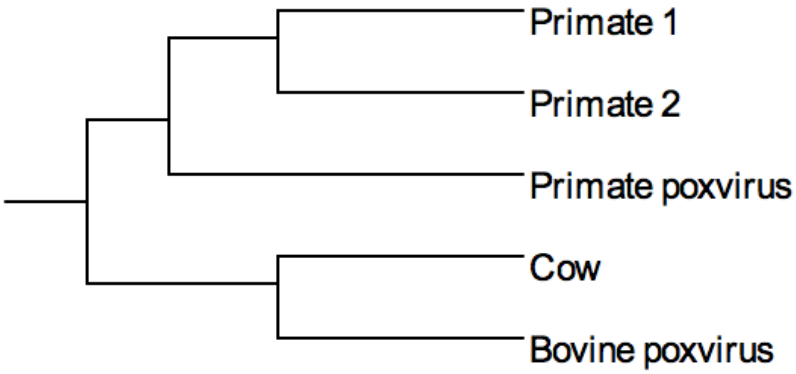


Figure 3.1: Sample phylogeny showing multiple HGT events. This tree shows a polyphyletic origin of poxvirus genes where the gene in question may have been transferred twice in the evolutionary history of the virus: once from a primate ancestor to a primate poxvirus and another time from a cow-like ancestor into a bovine poxvirus.

shared history of the HT gene in the virus genomes and several independent transfer events. This is a sensible approach in poxviruses, due to the conserved synteny in their genome cores, and has been recommended, albeit not specifically for viruses, by Podell and Gaasterland (2007).

3.1.2 Synteny

Poxviruses have extremely well conserved gene order, or synteny (Greek: on the same band or ribbon), particularly of the core families. This was noted very early on in the analysis of poxvirus genomes by Mackett and Archard (1979) (see also Section 1.3) and has become much clearer with the availability of more and longer genome sequences. Figure 3.2 is a visualisation of poxvirus synteny, which was first used by McLysaght *et al.* (2003), based on fewer genomes. Vertical lines in Figure 3.2 that do not cross over other vertical lines represent perfectly conserved local gene arrangements, see figure legend for a detailed description of this visualisation.

We estimated the overall level of synteny conservation throughout poxvirus genomes by considering the conservation of neighbouring genes

between orthologues in pox. For these purposes, we examined only genes and neighbours which had been grouped into families without paralogues and which were present in at least two complete genomes, as defined in Section 2.3. If a given gene always has the same two immediate neighbours in every genome in which it is found, then it is considered to have perfect neighbour conservation. By contrast, if the neighbours are always different, then there is zero conservation. This was measured using the following statistic:

$$Score = \frac{M - N}{M - 2}$$

where M is the maximum possible number of unique neighbours, i.e., different neighbors in every case, and N is the observed number of neighbours. This will always return a value between 0 and 1. 69% of orthologue families had a neighbour conservation value of at least 0.9 and 89% had at least 0.7 neighbour conservation. A similar measure of conserved synteny was used by Notebaart *et al.* (2005), when they looked at which gene of a pair of duplicates was the ancestral copy. These values once again confirm that comparison of the genomic location of HT genes between poxvirus genomes is a sensible approach.

3.2 Methods

3.2.1 Detection of Non-Poxviral Homologues

We used PSI-BLAST (Altschul *et al.*, 1997) to search all 17,815 poxvirus proteins of the 2006 dataset against the GenBank non-redundant database (May 2007; 4.9 million sequences). The *E*-value threshold for inclusion in the PSSM was E-20, with a maximum of five iterations. Only alignments of 70% identity or higher and at least 80% length of query sequence were kept as homologues. These parameters were selected after some experimentation

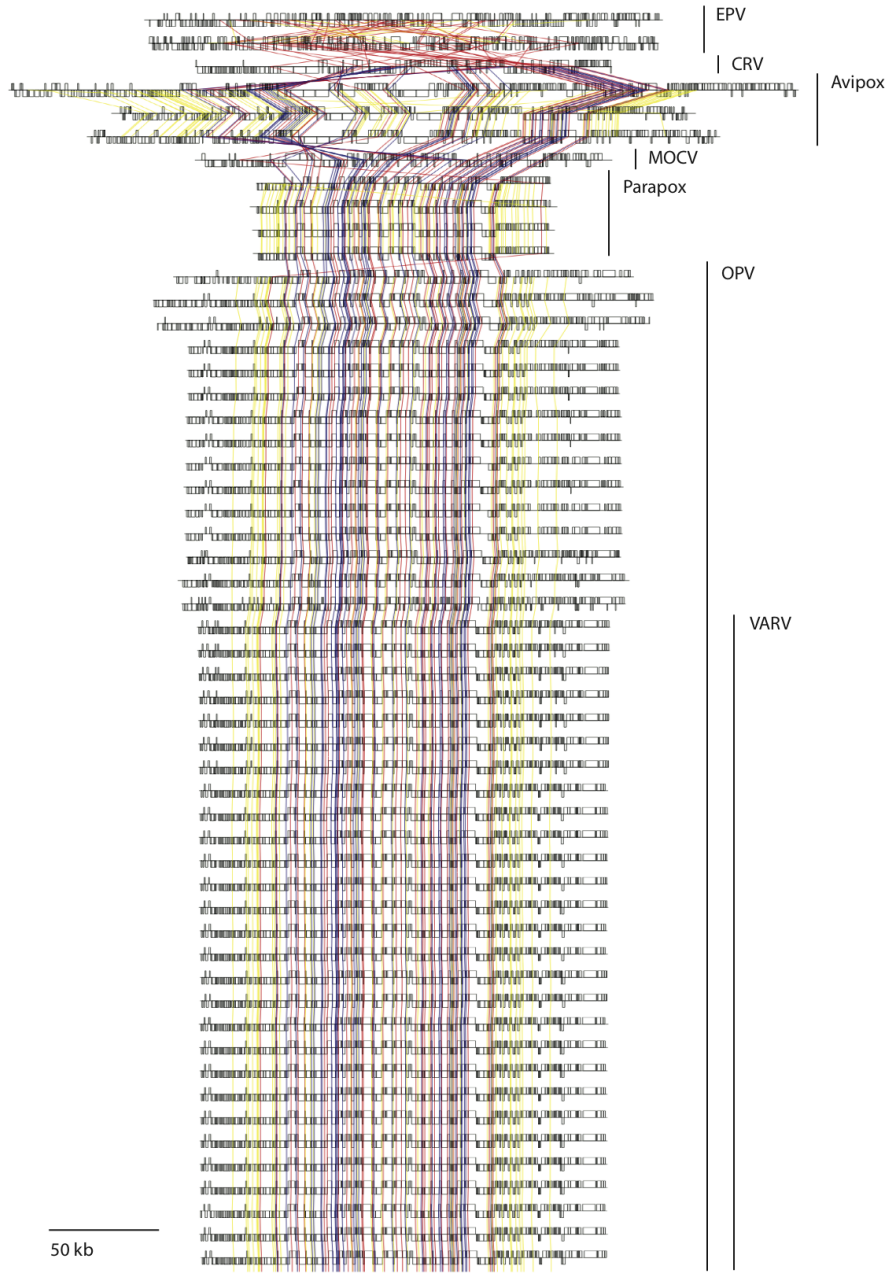


Figure 3.2: Illustration of poxvirus synteny, showing half of all PV genomes for easier viewing. Horizontal lines represent completely sequenced genomes with genes on the 5'-3' strand and 3'-5' strand represented as blocks above and below the line, respectively. Vertical lines connect genes in different poxviruses that belong to single-copy families conserved in all poxviruses ("core families" as described in Section 2.3, red lines), all chordopoxviruses (blue lines) and the genus in question (yellow lines). One major inversion between avipox and other chordopox is visible, as well as excessive rearrangements in the entomopox lineage, where synteny breaks down. In all other places, synteny is extremely well conserved between genomes.

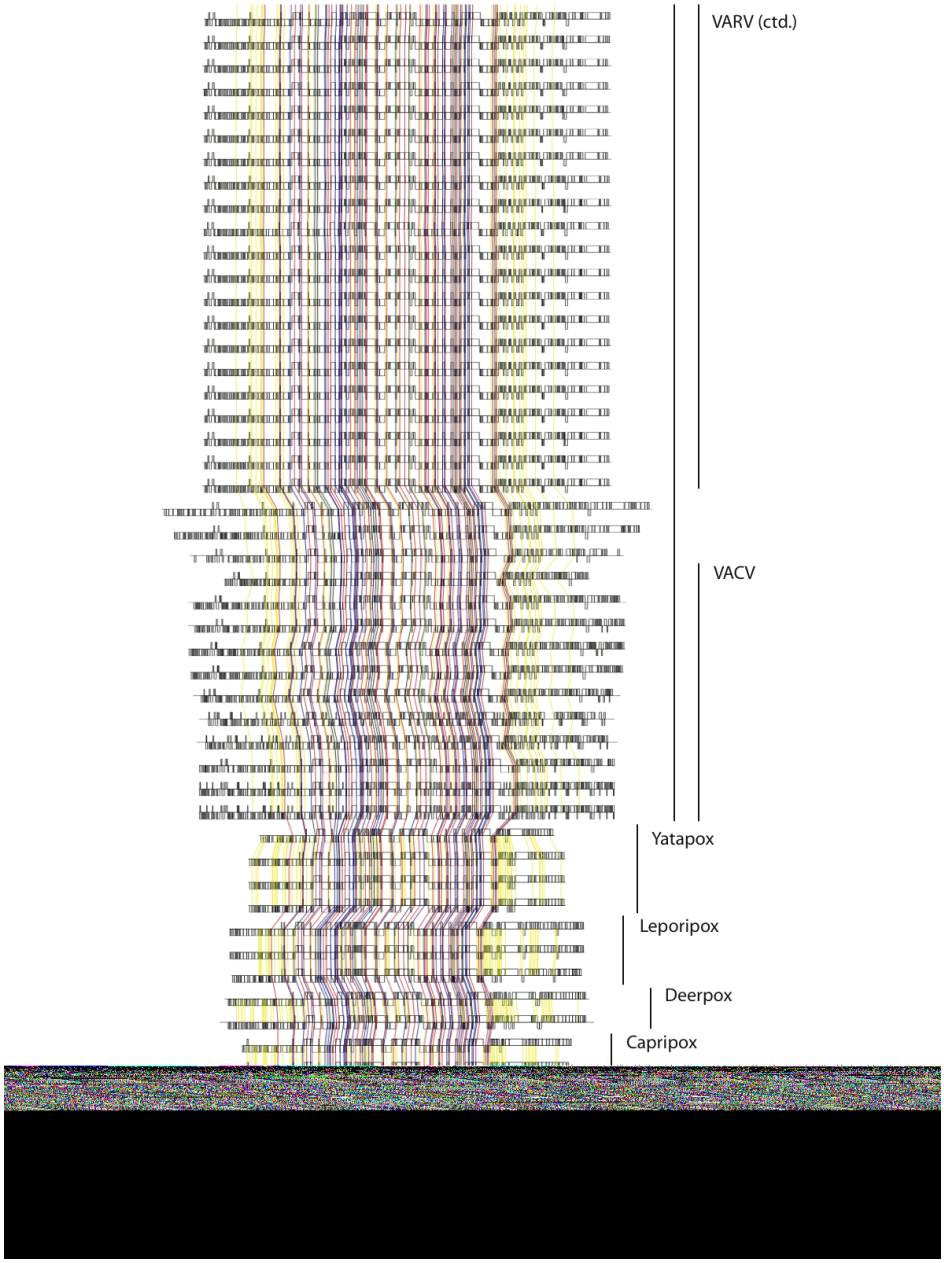


Figure 3.3: Illustration of poxvirus synteny, showing the second half of PV genomes.

with different values in order to retrieve only strong candidates for HGT and keep detection of false positives to a minimum. We excluded hits to synthetic constructs (e.g., artificial vectors). The resulting groups of tentative homologues were only considered for further analysis if they included at least one non-poxviral hit, as the aim was to identify possible host-virus gene transfers. Using PSI-BLAST to detect homologies between viral and cellular proteins has been successful for herpesviruses (Holzerlandt *et al.*, 2002).

3.2.2 Gene Phylogenies

The members of gene families with PSI-BLAST hits outside poxviruses were aligned with their non-poxviral homologues using MUSCLE (amino acid (aa) sequences, Edgar (2004)). NJ trees were drawn with ClustalW (Thompson *et al.*, 1994) and bootstrapped 1000 times. ML trees were drawn with PhyML (Guindon and Gascuel, 2003) using model JTT, proportion of invariable sites zero, six substitution rate categories and estimating the gamma distribution parameter.

3.2.3 Synteny Diagrams

Genes neighbouring each gene of interest and the positions of their family members based on MCL family classification (see Section 2.3) in all other genomes were noted and compared, and the synteny diagrams were drawn by a Perl program. A sample is shown in Figure 3.4. We made three kinds of illustrations of the genomic context of HT genes, to maximise information gained from this comparison, distinguished by the type of gene families that the HT gene is plotted alongside. In the first type, the HT gene location is plotted with respect to the core families only. There are only 33 of these genes (the 2006 data set was used for this chapter), so the distances within the diagram may be quite large, however, because these are shared by all pox genomes, this view permits a low-resolution comparison across all genera.

The second type of synteny diagram plots the HT gene with respect to orthologue families, those with one copy in two or more genomes, which are much greater in number. This increases the resolution of the plot and still avoids the confounding effect of paralogous genes. Finally, we plot the location of the HT gene with respect to any of the gene families ("ortho-paralogue families") and examine it with the caveat that paralogous genes may be misleading. Depending on the extent of gene order conservation around the HT gene between genomes, it may be possible to distinguish single from multiple transfer events, as indicated in panels A and B of Figure 3.4, respectively. If the gene of interest is present in different genomic environments in two poxvirus genomes, this was read as support for multiple independent horizontal gene transfer events.

If the local gene order is not well conserved then a single transfer followed by genome rearrangement cannot be excluded, even when the HT gene is present in a completely different context in different genomes, as shown in panel C. Within poxvirus genera this is a particular issue when comparing EPV gene order to that of ChPV, and to a certain degree also in comparison of avipox to other ChPV because of the high amount of rearrangement, as visible in Figure 3.2. It has been observed that blocks of homologous genes are often conserved between avipox and OPV, but the genomic locations of these blocks are quite different (Mockett *et al.*, 1992).

3.3 Results

We identified 14 pox gene families that included at least one member with a significant PSI-BLAST hit to a gene in a non-pox organism (Table 3.1; upper section). The search protocol we employed is conservative and we did not detect some previously reported non-poxviral homologues of poxvirus proteins (Anantharaman and Aravind, 2003; Hughes and Friedman, 2005). However, we initially wished to confine our gene order comparisons to genes

Figure 3.4: Model gene order comparisons. The relative order of genes around the gene of interest (GOI) is illustrated. Orthologous genes are shaded in the same colour and aligned vertically. A white box indicates that the gene is present in the genome, but not at the expected equivalent location. Horizontal lines indicate genomic segments. Thick horizontal lines indicate that the connected genes are neighbouring genes of the specified type (where GOI is present) or have fewer than three intervening genes of the specified type (where GOI is absent). Thin horizontal lines signify three to six intervening genes. Vertical lines separate discontiguous genomic regions. (A) The GOI is in an equivalent location in all genomes where it is found. This supports a single origin of this gene. (B) The GOI is found in two different genomic locations and the relative arrangement of other genes in each of these locations is conserved, i.e., there is no support for local rearrangement. This supports two independent transfers of this gene. (C) The GOI is found in two different locations but these regions are not conserved between genomes, so it is not possible to decide whether a single transfer followed by genome rearrangement, or two independent transfer events are more likely.

with convincing aa sequence similarity to genes found outside poxviruses. It was then decided to include HGT candidates identified elsewhere in our synteny analysis and they are listed in the lower section of Table 3.1. Of the initial 14 gene families, some did not warrant further investigation. The *gag*, *pol* and *env* open reading frames found in avipox are part of a reticuloendotheliosis virus (REV) sequence that is inserted into the genome of some fowlpox viruses and clusters with sequences from REV and other avian viruses. While this is HGT of sorts, it is a special case where the REV sequence can reform into infectious particles after poxvirus infection of a chicken host (Hertig *et al.*, 1997), and these genes were not investigated further here. The fusolin/enhancing factor (EF) family was found to be present only in incompletely sequenced entomopox genomes and one insect larva genome (*Pseudaletia separata*). The similarities between the EF and fusolin genes of different EPV and their detrimental effect on the host's resistance to the viruses have been reported (Hayakawa *et al.*, 1996), but it appears there has been an annotation mistake in the record of the insect sequence [GenBank: JC5185], as the paper it refers to (Hayakawa *et al.*, 1996) makes no mention of an insect sequence, and it is identical to that of the insect's EPV sequence [GenBank: BAA09138]. As we are unsure of the veracity of the candidate host sequence and could not contact the authors, we excluded this gene from further analysis. Ubiquitin is also found by our search. The large number of homologues makes further investigation difficult, and as this protein is shared by all eukaryotes, it is likely to be ancestral to all poxviruses. The remaining nine protein families, a further 15 candidate proteins from the literature and the HGT histories suggested by our analyses are summarised in Table 3.1.

In the cases of dUTPase and Uracil DNA glycolase, there was insufficient gene order conservation to allow comparisons; this reflects the fact that many of the transfers were directed into the poxvirus chromosome ends where the

Figure 3.5: IL-10 phylogenies including poxvirus sequences and representative non-pox homologues. (A) NJ tree based on amino acid alignment (B) ML tree based on amino acid alignment. Branch lengths are not to scale. (C) ML tree based on nucleotide alignment (gaps copied from amino acid alignment). Branch lengths are not to scale. Tip labels indicate GenBank GI and the common species name for each sequence. Numbers on branches are bootstrap values out of 1000 for NJ tree (A) and out of 100 for ML trees (B and C).

3.3.2 Thymidine Kinase

Thymidine kinase (TK) is present in many poxviruses (Upton and McFadden, 1986; Boyle *et al.*, 1987; Amano *et al.*, 1995). It phosphorylates a wide range of substrates and, surprisingly, has been found to increase susceptibility to antiviral drugs by selectively phosphorylating some thymidine analogue drugs to a greater degree than the host TK1 gene and thus increasing the efficacy of the antiviral drug (Prichard *et al.*, 2007). In this study, the TK family only contains human and poxviral members, as sequences from other organisms were too distantly related to be included. The human sequences cluster with OPV genes (bootstrap 35.4%) in a tree rooted by EPV sequences. Separate origins for avipox and orthopox TK have previously been suggested based on phylogenetic evidence (Koonin and Senkevich, 1992; Hughes, 2002) and this is supported here by the different genomic environments in avipox and other chordopox (Figure 3.8; and Binns *et al.* (1988)). The gene order is completely scrambled in entomopox, however, so these data do not distinguish between shared origins for entomopox and either the avipox or other chordopox TK genes and distinct origins (i.e., the data do not distinguish between two or three transfer events). The former scenario would require that the TK gene was present in the common ancestor of all pox and subsequently lost in either avipox or the other chordopox only to be regained independently. If the EPV and avipox TK gene are descended from the same HGT event, then this scenario would require three events (HGT into pox ancestor, loss from ChPV excluding avipox, HGT into another ChPV lineage). In a simplistic sense, this is equally parsimonious with three HGT events. A more biologically realistic analysis should include consideration of any selection to retain, lose or acquire this gene; properties that we do not currently know. The fact that this gene has been acquired and retained independently at least twice indicates that multiple HGT events are possible. Loss and re-acquisition in the ChPV lineage may indicate that significant changes in the environment

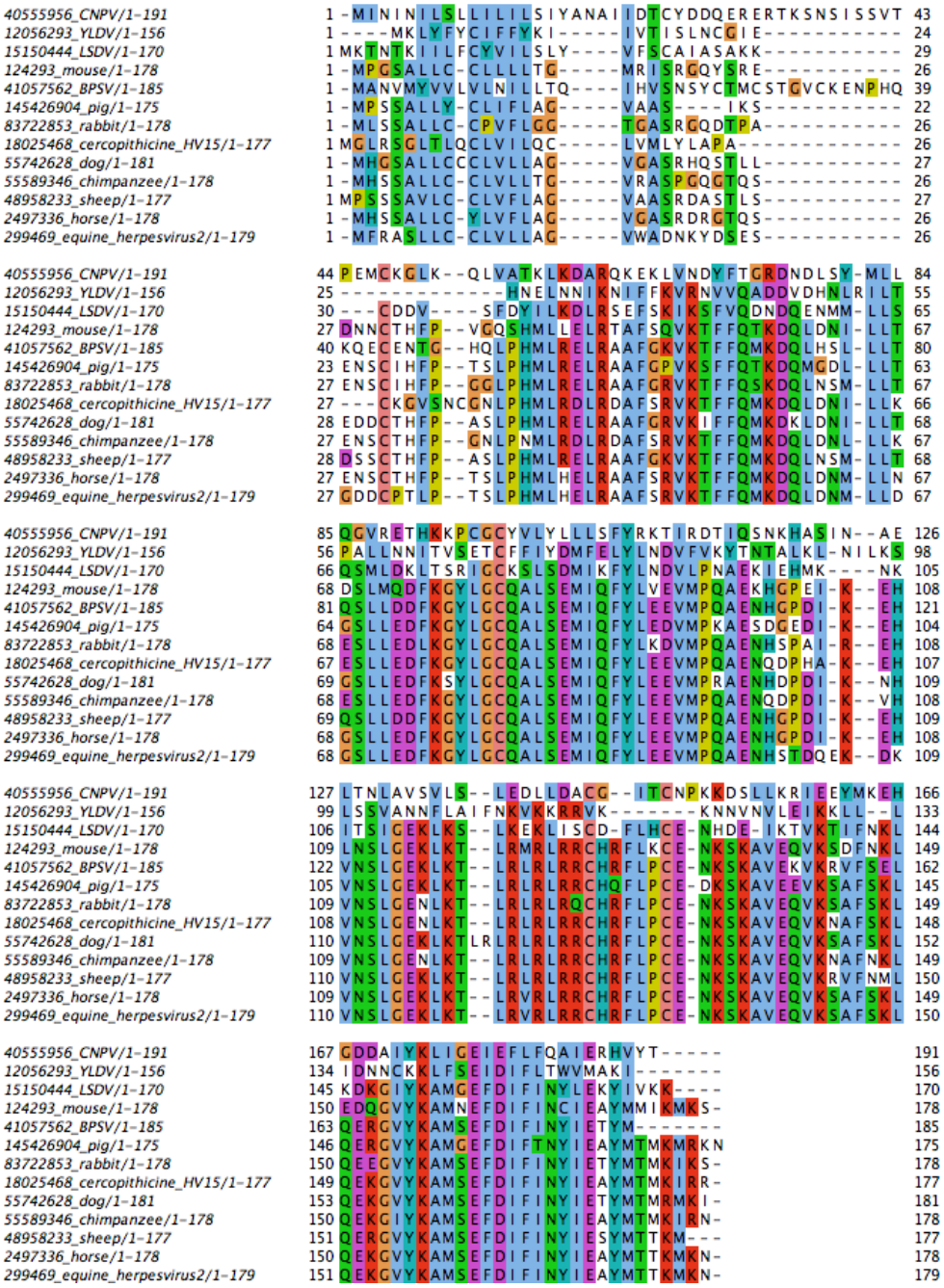


Figure 3.6: IL-10 multiple sequence alignment. Each sequence is labelled with its GenBank GI. The top two sequences are CNPV and YLDV, clearly not as similar to each other as the phylogenetic tree would suggest. This image was created using JalView (Clamp *et al.*, 2004).

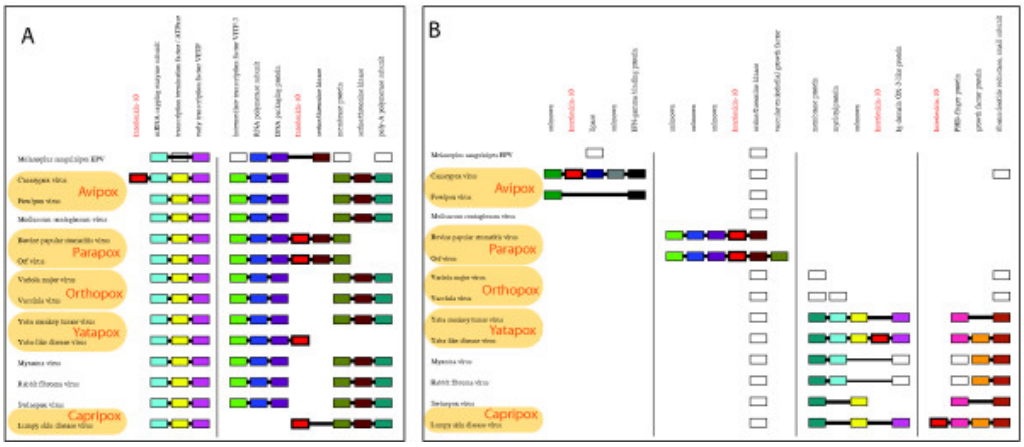


Figure 3.7: IL-10 synteny. The gene of interest is coloured red with a bold outline. A horizontal line that crosses a box indicates that this gene is not part of the contiguous segment. (A) Neighbourhood with respect to core families. The Serine/Threonine Kinase gene has experienced a local inversion in the parapox genomes and is shown in two columns for clarity; (B) Neighbourhood with respect to orthologue families.

have changed the nature of the selection acting on this gene.

In avipox, the TK gene and the adjacent ht-motif protein gene (host targeting) are in the location occupied by ribonucleotide reductase, large subunit (RRL) in orthopox and suipox. This, along with reported 15 bp direct repeats flanking TK and the ht-motif gene led to the suggestion that these genes transposed within avipox to this location in avipox and replaced RRL (Binns *et al.*, 1992). This suggestion is at odds with the experimental strategy employed here because if intra-genomic translocations are a common occurrence, it would mean that completely distinct genomic neighbourhoods are not suffcient to indicate distinct origins. However, there is no real support for this transposition hypothesis from the data now available: (i) The phylogenetic evidence supports an independent origin of avipox TK; (ii) There is no evidence that the RRL gene was ever present in avipox because it is not found outside orthopox and suipox (whose most recent common ancestor is not shared with avipox); (iii) Furthermore, we cannot identify the 15 bp repeat referred to in the 1992 paper (Binns *et al.*, 1992) but we

did identify a 12 bp direct repeat (starting nucleotide positions 87698 and 88647 in GenBank:AF198100). One of the repeat copies is inside the open reading frame of the I5L gene, which adds the constraint that if it is the result of an insertion-site duplication then it must not have disrupted the function of this gene. The relationship of the EPV TK gene to the other pox TK genes remains inconclusive, with a third independent transfer being one possibility. We conclude from the gene order data that there were at least two independent horizontal transfers of TK into poxviruses, and possibly a third (into entomopox). The independent transfer into the same locus of the TK gene in avipox and the RRL gene into the ancestor of orthopox and suipox suggests the intriguing possibility that this genomic location has properties that make it particularly receptive to gene transfers.

3.3.3 Ribonucleotide Reductase, Small Subunit

Ribonucleotide reductase (RR) functions as a heterodimeric tetramer and leads to the reduction of all four ribonucleotides to deoxyribonucleotides in DNA biosynthesis. Rajagopal *et al.* (1995) showed that vaccinia RR only becomes essential to the virus when the host copy is depleted. The small subunit (RRS) was sequenced in vaccinia and found to be much more closely related to mouse and clam RRS than to herpesvirus RRS, suggesting host gene capture (Slabaugh *et al.*, 1988). It was apparently transferred three separate times into *Baculoviridae*, another family of dsDNA viruses (Hughes and Friedman, 2003). RRS is present in most chordopox genera except parapox, molluscipox and some avipox. The avipox RRS gene is present in a different synteny block than the RRS gene in other chordopox (Figure 3.9). These data support two independent transfers of RRS into pox: one into the avipox lineage, and one into the ancestor of orthopox and clade II. The term "clade II" poxviruses is used for simplicity and refers to the group of yatapox, deerpox, capripox and suipox (Hughes and Friedman, 2005).

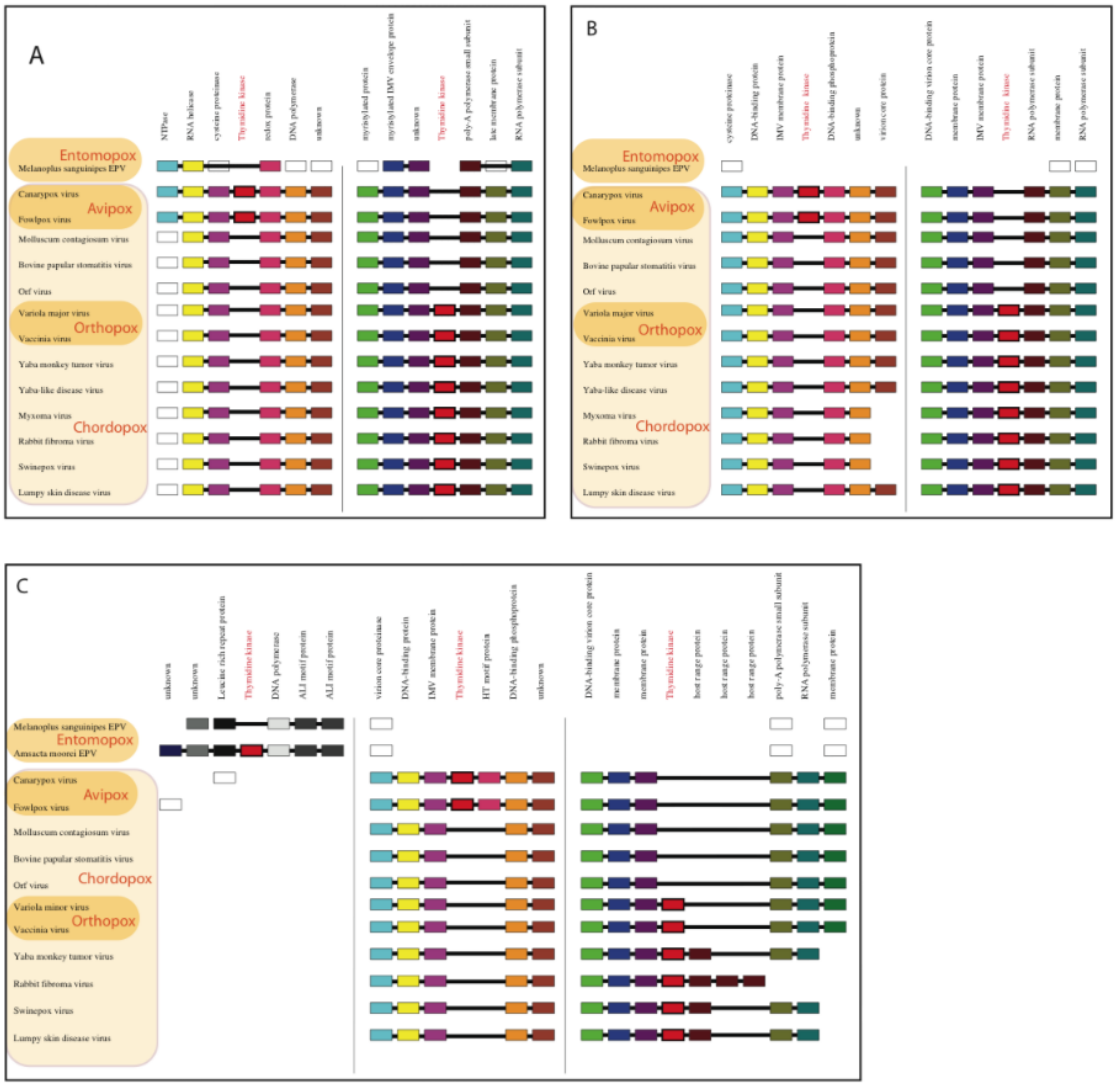


Figure 3.8: TK synteny. The gene of interest is coloured red with a bold outline. (A) Neighbourhood with respect to core families; (B) Neighbourhood with respect to orthologue families; (C) Neighbourhood with respect to ortho-para families.

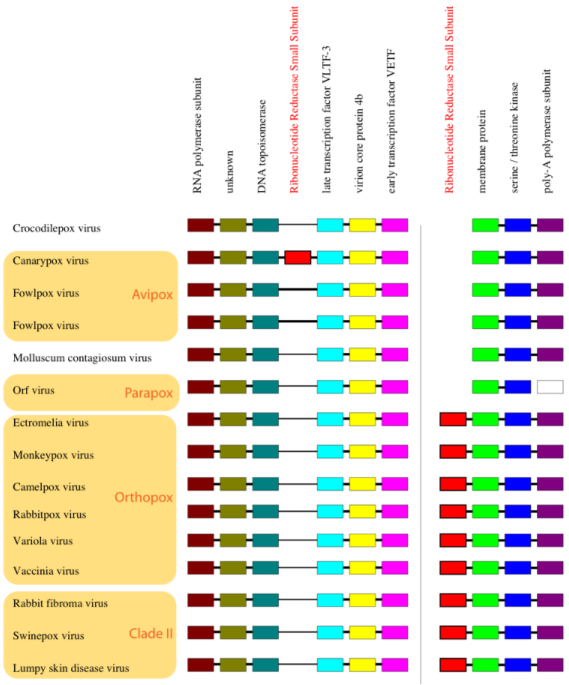


Figure 3.9: RRS synteny. The gene of interest is coloured red with a bold outline. Neighbourhood with respect to core families.

3.3.4 Glutathione Peroxidase

The viral glutathione peroxidase (GP) protects virus and infected cells against oxidative damage resulting from interaction with the immune system, which usually induces apoptosis (Shisler *et al.*, 1998; Moss *et al.*, 2000). It is not present in OPV and was first identified in Molluscum contagiosum virus (Senkevich *et al.*, 1996). GP is found in avipox and molluscipox and separate transfer events have previously been suggested (Hughes and Friedman, 2005). The synteny diagram with orthologue families clearly shows that the gene is in an equivalent location in all avipox genomes (Figure 3.10), but is in a distinct location in molluscipox and thus supports independent transfers of this gene into these two lineages.

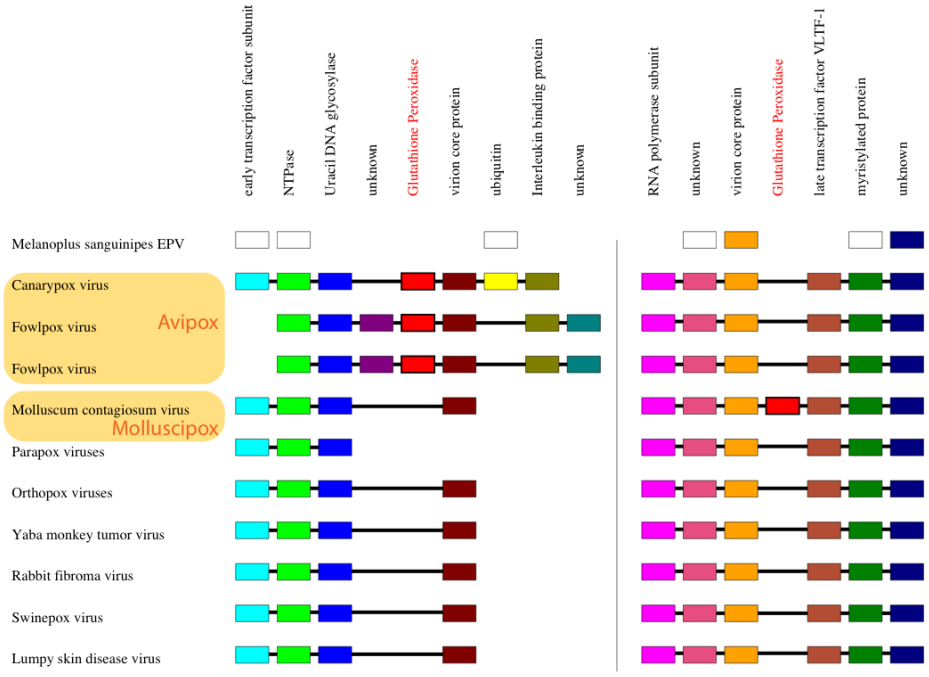


Figure 3.10: GP synteny. The gene of interest is coloured red with a bold outline. Neighbourhood with respect to orthologue families.

3.3.5 Deoxyribodipyrimidine Photolyase

Photolyases are enzymes which catalyse the repair of UV-induced DNA damage. In poxviruses, photolyases are implicated in the survival of the virus outside host cells in between infections (Srinivasan *et al.*, 2001). Deoxyribodipyrimidine photolyase (DP) is found in entomopox, avipox and leporipox. It was identified as a candidate HT gene through sequence similarity with genes in insect and vertebrate genomes by Hughes and Friedman (2005). The avipox gene lies in a region of conserved synteny with other ChPV, but the DP gene is not at the syntenous location in leporipox (Figure 3.11). This supports independent origins for the avipox and leporipox DP genes. It is not possible to compare the location of the EPV genes with each other or with ChPV due to insufficient synteny conservation. We conclude that there were at least two independent transfers of DP into poxviruses.

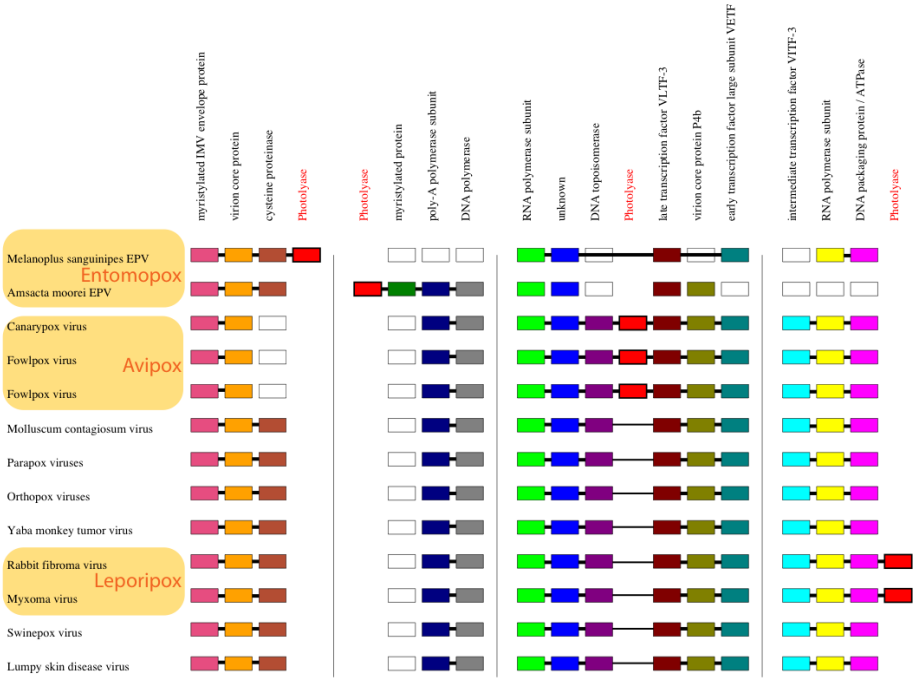


Figure 3.11: DP synteny. The gene of interest is coloured red with a bold outline. Neighbourhood with respect to core families.

3.4 Discussion

Here we systematically searched all known poxvirus proteins for evidence of HGT and suggest multiple transfer events in the gene histories of thymidine kinase, interleukin-10 and glutathione peroxidase. While searches for HGT from hosts to poxviruses have previously been carried out based on aberrant virus-host phylogenies (Hughes and Friedman, 2005), this is the first comprehensive study that goes beyond the comparison of the sequences of homologous genes and introduces comparative synteny data into the analysis of horizontal gene transfer. These data are completely independent of phylogenetic tree and sequence composition data. For the 24 genes with evidence of horizontal transfer analysed here, all but four had sufficiently conserved synteny between poxviral genomes to permit comparison. In five cases, the presence of the HT gene in a well-conserved, yet distinct, region in different poxvirus genomes supported a conclusion of multiple independent

transfers of a homologous gene into *Poxviridae*. In the other cases, this study confirmed that a single event gave rise to the HT gene in poxviruses.

This strategy is based on the assumption that there is no bias in the insertion site of a HT gene with respect to the neighbouring genes in the receiving genome. The only known bias in terms of the location of new genes in poxviruses is that they tend to be located towards the ends of chromosomes (McLysaght *et al.*, 2003; Filee *et al.*, 2008) and this is not expected to influence the neighbours for this analysis. More importantly, we interpret the conservation of synteny across poxvirus genomes as evidence against genomic rearrangement events in that region. This means that under these circumstances we can exclude the possibility that the new genomic location is caused by genome rearrangement within the pox lineage rather than independent transfer into the lineage.

A major strength of this method is that it does not rely on phylogenetic trees (and thus multiple sequence alignments) in order to interpret the number of transfer events. In previous studies, many HT genes were refractory to phylogenetic analysis due to difficulties rooting the trees or insignificant support for critical branches (e.g. Hughes and Friedman (2005) in pox, and Huang *et al.* (2004)). Phylogenetic trees are frequently plagued by problems which abolish their usefulness for analysis of HGT (Podell and Gaasterland, 2007). Because the conserved synteny data are independent of the phylogenetic tree for the HT gene, these results could be combined.

The HT genes with multiple origins identified here have very interesting biological properties. The presence of viral TK increases susceptibility to antiviral drugs, a surprising finding (Prichard *et al.*, 2007). The particular benefit to the virus, if any, of having this gene is not clear. Three other HT genes are known to improve the survival of the virus against the host immune system (IL-10) and environmental damage (GP and DP). In the case of these genes it is easy to speculate that they are particularly advantageous to the

virus, and the fact that they have been successfully transferred into poxvirus genomes on more than one occasion lends support to this hypothesis.

Considering both phylogeny and gene order can alter predictions about potential HGT events (Table 3.1). What previously looked like one transfer based on phylogenetic relationships is sometimes more parsimoniously due to multiple independent HGT events in the history of one gene, such as thymidine kinase and interleukin-10. The opposite was the case for the small subunit of ribonucleotide reductase, where badly supported phylogenetic trees suggested multiple transfers, but synteny around the transferred gene favoured one transfer with subsequent divergence of sequences.

Three trends for genes that have experienced HGT are reflected by this study: Firstly, HGT from host to viruses is often found in the history of genes involved in host-virus interaction as viruses use the cellular copies to their advantage to evade or modify the host's immune responses, here this is the case for IL-10, GP, S1R, EF and serpin. Secondly, essential genes are unlikely to have originated by recent HGT, and the gene families investigated here are mostly advantageous, rather than essential, to the virus. Thirdly, HGT involves mostly genes near the ends of the genome in poxviruses, and this, too, is observed in our study.

Horizontal gene transfer can allow a virus to mimic a host protein and thus gain a selective advantage as it avoids being detected as foreign material. It can also play a role in the evolution of the donor genome, as pathogen homologues of host defence proteins may promote diversity of this group of proteins (Murphy, 1993). Multiple transfers have a rather stronger implication than a single transfer, as repeated host gene capture emphasises the importance of this gene to the virulence or life cycle of the virus. We therefore recommend, particularly in the absence of a robust phylogenetic tree, that the gene order should be taken into account where possible when considering HGT, as it can provide important additional information on

the gene history, especially in organisms where gene order is as excellently conserved as in poxviruses.

Chapter 4

Predicting Protein-Protein Interactions in Poxviruses using Computational Methods

In order to understand how an organism works it is important not only to know its proteins but also to know the way in which these proteins interact with each other. Detecting previously unknown protein-protein interactions (PPI) can increase our knowledge of metabolic processes and protein complexes. In this chapter, we attempt to use computational methods to identify candidate protein pairs that are likely to be involved in poxvirus PPI.

4.1 Introduction

Data on presence and absence of a given gene family from related organisms have frequently been used to construct phylogenies (Montague and Hutchison, 2000; Herniou *et al.*, 2001; Hughes and Friedman, 2005; Wang *et al.*, 2007), and while these are sometimes found to concur with more rigidly defined phylogenies, this method is not as reliable as using sequence

conservation. Such information on presence and absence is also known as the phyletic (or phylogenetic) pattern of a gene and can be used in different ways in evolutionary genome analysis. Phyletic patterns of genes involved in several key features of apoptosis, for example, were used by Castro *et al.* (2008) to pinpoint the evolutionary origins of such features. A similar analysis by Boekhorst *et al.* (2008) yielded information about the evolution of phosphorylation in different eukaryotes.

Patterns of gene presence and absence can also indicate functions and dependencies of genes. Two or more genes that are always present together in the examined genomes can point to an interaction between them or a role in the same pathway. Complementary presence and absence of genes over a number of species' genomes may indicate functions in alternative pathways or similar roles that are not both necessary for the survival of the organism. This has been demonstrated, amongst others, by Rocha *et al.* (2005) for recombination systems in bacterial genomes, and Aravind *et al.* (2000), who found concerted loss of functionally connected gene groups in yeast. Co-gain and co-loss of genes lead to patterns of co-occurrence, and while multiple independent gain or loss events of pairs of genes make a strong statistical case for a functional link (Barker and Pagel, 2005), two co-occurring genes may be present together simply because the organisms that encode a copy of these genes are closely related, phylogenetically. In order to correct for phylogenetic distance when examining phyletic patterns, Barker and Pagel (2005) introduced a method of detecting independent instances of co-gain and co-loss of pairs of proteins, i.e., those not due to one co-event in a common ancestor, and find them to be strong predictors of functional linkage of yeast proteins. In a more recent study, the same group compared various methods of finding functional links based on co-gain and co-loss of proteins and introduced an algorithm that estimates the rate at which independent gene gain should be fixed for the particular data set used (Barker *et al.*,

2007).

A group of nucleo-cytoplasmic large DNA viruses (NCLDV), which include poxviruses, was recently studied in detail (Iyer *et al.*, 2006) and some co-occurrence and complementarity patterns were observed. For example, NAD^+- and ATP-dependent ligases are present in distinct subsets of poxviruses, which led the authors to conclude that they probably play equivalent roles in DNA replication. Photolyase and uracil DNA glycosylase are found to co-occur in poxviruses and mimivirus and are both involved in DNA repair. Although Iyer *et al.* (2006) only used five different poxvirus genomes, the method is valid and useful and when we now include sequences from all currently sequenced genomes, the phyletic patterns of the two ligases are still complementary. Ettema *et al.* (2001) combined events of co-occurring gene gains and losses with the genomic location of genes to make functional interaction predictions in *Pyrococcus*.

From these and other reports, it is clear that using phyletic patterns to gain evolutionary insight is an important area of research. There is a database collecting such patterns based on Ensembl proteins and organisms (Hulsen *et al.*, 2006, 2008).

There are many genes of unknown function in poxviruses. Even in vaccinia virus strain Copenhagen, often used as a reference genome and well studied, 50 out of 199 genes fulfill a function that is not known and have no domain or structural information listed.

Here, an analysis of all poxvirus genes is carried out. The aim of the first part of this research project was to investigate whether it is possible to predict protein-protein interactions in poxviruses by examining the phyletic patterns (presence or absence in all genomes in the study) of all proteins. For example, if some proteins known to be involved in a certain pathway are found to co-occur in those genomes that use this pathway, and if a further, perhaps unknown, gene is present with them and absent where they are

absent, this could suggest a role of the new gene in the known pathway, and could thus further our knowledge of poxvirus metabolic pathways and reduce the number of genes with unknown function. Instead of functioning in different parts of the same pathway, groups of co-occurring genes may also code for protein complexes, whose components similarly would need to be present in the same genome to function together. Conversely, if a group of genes exhibit a phyletic pattern complementary to one corresponding to a certain pathway, it is possible that these genes encode a complementary function to that pathway.

Pellegrini *et al.* (1999) used a similar approach to uncover unknown functions of bacterial genes. They clustered proteins with similar phyletic patterns and show that they tend to occur in the same pathways and are thus functionally linked, as was the case, for example, for a group of flagellar proteins. Taking the reverse approach, they also showed that proteins known to be functionally linked often have similar patterns of presence and absence.

In the second part of this chapter, I describe how this latter approach to linking phyletic patterns and protein-protein interactions was also carried out: noting the patterns of poxvirus genes whose protein products are known to interact with each other in order to investigate whether the observed patterns could be used to find further interacting partners. More than merely looking at the phyletic patterns to explore trends for interacting protein pairs, patterns of location, sequence conservation and temporal expression category were also taken into account.

4.2 Methods

The protein families identified as described in Chapter 2 were used to construct a matrix of presence and absence of every family in every genome. This matrix contains a column of ones (presence) and zeros (absence) for each protein family, with rows corresponding to genomes (Table 4.1). All

	Gene Family				
Organism	**1**	**2**	**3**	**4**	**5**
MsEPV	1	1	0	0	1
VARV	1	0	1	1	0
VACV	1	1	1	0	1
CPXV	1	0	1	1	0
MYXV	1	1	0	0	1

Table 4.1: Sample Matrix of Presence and Absence

104 complete genomes were included in this study (listed in Table 1.1 on page 25). The three definitions of protein families described in Section 2.3 were used for the following analyses, and numbers refer to set 1 (set2, set3), respectively.

4.2.1 From Patterns to PPI

In this part we attempted to infer PPI based on co-occurring protein families, or functional complementarity for protein families with complementary patterns.

Those protein families that occur in all 104 genomes are disregarded, as they are most likely essential for virus survival and spread and their co-occurrence with another family would not be of functional significance. For the following analyses, each possible pair of remaining families was examined for similar or dissimilar patterns of presence and absence.

4.2.1.1 Co-occurrence

Here, monophyletic families are also discarded, as co-occurrence of two proteins that are only present in, for example, all orthopox viruses may just reflect a genus-specific pair of proteins, rather than a functionally significant relationship. Different measures were used to assess the co-occurrence of a pair of proteins.

Perfect Co-occurrence The following equation is adapted from a paper on influenza H3N2 evolution (Du *et al.*, 2008),

$$Score = \frac{f(1_A 1_B)^2}{f(1_A) \times f(1_B)}$$

where $f(1_A)$ is the frequency of protein A, $f(1_B)$ is the frequency of protein B and $f(1_A 1_B)$ is the frequency of co-occurrence of protein A and protein B. The score will equal one if both proteins co-occur wherever one of them is present, and this, while allowing for genomes that code for neither protein, is considered perfect co-occurrence.

Simple Co-occurrence In this method, the evolutionary distance of two genomes is taken into account. A similarity matrix was obtained using the *protdist* program of the Phylip package (Felsenstein, 2005). The core families are concatenated and aligned and *protdist* calculates a similarity value between 0 and 1 for each pair of genomes based on this alignment. The distance score used below is defined as 1 minus this similarity value. The simple co-occurrence score is calculated by adding the distance scores for all genome pairs in which two proteins co-occur and dividing the sum by the number of genome pairs. For example, if proteins A and B co-occur in genomes 1, 3 and 4,

$$Score = \frac{Distance(1 \leftrightarrow 3) + Distance(1 \leftrightarrow 4) + Distance(3 \leftrightarrow 4)}{3}.$$

This measure can never give a score of 1 as no two genomes have a similarity score of 0, and it is therefore always smaller than the perfect co-occurrence score above. A third measure used was the product of the simple and perfect co-occurrence scores above and this gives the lowest score of the three for a given pair of proteins. These two distance-based measures were only used in preliminary analysis (i.e. the 2006 data set), and did not yield any

additional information to the first type mentioned above. They are therefore not discussed further here.

BayesTraits PPI were then predicted using the algorithms provided by Barker and Pagel (2005) and Barker *et al.* (2007). This method is based on maximum likelihood statistics and incorporates the phylogeny of the genomes examined. It compares the likelihoods of observing the given data on presence and absence of proteins under independent evolution and under correlated evolution, using Markov models.

Monophyletic protein families were not excluded for this analysis, as the algorithm is expected to recognise the lack of independent co-gain or co-loss between such families. The input needed included a phylogenetic tree and positive and negative sample data, i.e., protein pairs that do and do not interact. The NJ tree was drawn with Clustal W based on the 28 core families of set 1 (Section 2.3) and the positive and negative data sets were taken from the Poxvirus Bioinformatics Resource Center (PBRC, www.poxvirus.org) as described in Section 4.2.2.1 below. The negative set included proteins involved in unrelated functions, such as transcription-associated proteins and those involved in host defense modulation. However, they were then pruned to exclude minor ORFs (short open reading frames that are contained within longer ORFs on the complementary strand and do most likely not correspond to functional protein products), core and rare (only present in one genome) proteins, and then contained just 18 (29, 30) known interacting and 104 (182, 202) non-interacting protein pairs, respectively, in the three family sets. BayesTraits and the connected algorithm BMS_runner were run with varying input parameters (rate of gene gain r and likelihood ratio cutoff) and the predicted PPI compared. The algorithm outputs measures of sensitivity and specificity and these suggested that a rate of gene gain of 6 would yield the most reliable results, albeit with very low specificity values, due to our tiny training sets, as specificity is defined here as proportion of predicted

pairs that are positive pairs. However, it was decided that not restricting the rate of gene gain was biologically the most sensible here, as, unlike in eukaryotes, used in the study by Barker *et al.* (2007), multiple gains of the same gene in different parts of the phylogeny are plausible in viruses, and indeed have been inferred in Chapter 3.

In order to choose the likelihood ratio cutoff for this algorithm, several values were evaluated based on the number of protein links that would be predicted above this threshold value and the number of proteins involved in forming these links. Figure 4.1 shows how a threshold was chosen that predicts as many PPI proteins as possible without the number of interactions they form increasing too much, as the prediction of large amounts of unmanageable and potentially nonsensical interactions, for example, between large groups of monophyletic proteins, had posed a problem in preliminary analyses. The initial results produced very large PPI networks where everything seemed to be part of one big interaction system. We did not find these results credible and decided to examine the patterns within this large network to identify reliable sub-networks of PPI.

To refine the large PPI networks predicted by this method, the Markov Cluster Algorithm (van Dongen, 2000, MCL) was used to find small clusters of PPI, as described in Chapter 2. Furthermore, weighted linked pairs were used as input to MCL, so that each predicted PPI was given a weight, W, according to the percentage of neighbouring nodes (interacting proteins, referred to as "neighbours" in this analysis) they share, calculated by the following equation:

$$W = \frac{N_s}{\frac{(N_A-1)+(N_B-1)}{2}}$$

N_s is the number of shared neighbours, and N_A, N_B are the numbers of neighbours of protein A and B, respectively. This weighting scheme will assign more credibility to protein pairs that are part of a highly connected PPI cluster. For example, proteins A and B in Figure 4.2 share more of their

neighbours than do proteins K and L, and this is reflected in their weight. This measure is related to the clustering coefficient of a node, explained below.

After experimentation, the value of the MCL inflation parameter was picked based on the average clustering coefficient C of the protein nodes that result from the application of MCL. The clustering coefficient expresses how close a node and all its neighbours are to being a completely linked graph. It was defined, by Watts and Strogatz (1998), as the proportion of links between the nodes within its neighbourhood divided by the number of links that could possibly exist between them. It is expected to reflect the real biological relationship between the proteins. We chose here the highest average C of all nodes at different values of the inflation parameter. The resulting clusters were then manually pruned, as they included some large unmanageable clusters.

To check whether the large number of variola isolates confounded the PPI predictions thus reported, similar analyses were carried out excluding all but one variola isolate and another only including variola isolates. In the former data set we include only representatives of the other large OPV subgroups MPXV and VACV. Strains MPXV-USA44, VARV-BEN68 and VACV-WR were used as representatives, resulting in 38 genomes overall. In the latter data set, all variola strains were used, as well as TATV, as an outgroup, resulting in 50 genomes. Likelihood ratio cutoff and MCL inflation parameter were decided as described above.

4.2.1.2 Complementarity

The 2006 data set was also used for examining complementary phyletic protein patterns. This preliminary analysis indicated difficulties where large numbers of proteins present in mutually exclusive monophyletic groups of genomes were complementary to each other. Any functional significance

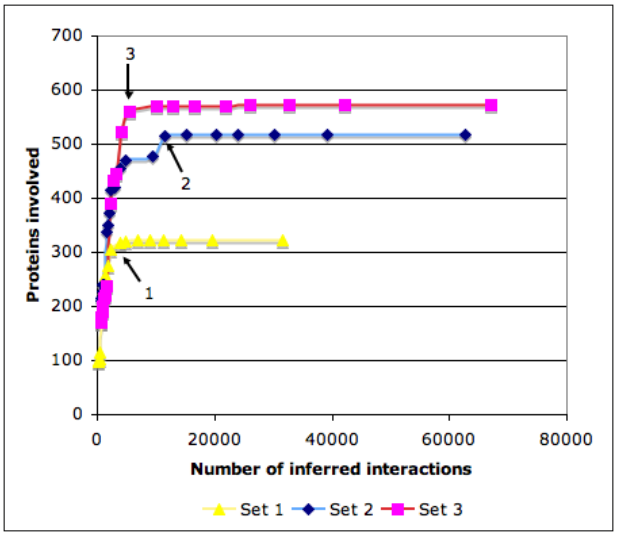


Figure 4.1: Choosing a Likelihood Ratio (LR) Threshold. Different cutoff values for the LR lead to varying numbers of proteins predicted to participate in varying numbers of protein-protein interactions. Arrows indicate the threshold that was chosen for each of the three family sets. Where the graph flattens out, the algorithm is inferring greater numbers of interactions between a fixed set of proteins and we interpreted this to be spurious. Set 1, 2 and 3 refer to the family classification sets.

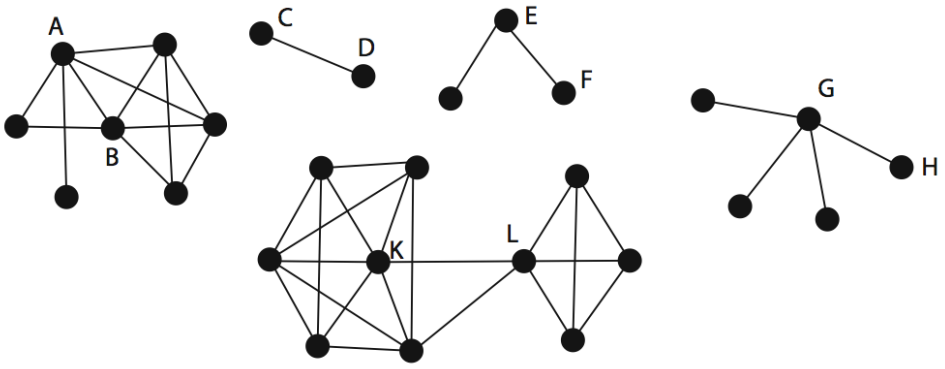


Figure 4.2: Sample PPI groups. According to the equation given on page 88, proteins A and B have a weight of 0.67, and proteins K and L have a weight of 0.22. The weight of a single isolated PPI like that between proteins C and D is defaulted to 1, as they share 100% of their neighbours. The weight of links like that between proteins E and F or G and H, i.e., where one partner only has one neighbour, is defaulted to 0.1.

would have been very difficult to spot and so this complementarity search was not repeated for the 2008 data set.

4.2.2 From PPI to Patterns

4.2.2.1 The PPI Dataset

Reported protein-protein interactions in vaccinia were taken from McCraith *et al.* (2000), which were identified by a yeast two-hybrid analysis, and also from the PBRC (www.poxvirus.org), where many vaccinia virus strain Copenhagen (VACV-COP) genes are annotated. A combination of these sources yielded a set of 81 PPI. 23 came from McCraith *et al.* (2000), excluding self-interactions (12) or those that involve minor ORFs (1) or unnamed protein products (1). The reasons for these exclusions are inherent in this project: self-interactions can never be predicted from the patterns of presence of two genes, which we are trying to do here. The unnamed gene could not be assigned to any gene used in this thesis. Table 4.2 lists the PPI used and the annotated functions of genes are shown where they are known.

As all of these PPI were annotated in VACV-COP, the family classifications described in Section 2.3 were used to identify the homologous genes in other genomes and thus the phyletic patterns. The patterns and locations of these interacting partners were then analysed in numerous ways, described below, in order to test whether they show any special characteristics that might allow us to predict other interacting pairs of proteins that share similar characteristics. Where a Z-score is mentioned, this was calculated using the equation

$$Z = \frac{x_i - \mu}{\sigma}$$

where x_i is the observed data point, μ is the mean and σ the standard deviation of the distribution. Associated p-values indicate the proportion of the normal curve that is more extreme than this calculated deviate.

Where random pairs are picked to investigate if the PPI pairs are a representative sample of all possible protein pairs, we pick only from those families that have an orthologue in VACV-COP.

Previous misannotations of vaccinia virus strain Tian Tan ORFs, reported by Upton *et al.* (2003), were corrected in our phyletic pattern matrix.

Table 4.2: Protein-Protein Interactions

Protein 1	**Protein 2**	**Function**
Identified by McCraith et al. *(2000)*		
A20R	H5R	DNA replication
A20R	D5R	DNA replication
A20R	D4R	DNA replication
A1L	G8R	transcription
F4L	I4L	ribonucleotide reductase subunits
A23R	A8R	VITF-3 subunits
A49R	H5R	transcription
D7R	A32L	transcription
F8L	L4R	transcription
G2R	H5R	transcription
H5R	B1R	transcription
A12L	A19L	unknown
A14L	A40R	unknown
A32L	A11R	unknown
A45R	A4L	unknown
J1R	A45R	unknown
A26L	A42R	unknown
K1L	C10L	unknown
E7R	A39R	unknown
Continued on next page . . .		

Protein 1	**Protein 2**	**Function**
A21L	A6L	unknown
E8R	A51R	unknown
F15L	D9R	unknown
I2L	A40R	unknown
Inferred from PBRC annotation		
E9L	A20R	DNA replication
G4L	E10R	SS-bond formation pathway
E10R	F9L	SS-bond formation pathway
E10R	L1R	SS-bond formation pathway
A10L	A4L	core protein subunits
E1L	J3R	poly(A) polymerase subunits
D1R	D12L	mRNA capping enzyme subunits
A17L	A14L	unknown
A14L	F10L	unknown
A14L	H1L	unknown
A33R	A36R	unknown
L5R	H2R	entry-fusion complex
L5R	A21L	entry-fusion complex
H2R	A21L	entry-fusion complex
G7L	J1R	virosomal uptake complex
G7L	A30L	virosomal uptake complex
J1R	A30L	virosomal uptake complex
D6R	A7L	VETF subunits
E4L	J4R	RNA polymerase subunits
E4L	J6R	RNA polymerase subunits
E4L	H4L	RNA polymerase subunits
E4L	D7R	RNA polymerase subunits

Continued on next page ...

Protein 1	**Protein 2**	**Function**
E4L	A5R	RNA polymerase subunits
E4L	A24R	RNA polymerase subunits
E4L	A29L	RNA polymerase subunits
J4R	J6R	RNA polymerase subunits
J4R	H4L	RNA polymerase subunits
J4R	D7R	RNA polymerase subunits
J4R	A5R	RNA polymerase subunits
J4R	A24R	RNA polymerase subunits
J4R	A29L	RNA polymerase subunits
J6R	H4L	RNA polymerase subunits
J6R	D7R	RNA polymerase subunits
J6R	A5R	RNA polymerase subunits
J6R	A24R	RNA polymerase subunits
J6R	A29L	RNA polymerase subunits
H4L	D7R	RNA polymerase subunits
H4L	A5R	RNA polymerase subunits
H4L	A24R	RNA polymerase subunits
H4L	A29L	RNA polymerase subunits
D7R	A5R	RNA polymerase subunits
D7R	A24R	RNA polymerase subunits
D7R	A29L	RNA polymerase subunits
A5R	A24R	RNA polymerase subunits
A5R	A29L	RNA polymerase subunits
A24R	A29L	RNA polymerase subunits
I3L	F4L	unknown
A34R	A36R	unknown
A34R	B5R	unknown

Continued on next page ...

Protein 1	Protein 2	Function
A56R	F13L	unknown
A27L	A14L	unknown
A27L	A17L	unknown
J3R	H4L	unknown
H2R	A28L	unknown
H5R	F10L	unknown
F10L	A30L	unknown
F10L	E8R	unknown
A30L	H1L	unknown

Potential Herpesvirus Homologues As poxviruses and herpesviruses are closely related dsDNA virus families, and herpesvirus genomes have been studied more extensively, there was a possibility of having further known PPI available by using poxvirus homologues of herpesvirus PPI. Uetz *et al.* (2006) published results of a yeast two-hybrid analysis, which included lists of 123 and 173 interactions between proteins from Kaposi sarcoma-associated herpesvirus (KSHV) and varicella zoster virus (VZV), respectively. When searching for poxvirus homologues of the proteins involved in these interactions, most cannot be found, some have no homologues in VACV-COP, a prerequisite for being included in our known PPI set, and one interaction, formed by the subunits of ribonucleotide reductase, is already included in the set listed above. This leaves only two potential additional interactions that could be included in our list through their homology with KSHV, but it was decided to forgo this limited set for the sake of keeping the definition of the known poxvirus PPI clearer.

4.2.2.2 Protein Conservation

At first, the level of conservation of a protein was categorised as either PV (present in all poxvirus genomes), ChPV (present in all ChPV), OPV (present in all OPV), VacVar (present in all vaccinia and variola strains), Vac or fewer (each family is at least present in VACV-COP). This is illustrated by the first line, definition 1, in Figure 4.3. Then, a second definition was also used, one naming the common ancestor of a group of genomes that encode the protein in question, as illustrated by definition 2 in Figure 4.3. For example, if a protein family is present in all vaccinia strains, one variola strain and ectromelia virus, it was labelled as Vac (all VACV) for definition 1, whereas for definition 2 it was labelled as OPV, as the protein family first originated in an orthopox virus at the latest (protein A in Figure 4.3). PV, in definition 2, denotes a protein family present in at least one EPV and one ChPV (protein B); ChPV, in definition 2, means it is present outside OPV but not outside ChPV (protein C) and so on.

4.2.2.3 Sequence Conservation

The evolutionary distance between genomes is determined using *protdist* of the Phylip software package (Felsenstein, 2005), which takes the concatenated alignment of all the core families described in Chapter 2 as input and gives a similarity table as output, where each pair of genomes has an associated score between 0 and 1, based on the level of sequence conservation of the core families within them. This serves as an approximation of the similarity between the entire genomes. Where an indication for the distance between a group of genomes is needed, the smallest similarity between any two genomes in the group is chosen to represent this. In order to show that the conservation of core families is a good indication of evolutionary distance between whole genomes, we plotted the *protdist* similarity value of any two poxvirus genomes against their average fraction of shared orthologues (in family set 1) and

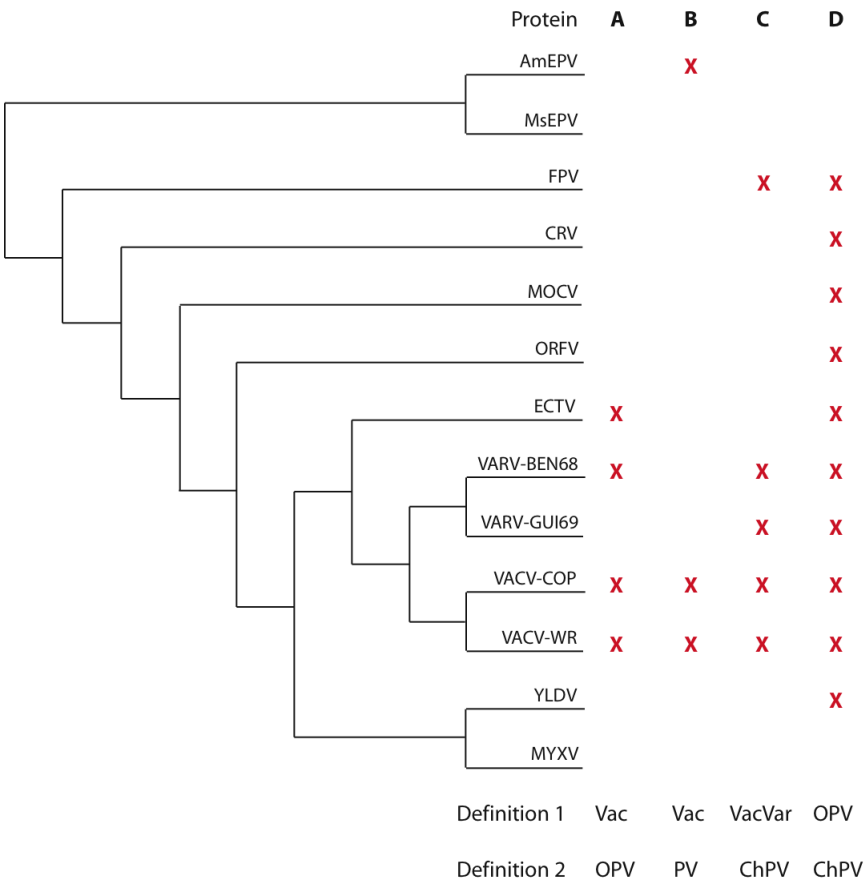


Figure 4.3: Definitions of conservation. Shown is a tree of representative poxviruses, and presence of a protein is marked with a red X. Definition 1 is based on the presence of a protein in all members of a clade, while the new definition reflects where the protein first originated.

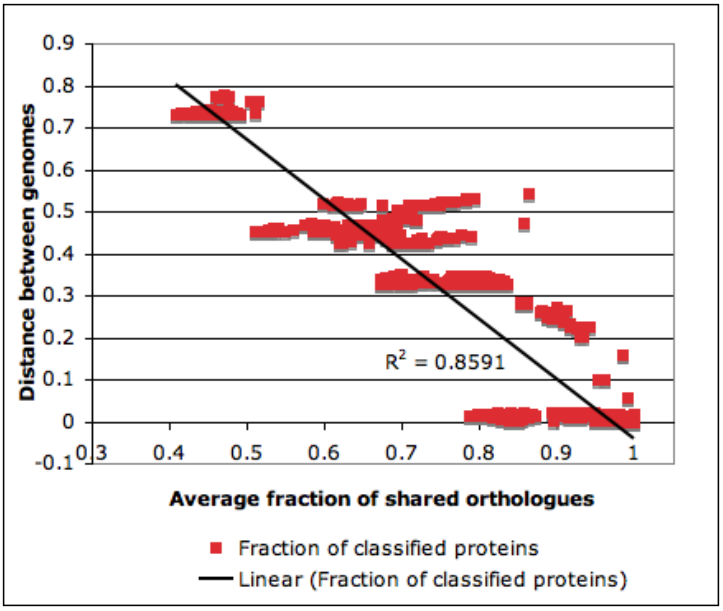


Figure 4.4: Shared orthologues as a measure of evolutionary distance. The x-axis value is the average of the proportion of shared orthologues (a fraction of all proteins that could be classified into families) in both genomes involved. The y-axis value is distance (i.e., 1 minus the similarity value of the two genomes as determined by *protdist*). R^2 is a measure of correlation and the high value of 0.86 indicates a good fit of the line to the data.

the resulting graph can be seen in Figure 4.4. Huynen and Bork (1998) and Snel *et al.* (1999) have shown that the fraction of shared orthologues is an acceptable measure of evolutionary distance and decreases as distance increases. The high correlation value, R^2, indicates that the similarity value is a good measure for evolutionary distance, too. The reason for the layered appearance of data points in Figure 4.4 is due to the groups of closely related genomes, as described in Section 4.3.3.2 below.

4.2.2.4 Location

The location of proteins in the genome is measured and expressed as a percentile location, because poxvirus genomes are of varying length. For example, where the protein in question is encoded by gene number 40 (from the 5' end) out of 200 genes, its percentile location is 0.2. Where a measure

of the location of many family members over a range of genomes is needed, the average of these percentile measurements is taken. The distance between two proteins was measured as the number of intervening genes plus 1 and also as distance in kb.

4.2.2.5 Expression Category

Poxvirus genes are expressed in three distinct phases: early, intermediate and late, as described in Section 1.2.3. For 177 VACV-COP genes, the PBRC lists some evidence showing which temporal class they belong to, and these genes were used to determine whether the proteins involved in PPI belong to the different classes in the same proportions as all VACV-COP genes, and whether PPI were preferentially formed between genes expressed during the same phase. The PBRC website lists this information either as "known" - based on published experimental results - or "predicted" - based on prediction of early and late transcriptional promoters. Where both types of information are available, I have chosen to use only the "known" expression category. A χ^2-test and a Z-score test were carried out to analyse this expression data.

4.3 Results

One of the first steps in this analysis of phyletic patterns was to check for proteins that are present only in genomes that infect one particular host or group of hosts, even if the viruses are not closest phylogenetic neighbours. The idea was that these genes, perhaps, are required in order to invade the immune system of this particular host. When investigating this for VARV and MOCV (human host), CPXV and BPSV (bovine host), ORFV, SPPV and GTPV (caprinae host), MPXV and yatapox (monkey host) and RPXV and leporipox (rabbit host), there are no proteins that are exclusive to these groups of genomes.

4.3.1 Co-occurrence

When searching for perfect co-occurrence of proteins, 7 (24, 20) groups in family set 1 (2, 3) of co-occurring proteins were found and examined for biologically interesting connections. The results for family set 1 are summarised in Table 4.3. Only 6 of the 12 interactions from set 1 were found in set 2, and only 2 in set 3. This is due to the fact that the families are smaller and more conservative in sets 2 and 3 so that more perfectly co-occurring groups are found, and many of the large set 1 families are split in sets 2 and 3.

Predicted functions were taken from the PBRC, annotations of any members of the family in question and BLAST searches for annotated homologues. Sometimes, the non-monophyletic co-occurring pattern is due to concerted loss of the partners in one genome only. In such cases, where possible, annotation mistakes were excluded as an explanation for this pattern by examining the nucleotide (nt) sequence at the relevant location, indicated by the conserved syntenic environment of closely related genomes. Often, such examination of the nt sequence showed how differing annotation standards, e.g., a difference in threshold size for an ORF to be considered a likely gene, can cause confounding phyletic patterns. The genomes of the monkeypox clade, for example, having been sequenced by two different groups, in several cases, on closer inspection, proved to be not as different in gene content as genome annotation would suggest. Some proteins thus show unusual co-occurring patterns, which is often due to annotation practices rather than biological relevance but some examples of multiple losses in a single genome leading to the observed pattern remain.

We examined the functions of the genes with predicted interactions to gauge the plausibility of the inference. The first prediction of interaction is between the viral chemokine binding protein (ChK-BP) C23L (as well as its ITR duplicate B29R) and the protein D8L. C23L prevents chemokines from

mediating migration of immune cells into sites of inflammation. The enzyme p38 MAP kinase has been implicated in stabilising both chemokines and carbonic anhdyrase CA2 in humans (Frevel *et al.*, 2003). Some homologues of D8L, co-occurring with C23L/B29R, are annotated as CA-like, and D8L itself is an IMV membrane protein that binds cell surface chondroitin and may affect viral entry. CAs are involved in a wide range of processes like pH regulation, respiration, secretion of electrolytes, carcinogenesis, signalling, and, in humans, macrophage activation. It is plausible that ChK-BP and CA could interact in some of these processes and this hypothesis is worthy of further examination.

Another predicted interaction that is potentially interesting is that between CNPV049, CNPV024 and CNPV096, which encode a leucine rich repeat (LRR) protein containing the $C_3H_2C_3$ RING finger motif, lipase, and ubiquitin, respectively. Both LRR proteins and ubiquitin are involved in the ubiquitination pathway, which targets proteins for degradation. This occurs via an F-box protein, which recognises a substrate and links it to a ubiquitin ligase, which then transfers ubiquitin to the target protein. F-box proteins often include an LRR protein-binding domain (van Buuren *et al.*, 2008). As viruses have been shown to exploit the cellular ubiquitination system, (Nerenberg *et al.*, 2005; Iyer *et al.*, 2006), for example by encoding RING finger proteins acting as ubiquitin ligases (Zhang *et al.*, 2009b), it is plausible that these proteins do indeed work together and that this is a genuine case of co-occurrence due to functional interaction.

There is currently not enough information available to make similar statements for the other interactions in Table 4.3.

The methods based on algorithms by Barker *et al.* (2007) predicted pairs as summarised in Table 4.4. The predicted PPI for family set 1 are visualised in Figure 4.5. Of the 78 proteins involved in the predicted PPI, 28 have no functional information available. This highlights the

difficulty in this approach: Trying to infer functional relationships when many proteins carry out an unknown function is simply unrealistic. There is some overlap between the strict co-occurrence pairs listed in Table 4.3, and this set of protein families with correlated independent loss and gain in their history. For family set 1, six protein pairs are common to both sets, 55% of the strict set, perhaps an indication of the relative success of these methods, despite the lack of substantial claims of functional relatedness made. These potential interactions are highlighted in red in Figure 4.5. **Looking at these predicted pairs in more detail reveals some potentially interesting interactions. Most notably, three host defense modulators are connected through their presence in OPV and some clade II genomes: Toll-like receptor N1L, IFN-γ-receptor B8R and IFN-α/β-receptor B19R. As the mechanisms in which these proteins are involved are rather well understood, however, it is likely that their links here merely represent similar strategies of OPV and clade II genomes of modulating the host's immune system. It has to be remembered here that some proteins in Figure 4.5 are connected through co-gain and co-loss that has led to complementary patterns, rather than co-occurrence.**

The large number of variola virus strains does not seem to make a huge impact on the predictions made. The main reason for this is that the closely related variola genomes share most of their proteins. In family set 1, 155 families are present in all variola strains (two of them exclusively in variola), while 259 families are absent from all variola strains, and only 23 families have mixed patterns of presence and absence in variola, mostly due to single loss or gain of a protein in one lineage.

Names	Functions	Present in
A41L/B29R/C23L, D8L	Secreted chemokine binding protein/host defense modulator, IMV membrane protein/carbonic anhydrase	OPV, leporipox
H1L, E2L, A20R	tyr/ser phosphatase, unknown, DNA pol processing factor	all but MsEPV
D2L, A6L	virion core protein, morphogenesis	all but AmEPV
CNPV049, CNPV024, CNPV096	LRR protein, lipase, ubiquitin	EPV, CNPV
SPV003/148, SPV123	TNF-α BP/MHC-I α chain-like protein, EEV glycoprotein	yata, sui and deerpox
MC027L, MC043L	both containing large non-globular domain	MOCV, parapox
FPV001/260, FPV066	C-type lectin, unknown	FPV-I, CNPV

Table 4.3: Predicted PPI based on strict co-occurrence. Names are those of the VACV-COP orthologue, where present, and of indicated genome, where not.

Set	LR cutoff	MCL -I	Proteins Involved	Pairs Formed	Clusters
Including all genomes					
1	8	5.0	78	108	22
2	7	2.2	85	260	15
3	9	3.7	130	251	25
Including only representative VARV, VACV, MPV					
1	7	5.0	308	2591	37
2	8	2.2	508	7229	31
3	8	4.2	546	7510	41
Including only VARV, TATV genomes					
1	2	1.2	19	107	2
2	2	5.0	53	542	7
3	2	1.2	57	613	7

Table 4.4: Numbers of PPI Predicted by BayesTraits

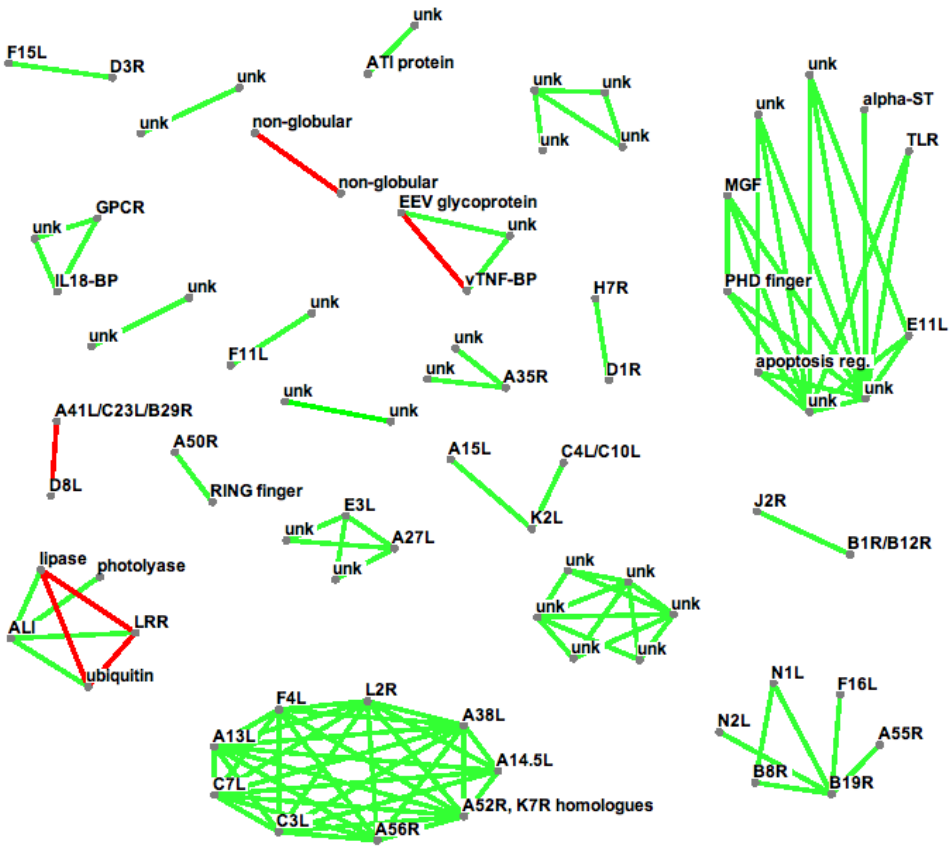


Figure 4.5: PPI clusters predicted by BMS_runner and BayesTraits as described in methods, for family set 1. Nodes are labelled according to vaccinia homologue gene name (e.g. E3L) if known or protein function if it is not. A node is labelled *unk* (unknown) if no functional information is available at all. Red interactions are those also found by strict co-occurrence methods. Graph was created with Osprey (Breitkreutz *et al.*, 2003). Abbreviations: GPCR - G-protein coupled receptor; IL18-BP - IL18 binding protein; vTNF-BP - viral tumor necrosis factor binding protein; MGF - myxoma growth factor; alpha-ST - α-2,3-sialyltransferase; ATI - A type inclusion; TLR - Toll-like receptor

4.3.2 Complementarity

From the 2006 data set, the two groups that exhibited perfect complementarity are described below. With a more lenient threshold for complementarity, more groups were found to show complementary patterns. As many of the

complementary groups fell into alternative clades (e.g. EPV vs. ChPV proteins) biological significance was hard to detect and this analysis was not pursued further.

4.3.2.1 RING Finger vs. Serine/Threonine Kinase

Six genomes (parapox, crocodilepox and MOCV) code for a family of RING finger domain proteins, while all other genomes encode a serine/threonine kinase, i.e., these proteins are found in a complementary pattern. RING finger domain proteins are a type of zinc finger protein and have been linked to the ubiquitination pathway (Barinaga, 1999). While their interleaved patterns of presence and absence are interesting, it is not clear how a RING finger protein would fulfill a function complementary to the phosphotransferase. In mimivirus, LRR proteins, which are often fused to serine/threonine kinases, are known to recruit RING finger E3 ligases as a way of interfering with ubiquitin signalling (Iyer *et al.*, 2006).

4.3.2.2 NAD^+-Dependent DNA Ligase vs. Transmembrane Protein

Three genomes (EPV and CRV) encode a NAD^+-dependent DNA ligase, responsible for joining breaks during DNA replication. This protein is absent in all other poxviruses, and they encode a different protein of unknown function. BLAST searches of members of this family of unknown function indicate homology to transmembrane (TM) proteins found in the IMV stage of the viral cycle. As this pattern of presence and absence can be achieved by one gene loss (ligase) and one gene gain (TM protein) after the divergence of crocodilepox and before the divergence of avipox (see phylogenetic tree in Figure 2.2 on page 51), there is not much convincing evidence to suggest that these proteins carry out a complementary function.

4.3.3 PPI to Patterns

When the above analyses failed to yield robust and convincing results, a different approach was tried. Our aim now was to find a special characteristic of proteins that are known to be involved in PPI; their phyletic pattern, sequence conservation, genomic location or time of expression. Perhaps by thus defining PPI partners, it would be possible to draw conclusions about any other proteins that might interact. A secondary objective was to discover how the known PPI pairs compared with those predicted by our methods and why these sets were not congruent. The known PPI interactions are shown in Figure 4.6, where the large RNA polymerase (RPO) cluster is clearly visible.

4.3.3.1 Protein Conservation

A summary of the phyletic patterns for the proteins involved in PPI is shown in Figure 4.7, where red rows indicate that both proteins of the PPI pair are encoded by core genes - they are present in all genomes. This is the first reason why many known PPI pairs were missed by our predictions, as we deliberately excluded core genes from the analysis due to their patterns being uninformative. This and the following figures relate to the PPI pairs and their protein families as defined by family set 1. Except where specifically stated, sets 2 and 3 show very similar results but as set 1 is the most inclusive, it has the greatest tendency towards showing families as highly conserved. The dataset obtained lists 81 PPI, however, one of the interactions is between E10R and a pair of paralogues (F9L and L1R), which are members of the same gene family in this analysis.

To check whether these 80 PPI pairs were a representative sample of all possible protein pairs or, in fact, there were any trends within this set of known PPI, we simulated picking the same number of pairs at random 1000 times from the pool of all protein families containing an orthologue in VACV-COP (n=213, 238, 242). From the results, summarised in Table

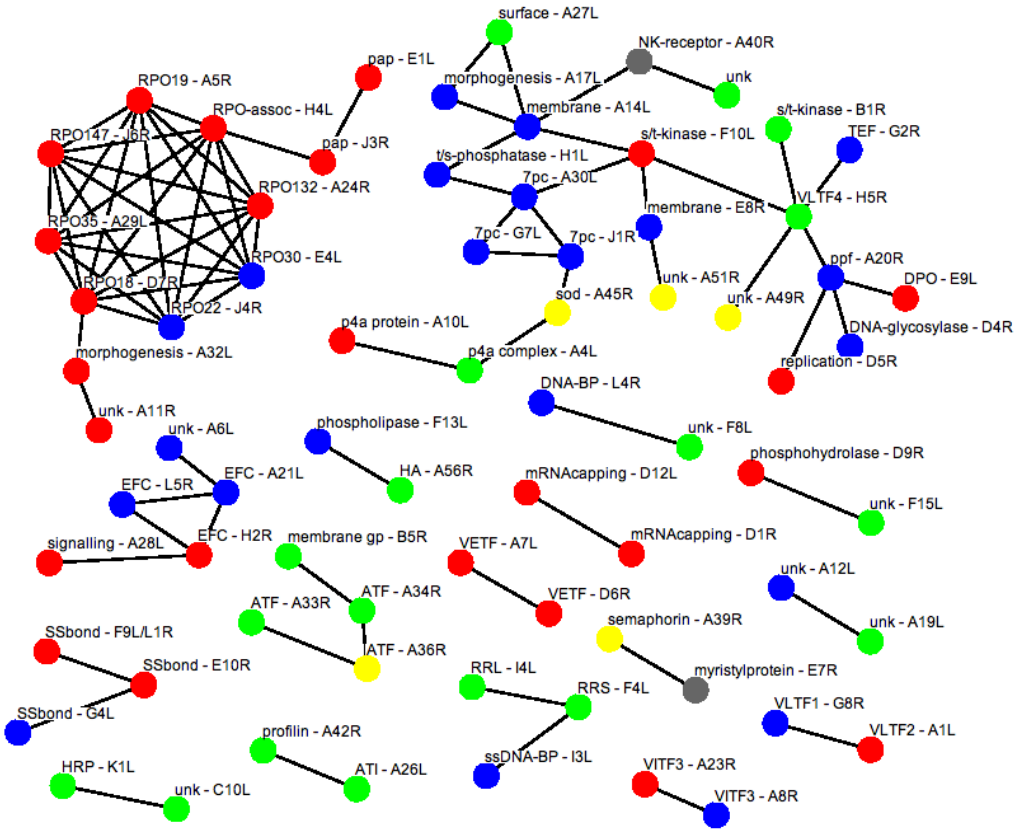


Figure 4.6: Osprey graph of known PPI. Nodes are coloured according to conservation of the protein: red for all PV, blue for ChPV, green for OPV, yellow for VacVar, grey for Vac. Abbreviations: EFC - entry fusion complex; RPO - RNA polymerase; ATI - A type inclusion protein; HRP - host range protein; ATF - actin tail formation; gp - glycoprotein; VETF - viral early transcription factor; VITF - viral intermediate transcription factor; VLTF - viral late transcription factor; ssDNA-BP - single stranded DNA binding protein; HA - haemagglutinin; sod - superoxide dismutase; 7pc - protein complex for virosomal uptake; ppf - polymerase processivity factor; DPO - DNA polymerase; TEF - transcription elongation factor; pap - poly(A) polymerase

4.5, it is clear that pairs where one or both proteins are highly conserved are less frequent than the PPI data suggest and similarly, rare pairs are more common. Specifically, pairs where both proteins are present in all PV (five red boxes), one in PV/one in ChPV (four red, one orange) and both in ChPV (four red, one white) are significantly over-represented in the PPI set as determined by a Z score significant at the $p < 0.0001$ level, while pairs where neither partner is conserved in all VACV (all white boxes) or one partner conserved in VACV and VARV (two orange, three white) are significantly under-represented in the PPI set ($p < 0.0001$). This simulation shows that the PPI pairs tend to consist of highly conserved proteins, and this is not as a result of the general distribution of highly conserved proteins. Below, and in Table 4.5, a numerical representation of patterns is used, where five digits correspond to the five boxes used in Figure 4.7. The number 2 indicates that both partners are present in the group of genomes, 1 means only one is present, 0 neither. 22222, or five red boxes, is therefore a pattern where both genes are core genes.

These analyses were then repeated, using the point of the common ancestor of each protein family as an indicator of conservation, rather than the current distribution pattern, as shown in definition 2 in Figure 4.3 on page 97. Simulations of pairs once again showed that patterns 22222, 22221 and 22220 were significantly more common among the PPI set than among randomly picked pairs, while 22200 (common ancestor of both proteins is an OPV outside VACV and VARV) was always significantly less common in the PPI data, confirming that known interacting pairs tend to have been present for more of the evolutionary history of poxviruses than expected.

The next step was to find out whether PPI tend to be formed between proteins with similar degree of conservation, i.e., with a common ancestor from the same point in the phylogenetic tree: OPV, ChPV or PV (there were no proteins with a common ancestor after the divergence of variola or

vaccinia). Given the conservation of the proteins involved in PPI and their interaction structure (as shown in Figure 4.6), the distribution of proteins in the PPI network was simulated 1000 times and the average occurrence of possible pair (PV-PV, PV-ChPV,ChPV-ChPV etc.) and triplet type (three proteins, all connected to each other) were calculated and compared to the actual PPI network. The only significant difference (at $p < 0.001$, based on Z score) between what is observed and what is expected is an increased number of PPI between PV and ChPV proteins in the real interaction data. When looking at triplet PPI motifs of three interacting proteins, there are no significant differences between the real data and the shuffled simulations. These results seem to suggest that proteins of similar conservation do not preferably interact with one another. One factor influencing the simulation data is the large RPO protein complex cluster. Within that cluster are contained 32 out of the 80 PPI, and 56 of the 59 triplet motifs, so that the proteins that are randomly chosen to participate in this cluster have a very large influence on the interaction pattern, and this leads to high standard deviation (SD) values for the simulations, making a highly significant p-value unlikely.

4.3.3.2 Sequence Conservation

Another approach to finding special characteristics of PPI partners was to look at the conservation of their sequence throughout the phylogenetic history of poxviruses, rather than just their presence. A plot of sequence conservation of a family versus evolutionary distance of the genomes can be seen in Figure 4.8. As well as the families involved in PPI, all comparable protein families are plotted, i.e., all with an orthologue in VACV-COP, to see whether these two sets form distinct groups on the graph. Distance is calculated as described in Section 4.2.2.3, conservation is calculated in two different ways as described in the figure legends. The evolutionary distances

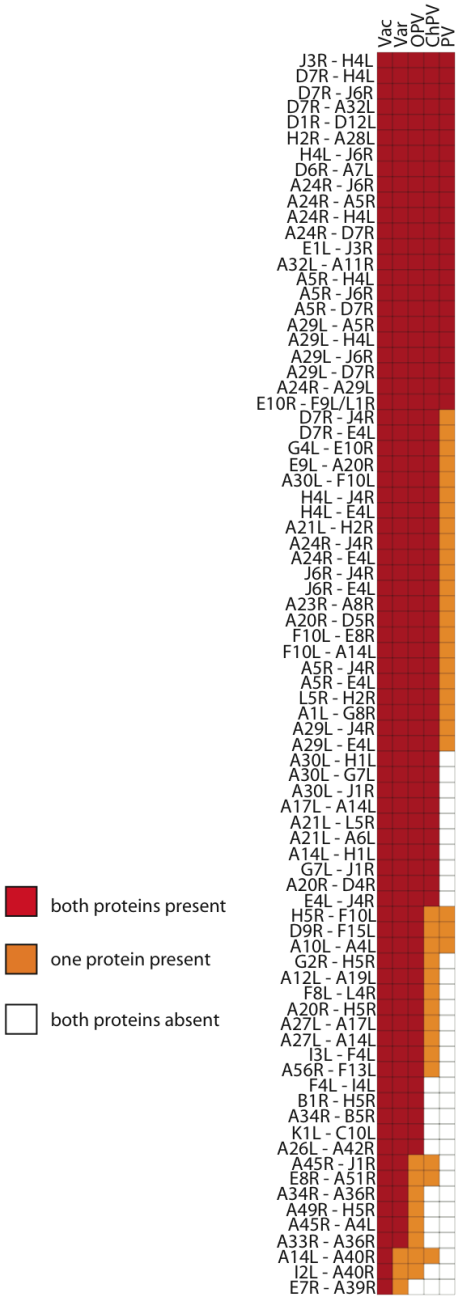


Figure 4.7: Conservation of PPI pairs in set 1. Each row corresponds to a pair and each column to a group of genomes: VACV, VARV, OPV, ChPV, PV. The colour of each block indicates whether the members of the PPI pair are conserved in this group of genomes: red means both proteins are present, orange means one of the pair is present, white neither.

	Set 1			Set 2			Set 3		
Pattern	**PPI**	**Mean**	**SD**	**PPI**	**Mean**	**SD**	**PPI**	**Mean**	**SD**
00000	0	9.6	2.28	0	12.82	2.52	0	13.62	2.65
01000	0	0	0	0	0	0	0	0	0
02000	0	0.09	0.29	0	0.12	0.34	0	0.12	0.34
10000	0	0	0	0	0	0	0	0	0
11000	0	20.73	3.3	0	22.02	3.56	1	23.52	3.56
11100	0	0	0	0	0	0	0	0	0
11110	0	0	0	0	0	0	0	0	0
11111	0	8.19	2.49	0	6.33	2.09	0	5.23	1.94
12000	0	0.33	0.55	0	0.38	0.61	0	0.43	0.62
12100	0	1.3	1.01	0	1.72	1.17	0	1.77	1.23
12110	0	0.77	0.82	0	1.02	0.93	0	1.05	0.93
12111	0	0.97	0.9	0	0.89	0.92	0	0.68	0.77
20000	0	0.12	0.34	1	0.38	0.6	1	0.38	0.6
21000	1	0.4	0.61	0	0.65	0.76	0	0.68	0.79
21100	1	1.8	1.2	3	2.97	1.55	2	3.15	1.58
21110	1	1	0.94	0	1.76	1.27	0	1.76	1.2
21111	0	1.25	1.04	0	1.56	1.17	0	1.17	1.05
22000	0	0.23	0.47	0	0.24	0.48	0	0.27	0.5
22100	4	2.37	1.4	5	2.35	1.38	5	2.49	1.35
22110	2	1.32	1.06	2	1.31	1.11	2	1.48	1.13
22111	0	1.66	1.2	0	1.19	1.02	0	0.91	0.9
22200	5	5.3	1.87	5	5.29	1.81	8	5.85	1.98
22210	8	6.24	2.12	13	6.3	2.09	10	6.62	2.21
22211	3	7.54	2.34	5	5.35	2.07	5	4.54	1.79
22220	10	1.8	1.2	15	1.81	1.19	16	1.86	1.2
22221	22	4.37	1.89	22	3.17	1.64	24	2.6	1.46
22222	23	2.64	1.4	10	1.37	1.09	7	0.83	0.84

Table 4.5: Results of pair simulation. Number patterns correspond to colours in Figure 4.7 where 2 is red, 1 is orange and 0 white. The PPI column notes the numbers of PPI pairs found to observe each pattern while Mean and SD correspond to results from picking 80 (or 81) pairs at random 1000 times. Rows in grey show a significant difference between observed (PPI) and expected (simulated mean) pair numbers, according to a Z test, $p < 0.0001$.

between families appear to be separated into four or five layers in the figures, which roughly correspond to core families (top layer), ChPV families (layers around 0.5 distance), families present in OPV and clade II viruses and, finally, families present only in OPV genomes (bottom layer), whose core families are very well conserved and thus have very small distance measurements. The conservation of families in Figure 4.8, panel B ranges from 0 to 1, whereas it is concentrated between 0.5 and 1 in panel A, due to taking the average percent identity. Both figures show how the PPI families lie among all control families. Instead of considering an entire family's sequence conservation, we also looked at pairwise comparisons of any two genes in a family (excluding paralogues) and plotted the similarity of the two sequences versus the distance of the genomes which encode them. A sample of these pairwise comparisons (n=1000) is shown in Figure 4.9.

χ^2-tests were carried out to identify whether the distribution of PPI families was significantly different to the distribution of all families. These tests show that the distribution of families into the layers of genome distance is not independent of PPI status: more than expected PPI families tend to be present in genomes that are distantly related and fewer in those that are closely related. This shows, once again, that PPI families are present in many, distantly related, genomes. When making categories for degree of conservation (bins of 10%), and comparing PPI to all VACV-COP-families, the degree of conservation also seems significantly dependent on PPI status, as more PPI families than expected are conserved to a degree of 80 - 89%, while fewer are in the 90 - 99% bin. These distinctions, however, are probably biologically meaningless, as the other observed numbers were very close to expected values and the categories were decided arbitrarily.

The χ^2-calculations corresponding to Figure 4.8, panel A are presented in Table 4.6 and all other calculations yielded very similar results.

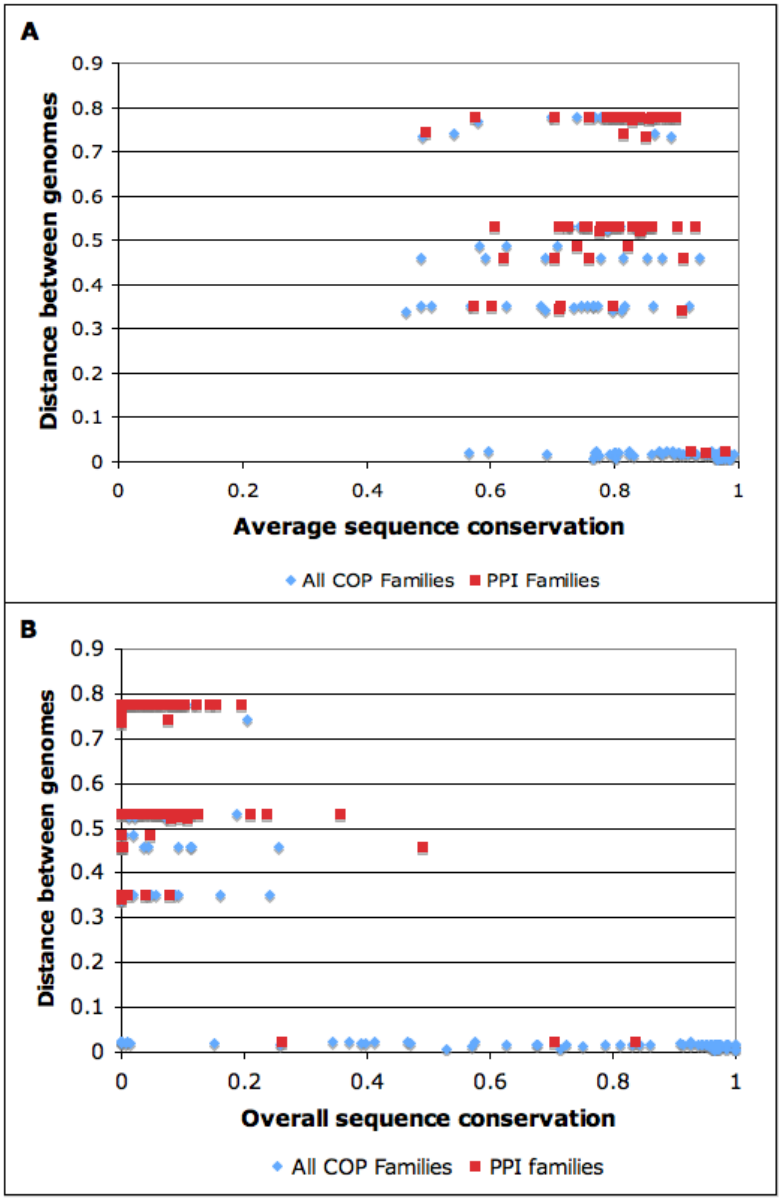


Figure 4.8: Genome distance vs. family conservation. In panel A, family conservation is the average pairwise amino acid identity between all proteins included in this family. In panel B, family conservation is the percent amino acid identity of the entire multiple sequence alignment.

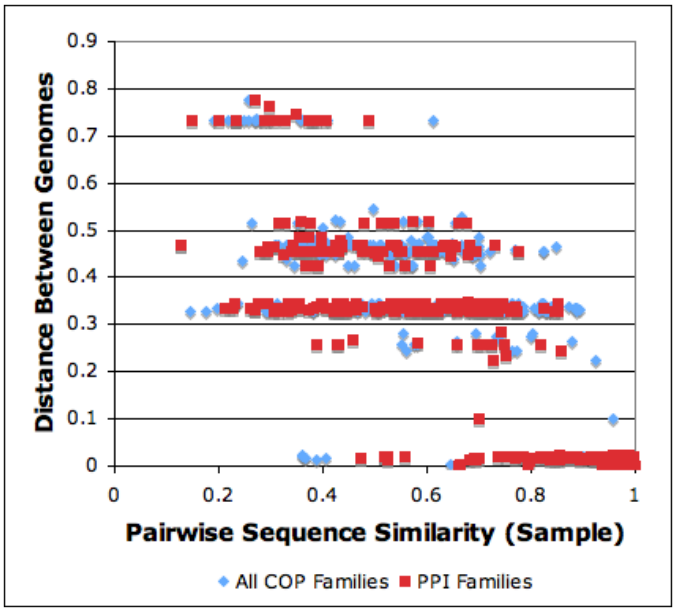


Figure 4.9: Sample pairwise sequence similarities showing genome distance versus sequence conservation.

	Layer 1 0.7-0.79	**Layer 2 0.4-0.59**	**Layer 3 0.3-0.39**	**Layer 4 0-0.1**	**Total**
All fams	26 (38.4)	25 (37.1)	18 (15.9)	72 (49.6)	141
PPI fams	32 (19.6)	31 (18.9)	6 (8.1)	3 (25.4)	72
Total	58	56	24	75	213
$\chi^2 = 54.1$			$\nu = 3$		$p < 0.001$

Table 4.6: χ^2-test for distribution of distance among PPI and all VACV-COP families. Categories were defined in terms of the clearly visible layers of genome distance in Figure 4.8, panel A. The statistically significant p-value is due to more PPI families than expected being present in genomes with high evolutionary distance to each other, and fewer in ones with small distance.

4.3.3.3 Location

Another potentially interesting characteristic of PPI families is their genomic location. Our hypothesis was that perhaps interacting protein families were preferably located close together or in a particular region of the genome.

When plotting the average relative position of all members of a family in their genomes, and comparing that to the average relative locations of the control set of all families that include a VACV-COP orthologue, it seems that PPI families are preferentially located in the centre of the genome (Figure 4.11). Of course, as those genes present over a broad range of poxviruses tend to be located near the centre of the genome (Mackett and Archard, 1979), and as we have already seen that PPI families tend to be among them, this result makes intuitive sense but unfortunately does not add any new insights into the genomic location of genes encoding proteins involved in PPI. A χ^2-test on the data presented in Figure 4.11 shows that the observed difference between VACV-COP and PPI family distributions is not statistically significant. A further illustration of which VACV-COP genes are involved in PPI is given by Figure 4.12, highlighting the stronger conservation and central location of PPI genes. The topmost horizontal line has one box for each VACV-COP gene, in the correct genomic order, coloured red if the corresponding protein is involved in a PPI. Each further horizontal line represents another poxvirus genome, ordered by increasing distance from VACV-COP, with each red box indicating that the VACV-COP gene above is present in this genome.

With several lines of evidence pointing to an ancestral presence and high sequence conservation of genes found to be involved in pox-pox PPI, there is the possibility that genes located in the periphery of the genome, which tend to be more recently acquired, are forming interactions with host proteins. Indeed, in a recent study by Zhang *et al.* (2009a), who carried out a yeast two-hybrid analysis searching for PPI between VACV-WR and human proteins, a trend for viral proteins involved in these interactions to be located away

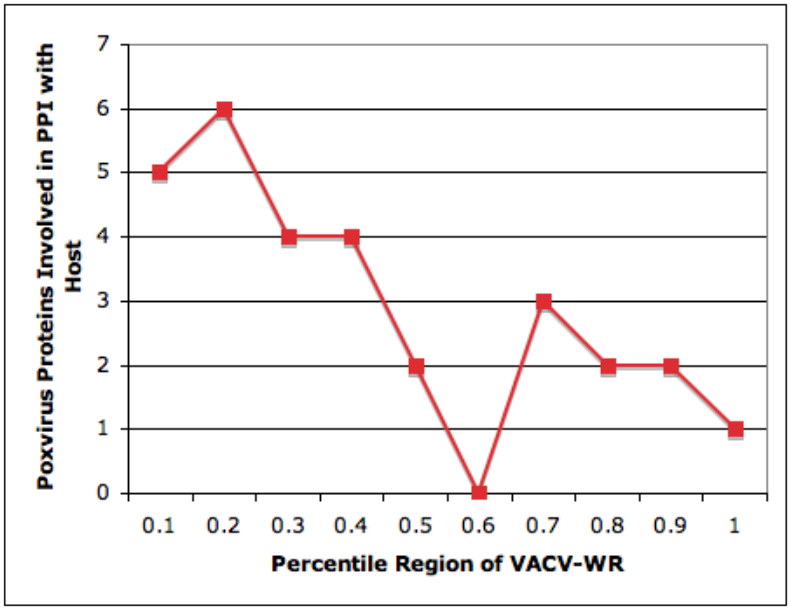


Figure 4.10: Location of genes interacting with host proteins, as detected through yeast two-hybrid analysis by Zhang *et al.* (2009a)

from the centre of the genome was observed (Figure 4.10).

Another hypothesis was that the location of PPI proteins was more conserved among different members of the family, as these proteins' interactions might be disturbed if they were involved in genome rearrangement. The standard deviations of the average percentile location of family members, as a measure of how conserved the mean location in the genome is, was therefore plotted and compared with the control families (data not shown). There were no significant differences, as determined by a χ^2-test, between these standard deviations, which suggests that the genomic location of PPI families is no more conserved across genomes than that of any family present in VACV-COP.

Perhaps the genomic distance between two proteins involved in a PPI shows a trend dissimilar to the distance between other proteins, for example being located close together? Or perhaps this distance is more conserved between PPI proteins in different genomes? To avoid confusion, we considered only those proteins that were not present in duplicate in any

of the genomes that encoded them, leaving 45 (65, 66) pairs. The average distance between two PPI proteins was then plotted, as well as the average of random same-sized sets of protein pairs (Figure 4.13). The average of distances is lower in PPI pairs, but not significantly so, as determined by a Z-score test. However, when considering individual bins of distance as shown in Figure 4.13, significantly more PPI pairs have smaller distances between them (1-30 intervening genes), than would be expected from randomly picked pairs. When protein families with paralogues are included in the analysis (choosing the shortest genomic distance between two representative genes), similar results were obtained. The standard deviation of distances, i.e., how conserved the distance between PPI partners is across genomes, was also plotted (Figure 4.14). Here, the standard deviation of random pairs was found to be similar, except for low SD (between 2 and 3) being significantly more common in random pairs.

These results indicate that interacting partners tend to be encoded by genes that are relatively close in the genome, closer than would be expected by chance. This is an interesting trend, but can be partially explained by the observation made above, that genes encoding PPI proteins are highly conserved among genomes, and such genes are mostly present near the centre of the genome. This naturally makes it more likely for two interacting partners to be located close together. In addition, read-through transcription for neighbouring genes in poxviruses is almost unheard of, let alone for genes located up to 30 genes apart, and thus it is difficult to imagine a biological reason for PPI partners to be located close together. No PPI partners are immediate genomic neighbours in VACV-COP.

The same tests were carried out using only the distance of the PPI partners in vaccinia, and measuring all previous distances in kb, rather than intervening genes. In all analyses, the results are very similar to the ones described above.

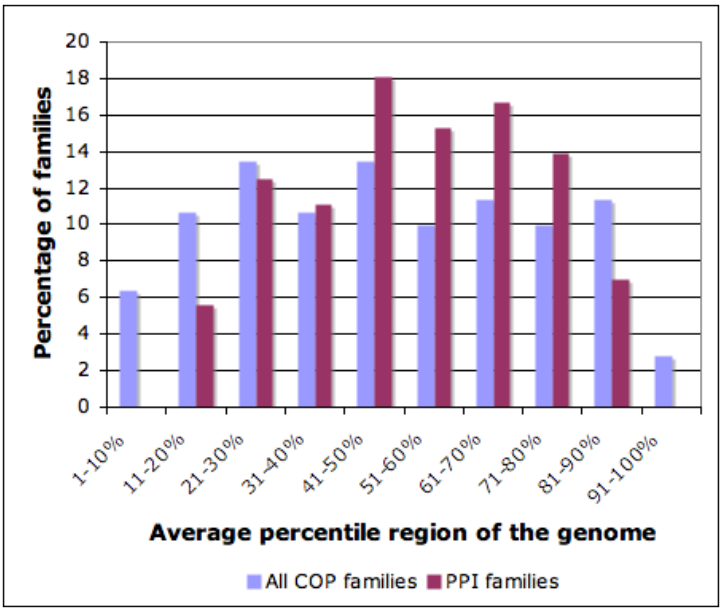


Figure 4.11: Location of PPI families expressed in percentage of families present, on average, within a certain percentile. As a comparison, all control families are presented by blue bars. PPI families seem to account for a larger proportion of central families and fewer in the terminal regions. These differences are not statistically significant.

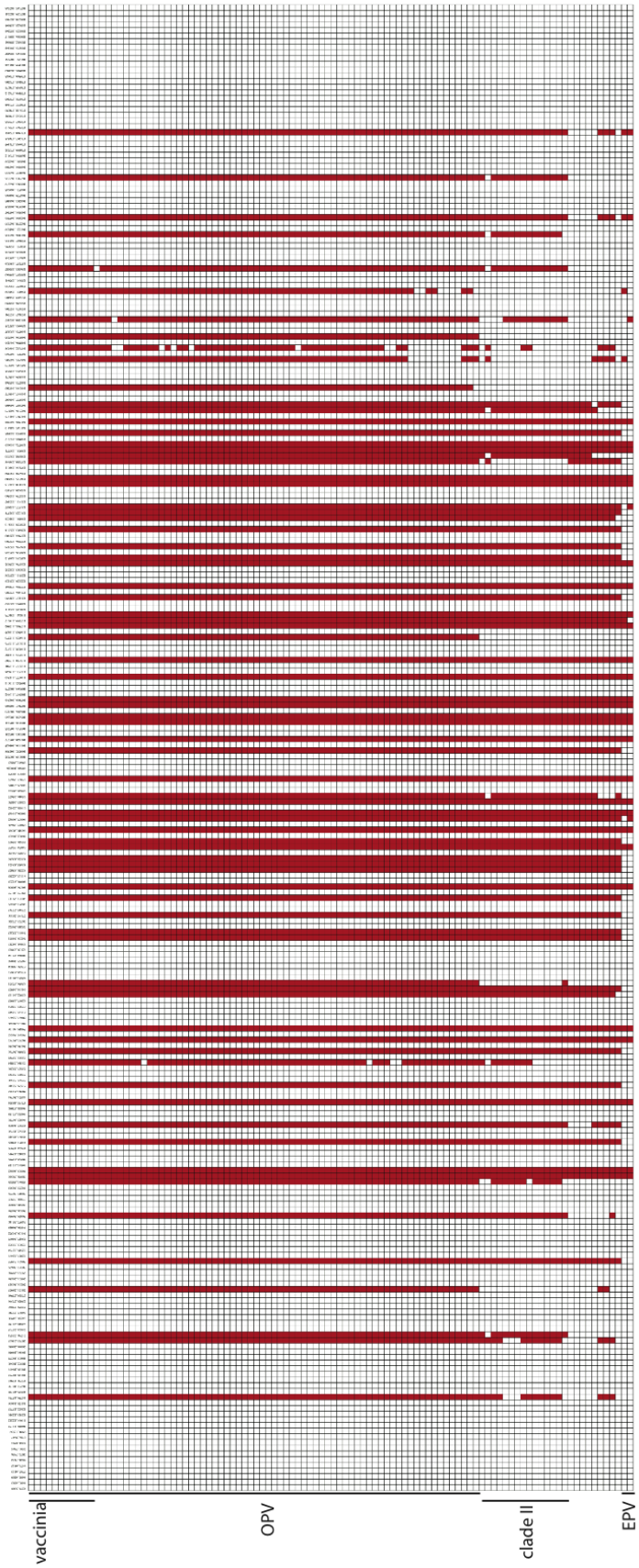


Figure 4.12: A grid-like representation of the location of PPI genes in the VACV-COP genome and their conservation among other poxviruses.

4.3.3.4 Expression Category

The 177 VACV-COP genes, for which an expression class was known or predicted, are represented in Figure 4.15. Most proteins are categorised as early or late but many have joint expression patterns, for example, where there is evidence for promoters being active for both early and late or both early and intermediate transcription. It is clear from this Venn diagram, and a χ^2-test, that the distribution of expression classes is not significantly different ($0.05 < p < 0.1$) among PPI proteins and other VACV-COP proteins. We expect that PPI should be between genes expressed at the same stage. The next step was to examine whether the interacting partners are more frequently expressed in the same temporal class than a sample of random protein pairs, as might be expected. Resampling was allowed when picking the random pairs, as many proteins involved in PPI have several interacting partners. Random pairs were picked 1000 times and average and standard deviation were determined and used in a Z-score test. This analysis showed that PPI partners indeed tend to be expressed in the same phase, as shown in Table 4.7. Early-early and late-late interactions were significantly more common in the PPI set than in random pairs, at $p < 0.001$. Biologically, this can be explained by the fact that interaction is often a physical bond, for which proteins need to be active in the cell at the same time. An early-late PPI would require stability of the early protein so that it is not degraded before the late stage. Expression at the same stage should be favoured because it does not require extra measures to avoid protein degradation. The finding that interacting proteins tend to be co-expressed adds to the credibility of these PPI defined outside poxviruses in yeast two-hybrid screens (McCraith *et al.*, 2000). Of the 16 early-early interactions, 13 are between subunits of a protein complex, and the same is the case for 8 of the 25 late-late interactions. The others are engaged in unknown interactions or function in the same pathway or process.

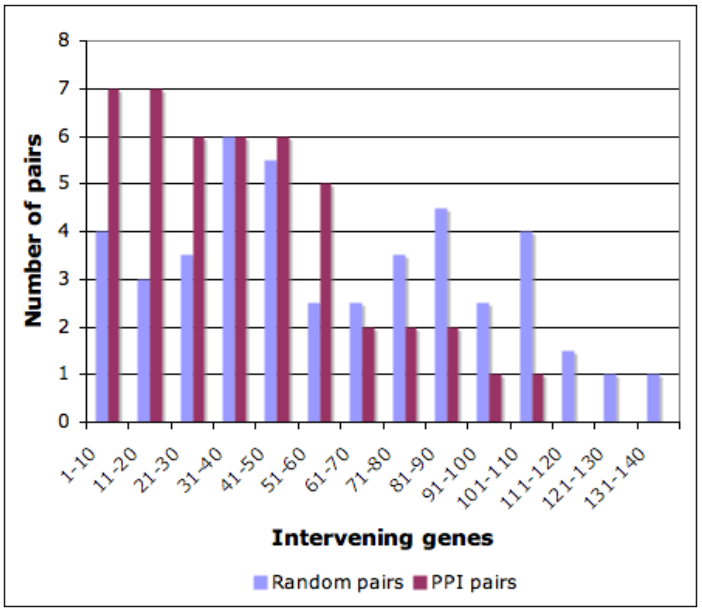


Figure 4.13: Distance between PPI partners expressed in intervening genes plus one. The blue bars correspond to the distances between simulated pairs.

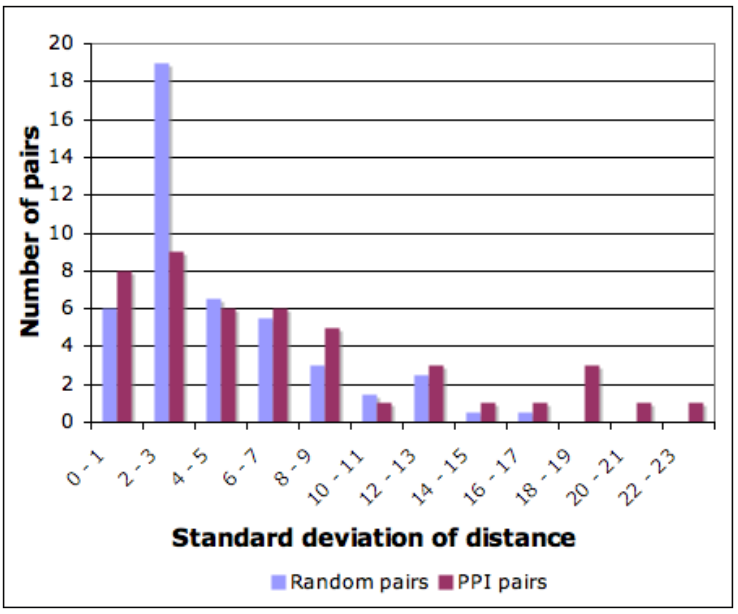


Figure 4.14: SD of mean distance between PPI partners (red) and random partners (blue).

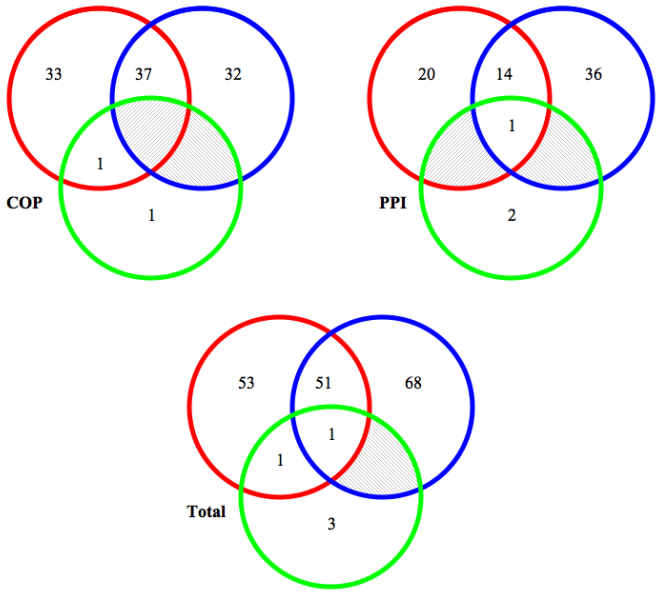


Figure 4.15: Venn diagrams of expression classes. Red circles represent early genes, blue late and green intermediate genes. The PPI part of the diagram shows those VACV-COP genes that are involved in PPI and for which an expression class is known, the COP part shows all other COP genes for which an expression pattern is known and Total combines the two parts.

This analysis of PPI pairs with respect to their expression category suggests that there might be a way of verifying predicted PPI, or at least supporting or weakening such a prediction. It is, however, likely not useful for predicting unknown PPI, as we cannot expect all co-expressed genes to directly interact. There are 1362 possible interactions between the VACV-COP proteins that are expressed early, and 2253 possible late-late interactions. These numbers exclude known PPI and self-interactions and are too large to make any meaningful predictions about which pairs might be good candidates for previously unknown protein-protein interactions.

		Random pairs		
Classes	**PPI pairs**	**Mean**	**S.D.**	**p-value**
E - E	16	7.221	2.56	0.0003
E - EIL	1	0	n/a	n/a
E - EL	14	14.021	3.31	0.4960
E - L	10	18.697	3.75	0.0102
EL - EL	4	6.56	2.49	0.1515
EL - L	10	18.098	3.78	0.0162
I - I	1	0	n/a	n/a
L - L	25	12.02	3.25	< 0.0001

Table 4.7: Expression classes of interacting partners. Shown are only interactions between classes that are actually observed in the PPI set. E=early, I=intermediate, L=late.

4.4 Discussion

Predicting protein-protein interactions proved to be very difficult in poxviruses. The phyletic patterns, sequence conservation, genomic location and expression patterns of poxvirus proteins were all examined in an attempt to identify candidate PPI pairs. Due to the large number of proteins that fulfill an unknown function, however, even robust co-occurrence and frequent independent co-gain and co-loss of proteins have not led to any clear predictions of new protein-protein interactions.

In the first part of this chapter we tried to predict PPI based on the co-occurrence of two proteins after multiple independent co-gains and co-losses, i.e., co-occurrence not due to phylogenetic relatedness of the genomes. It is now clear these attempts at PPI prediction did not yield any of the known PPI pairs. In the strict co-occurrence prediction we deliberately excluded core and other monophyletic proteins from the data set in order to avoid many false positive predictions. Pruning the sets of known PPI families thus leaves 6 (12, 11) potential pairs that could have been found by our methods but were not. The reason for missing them is that none of them co-occur perfectly. In the BMS_runner / BayesTraits analysis, monophyletic families were not excluded from the data set but the algorithm is expected to recognise

that their gains and losses are not independent. It failed to turn up 5 (10, 8) of the pairs because their likelihood ratio was below the thresholds we chose as described in Section 4.2.1.1 above. The training sets used in this algorithm should have been much larger than what we had available, with, ideally, 50 times more negative pairs than positive pairs (Daniel Barker, personal communication), which explains the lack of sensitivity and specificity. The remaining 1 (2, 3) pairs were pruned by the additional MCL algorithm that was applied to the predicted interaction network.

In the second part of this chapter, known poxvirus PPI were used in order to find some generalisms marking the proteins involved in these interactions. Characteristics under investigation were the protein conservation, sequence conservation, location and expression class of interacting proteins. Each of these analyses indicated that proteins that are known to interact with other proteins tend to be present in a wide range of poxviruses: they frequently involve core families, their representatives in VACV-COP are biased towards location near the centre of the genome and the distance between the partners tends to be smaller than between random pairs of genes, probably due to the preferential location near the genomic centre.

The conserved nature of the interacting partners is supported by their functions, where they are known. They are listed in Table 4.2 and involve basic processes of transcription, DNA replication and virus maintenance. In addition, interacting partners are expressed in the same temporal phase more often than would be expected by chance. While each of these characteristics - same expression class and high conservation - are not specific enough to produce a list of potential new interactions likely to be meaningful, combining these requirements might. Figure 4.16 shows how many new predictions can be made taking these three aspects into account: partners must come from fully conserved core families, must be expressed during the same phase as each other, and must be located near the centre of the VACV-COP genome.

As there is no clear convention for which ORFs are included in the "centre" of a poxvirus genome, Figure 4.16 shows how the numbers of predictions vary depending on the region categorised "centre". There is a maximum of 267 potential new PPI predicted, where the centre is considered to include all 262 VACV-COP ORFs, i.e., where the location is ignored and only conservation and expression are taken into account. The same interactions are found if only the central 152 ORFs are considered, suggesting that the outside 55 ORFs at each end of the VACV-COP genome do not contribute any predicted interactions. These 267 interactions exist between just 32 proteins and are arranged in two clusters, as shown in Figure 4.17. Many of the proteins are involved in the set of 81 known interactions, connected here to new partners, and again, many of them function in basic processes like transcription and DNA replication. While it is not clear whether any of these predicted PPI have a real biological grounding, the research carried out here provides a thorough analysis of characteristics of poxvirus proteins that might be useful in PPI prediction.

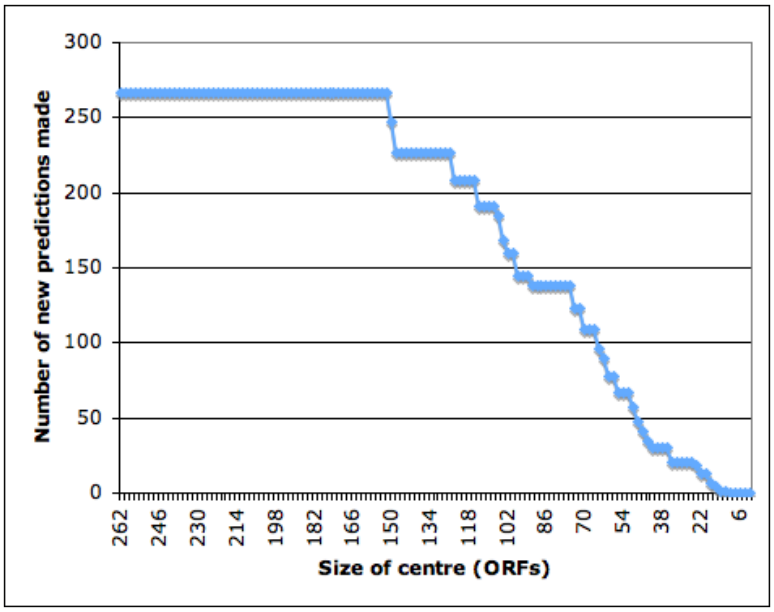


Figure 4.16: Numbers of new PPI predicted between core families that are expressed during the same phase and located in the genomic centre, at varying definitions of this "centre".

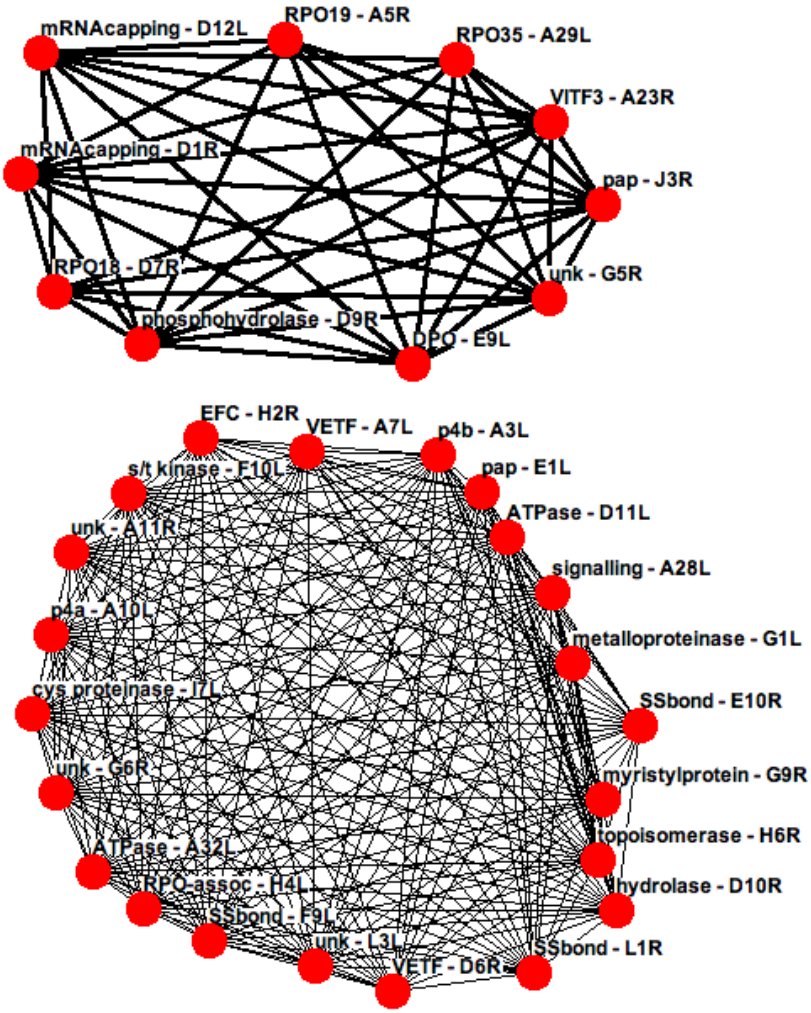


Figure 4.17: New PPI predicted, visualised by Osprey (Breitkreutz *et al.*, 2003). Abbreviations and colours are the same as in Figure 4.6. The upper cluster consists of proteins expressed early, the lower of those expressed late.

Chapter 5

Gene Origins in Vaccinia Virus Strain Copenhagen

In this chapter, we wanted to achieve an overview of the history of each gene in a reference genome, vaccinia virus strain Copenhagen (VACV-COP), indicating when in the evolution of this virus it originated, and through which processes.

5.1 Introduction

Long *et al.* (2003) wrote a comprehensive review on the analysis of new genes and split the interest in the origin of new genes into two questions: What are the initial molecular mechanisms that generate new gene structures? And how often, at the genome level, do new genes originate?

Examining the genome for clues as to how genes or other elements evolved has become very important, as finding the mechanisms can provide clues about the function of the gene. Enright *et al.* (1999), for example, show that genes that originated through fusion are likely to interact with each other in their ancestral, unfused form. Genes that originated through gene duplication start off with the same function and in some cases can become

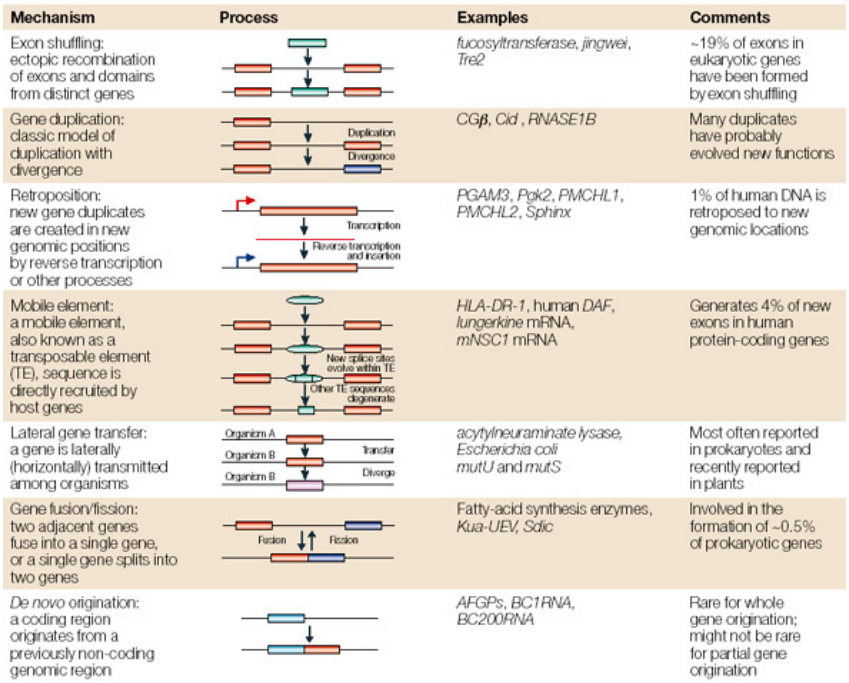


Figure 5.1: Mechanisms of Creating New Genes, taken from Long *et al.* (2003).

subfunctionalised, where their combined complementary functions perform the same actions as the ancestral gene (Lynch and Conery, 2000). Studying the mechanisms by which genes have evolved can also unveil unexpected combinations of processes that have worked together to shape a gene, as in the case of *jingwei*, a *Drosophila* gene, which arose through a combination of retroposition, gene duplication and recombination (Long *et al.*, 2003).

Where one is interested in a particular gene or group of genes, examining their origins can provide a greater insight into their function. In a forthcoming paper by Rothenburg *et al.* (in preparation, on which I am a co-author) this has been done for a group of poxvirus host range genes and shows how the analyses described in this chapter can be usefully applied to improve our understanding of pathogen virulence.

Figure 5.1 was taken from Long *et al.* (2003) and summarises the mechanisms that have been shown to create new genes.

Gene Duplication This has been the most widely studied and probably is the process which creates the majority of new genes in many organisms (Lynch and Conery, 2000; Katju and Lynch, 2006). While whole-genome duplication plays a major role in many eukaryotes, this mechanism does not apply to viruses. Duplication of single genes, however, is common in poxviruses.

Exon Shuffling This process is also known as domain shuffling, as domains that are contained within one exon of eukaryotic genes can be rearranged through non-homologous recombination. As poxviruses do not have introns or exons this does not apply as such, but recombination, the underlying mechanism, is still possible even in intronless genes.

Retroposition This creates duplicate genes in new positions in the genome by reverse transcription of expressed genes. Often, new regulatory elements have to be recruited for the new gene to become functional.

Mobile Elements These can be integrated into the genome and generate new functions if inserted into a transcribed gene. This has been shown, via primate-specific Alu elements, in the human genome. No mobile elements have been observed in poxviruses.

Horizontal Gene Transfer As introduced in Section 3.1.1, horizontal gene transfer (HGT) is any transfer of genetic material from one organism to another, which is not its offspring. This mechanism of gaining genes is commonly observed in bacteria (Ochman *et al.*, 2000) and is thought to be responsible for making a tree of life impossible to infer, creating a web-like structure of organisms across all domains of life contributing genetic material to each other (Doolittle, 1999).

Gene Fusion / Fission A mutation that removes a stop codon or transcription termination signal may lead to the fusion of two adjacent genes, which will then be transcribed and translated as one. This has often been observed in prokaryotes and genes that have been fused together are often functionally related, and benefit from the common regulatory sequence so that selective pressure keeps them maintained throughout evolution (Yanai *et al.*, 2002). Indeed, searching for fusion gene products has been used in order to find interacting proteins (Enright *et al.*, 1999). Fission is the opposite mechanism, where one gene is split in two by the insertion of a stop codon and the beginning of independent transcription for the two new genes. This will only yield two functional products if both parts are able to function independently and often leads to pseudogenisation of the downstream ORF which does not retain the promoter sequence.

De Novo Formation This is a very rare process through which a non-coding region mutates into an ORF and acquires the regulatory elements necessary to become a functional gene. A recent paper by Knowles and McLysaght (2009) reports three cases of protein-coding genes having originated from non-coding sequence in the human lineage and examples are known from other eukaryotes like *Drosophila* (Zhou *et al.*, 2008) but no evidence of de novo gene formation has been found in viruses.

Here we take each open reading frame (ORF) of vaccinia virus Copenhagen (VACV-COP) in turn and try to determine on the one hand where it first appears in the history of poxvirus genomes and on the other hand the mechanism by which it came to be present in poxvirus genomes. It may be a core gene, present in all poxviruses and therefore part of the ancestral poxvirus genome. If, however, it is only encoded by a subset of poxvirus genomes, our expectation was that there will still remain some evidence of how it originated. Figure 5.2 is a summary of our findings and will be referred to throughout this chapter.

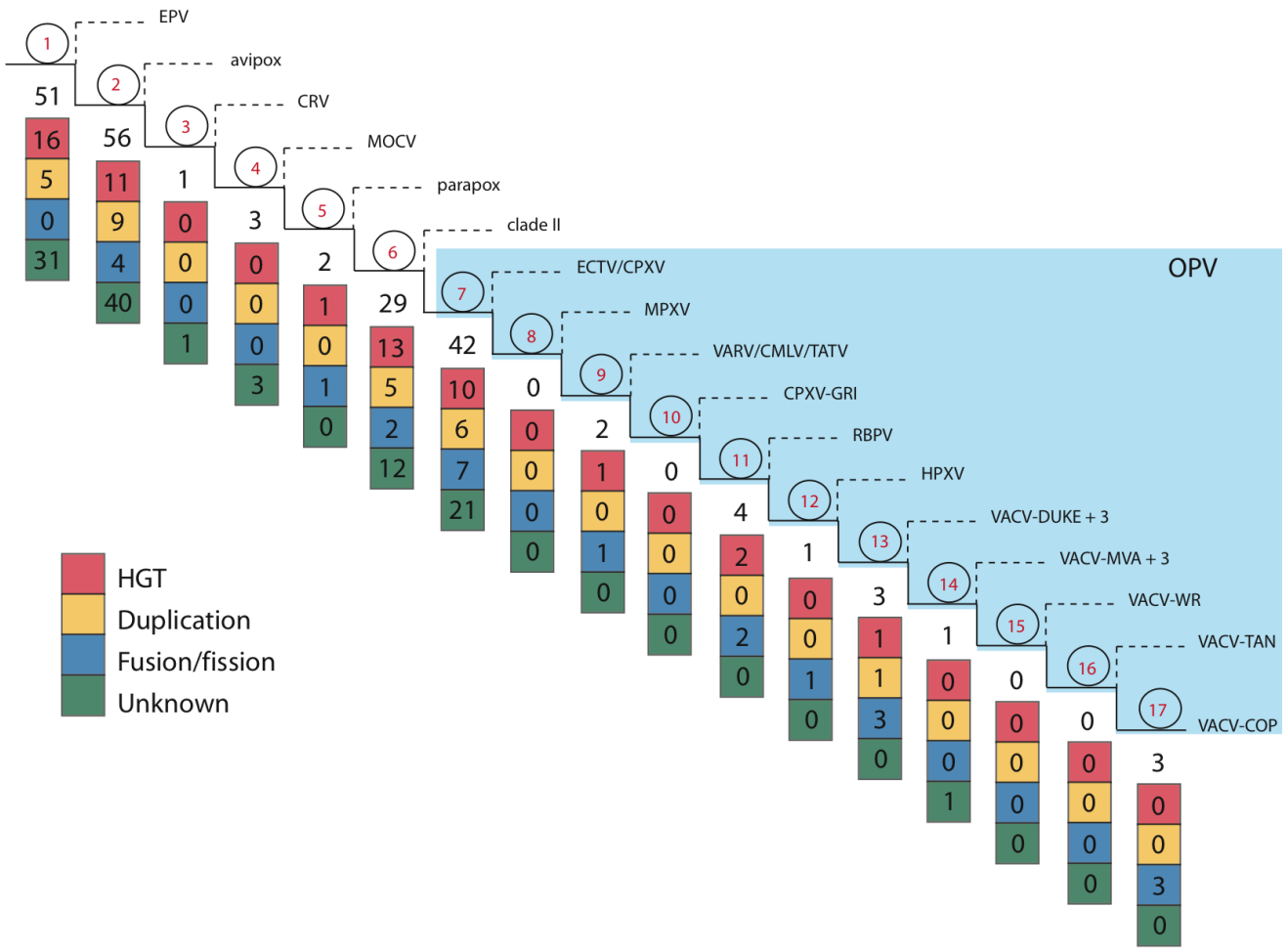


Figure 5.2: Genes gained throughout VACV-COP evolution. Nodes are labelled with red numbers. Black numbers show how many genes we found to have been gained at each node and by which mechanisms. The blue box incorporates all orthopox viruses.

5.2 Methods

5.2.1 Family Quality Control

As this analysis relies heavily on family classifications and which genes we consider to be orthologues, a quality control of our family definitions, as described in Chapter 2, was deemed valuable. Set 1 (S1), the most inclusive set with the largest families, was chosen to be the starting point. We then asked, for each family in this set, do all members similarly belong to one family in set 2 (S2)? If so, this family will be considered robust, as the BLAST hits between its members are identified as significant by clustering algorithm MCL at two different levels of granularity (see Section 2.3).

If this is not the case, the quality of the BLAST hits, the genomic location of the proteins in question and any available annotation is checked for further clues as to which family definition is the more sensible and which members of the S1 family should perhaps be disregarded as unlikely homologues. Where a protein is annotated as being a homologue of the VACV-COP ORF under scrutiny, or a homologue to an ORF in another genome that is clearly identified as being a homologue to the VACV-COP gene, this is considered evidence of true homology. Similarly, where proteins are members of the same superfamily, as annotated by NCBI, this makes them homologues for our purposes. On the other hand, where only a domain is shared between two proteins, this is not enough evidence to definitely label these proteins homologues, as domains can be exchanged between proteins through recombination. Where there are two or more copies of a gene in one genome this indicates gene duplication.

This approach can only lead to smaller families, as S1 families are larger than those of S2. In order to check if further protein members should be included in a given family, a DNA level homology search is also carried out, as described below. Family set 3 is ignored here as preliminary tests showed that

it favours splitting truly orthologous groups into separate, smaller families.

5.2.2 DNA Similarity Search

Where a given VACV-COP protein is not annotated in other poxvirus genomes, it is either truly absent from this genome or its encoding ORF may have been disabled by mutations leading to, for example, a premature stop codon. Alternatively, a non-coding sequence may be present in this genome, which resembles the coding sequence of VACV-COP and may be the source from which the protein-coding region was formed de novo, by the acquisition of fortuitous mutations. Which scenario is more likely, a disabling in the genome lacking the protein in question, or an enabling (de novo formation) in those that encode it, can be determined by looking at the number of genomes in which such a similar but non-coding sequence is present and the kind of differences these sequences contain. To find homologous sequences at the DNA level, we used *ssearch* from the FASTA package (Pearson and Lipman, 1988). Each DNA sequence corresponding to a VACV-COP ORF was searched against the entire DNA sequence of each poxvirus genome with an E-value threshold of 1. The 198 non-minor VACV-COP ORFs have between 32 and 358 homologous sequences in all poxvirus genomes. To determine which ones are reliable hits due to true homology, a threshold stricter than 1 was imposed and individually determined for each VACV-COP ORF as follows. The reason for choosing a variable E-value threshold is that the protein families are present in groups of genomes of different evolutionary distance. For very distant genomes, the E-values between DNA sequences coding for a protein in a given family might already be quite high, making an even higher value for homologous non-coding sequences likely.

We determined the highest E-value that was found for a hit between the VACV-COP ORF and any sequence coding for a member of its protein family. Any hits with an E-value that was $1e^5$ (100,000) times higher than this value

were also accepted as relevant, as some variation is expected, especially when the genome in question is more distant. Where this value was still very low (lower than $1e^{-50}$), $1e^{-50}$ was chosen as the E-value threshold instead, as this enables us to find all truly homologous sequences while still allowing for more distant homologies. Further non-coding homologous sequences are found in 41 out of 157 families.

To evaluate this method of choosing individual thresholds, we compared it to one where a threshold of $1e^{-5}$ was applied to all families. Extra homologous hits are found in 35 of the 41 families. In seven families, homologous sequences are found that are not found by the variable threshold method. On closer inspection, the sequences that are thus missed by our method do not provide any further evidence towards de novo gene formation, as none of them are in outgroup genomes.

All potentially homologous sequences ($E < 1$) were kept in a database, which was consulted where no evidence for any other mechanism of gene formation was present.

5.2.3 Gene Fusion/Fission Search

A successful method of detecting genes that have been fused together, or, conversely, are the result of gene fission, was introduced by Enright *et al.* (1999). It is based on sequence comparisons between complete genomes of bacteria and, in a later paper, eukaryotes (Enright and Ouzounis, 2001). We apply this method to poxviruses and carry out the following search to look for fission or fusion in VACV-COP. The first step was to find two candidate component proteins A and B in VACV-COP, that display similarity to a protein C in any other poxvirus genome, but that do not have any significant similarity to each other, as determined by a BLASTP search as described in Chapter 2. In the interest of not missing too many candidate pairs, we also accepted those cases where there was only a unidirectional homology hit

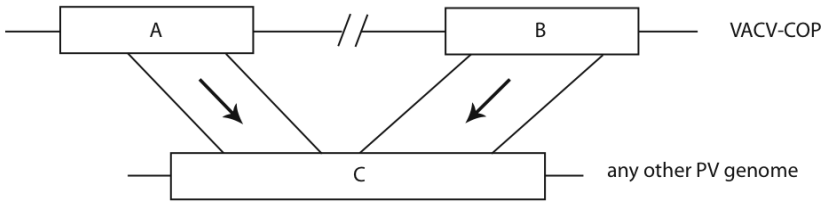


Figure 5.3: Fusion Detection. Where A and B both have homology to C but not to each other, fusion or fission is likely.

between A and B or where one of a pair of hits (A to B and B to A) is of poor quality ($E <= 1e^{-5}$). The E-value threshold chosen for hits between A/B and C was $1e^{-4}$, after some experimentation. Another restriction is then imposed on the candidate pair: the regions where A and B hit C may not overlap by more than 10% in protein C. Figure 5.3 illustrates this method. The reverse search was then carried out to identify potential fusion events in VACV-COP, where two genes in another genome are homologous to a gene in vaccinia, but not each other. This search was done in a representative set of poxvirus genomes: 3 VACV, HPXV, RBPV, CPXV-GRI, VARV-AFG70, CMLV-CMS, TATV, MPXV-Z96, CPXV-BR, ECTV, YLDV, MYXV-LAU, DPV-W83, LSDV-2490, SWPV, ORFV-I, MOCV, CNPV, FWPV-I, CRV.

Looking at the genomes where the proteins are present as two (component proteins) and where they are present as one (composite protein) will provide evidence for the mechanism that has taken place: an ancestral gene that was split in two due to a premature stop codon and use of a downstream start codon (fission) or fusion of two ancestral (not necessarily neighbouring) genes into one.

5.2.4 Horizontal Gene Transfer Search

Where VACV-COP ORFs do not show any signs of originating through intra-genomic mechanisms (e.g., fusion or de novo formation), we wanted to look

for signs that this gene may have come from the host through horizontal gene transfer. While a rigorous search was carried out in Chapter 3, the search needed here was much more lenient as a host origin is suggested by the absence of any other evidence. A BLASTP search of all poxvirus proteins encoded by completely sequenced genomes (as listed in Table 1.1 on page 25) was carried out against the GenBank non-redundant database, downloaded in August 2009. For details on BLAST, please see Section 2.1.

5.2.5 Workflow

This analysis was based on input from various comparative genomics techniques, which were used to pinpoint the most likely scenario of evolution for each ORF in VACV-COP. Figure 5.4 illustrates the workflow that led to the results described below. While the same checks were performed for each ORF, the analysis was not fully automated, however, as it proved to be more useful to employ manual editing at various stages. For example, in the family quality control step described above, in order to determine whether an S1 family that was split in two in S2 was really a group of homologues, the position of these genes in all genomes was checked for common synteny, where annotation was lacking. In the DNA homology search step, non-coding homologous sequences were checked in unclear cases, even when they did not have an *E*-value hit below the threshold, to avoid missing genuine homologies.

5.3 Results

5.3.1 Family Quality Control

Of the 262 VACV-COP ORFs, 258 are members of one of our 437 (in S1) protein families. The unclassified four are all minor ORFs. Because 19 families include more than one VACV-COP protein (i.e. they contain gene duplicates within VACV-COP or ORFs split in VACV-COP), VACV-

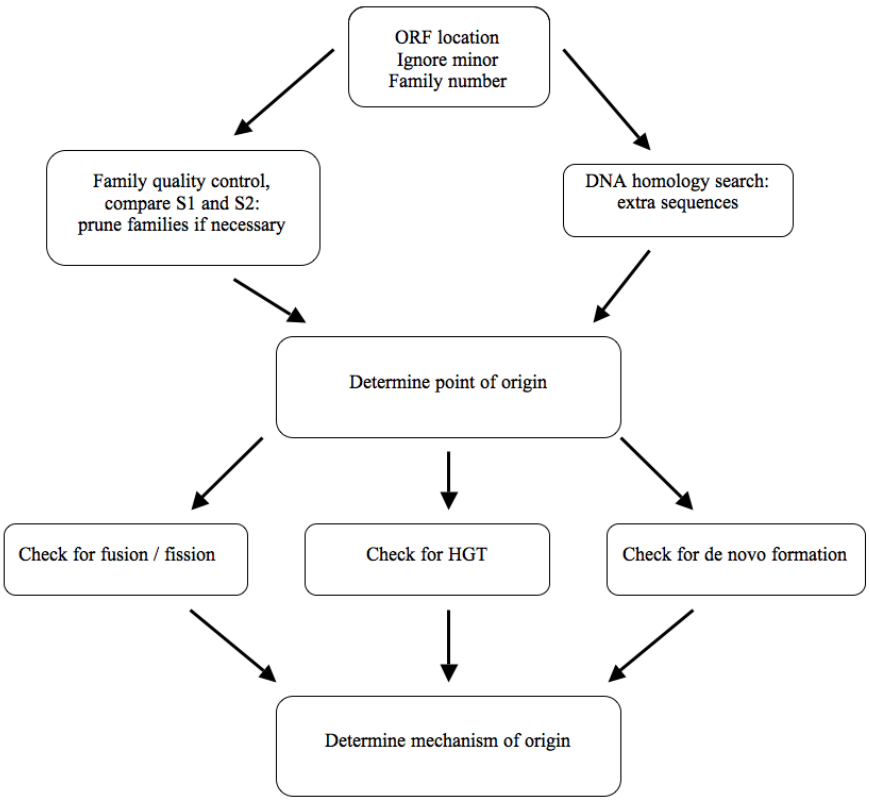


Figure 5.4: Workflow to determine point of origin and mechanism of origin for each VACV-COP ORF.

COP ORFs are members of 213 different families. Of those, 157 include VACV-COP ORFs that are not minor ORFs. Minor ORFs, once again, are excluded from our analyses, as they are located opposite and inside bigger and more plausible ORFs, whose analyses will cover any information that might be gained from looking at inconsistently annotated minor ORFs. Of the 157 S1 families, 98 (62.4%) have a clearly corresponding family in S2 and are therefore very good robust families. A further 46 families have two (in 28 cases) or more (in 18 cases) corresponding S2 families but on closer inspection, the S1 family seems more realistic, as, for example, all members are annotated as homologs or are located within the same syntenic environment in their genomes. The remaining 13 families need to be curated in some way. Most of these are easily pruned with clear homologies, but in two cases, deciding which proteins should belong to one family proved contentious. These are the largest families in S1 and we therefore decided to assign homologies based on the S2 family classification. One of these families contained 1132 members in S1 (17 of them in VACV-COP), based on the presence in all members of an ankyrin-like domain. **The other one contained 479 members (6 in VACV-COP), mostly annotated as being part of the Pox A46 superfamily in the NCBI conserved domain database, although A46R is not part of this family, probably due to the presence of a TIR domain that is absent in A52R, K7R and the other members of this protein family (Stack *et al.*, 2005). It is likely that a combination of independent acquisitions (e.g. through multiple HGT events) and duplication and domain shuffling led to the spread of these families.**

Apart from these two families there are ten others with potential paralogues in VACV-COP. As some of them are known to share a domain it is of interest to know if the homology of the genes is only due to this shared domain or if they are homologous across the entire sequence. In

the latter case, the evidence for paralogy is strong, while the former offers recombination and domain shuffling as alternative explanations. A dot plot analysis of the DNA alignment produces graphs such as the ones in Figure 5.5. Here, C4L and C10L are shown to share homologous regions all along the length of both genes, signified by diagonal black lines, while C23L (and its identical ITR duplicate B29R) and A41L are similar only in small areas. Table 5.1 lists all potential paralogues. This list does not include ITR duplicates, which are strictly speaking also paralogues but due to a special poxvirus mechanism where the ends of the genome overlap and frequently exchange genetic material. There are nine pairs of ITR duplicate genes, C23L - C15L and B29R - B21R.

Another explanation for the presence of similar genes in a genome is repeated HGT into the poxvirus lineage from some cellular organism, but this is very difficult to test.

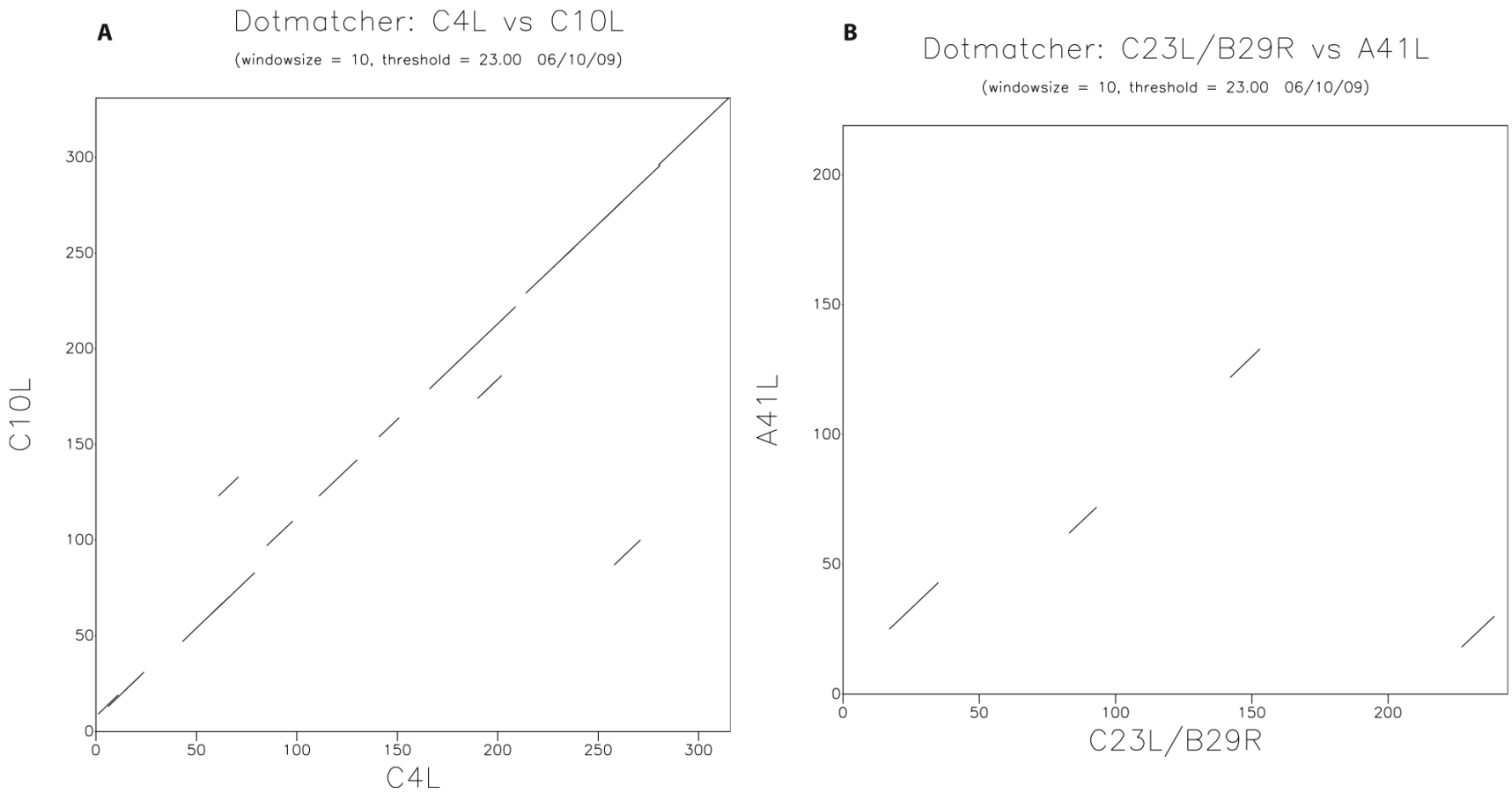


Figure 5.5: Dot plot of potential paralogues. A: C10L and C4L, which appear to share homologous regions across the lengths of both genes, as shown by the diagonal black lines. B: C23L/B29R and A41L, which are only similar to each other in small isolated regions. Drawn using *Dotmatcher* from the EMBOSS package (Rice *et al.*, 2000), with default parameters as shown.

5.3.2 DNA Similarity Search

One problem with searching for homology at the DNA level is the varying GC-content of poxviruses: as listed in Table 1.1 on page 25 it varies from 18-65%. Specifically, crocodilepox, parapox and MOCV have a GC-content of over 60%, compared to VACV-COP, whose GC-content is 33.4%. Out of 92 VACV-COP / MOCV homologue pairs, only 59 are linked through ssearch, indicating how unlikely it is to find homologous non-coding DNA sequences, which are likely to be more different from each other than functional homologues. However, MOCV and parapox genomes are also much shorter than those of OPV, meaning that we do not expect to find many traces of homology within non-coding sequences.

Sixty-one VACV-COP ORFs show similarity to DNA in other genomes that is not annotated as a gene. Of these potential homologues, most (87%) are located in genomes internal to the group in which the family is present, as illustrated in panel A in Figure 5.6. Many of these cases, incidentally, appear to be due to mis-annotation rather than loss of a gene. Some, however, (13%) are found in outgroup genomes, as shown in panel B, or both types of genomes. The latter are, in all cases, VACV-COP ORFs that belong to ChPV families, with the extra homologous sequences being encoded by EPV. This shows that there is a dearth of evidence for de novo origination of vaccinia genes. If a coding gene had originated somewhere in the poxvirus lineage leading to VACV-COP from non-coding sequence, we would expect to still find traces of the original non-coding sequence in ancestral genomes through a DNA homology search. While it is certainly possible that such de novo genes originated in an ancestral ChPV, through mutations in a sequence still present now in EPV genomes, it is difficult to determine the veracity of such a speculation, as synteny is not conserved between EPV and ChPV and it cannot be known whether an EPV sequence found to be homologous to a ChPV gene is in an equivalent location and therefore likely to have

Genes	**Note**	**Genomes with Paralogues**	**Homologous Area**
C2L, F3L, A55R, B10R	Share kelch, BTB and BACK motifs/domains	some OPV	Similarity in many small regions
C23L, A41L, B29R	Members of OPV 35 kD superfamily	OPV and clade II	Similarity in some small regions
C12L, K2L, B13R/B14R	Members of serpin superfamily	many ChPV	Similarity across whole gene for C12L and B13R/B14R, K2L is much weaker
K4L, F13L	Share phospholipase D domains	OPV	Similarity in many small regions across genes
C22L, A53R, B28R	Share tumor necrosis factor domain	OPV	Similarity over most of the gene, very short sequences
F9L, L1R	Share L1R-F9L lipid membrane protein domain, not neighbouring	all PV	Similarity over most of the gene
C3L, B5R	Share complement control protein domain, not neighbouring	OPV	Similarity in small regions across whole gene
D9R, D10R	Share a Nudix hydrolase domain, neighbouring	ChPV - parapox	Similarity in some small regions
C10L, C4L	Members of C4L/C10L superfamily, not neighbouring	OPV and clade II	Similarity across whole gene
B1R, B12R	Both are serine/threonine kinases, not neighbouring	OPV, avipox, EPV	Similarity across whole gene

Table 5.1: Potential Paralogues in VACV-COP

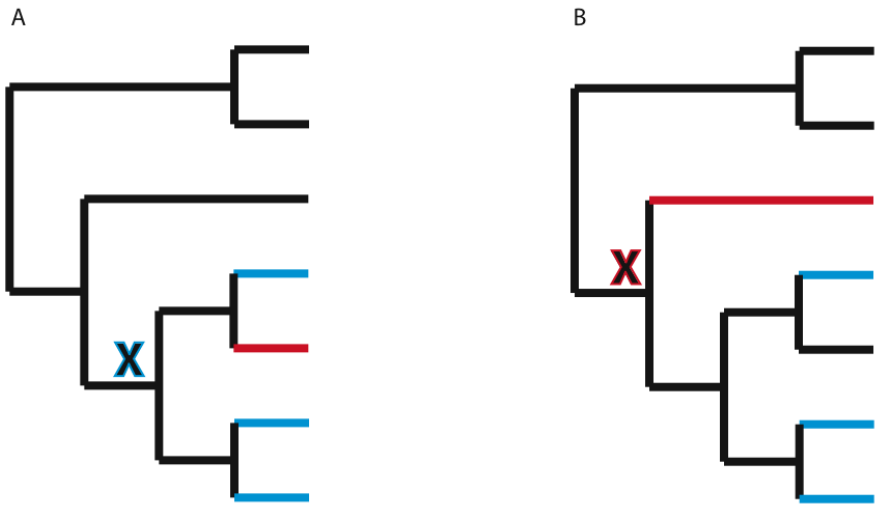


Figure 5.6: Effect of finding non-coding homologous sequences. Proteins are present in the genomes represented by the blue branches; a homologous non-coding DNA sequence is present in the genome represented by the red branch. In panel A, this does not change the node in the tree in which the protein-coding sequence first originated (symbolised by X), while in panel B, the genome with the non-coding homologous sequence is an outgroup to those with the proteins, implying a de novo origin of this protein-coding sequence.

led to the coding gene. One also cannot exclude the alternative possibility of gene inactivation EPV, as a third, ancestral branch would be needed to distinguish between de novo origination in ChPV and inactivation in EPV. Furthermore, in most cases the significance of the BLASTP hit suggesting homology is very low, with the most similar hit to an EPV sequence having an E-value of $E > 0.1$ in six out of eight cases.

The results of this search are noted in the column "DNA Similarity" in Table 5.4 on page 158.

5.3.3 Gene Fusion/Fission Search

Based on the criteria described in the methods Section 5.2.3 above, we find 33 pairs of VACV-COP ORFs, that have potentially undergone fusion or fission at some point in the poxvirus evolutionary lineage. Eleven pairs are direct neighbours on the same strand of the genome, which makes them more likely to have undergone fusion or fission, but genome rearrangements can lead to cases where the component proteins are no longer located next to each other. Table 5.2 lists all candidate component protein pairs found in VACV-COP and where the potential composite protein was found ("Composites"). Nineteen of these potential fusion events are between ORFs annotated as ankyrin-like proteins, members of the large S1 family mentioned above. Three others shared other domains or motifs, which may create the appearance of fusion and make a clear conclusion very difficult. Nevertheless, all pairs were examined and we find five likely cases of fission in groups of genomes that include VACV-COP, which are discussed below.

C14L and C13L An 11 bp-deletion leads to a frameshift and premature stop codon in the VACV-COP genome with a downstream ORF encoding the second part of the ancestral homologue.

K5L and K6L A 20 bp-deletion in VACV and RBPV causes a premature stop codon compared to HPXV, CPXV-GRI and other OPV. This supports an alternative phylogeny to the one presented in Figure 2.4 on page 53, one where RBPV and the VACV group are monophyletic, with HPXV as an outgroup. The downstream K5L has become a pseudogene.

B2R and B3R A 19 bp-deletion in VACV, HPXV and RBPV leads to one ORF with a premature stop codon and a second, believed to be a pseudogene.

B13R and B14R A 14 bp-deletion in most VACV genomes leads to a premature stop codon and two genes with partial homology to CPXV-GRI gene B12R. VACV-WR, interestingly, retains the 14 bp segment and thus encodes one full-length ORF. As VACV-WR is not generally considered the outgroup to all other VACV genomes, this contradicts the VACV phylogeny in Figure **??**. For B13R/B14R, the gene history seems to disagree with the species history.

A25L, A26L and A27L A27L is conserved throughout OPV, parapox and clade II genomes but there is some evidence for fission after the divergence of MOCV (node 5 in Figure 5.2 on page 132), as A26L and A27L both have homologous regions to MC131L and MC133L (potential paralogues) in MOCV.

A25L and A26L represent fragments of CPXV-GRI protein A26L, an A-type inclusion (ATI) protein. Most VACV strains, along with HPXV and RBPV, have a very similar arrangement of ORFs in this region (see Figure 5.7), while VACV-COP and VACV-MVA have experienced large independent deletions between A24R and A27L (roughly 4 kb in size). This deletion in VACV-COP covers two ORFs in most other VACV strains and HPXV (VACV-WR 147 and VACV-WR 148) and leads to a fusion of parts of VACV-WR genes 146 and 149 in VACV-COP ORF A26L. In CPXV-GRI

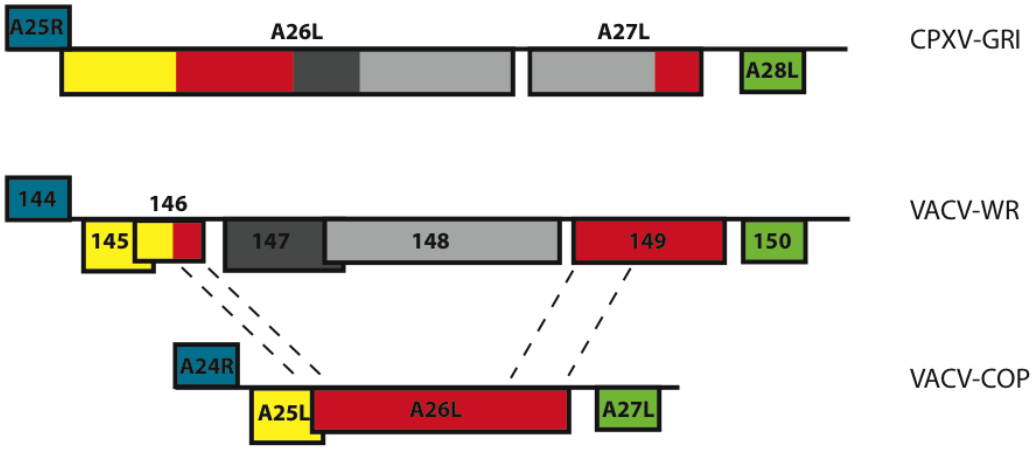


Figure 5.7: Large deletion in VACV-COP. Boxes of the same colour represent homologous gene regions. VACV-WR represents most other VACV strains and HPXV, which have the same arrangement of ORFs. The large CPXV-GRI A26L is homologous to many VACV-WR and VACV-COP genes. VAC-WR genes 146 and 149, and CPXV genes A26L and A27L have partial homology to VACV-COP A26L, due to a fusion event (dashed lines) having taken place in VACV-COP. VACV-COP A25L and A26L both have homology to CPXV A26L (and VACV-WR 146), giving the appearance of fission. This is most likely a real fission event originating several times, or in several stages, over the VACV lineage.

and other ancestral sequences, the ORF named A26L covers the entire region corresponding to VACV-WR 145-148, which in VACV-COP includes A25L and A26L, so we infer that these two genes are fission products. This fission probably happened after the divergence of CPXV-GRI (node 11 in Figure 5.2 on page 132) and possibly multiple independent times since then, as it is shared by most VACV, HPXV and RBPV, but is not identical in all these genomes. This region is shown in Figure 5.7.

The Poxvirus Bioinformatics Resource Center (PBRC) further lists the following ORFs that have been split into parts in VACV-COP: C22L compared to CPXV-GRI D2L, B16R compared to CPXV-GRI B14R and

B28R compared to CPXV-GRI I4R. All of these have lost the other half of the ORF due to pseudogenisation, and can therefore not be found by our methods, and do not alter our results, as the origin of the initial ORF can still be determined through looking at the presence of homologues.

When searching for cases where there might have been fusion in VACV-COP or a group of genomes that includes VACV-COP, or alternatively fission in other genomes, we find 33 VACV-COP genes that have similarity to two mutually non-homologous genes in at least one other poxvirus genome. Table 5.3 lists all candidate VACV-COP composite proteins with component proteins. While representative genomes were used for the initial search of fusion candidates, all genomes were then inspected closely to determine the most likely scenario for each candidate trio of genes. In most cases, the presence of the composite protein in VACV-COP was shown to be due to a gene fission in another poxvirus lineage, rather than a fusion event in the VACV-COP lineage. An illustration of how this was determined is shown in Figure 5.8, using the gene C1L. A look at the synteny of the ORF in question, as well as the genomic region in outgroup genomes provides more information about the fusion/fission event. After examining all 33 candidate cases, there remains only one clear case of gene fusion in VACV-COP, that of ORF A26L, which is described above and illustrated in Figure 5.7. All cases of fusion or fission in the lineage leading to VACV-COP are listed in column "Fusion/Fission" in Table 5.4 on page 158.

Ankyrin proteins The similarity between domains of the same family (ankyrin, in most cases), which is often the reason for detecting potential fusion or fission through the methods described above, may instead be due to repeated horizontal transfers of ankyrin-containing genes, or domain shuffling, a form of recombination between genes. Because of the large number of genes containing ankyrin domains, however, the traces left behind by the events are not unambiguous, as gene locations have changed and it is

Gene A	Gene B	Neighbours	Composites	Note
C21L	C20L	y	OPV - VACV, RBPV	Plus ITR hit, shared ankyrin motif, fission in VACV-COP
C21L	C19L	n	OPV - some VACV	Plus ITR hit, shared ankyrin motif, C19L-C21L are one gene in CPXV-GRI
C21L	C15L	n	OPV - VACV, RBPV	Plus ITR hit, shared ankyrin motif
C21L	M1L	n	CPXV-BR	Plus ITR hit, shared ankyrin motif
C20L	C19L	y	OPV - VACV, RBPV	Plus ITR hit, shared ankyrin motif, fission in VACV
C20L	C15L	n	OPV - VACV, RBPV	Plus ITR hit, shared ankyrin motif
C19L	C9L	n	CPXV-GER, VARV-MIN	Plus ITR hit, shared ankyrin motif
C18L	C17L	y	CPXV, TATV, MPXV, CMLV, some VACV	Plus ITR hit, shared ankyrin motif
C18L	K1L	n	CNPV	Plus ITR hit, shared ankyrin motif
C18L	M1L	n	CNPV	Plus ITR hit, shared ankyrin motif
C18L	B18R	n	CNPV, CPXV, MPXV	Plus ITR hit, shared ankyrin motif
C17L	B4R	n	MPXV, TATV, CPXV, some VACV	Plus ITR hit, shared ankyrin motif
C17L	B18R	n	MPXV, TATV, CPXV	Plus ITR hit, shared ankyrin motif
C15L	C9L	n	CPXV-GRI	Plus ITR hit, shared ankyrin motif
C15L	M1L	n	CPXV-BR	Plus ITR hit, shared ankyrin motif
C15L	B4R	n	MPXV, HPXV, CPXV-GRI	Plus ITR hit, shared ankyrin motif
C14L	C13L	y	OPV - some VACV, VARV	Fission in VACV-COP
C9L	M1L	n	avipox	Shared ankyrin motif
C9L	B18R	n	leporipox	Shared ankyrin motif
M1L	K1L	n	avipox, DPV-W84	Shared ankyrin motif
M1L	B18R	n	CNPV, parapox	Shared ankyrin motif
K1L	B4R	n	FWPV	Shared ankyrin motif
K1L	B18R	n	CNPV	Shared ankyrin motif
K2L	B13R	n	OPV - some VACV	Shared serpin domain
K5L	K6L	y	ECTV, MPXV, TATV, CPXV-GRI, HPXV, yatapox	Fission in VACV and RBPV
F6L	F7L	y	CMLV	Due to overlap in CMLV
A25L	A26L	y	CPXV, some VACV	Many fission events in VACV lineage (see main text)
A26L	A27L	y	avipox, MOCV	Possible fission in parapox, OPV, clade II
A39R	A40R	y	TATV, some VARV	Due to overlap in VARV and TATV
A55R	B10R	n	CMLV, CPXV-BR, CPXV-GER	Shared kelch motif
B2R	B3R	y	ECTV, CPXV, MPXV, TATV, CMLV	Fission in VACV, RBPV, HPXV
B13R	B14R	y	OPV - some VACV; clade II - suipox	Shared serpin domain, fission in most VACV
B18R	B20R	n	OPV - ECTV, TATV, some VACV	Shared ankyrin motif

Table 5.2: Component Proteins in VACV-COP

difficult to locate initial homologues. In some genes it is possible to construct a history, as in the case of C19L-C21L, which correspond to one gene in CPXV-GRI. Nevertheless, the number of ankyrin-like genes in poxviruses is remarkable. Ankyrin repeats mediate protein-protein interactions and are often found in a class of F-box protein that may be involved in the degradation of proteins via the ubiquitin pathway (Sonnberg *et al.*, 2008). They are present across all phyla and, according to Bork (1993), this suggests domain shuffling and horizontal gene transfer as mechanisms for spreading these proteins. In poxviruses, Bork (1993) examines genes neighbouring those with ankyrin repeats and notes a high number of receptors for interleukin-1 and tumor necrosis factor (TNF) proteins among them. He suggests that, since IL-1 and TNF both initiate NF-κB-directed transcription, ankyrin proteins may decrease transcription through competition for IκB and thus contribute to virus spread within the host. Shchelkunov *et al.* (1993) examined ankyrin repeat proteins in vaccinia and variola and while noting the length and homology of different proteins they similarly fail to produce a scenario of ankyrin protein evolution within poxviruses.

5.3.4 Horizontal Gene Transfer Search

The BLASTP search yielded homologous non-poxviral hits for 146 out of 198 non-minor VACV-COP ORFs with the very high, relaxed *E*-value threshold of 1. However, as these do not necessarily have strong significance, we decided, after some experimentation, on an *E*-value threshold of $1e^{-4}$. Using this threshold, 63 non-minor VACV-COP ORFs had similarity to non-poxviral hits. These are listed in Table 5.4 (page 158) in the "Putative Cellular Homologues" column and are summarised into groups: mammals, other vertebrates (excluding mammals), insects, plants, fungi, other eukaryotes (excluding all previous groups), bacteria, archaea, and viruses. They are listed in the order of most to least similar There are 46

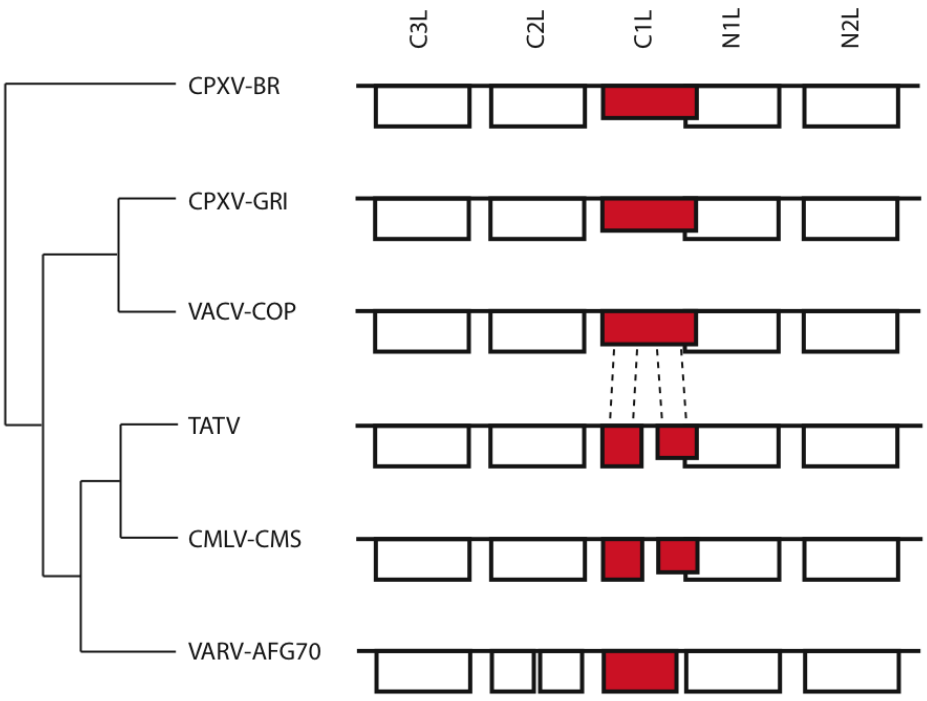


Figure 5.8: Fission event in TATV, illustrating tests undertaken for each candidate. Synteny confirms homology, and a look at outgroup genomes confirms where the fission/fusion event has taken place. Here, the ancestral C1L ORF, present in CPXV and VACV-COP has been split in TATV and CMLV, with one of the resulting ORFs having been lost or pseudogenised in CMLV. The dotted lines indicate the homologous regions between VACV-COP and TATV.

Gene C	Components	Neighbours	Note
C23L	TATV	y	Fission in TATV; plus ITR hit
C19L	VACV-MVA	n	Plus ITR hit, shared ankyrin motif
C17L	VACV-MVA, CMLV	n	Plus ITR hit, shared ankyrin motif
C10L	TATV, CMLV	y	Fission in CMLV and TATV
C9L	VACV, HPXV, VARV, CMLV	Some	Shared ankyrin motif
C5L	HPXV, CPXV-BR	y	Fission in HPXV and CPXV-BR
C4L	VACV-DUKE, VACV-AC, HPXV	y	Fission in VACV-DUKE and HPXV
C2L	VARV, MPXV	y	Fission in VARV and MPXV
C1L	TATV	y	Fission in TATV
M1L	HPXV	Some	Shared ankyrin motif
K1L	VACV-TAN, VARV, TATV	y	Fission in VARV, TATV, VACV-TAN
F5L	VACV-MVA	y	Fission in VACV-MVA
F11L	VACV-MVA	y	Fission in VACV-MVA
E3L	VACV-TAN	y	Fission in VACV-TAN
O1L	VACV-MVA, TATV, CMLV	y	Fission in VACV-MVA, TATV, CMLV
I4L	HPXV	y	Fission in HPXV
H4L	VACV-TAN	y	Fission in VACV-TAN
D10R	VACV-TAN	y	Fission in VACV-TAN
D11L	VACV-TAN	y	Fission in VACV-TAN
A26L	some VACV, HPXV, VARV	y	Fusion in VACV-COP (see main text)
A33R	VACV-TAN	y	Fission in VACV-TAN
A39R	some VACV, VARV	y	Fission in some VACV, VARV
A44L	VARV, TATV	y	Fission in VARV, TATV
A47L	CMLV	y	Fission in CMLV
A49R	TATV	y	Fission in TATV
A51R	some VACV, HPXV	y	Fission in some VACV and HPXV
A55R	VARV, CMLV, TATV, MPXV	Some	Shared kelch motif
B4R	avi	n	Shared ankyrin motif
B6R	HPXV, VARV	y	Fission in HPXV and VARV
B12R	TATV	y	Fission in TATV
B16R	CMLV, VARV	y	Fission in CMLV, VARV
B17L	CMLV	y	Fission in CMLV
B18R	Some VACV, HPXV, CMLV, TATV	Some	Shared ankyrin motif

Table 5.3: Composite Proteins in VACV-COP

genes that show similarity to some mammalian protein and a further three that have a putative homologue in a non-mammalian vertebrate, indicating a potential transfer of this gene from the current or ancestral host into poxviruses. Most of these genes (23) have their strongest non-poxviral hits to mammalian genes. Odom *et al.* (2009) recently performed a similarity search of all poxvirus-encoded proteins versus different GenBank protein data sets and also find that ChPV proteins have greater similarity to eukaryotic proteins than to bacteria- or virus-encoded ones.

If similar proteins are only present in non-host organisms like bacteria, archaea, viruses or non-vertebrate eukaryotes this does not rule out HGT. It is possible for HGT to proceed via an intermediate host and in fact, it has been suggested that poxviruses may be vectors for transfer of genetic material between reptiles and mammals (Piskurek and Okada, 2007). Alternatively, HGT may have have taken place between co-infecting parasites or a sequence that was ancestral to many life forms has been lost in vertebrates or the donor sequence in a host organism has been lost since the transfer.

5.4 Discussion

Table 5.4 provides a summary of all analyses for each VACV-COP gene. The first and last nine proteins (C23L-C15L and B21R-B29R) are ITR homologues of each other and their results are therefore identical (except for slightly different *E*-values in the DNA homology search). The number listed under "Node of Origin" corresponds to the node number in Figure 5.2 and pinpoints the common ancestor of the genomes that encode this gene, based on our family information and annotation.

Where a domain name is listed in the "Fusion/Fission" column there is some evidence of composite or complement proteins in VACV-COP, respectively, but it is not possible to deduce clearly what combination of events has taken place. "Putative Cellular Homologues" lists similarities to

any type of cellular genome or other viruses. Under "DNA Similarity", only those genomes are mentioned which are outgroups to the genomes encoding this gene. Where the similarity is marked "weak" this indicates an E-value higher than the threshold determined as described in Section 5.2.2 above but is mentioned here for completeness' sake. The E-value listed in the table is the lowest (most significant) one found to any sequence in the given genome.

The "Verdict" column then summarises all results and gives the most likely scenario of gene evolution. Where "ankyrin" is mentioned, this points to the complicated evolutionary history of this gene family and that not all evidence for fission and fusion can be followed up satisfactorily. Where "A46 sf" is mentioned, this indicates a homology to the large Pox A46 superfamily, whose members are similarly plentiful and have evolved through many events, now difficult to decipher. Figure 5.2 on page 132 shows how many genes we believe to have originated at which nodes in the lineage to VACV-COP, and through which mechanism. In some cases, more than one mechanism is involved. Thirty-one of the genes for which the mechanism is listed as unknown are core genes and the lack of evidence is therefore not surprising, as these genes have been in the poxvirus lineage since its start.

To summarise Table 5.4, 63 genes have potentially originated through HGT, 26 are possibly paralogues of another VACV-COP gene (not counting ITR duplicates) and 17 are implicated in fusion or fission events (not counting ankyrin genes).

After all analyses, there are still 77 VACV-COP genes that just appear at some point in the poxvirus lineage with no evidence for how they got there. As every sequence must have its origin somewhere, it is possible that they did come from a cellular genome, which is not fully sequenced or it has since diverged so much as to be unrecognisable. It is also possible that they did originate via de novo formation of a protein-coding gene, but we do not have the evidence to distinguish between these possibilities. Sixteen of

these genes find weak similarities in cellular or other virus genomes when the *E*-value threshold is raised to 0.1, but none of these is in vertebrates.

Similarly, if there is no HGT evidence and no traces outside the genomes in which it is present in pox, it may still have originated de novo but mutations in either the remnant genomes or the coding sequence have made the homology undetectable.

There are 51 genes that have been categorised "core genes", a significantly larger number than the 28 named in Table 2.1 on page 46. This is due to families with paralogues being included here, as well as those where an origin at the core has been inferred, because of presence of an EPV homologue, despite some genomes lacking homologues.

It is noticeable, that no traces of similarity are found in crocodilepox, parapox or MOCV, probably due to their very high GC-content.

Figure 5.9 shows how different regions of the genome have gained new genes at different time points, with the central region containing mostly older genes that originated at nodes 1 or 2, while the peripheral regions of the genome have gained genes throughout poxvirus evolution.

Figure 5.10 gives another indication that there is more detectable genomic activity towards the ends of the genome, i.e., fusion/fission and duplication events are more commonly found in the peripheral regions of VACV-COP genome and HGT, while found across the whole genome, also seems to be more common in those areas.

Thus, while an analysis of the origins of all genes in vaccinia virus has not revealed a clear picture for each gene, this was to be expected as viruses have a high rate of evolution and therefore traces of a gene's history may be clouded after only a short period of time. Interestingly, the proportion of genes that appear to have experienced horizontal gene transfer is perhaps higher than expected, at 32%.

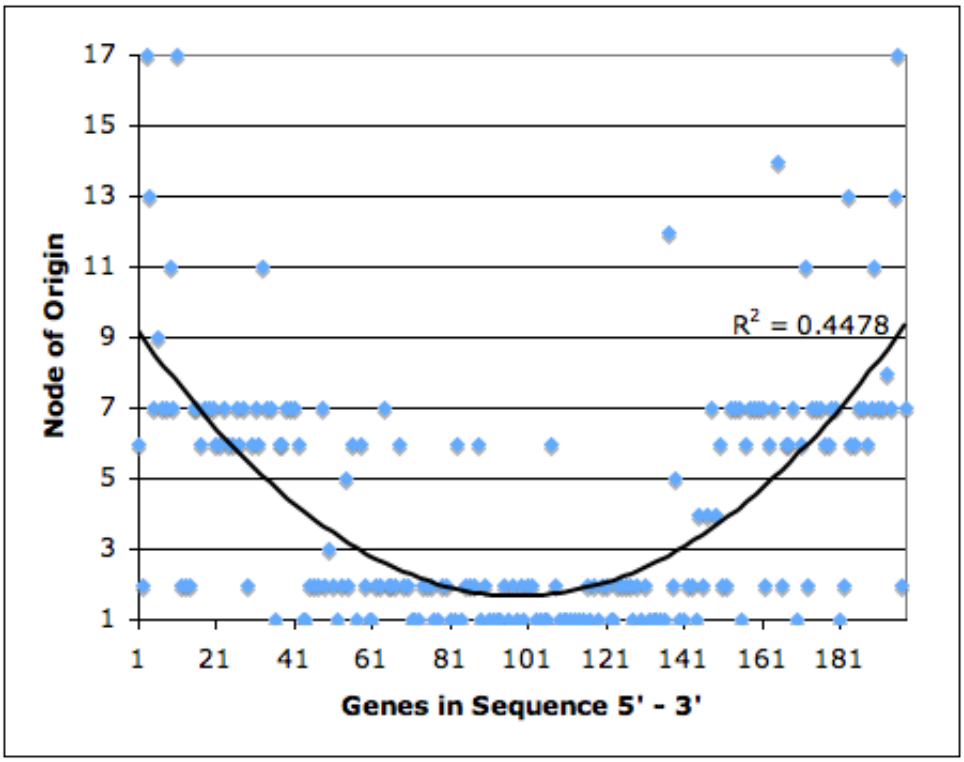


Figure 5.9: Gene gain across the VACV-COP genome. Genes are listed according to their position in the genome and the node at which they were gained is taken from Table 5.4. A polynomial trendline (order 2) is shown.

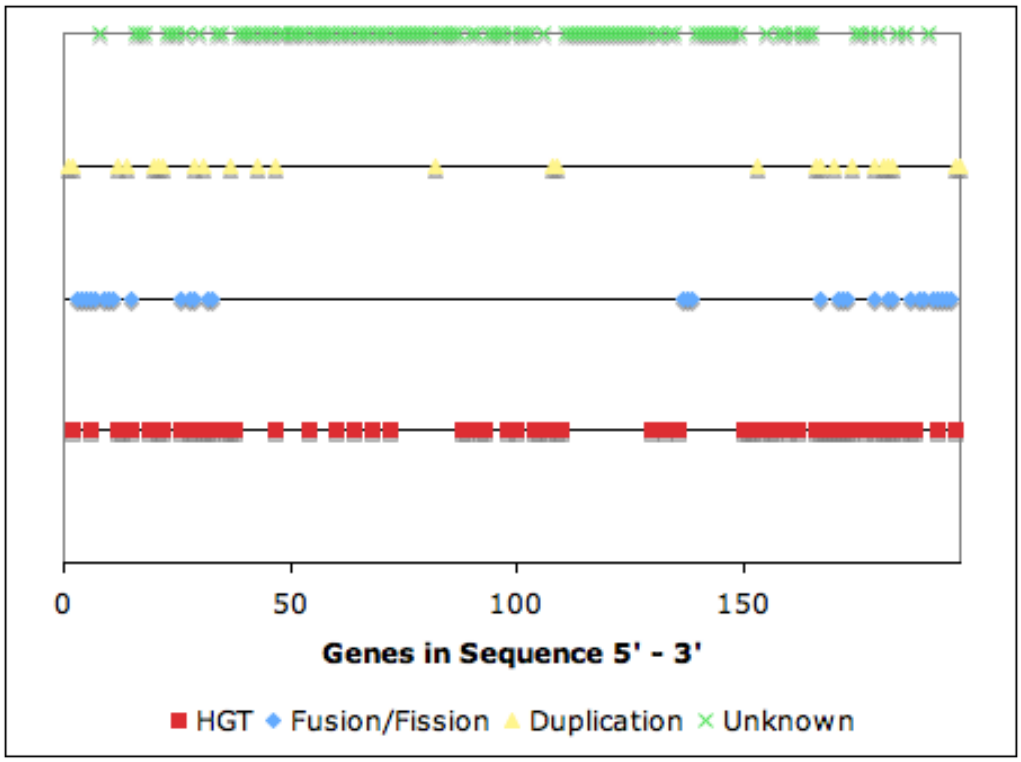


Figure 5.10: Events shaping regions of the VACV-COP genome. Genes are listed according to their position in the genome. Red, blue and yellow labels denote genes for which we have inferred HGT, fusion/fission and duplication events, respectively, green labels are assigned to genes without strong information.

Table 5.4: Evidence for Origins of VACV-COP Genes

Name	**Node of Origin**	**Fusion/Fission**	**Putative Cellular Homologues**	**DNA-level Similarity**	**Verdict**
C23L	6	none	none	none	ITR duplicate, potential A41L duplicate
C22L	2	none	other vertebrates ($7e^{-5}$)	none	HGT, ITR duplicate, potential A53R duplicate
C21L	17	With C20L, fission in VACV-COP	none	other OPV ($1.1e^{-56}$)	Fission, ankyrin, ITR duplicate
C20L	13	With C19L, fission in VACV; with C21L fission in VACV-COP	none	other OPV ($4.3e^{-37}$)	Fission, ankyrin, ITR duplicate
C19L	7	With C20L, fission in VACV, ankyrin fusion	none	weak CNPV (0.98), weak YMTV (1)	Fission, ankyrin, ITR duplicate
C18L	9	Ankyrin fission	bacteria ($3e^{-5}$)	other OPV ($3.5e^{-80}$), weak avipox (0.23)	HGT, ankyrin, ITR duplicate
C17L	7	Ankyrin fusion, Ankyrin fission	none	weak CNPV (0.55)	Ankyrin, ITR duplicate
C16L	7	none	none	none	A46 sf, ITR duplicate
C15L	11	Ankyrin fission	none	other OPV ($2.5e^{-42}$), weak clade II (0.83), weak AmEPV (0.98)	Ankyrin, ITR duplicate
C14L	7	With C13L fission in VACV-COP	none	none	Mechanism unknown, later fission
C13L	17	With C14L fission in VACV-COP	none	none	Fission

Continued on next page ...

Name	Node of Origin	Fusion/Fission	Putative Cellular Homologues	DNA-level Similarity	Verdict
C12L	2	none	mammals ($4e^{-49}$), insects ($1e^{-45}$), other vertebrates ($7e^{-43}$), bacteria ($3e^{-40}$)	weak EPV ($4.3e^{-4}$)	HGT, duplicate of B13R/B14R, potential K2L duplicate
C11R	2	none	mammals ($1e^{-8}$), other vertebrates ($3e^{-7}$)	none	HGT
C10L	2	none	none	weak EPV ($4.4e^{-6}$)	Mechanism unclear, later duplication: C4L
C9L	7	Ankyrin fusion, Ankyrin fission	other eukaryotes ($1e^{-4}$)	weak EPV ($2.3e^{-23}$), weak avipox ($1.4e^{-9}$), weak clade II ($1.8e^{-14}$)	HGT, ankyrin
C8L	7	none	none	weak EPV (0.067), weak CNPV (0.29), weak clade II (0.14)	Mechanism unclear
C7L	6	none	none	weak MsEPV (0.24)	Mechanism unclear
C6L	7	none	none	weak EPV (0.081), weak capripox (0.36)	Mechanism unclear, A46 sf
C5L	7	none	other eukaryotes ($6e^{-7}$), mammals ($4e^{-6}$), other vertebrates ($2e^{-5}$), insects ($1e^{-4}$)	weak EPV ($1.5e^{-4}$), weak clade II ($2.5e^{-4}$), weak avipox (0.0034)	HGT
C4L	7	none	none	weak EPV ($2.4e^{-7}$), weak swinepox (0.0022)	C10L duplicate
C3L	6	none	mammals ($5e^{-44}$), herpesviruses ($9e^{-40}$), other vertebrates ($5e^{-36}$)	weak AmEPV (0.38), weak avipox (0.21)	HGT, potential B5R duplicate

Continued on next page ...

Name	**Node of Origin**	**Fusion/Fission**	**Putative Cellular Homologues**	**DNA-level Similarity**	**Verdict**
C2L	6	none	other eukaryotes ($2e^{-25}$), insects ($9e^{-24}$), other vertebrates ($2e^{-22}$), mammals ($3e^{-22}$)	weak avipox (0.033)	HGT, potential F3L, A55R, B10R duplicate
C1L	7	none	none	weak EPV ($2e^{-6}$), weak avipox (0.13), weak capripox (0.36)	Mechanism unclear
N1L	6	none	none	weak MsEPV (0.098)	Mechanism unclear
N2L	6	none	none	weak EPV (0.032)	Mechanism unclear
M1L	7	Ankyrin fusion, Ankyrin fission	other eukaryotes ($1e^{-18}$), insects ($3e^{-17}$), archaea ($1e^{-16}$), bacteria ($9e^{-15}$), plants ($4e^{-13}$), fungi ($5e^{-13}$), other vertebrates ($2e^{-12}$), mammals ($2e^{-11}$)	weak MsEPV ($4.7e^{-4}$), weak avipox ($1e^{-4}$), weak clade II (0.002)	HGT, ankyrin
M2L	6	none	none	weak MsEPV (0.65)	Mechanism unclear
K1L	7	Ankyrin fission	other eukaryotes ($2e^{-10}$), bacteria ($3e^{-9}$), fungi ($2e^{-8}$), insects ($6e^{-8}$), other vertebrates ($1e^{-7}$), mammals ($2e^{-7}$), archaea ($3e^{-6}$), plant ($6e^{-6}$)	weak EPV (0.24), weak CNPV (0.62), weak clade II (0.13)	HGT, ankyrin
K2L	2	Serpin fission	mammals ($1e^{-35}$), other vertebrates ($5e^{-34}$), bacteria ($1e^{-27}$), archaea ($6e^{-26}$)	weak EPV (0.0015)	HGT, potential C12L, B13R/B14R duplicates
K3L	6	none	none	AmEPV (0.5)	Mechanism unclear

Continued on next page ...

Name	**Node of Origin**	**Fusion/Fission**	**Putative Cellular Homologues**	**DNA-level Similarity**	**Verdict**
K4L	7	none	mammals ($9e^{-101}$), other vertebrates ($6e^{-92}$), insects ($4e^{-81}$), other eukaryotes ($7e^{-80}$), bacteria ($3e^{-29}$), plants ($5e^{-29}$)	weak EPV (0.043), weak avipox ($4.4e^{-4}$)	HGT, potential F13L duplicate
K5L	6	With K6L fission in VACV, RBPV	mammals ($5e^{-14}$), other vertebrates ($5e^{-13}$), other eukaryotes ($2e^{-12}$), bacteria ($9e^{-10}$), insects ($1e^{-5}$), plants ($7e^{-5}$)	none	HGT, later fission
K6L	11	With K5L fission in VACV, RBPV	mammals ($2e^{-19}$), other vertebrates ($2e^{-18}$), other eukaryotes ($1e^{-12}$), bacteria ($4e^{-10}$), insects ($4e^{-7}$), plants ($3e^{-6}$), archaea ($5e^{-5}$)	none	HGT, later fission
K7R	7	none	none	weak EPV (0.32)	Mechanism unclear, A46 sf
F1L	7	none	none	weak EPV ($3.7e^{-4}$), weak avipox (0.34), weak clade II (0.0032)	Mechanism unclear
F2L	1	none	mammals ($1e^{-38}$), other vertebrates ($5e^{-38}$), other eukaryotes ($1e^{-35}$), insects ($4e^{-33}$), bacteria ($6e^{-31}$), viruses ($2e^{-29}$), fungi ($3e^{-27}$), plants ($2e^{-25}$)	none	HGT, possibly multiple times

Continued on next page ...

Name	Node of Origin	Fusion/Fission	Putative Cellular Homologues	DNA-level Similarity	Verdict
F3L	6	none	mammals ($2e^{-26}$), insects ($3e^{-25}$), other eukaryotes ($1e^{-23}$), other vertebrates ($2e^{-22}$)	weak MsEPV ($2.7e^{-5}$), weak avipox (0.018)	HGT, potential C2L, A55R, B10R duplicate
F4L	6	none	other vertebrates ($3e^{-141}$), mammals ($8e^{-141}$), other eukaryotes ($2e^{-134}$), insects ($2e^{-134}$), fungi ($5e^{-124}$), viruses ($7e^{-124}$), bacteria ($7e^{-124}$), plants ($5e^{-122}$)	weak EPV (0.25)	One of multiple HGT events
F5L	7	none	none	weak EPV (0.0057), weak clade II (0.21)	Mechanism unclear
F6L	7	none	none	weak MsEPV (0.68)	Mechanism unclear
F7L	7	none	none	weak EPV (0.0075), weak avipox (0.34), weak clade II (0.082)	Mechanism unclear
F8L	6	none	none	weak AmEPV (0.56)	Mechanism unclear
F9L	1	none	none	none	Potential L1R duplicate, core gene
F10L	1	none	none	none	core gene
F11L	2	none	none	weak MsEPV (0.13)	Mechanism unclear
F12L	2	none	none	weak EPV ($1.7e^{-4}$)	Mechanism unclear
F13L	2	none	mammals ($2e^{-24}$), other vertebrates ($1e^{-23}$), other eukaryotes ($3e^{-21}$), insects ($1e^{-20}$), bacteria ($4e^{-11}$)	weak EPV (0.72)	HGT, potential K4L duplicate

Continued on next page ...

Name	**Node of Origin**	**Fusion/Fission**	**Putative Cellular Homologues**	**DNA-level Similarity**	**Verdict**
F14L	7	none	none	weak deerpox (0.37)	Mechanism unclear
F15L	2	none	none	weak EPV ($6.5e^{-5}$)	Mechanism unclear
F16L	3	none	none	weak EPV ($2.8e^{-4}$), weak CNPV (0.59)	Mechanism unclear
F17R	2	none	none	none	Mechanism unclear
E1L	1	none	none	none	core gene
E2L	2	none	none	weak MsEPV ($2.6e^{-17}$)	Mechanism unclear
E3L	5	none	other eukaryotes ($1e^{-6}$), other vertebrates ($1e^{-5}$), mammals ($4e^{-5}$)	none	HGT
E4L	2	none	none	weak EPV ($3.9e^{-4}$)	Mechanism unclear
E5R	6	none	none	weak EPV ($5.3e^{-7}$), weak avipox (0.027)	Mechanism unclear
E6R	1	none	none	none	core gene
E7R	6	none	none	weak EPV (0.0023)	Mechanism unclear
E8R	2	none	none	weak MsEPV (0.0043)	Mechanism unclear
E9L	1	none	viruses ($9e^{-27}$), other eukaryotes ($2e^{-23}$), fungi ($4e^{-23}$), archaea ($7e^{-21}$), plants ($3e^{-19}$), insects ($9e^{-19}$), other vertebrates ($2e^{-17}$)	none	HGT, core gene
E10R	1	none	none	none	core gene
E11L	2	none	none	none	Mechanism unclear
O1L	2	none	none	weak EPV ($6e^{-10}$)	Mechanism unclear

Continued on next page . . .

Name	Node of Origin	Fusion/Fission	Putative Cellular Homologues	DNA-level Similarity	Verdict
O2L	7	none	mammals ($1e^{-21}$), other vertebrates ($4e^{-21}$), other eukaryotes ($9e^{-16}$), insects ($2e^{-10}$), fungi ($7e^{-10}$), plants ($1e^{-8}$), bacteria ($3e^{-7}$)	weak deerpox (0.92)	HGT
I1L	2	none	none	weak EPV ($3.5e^{-4}$)	Mechanism unclear
I2L	2	none	none	MsEPV (0.9)	Mechanism unclear
I3L	2	none	none	weak EPV (0.3)	Mechanism unclear
I4L	6	none	mammals (0), other vertebrates (0), other eukaryotes (0), insects (0), fungi (0), bacteria (0), plants (0), viruses (0)	weak EPV (0.077), weak CNPV (0.17)	HGT
I5L	2	none	none	none	Mechanism unclear
I6L	2	none	none	weak EPV ($5.3e^{-5}$)	Mechanism unclear
I7L	1	none	none	none	core gene
I8R	1	none	insects ($2e^{-11}$), other eukaryotes ($8e^{-10}$), mammals ($2e^{-9}$), bacteria ($4e^{-8}$), plants ($6e^{-8}$), fungi ($6e^{-8}$), other vertebrates ($6e^{-7}$)	none	HGT, core gene
G1L	1	none	none	none	core gene
G3L	2	none	none	EPV (0.11)	Mechanism unclear
G2R	2	none	none	weak EPV (0.0016)	Mechanism unclear
G4L	2	none	none	none	Mechanism unclear
G5R	1	none	none	none	core gene
G6R	1	none	none	none	core gene

Continued on next page ...

Name	**Node of Origin**	**Fusion/Fission**	**Putative Cellular Homologues**	**DNA-level Similarity**	**Verdict**
G7L	2	none	none	weak EPV ($1.8e^{-4}$)	Mechanism unclear
G8R	2	none	none	weak EPV (0.22)	Mechanism unclear
G9R	1	none	none	none	core gene
L1R	1	none	none	none	Potential F9L duplicate, core gene
L2R	6	none	none	weak AmEPV (0.28), weak FWPV (0.85)	Mechanism unclear
L3L	1	none	none	none	core gene
L4R	2	none	none	weak EPV (0.081)	Mechanism unclear
L5R	2	none	none	weak EPV (0.19)	Mechanism unclear
J1R	2	none	none	EPV (0.43)	Mechanism unclear
J2R	6	none	mammals ($5e^{-59}$), other vertebrates ($4e^{-57}$), other eukaryotes ($2e^{-53}$), insects ($9e^{-42}$), plants ($1e^{-40}$), viruses ($7e^{-31}$), bacteria ($2e^{-26}$)	none	One of multiple HGT events
J3R	1	none	other eukaryotes ($7e^{-9}$)	none	HGT, core gene
J4R	2	none	none	EPV (0.0015)	Mechanism unclear
J5L	1	none	none	none	core gene
J6R	1	none	archaea ($4e^{-54}$), other eukaryotes ($3e^{-50}$), other vertebrates ($4e^{-49}$), mammals ($1e^{-48}$), insects ($5e^{-45}$), fungi ($5e^{-45}$), viruses ($2e^{-41}$), plants ($1e^{-40}$)	none	HGT, core gene

Continued on next page ...

Name	Node of Origin	Fusion/Fission	Putative Cellular Homologues	DNA-level Similarity	Verdict
H1L	1	none	other vertebrates ($1e^{-10}$), other eukaryotes ($2e^{-10}$), insects ($6e^{-10}$), mammals ($1e^{-9}$), plants ($2e^{-8}$), viruses ($6e^{-8}$), fungi ($2e^{-7}$)	none	HGT, core gene
H2R	1	none	none	none	core gene
H3L	2	none	none	weak EPV (0.32)	Mechanism unclear
H4L	1	none	none	none	core gene
H5R	2	none	none	weak MsEPV (0.39)	Mechanism unclear
H6R	1	none	viruses ($2e^{-18}$), bacteria ($2e^{-13}$), other eukaryotes ($2e^{-8}$), plants ($4e^{-7}$)	none	HGT, core gene
H7R	2	none	none	weak AmEPV (0.22)	Mechanism unclear
D1R	1	none	plants ($1e^{-6}$), other eukaryotes ($5e^{-6}$), mammals ($1e^{-5}$), fungi ($8e^{-5}$)	none	HGT, core gene
D2L	2	none	none	weak AmEPV (0.093)	Mechanism unclear
D3R	2	none	none	weak EPV ($1.1e^{-4}$)	Mechanism unclear
D4R	1	none	none	none	core gene
D5R	1	none	bacteria ($8e^{-7}$), viruses ($2e^{-6}$)	none	HGT, core gene
D6R	1	none	viruses ($1e^{-18}$), fungi ($2e^{-9}$), other eukaryotes ($7e^{-5}$)	none	HGT, core gene
D7R	1	none	none	none	core gene
D8L	6	none	mammals ($1e^{-32}$), other vertebrates ($5e^{-30}$), insects ($2e^{-27}$)	weak EPV (0.0016)	HGT
D9R	2	none	none	none	Potential D10R duplicate

Continued on next page ...

Name	**Node of Origin**	**Fusion/Fission**	**Putative Cellular Homologues**	**DNA-level Similarity**	**Verdict**
D10R	1	none	viruses ($4e^{-5}$)	none	Potential D9R duplicate, HGT, core gene
D11L	1	none	viruses ($2e^{-29}$), plants ($4e^{-11}$), bacteria ($1e^{-7}$), other eukaryotes ($1e^{-6}$), fungi ($1e^{-6}$), mammals ($2e^{-6}$), insects ($2e^{-6}$), other vertebrates ($3e^{-6}$)	none	HGT, core gene
D12L	1	none	none	none	core gene
D13L	1	none	none	none	core gene
A1L	1	none	none	none	core gene
A2L	1	none	none	none	core gene
A3L	1	none	none	none	core gene
A4L	2	none	none	none	Mechanism unclear
A5R	1	none	none	none	core gene
A6L	2	none	none	weak AmEPV ($3.9e^{-8}$)	Mechanism unclear
A7L	1	none	none	none	core gene
A8R	2	none	none	EPV ($2.4e^{-5}$)	Mechanism unclear
A9L	2	none	none	weak AmEPV (0.34)	Mechanism unclear
A10L	1	none	none	none	core gene
A11R	1	none	none	none	core gene
A12L	2	none	none	none	Mechanism unclear
A13L	2	none	none	none	Mechanism unclear
A14L	2	none	none	none	Mechanism unclear
A15L	2	none	none	none	Mechanism unclear
A16L	1	none	none	none	core gene
A17L	2	none	none	weak EPV (0.0026)	Mechanism unclear

Continued on next page ...

Name	**Node of Origin**	**Fusion/Fission**	**Putative Cellular Homologues**	**DNA-level Similarity**	**Verdict**
A18R	1	none	bacteria ($4e^{-14}$), other eukaryotes ($3e^{-6}$), archaea ($2e^{-5}$)	none	HGT, core gene
A19L	2	none	none	none	Mechanism unclear
A21L	1	none	none	none	core gene
A20R	1	none	other eukaryotes ($3e^{-5}$)	none	HGT, core gene
A22R	1	none	none	none	core gene
A23R	1	none	none	none	core gene
A24R	1	none	fungi ($2e^{-61}$), archaea ($2e^{-58}$), other eukaryotes ($3e^{-56}$), plants ($8e^{-56}$), insects ($2e^{-51}$), mammals ($5e^{-51}$), other vertebrates ($6e^{-51}$), viruses ($1e^{-50}$)	none	HGT, core gene
A25L	12	With A26L many fission events in VACV lineage (see main text)	none	none	Fission
A26L	2	With A25L many fission events in VACV lineage, with A27L fission at node 5, fusion in VACV-COP	none	none	Sequence present in all ChPV, fission at node 5, fusion in VACV-COP
A27L	5	With A26L fission at node 5	none	weak MsEPV (0.27), weak avipox (0.01)	Fission
A28L	1	none	none	none	core gene
A29L	1	none	none	none	core gene
A30L	2	none	none	AmEPV (0.15)	Mechanism unclear
A31R	2	none	none	weak EPV (0.0098)	Mechanism unclear

Continued on next page . . .

Name	Node of Origin	Fusion/Fission	Putative Cellular Homologues	DNA-level Similarity	Verdict
A32L	1	none	none	none	core gene
A33R	4	none	none	weak AmEPV (0.91)	Mechanism unclear
A34R	2	none	none	weak EPV (0.0033)	Mechanism unclear
A35R	4	none	none	weak EPV (0.0073), weak CNPV (0.93)	Mechanism unclear
A36R	7	none	none	weak EPV ($1.1e^{-4}$), weak avipox (0.41), weak clade II (0.0016)	Mechanism unclear
A37R	4	none	none	weak EPV (0.059)	Mechanism unclear
A38L	6	none	mammals ($2e^{-24}$), other vertebrates ($9e^{-9}$)	weak EPV ($5.1e^{-5}$), weak avipox (0.067)	HGT
A39R	2	none	viruses ($1e^{-32}$), mammals ($5e^{-29}$), other vertebrates ($9e^{-26}$), other eukaryotes ($6e^{-11}$), insects ($5e^{-10}$)	weak AmEPV (0.011), weak clade II (0.2)	HGT
A40R	2	none	mammals ($4e^{-8}$), other eukaryotes ($8e^{-7}$), other vertebrates ($2e^{-6}$), viruses ($5e^{-6}$)	weak EPV (0.064)	HGT
A41L	7	none	none	weak EPV (0.075), weak clade II (0.13)	Potential C23L/B29R duplicate
A42R	7	none	mammals ($4e^{-11}$), bacteria ($1e^{-10}$), other vertebrates ($6e^{-6}$)	none	HGT
A43R	7	none	none	weak avipox (0.002), weak clade II ($9.5e^{-4}$)	Mechanism unclear

Continued on next page ...

Name	**Node of Origin**	**Fusion/Fission**	**Putative Cellular Homologues**	**DNA-level Similarity**	**Verdict**
A44L	1	none	other vertebrates ($3e^{-74}$), mammals ($2e^{-73}$), viruses ($9e^{-64}$), other eukaryotes ($1e^{-48}$), bacteria ($2e^{-26}$), plants ($2e^{-26}$), fungi ($5e^{-26}$), insects ($3e^{-20}$)	none	HGT, core gene
A45R	6	none	mammals ($4e^{-6}$), insects ($6e^{-6}$), other eukaryotes ($1e^{-4}$)	weak MsEPV (0.027)	HGT
A46R	7	none	none	weak EPV (0.0017), weak avipox (0.049), weak clade II (0.039)	Mechanism unclear
A47L	7	none	none	none	Mechanism unclear
A48R	7	none	other vertebrates ($8e^{-44}$), mammals ($9e^{-40}$), insects ($5e^{-39}$), fungi ($6e^{-36}$), other eukaryotes ($2e^{-35}$), plants ($2e^{-35}$), viruses ($3e^{-34}$), archaea ($1e^{-16}$), bacteria ($2e^{-10}$)	weak EPV ($8.8e^{-7}$), weak deerpox (0.36)	HGT, possibly multiple times
A49R	7	none	none	weak AmEPV ($2.9e^{-4}$), weak avipox (0.019), weak clade II (0.0018)	Mechanism unclear

Continued on next page ...

Name	Node of Origin	Fusion/Fission	Putative Cellular Homologues	DNA-level Similarity	Verdict
A50R	2	none	mammals ($3e^{-176}$), other vertebrates ($3e^{-174}$), other eukaryotes ($2e^{-161}$), insects ($2e^{-156}$), viruses ($1e^{-62}$), bacteria ($4e^{-46}$), archaea ($3e^{-44}$), fungi ($7e^{-44}$), plants ($5e^{-37}$)	weak EPV ($6.5e^{-7}$)	HGT, possibly multiple times
A51R	6	none	none	weak EPV ($8.8e^{-4}$), weak CNPV (0.2)	Mechanism unclear
A52R	7	none	none	weak EPV (0.019), weak avipox (0.34)	A46 sf
A54L	14	none	none	other OPV ($1.1e^{-45}$), weak EPV (0.0012), weak avipox (0.14), weak clade II (0.57)	Sequence present in all other OPV, mostly as ORF, lack of annotation
A53R	2	none	mammals ($3e^{-7}$), other vertebrates ($1e^{-5}$)	none	HGT, potential C22L/B28R duplicate or homology through repeated HGT
A55R	6	Kelch fusion, Kelch fission	mammals ($8e^{-50}$), other vertebrates ($5e^{-49}$), other eukaryotes ($2e^{-44}$), insects ($3e^{-43}$)	weak EPV ($1.4e^{-8}$), weak avipox (0.023)	HGT, potential C2L, F3L, B10R duplicate
A56R	6	none	insects ($2e^{-5}$)	weak EPV (0.0077), weak CNPV (0.31)	HGT

Continued on next page ...

Name	**Node of Origin**	**Fusion/Fission**	**Putative Cellular Homologues**	**DNA-level Similarity**	**Verdict**
A57R	7	none	mammals ($6e^{-35}$), other vertebrates ($2e^{-31}$), insects ($8e^{-27}$), other eukaryotes ($5e^{-26}$), fungi ($1e^{-22}$), plants ($2e^{-22}$)	weak deerpox (0.73)	HGT
B1R	1	none	other vertebrates ($3e^{-63}$), mammals ($8e^{-63}$), other eukaryotes ($2e^{-56}$), insects ($3e^{-53}$), viruses ($2e^{-47}$), plants ($2e^{-24}$), fungi ($2e^{-24}$)	none	HGT, B12R duplicate, core gene
B2R	6	With B3R, fission in RBPV, HPXV, VACV	none	none	Mechanism unclear possibly HGT (see B3R), later fission
B3R	11	With B2R, fission in RBPV, HPXV, VACV	mammals ($1e^{-6}$)	none	Initially HGT, fission
B4R	2	Ankyrin fusion, Ankyrin fission	other eukaryotes ($2e^{23}$), bacteria ($6e^{-18}$), insects ($2e^{-16}$), fungi ($4e^{-14}$)	weak EPV ($1.4e^{-13}$)	HGT, ankyrin
B5R	7	none	other eukaryotes ($5e^{-17}$), mammals ($4e^{-16}$), other vertebrates ($3e^{-15}$), viruses ($6e^{-15}$)	weak MsEPV (0.15), weak clade II (0.036)	HGT, potential C3L duplicate
B6R	7	none	none	weak EPV (0.22), weak deerpox (0.16)	Mechanism unclear
B7R	7	none	none	weak EPV ($1.5e^{-5}$), weak avipox (0.084), weak clade II ($8.9e^{-4}$)	Mechanism unclear
B8R	6	none	mammals ($1e^{-6}$)	none	HGT

Continued on next page ...

Name	Node of Origin	Fusion/Fission	Putative Cellular Homologues	DNA-level Similarity	Verdict
B9R	6	none	none	none	Mechanism unclear
B10R	7	Kelch fission	bacteria ($7e^{-5}$)	weak SWPV (0.74)	HGT, potential C2L, F3L, A55R duplicate
B11R	7	none	none	weak EPV ($1.7e^{-5}$), weak avipox (0.025), weak clade II (0.0041)	Mechanism unclear
B12R	1	none	mammals ($2e{-}35$), other vertebrates ($1e^{-34}$), other eukaryotes ($4e^{-28}$), insects ($7e^{-28}$), viruses ($4e^{-26}$), plants ($2e^{-18}$)	none	HGT, B1R duplicate, core gene
B13R	2	With B14R fission in VACV	mammals ($4e^{-8}$), bacteria ($6e^{-7}$), other vertebrates ($1e^{-6}$)	none	HGT, later fission, with B14R C12L duplicate, potential K2L duplicate
B14R	13	With B13R fission in VACV	mammals ($1e^{-29}$), other vertebrates ($1e^{-27}$), other eukaryotes ($1e^{-27}$), bacteria ($7e^{-26}$), insects ($3e^{-24}$)	none	Initial HGT, fission, with B13R C12L duplicate, potential K2L duplicate
B15R	6	none	none	none	A46 sf
B16R	6	none	mammals ($4e^{-30}$), other vertebrates ($6e^{-25}$), other eukaryotes ($3e^{-7}$)	weak EPV (0.026)	HGT
B17L	7	none	none	weak EPV ($1.2e^{-5}$), weak avipox (0.019), weak clade II (0.011)	Mechanism unclear

Continued on next page ...

Name	Node of Origin	Fusion/Fission	Putative Cellular Homologues	DNA-level Similarity	Verdict
B18R	7	Ankyrin fusion, Ankyrin fission	other eukaryotes ($1e^{-26}$), insects ($6e^{-13}$), bacteria ($9e^{-12}$), other vertebrates ($2e^{-9}$), fungi ($9e^{-9}$)	weak EPV ($1.5e^{-15}$), weak avipox ($6e^{-6}$), weak clade II ($1.2e^{-7}$)	HGT, ankyrin
B19R	6	none	mammals ($4e^{-6}$), insects ($2e^{-5}$)	weak EPV (0.08), weak avipox (0.17)	HGT
B20R	7	Ankyrin fission	none	weak EPV (0.0023), weak avipox (0.021), weak clade II (0.026)	Ankyrin
B21R	11	Ankyrin fission	none	other OPV ($2.1e^{-43}$), weak AmEPV (0.91), weak clade II (0.77)	Ankyrin, ITR duplicate
B22R	7	none	none	none	A46 sf, ITR duplicate
B23R	7	Ankyrin fusion, ankyrin fission	none	weak CNPV (0.38), weak YMTV (0.79)	Ankyrin, ITR duplicate
B24R	9	Ankyrin fission	bacteria ($3e^{-5}$)	other OPV ($4.5e^{-83}$), weak avipox (0.19)	HGT, ankyrin, ITR duplicate
B25R	7	With B26R, fission in VACV, ankyrin fusion	none	none	Fission, ankyrin, ITR duplicate
B26R	13	With B27R, fission in VACV-COP; with B25R, fission in VACV	none	other OPV ($2.1e^{-61}$)	Fission, ankyrin, ITR duplicate

Continued on next page ...

Name	**Node of Origin**	**Fusion/Fission**	**Putative Cellular Homologues**	**DNA-level Similarity**	**Verdict**
B27R	17	With B26R, fission in VACV-COP	none	other OPV ($3.7e^{-71}$)	Fission, ankyrin, ITR duplicate
B28R	2	none	other vertebrates ($7e^{-5}$)	none	HGT, ITR duplicate, potential A53R duplicate
B29R	7	none	none	none	ITR duplicate, potential A41L duplicate

Chapter 6

Conclusions

6.1 Overview

Poxviruses are pathogens of vertebrates and insects and cause a range of devastating illnesses in humans and other animals. Any insights into their life cycle, evolution and genomic properties can only enhance our understanding and, ideally, predict viral gene functions and survival strategies. Furthermore, some features of eukaryotic DNA replication, like mRNA capping, are very similar to those found in poxviruses and may even have originated in the viral world (Takemura, 2001), and thus investigating pox evolution is relevant to the study of eukaryotic evolution.

Finding the correct orthologues is crucial in any comparative genomic analysis. In Chapter 2, protein sequences and the results of BLAST similarity searches were used to find groups of homologues across 104 closely related poxvirus genomes. Based on these families, a robust phylogeny of poxviruses was inferred, that agreed with previously published phylogenies of fewer poxvirus genomes.

Chapter 3 describes an in-depth analysis of horizontal gene transfer into poxviruses from their hosts, with a focus on distinguishing between single and multiple independent HGT events. This distinction relies heavily on

the conserved gene neighbourhood (CGN) present in most poxviruses. Being able to use CGN information is a huge advantage and many studies now take it into account when determining orthologues (Knowles and McLysaght, 2009; Tian *et al.*, 2009). The genes found in this analysis to have undergone multiple HGT events are involved in drug susceptibility (TK), host immune manipulation (IL-10) and response to environmental damage (GP and DP), lending plausibility to the idea that genes transferred multiple times may be of particular advantage to virus survival. If the viral eukaryogenesis theory is true, which suggests that a DNA virus was the ancestor of the nucleus (Bell, 2001), then some genes may have been ancestral to DNA viruses and what was to become eukaryotes.

In Chapter 4, I attempted to use bioinformatics methods to predict the function of poxvirus genes by inferring protein-protein interactions (PPI) based on phylogenetic patterns. When it remained difficult to predict PPI computationally, known interactions were used to find any commonalities in the genes involved. Location, conservation and expression were examined, yet what emerges from this study is only one clear pattern: that proteins involved in PPI tend to be highly conserved among poxviruses and these proteins are often located near the genomic centre.

Chapter 5 provides a look at every gene in a vaccinia virus reference genome with the aim to find its point of origin on the poxvirus phylogenetic tree, and the mechanisms by which it originated. Specifically, gene fusion/fission, HGT and de novo origins were the mechanisms tested. Unfortunately, for 38% of genes no convincing evidence for any particular mechanism of origination could be found. Roughly 32% exhibit homology to some cellular or (non-pox)viral genes and are therefore implicated in HGT, more than expected after the much stricter analysis of Chapter 3. Where no traces of the mechanism of origin can be found, we have to speculate; it is possible that many more genes have reached the poxvirus genomes through

horizontal transfer but their donor genomes have lost the sequence, it has diverged beyond recognition or has not yet been sequenced. De novo gene formation appears to be rare, and this is to be expected in a group of genomes that use genomic space very efficiently with few non-coding regions.

6.2 Future work

This thesis describes different approaches of gaining information about viral gene functions and viral strategies from looking at the viral genome.

Another type of genomic feature on which preliminary work was carried out was the presence of micro RNAs (miRNA) in poxviruses. MiRNAs are short segments of the genome that do not code for a protein and instead are complementary to, and bind to, the 3' UTR of other genes, where they downregulate transcription. One miRNA can regulate several different target sites and one gene can be co-regulated by several miRNAs (Hua *et al.*, 2006). Animal miRNAs are transcribed from so-called pri-miRNAs, which are processed to pre-miRNAs, approximately 70 nt in length, capable of forming stem-loop (SL) structures, and then to mature miRNAs, only 21-23 nt long, which include the region complementary to the target gene, with complementarity of nucleotides 2-8 (the seed) being most crucial to binding (reviewed by Du and Zamore (2007)). Their presence has been shown in viruses, which opens many new areas for investigation: Do all types of viruses encode them? Do they function the same way as those in their host genomes? How did they get into the viral genome and how did they evolve? What are their target genes in the virus and the host genome? Does this lead to any new potential drug target genes?

Encoding miRNAs is clearly advantageous for viruses as they provide a lot of regulation potential in a very small space and are non-antigenic.

Much is now known about presence and target genes of miRNAs in polyomavirus (Sullivan *et al.*, 2005), HIV (Bennasser *et al.*, 2004),

herpesviruses like cytomegalovirus (Stern-Ginossar *et al.*, 2009) and KSHV (Cai *et al.*, 2005), and other large DNA viruses (Pfeffer *et al.*, 2005), among others. Sarnow *et al.* (2006) have reviewed the different known instances of viruses encoding miRNA and, vice versa, being regulated by host miRNA. The first step in finding viral miRNAs is usually a computational prediction of candidate regions. They are difficult to identify through cross-species sequence conservation as they tend not to have close homologues, but instead are found based on structural information on mammalian miRNAs, and many different methods have been introduced (Pfeffer *et al.*, 2004; Thomassen *et al.*, 2006; Grundhoff *et al.*, 2006; Pan *et al.*, 2007).

Conceptual difficulties arise when dealing with poxviruses that remain in the host's cytoplasm, as normal miRNA processing involves enzymes like Drosha, located in the nucleus. However, a Drosha-less route of miRNA processing has been described by Ruby *et al.* (2007) for miRNAs encoded within introns, which suggests that alternative pathways are yet to be discovered.

Our aim was to find similarities between potential poxvirus miRNAs and known animal miRNAs and investigate the evolutionary processes that led to their divergence, with the hypothesis that some viral miRNAs may regulate host genes. We used the candidate poxviral miRNAs predicted by Li *et al.* (2008), listed in the Vir-Mir database (alk.ibms.sinica.edu.tw/). At the time of access (October 2007) this included 1015 unique poxviral predicted SL structures, the vast majority from parapox viruses. These were used in a search (using *ssearch*) against the roughly 5000 mature miRNAs from the miRBase database (www.mirbase.org, version 10.0, data 08/07), created by Griffiths-Jones (2006). It was reasonable to assume that virus miRNAs would regulate the same targets as a homologous animal miRNAs. After applying various restriction criteria, 14 groups of potential homologous miRNAs were found, with most of the viral sequences being encoded within

the region of a gene, which adds some uncertainty about the plausibility of these miRNA candidates. If future work was to be carried out into the evolution of poxvirus miRNAs, it would be interesting to examine the functions of the homologous animal miRNA target genes. Only 19 of these had known targets with annotated gene ontology (GO) terms in miRBase in 2007. Perhaps more are available now, and one hypothesis to investigate would be an overrepresentation of immune-related functions amongst these target genes. With more mature miRNAs now present in miRBase and the possibility to find SL candidates in other poxviruses, a more comprehensive analysis could be carried out.

Homologous matches between the stem part of a candidate SL structure and a mature animal miRNA, particularly if they involve the seed region, imply a transfer of non-coding genetic material between host and virus, with the potential for the virus to regulate the same targets as the host miRNA, opening yet another avenue for the virus to disrupt and manipulate the host immune system. Finding these homologous pairs computationally, and confirming the target regulation experimentally would be a very interesting step forward in poxvirus genomics and our preliminary analyses suggest that this might be a fruitful endeavour.

6.3 Summary

In summary, this thesis elucidates various things about poxvirus genomes. A robust, up-to-date phylogeny is inferred based on core families present in all 104 genomes. This phylogeny and the conserved synteny of chordopox virus genes are then used to identify single and multiple instances of horizontal transfer of genes from a host to a poxvirus. For five genes, multiple HGT origins are implied. Next, phylogenetic profiles, conservation, location and expression data of poxvirus genes are used to gain insight into protein-protein interactions between proteins encoded by vaccinia virus Copenhagen, with

the main conclusion that such interactions tend to take place between genes that have been present for nearly the entire poxvirus lineage. Lastly, each gene in the vaccinia virus Copenhagen genome is examined for evidence of its point and mechanism of origin, with a third of genes seemingly originating via horizontal transfer.

Bibliography

Afonso, C. L., Tulman, E. R., Lu, Z., Zsak, L., Kutish, G. F., and Rock, D. L. (2000). The genome of fowlpox virus. *J Virol*, **74**(8), 3815–31.

Afonso, C. L., Delhon, G., Tulman, E. R., Lu, Z., Zsak, A., Becerra, V. M., Zsak, L., Kutish, G. F., and Rock, D. L. (2005). Genome of deerpox virus. *J Virol*, **79**(2), 966–77.

Afonso, C. L., Tulman, E. R., Delhon, G., Lu, Z., Viljoen, G. J., Wallace, D. B., Kutish, G. F., and Rock, D. L. (2006). Genome of crocodilepox virus. *J Virol*, **80**(10), 4978–91.

Alcami, A. and Smith, G. L. (1995a). Cytokine receptors encoded by poxviruses: a lesson in cytokine biology. *Immunol Today*, **16**(10), 474–8.

Alcami, A. and Smith, G. L. (1995b). Vaccinia, cowpox, and camelpox viruses encode soluble gamma interferon receptors with novel broad species specificity. *J Virol*, **69**(8), 4633–9.

Alcami, A., Symons, J. A., Collins, P. D., Williams, T. J., and Smith, G. L. (1998). Blockade of chemokine activity by a soluble chemokine binding protein from vaccinia virus. *J Immunol*, **160**(2), 624–33.

Alibek, K. (1999). *Biohazard*. Arrow Books.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, **215**(3), 403–10.

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**(17), 3389–402.

Amano, H., Ueda, Y., and Miyamura, T. (1995). Identification and characterization of the thymidine kinase gene of yaba virus. *J Gen Virol*, **76**(Pt 5), 1109–15.

Anantharaman, V. and Aravind, L. (2003). Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes. *Genome Biol*, **4**(2), R11.

Aravind, L., Watanabe, H., Lipman, D. J., and Koonin, E. V. (2000). Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc Natl Acad Sci U S A*, **97**(21), 11319–24.

Assarsson, E., Greenbaum, J. A., Sundstrom, M., Schaffer, L., Hammond, J. A., Pasquetto, V., Oseroff, C., Hendrickson, R. C., Lefkowitz, E. J., Tscharke, D. C., *et al.* (2008). Kinetic analysis of a complete poxvirus transcriptome reveals an immediate-early class of genes. *Proc Natl Acad Sci U S A*, **105**(6), 2140–5.

Babiuk, S., Bowden, T. R., Boyle, D. B., Wallace, D. B., and Kitching, R. P. (2008). Capripoxviruses: an emerging worldwide threat to sheep, goats and cattle. *Transbound Emerg Dis*, **55**(7), 263–72.

Bajszar, G., Wittek, R., Weir, J. P., and Moss, B. (1983). Vaccinia virus thymidine kinase and neighboring genes: mRNAs and polypeptides of wild-type virus and putative nonsense mutants. *J Virol*, **45**(1), 62–72.

Baldo, A. M. and McClure, M. A. (1999). Evolution and horizontal transfer of dUTPase-encoding genes in viruses and their hosts. *J Virol*, **73**(9), 7710–21.

Ball, L. A. (1987). High-frequency homologous recombination in vaccinia virus DNA. *J Virol*, **61**(6), 1788–95.

Bamford, D. H., Burnett, R. M., and Stuart, D. I. (2002). Evolution of viral structure. *Theor Popul Biol*, **61**(4), 461–70.

Banadyga, L., Gerig, J., Stewart, T., and Barry, M. (2007). Fowlpox virus encodes a Bcl-2 homologue that protects cells from apoptotic death through interaction with the proapoptotic protein Bak. *J Virol*, **81**(20), 11032–45.

Barinaga, M. (1999). A new finger on the protein destruction button. *Science*, **286**(5438), 223, 225.

Barker, D. and Pagel, M. (2005). Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput Biol*, **1**(1), e3.

Barker, D., Meade, A., and Pagel, M. (2007). Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics*, **23**(1), 14–20.

Barnes, D., Kunitomi, M., Vignuzzi, M., Saksela, K., and Andino, R. (2008). Harnessing endogenous mirnas to control virus tissue tropism as a strategy for developing attenuated virus vaccines. *Cell Host Microbe*, **4**(3), 239–48.

Barry, M. and McFadden, G. (1997). Virus encoded cytokines and cytokine receptors. *Parasitology*, **115 Suppl**(100), S89–100.

Bartlett, N. W., Dumoutier, L., Renauld, J. C., Kotenko, S. V., McVey, C. E., Lee, H. J., and Smith, G. L. (2004). A new member of the interleukin 10-related cytokine family encoded by a poxvirus. *J Gen Virol*, **85**(Pt 6), 1401–12.

Bell, P. J. (2001). Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus? *J Mol Evol*, **53**(3), 251–6.

Bennasser, Y., Le, S. Y., Yeung, M. L., and Jeang, K. T. (2004). HIV-1 encoded candidate micro-RNAs and their cellular targets. *Retrovirology*, **1**(43), 43.

Bennink, J. R. and Yewdell, J. W. (1990). Recombinant vaccinia viruses as vectors for studying T lymphocyte specificity and function. *Curr Top Microbiol Immunol*, **163**, 153–84.

Binns, M. M., Tomley, F. M., Campbell, J., and Boursnell, M. E. (1988). Comparison of a conserved region in fowlpox virus and vaccinia virus genomes and the translocation of the fowlpox virus thymidine kinase gene. *J Gen Virol*, **69 (Pt 6)**(Pt 6), 1275–83.

Binns, M. M., Boursnell, M. E., and Skinner, M. A. (1992). Gene translocations in poxviruses: the fowlpox virus thymidine kinase gene is flanked by 15 bp direct repeats and occupies the locus which in vaccinia virus is occupied by the ribonucleotide reductase large subunit gene. *Virus Res*, **24**(2), 161–72.

Bisht, H., Weisberg, A. S., and Moss, B. (2008). Vaccinia virus L1 protein is required for cell entry and membrane fusion. *J Virol*, **82**(17), 8687–94.

Boekhorst, J., van Breukelen, B., Heck, A. J., and Snel, B. (2008). Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. *Genome Biol*, **9**(10), R144.

Bork, P. (1993). Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally? *Proteins*, **17**(4), 363–74.

Bowie, A., Kiss-Toth, E., Symons, J. A., Smith, G. L., Dower, S. K., and O'Neill, L. A. (2000). A46R and A52R from vaccinia virus are antagonists of host IL-1 and toll-like receptor signaling. *Proc Natl Acad Sci U S A*, **97**(18), 10162–7.

Bowie, A. G., Zhan, J., and Marshall, W. L. (2004). Viral appropriation of apoptotic and NF-kappaB signaling pathways. *J Cell Biochem*, **91**(6), 1099–108.

Boyle, D. B., Coupar, B. E., Gibbs, A. J., Seigman, L. J., and Both, G. W. (1987). Fowlpox virus thymidine kinase: nucleotide sequence and relationships to other thymidine kinases. *Virology*, **156**(2), 355–65.

Bratke, K. A. and McLysaght, A. (2008). Identification of multiple independent horizontal gene transfers into poxviruses using a comparative genomics approach. *BMC Evol Biol*, **8**(67), 67.

Bray, M. and Buller, M. (2004). Looking back at smallpox. *Clin Infect Dis*, **38**(6), 882–9.

Breitkreutz, B. J., Stark, C., and Tyers, M. (2003). Osprey: a network visualization system. *Genome Biol*, **4**(3), R22.

Breman, J. G. and Henderson, D. A. (2002). Diagnosis and management of smallpox. *N Engl J Med*, **346**(17), 1300–8.

Brown, J. R. (2003). Ancient horizontal gene transfer. *Nat Rev Genet*, **4**(2), 121–32.

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

Burke, D. S., Epstein, J. M., Cummings, D. A., Parker, J. I., Cline, K. C., Singa, R. M., and Chakravarty, S. (2006). Individual-based computational modeling of smallpox epidemic control strategies. *Acad Emerg Med*, **13**(11), 1142–9.

Burnett, R. M. (2006). More barrels from the viral tree of life. *Proc Natl Acad Sci U S A*, **103**(1), 3–4.

Cabirac, G. F., Mulloy, J. J., Strayer, D. S., Sell, S., and Leibowitz, J. L. (1986). Transcriptional mapping of early RNA from regions of the shope fibroma and malignant rabbit fibroma virus genomes. *Virology*, **153**(1), 53–69.

Cai, X., Lu, S., Zhang, Z., Gonzalez, C. M., Damania, B., and Cullen, B. R. (2005). Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. *Proc Natl Acad Sci U S A*, **102**(15), 5570–5.

Castro, M. A., Dalmolin, R. J., Moreira, J. C., Mombach, J. C., and de Almeida, R. M. (2008). Evolutionary origins of human apoptosis and genome-stability gene networks. *Nucleic Acids Res*, **2**, 2.

Chan, A., Baird, M., Mercer, A. A., and Fleming, S. B. (2006). Maturation and function of human dendritic cells are inhibited by orf virus-encoded interleukin-10. *J Gen Virol*, **87**(Pt 11), 3177–81.

Charrel, R. N., de Lamballerie, X., and Emonet, S. (2008). Phylogeny of the genus arenavirus. *Curr Opin Microbiol*, **16**, 16.

Choi, I. G. and Kim, S. H. (2007). Global extent of horizontal gene transfer. *Proc Natl Acad Sci U S A*, **104**(11), 4489–94.

Ciulla, E., Emery, A., Konz, D., and Krushkal, J. (2005). Evolutionary history of orthopoxvirus proteins similar to human complement regulators. *Gene*, **355**, 40–7.

Clamp, M., Cuff, J., Searle, S. M., and Barton, G. J. (2004). The Jalview Java alignment editor. *Bioinformatics*, **20**(3), 426–7.

Condit, R. C. and Niles, E. G. (1990). Orthopoxvirus genetics. *Curr Top Microbiol Immunol*, **163**, 1–39.

Condit, R. C. and Niles, E. G. (2002). Regulation of viral transcription elongation and termination during vaccinia virus infection. *Biochim Biophys Acta*, **1577**(2), 325–36.

Crosby, A. W. (1976). Virgin soil epidemics as a factor in the aboriginal depopulation in america. *William Mary Q*, **33**, 289–99.

Da Silva, M. and Upton, C. (2005). Host-derived pathogenicity islands in poxviruses. *Virol J*, **2**(30), 30.

Darwin, C. (1839). Journal and remarks.

Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genet*, **2**(5), e68.

d'Herelle, F. (1924). *Immunity in natural infectious disease*. Baltimore: Williams and Wilkins.

Diamond, J. (1997). *Guns, germs and steel*. Vintage, London.

Doolittle, W. F. (1999). Lateral genomics. *Trends Cell Biol*, **9**(12), M5–8.

Du, T. and Zamore, P. D. (2007). Beginning to understand microRNA function. *Cell Res*, **17**(8), 661–3.

Du, X., Wang, Z., Wu, A., Song, L., Cao, Y., Hang, H., and Jiang, T. (2008). Networks of genomic co-occurrence capture characteristics of human influenza a (H3N2) evolution. *Genome Res*, **18**(1), 178–87.

Eaton, H. E., Metcalf, J., and Brunetti, C. R. (2008). Characterization of the promoter activity of a poxvirus conserved element. *Can J Microbiol*, **54**(6), 483–8.

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**(5), 1792–7.

Elde, N. C., Child, S. J., Geballe, A. P., and Malik, H. S. (2008). Protein kinase R reveals an evolutionary model for defeating viral mimicry. *Nature*, **30**, 30.

Enright, A. J. and Ouzounis, C. A. (2001). Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol*, **2**(9), 34.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**(6757), 86–90.

Esposito, J. J., Sammons, S. A., Frace, A. M., Osborne, J. D., Olsen-Rasmussen, M., Zhang, M., Govil, D., Damon, I. K., Kline, R., Laker, M., *et al.* (2006). Genome sequence diversity and clues to the evolution of variola (smallpox) virus. *Science*, **313**(5788), 807–12.

Ettema, T., van der Oost, J., and Huynen, M. (2001). Modularity in the gain and loss of genes: applications for function prediction. *Trends Genet*, **17**(9), 485–7.

Everett, H. and McFadden, G. (2002). Poxviruses and apoptosis: a time to die. *Curr Opin Microbiol*, **5**(4), 395–402.

Felsenstein, J. (2005). Phylip (phylogeny inference program).

Fenner, F., Henderson, D., Arita, I., Jezek, Z., and Ladnyi, I. (1988). *Smallpox and its eradication*. World Health Organization.

Filee, J., Pouget, N., and Chandler, M. (2008). Phylogenetic evidence for extensive lateral acquisition of cellular genes by nucleocytoplasmic large DNA viruses. *BMC Evol Biol*, **8**(1), 320.

Fleming, S. B., McCaughan, C. A., Andrews, A. E., Nash, A. D., and Mercer, A. A. (1997). A homolog of interleukin-10 is encoded by the poxvirus orf virus. *J Virol*, **71**(6), 4857–61.

Fleming, S. B., Anderson, I. E., Thomson, J., Deane, D. L., McInnes, C. J., McCaughan, C. A., Mercer, A. A., and Haig, D. M. (2007). Infection with recombinant orf viruses demonstrates that the viral interleukin-10 is a virulence factor. *J Gen Virol*, **88**(Pt 7), 1922–7.

Forterre, P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res*, **117**(1), 5–16.

Frevel, M. A., Bakheet, T., Silva, A. M., Hissong, J. G., Khabar, K. S., and Williams, B. R. (2003). p38 mitogen-activated protein kinase-dependent and -independent signaling of mRNA stability of AU-rich element-containing transcripts. *Mol Cell Biol*, **23**(2), 425–36.

Gadagkar, S. R., Rosenberg, M. S., and Kumar, S. (2005). Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J Exp Zoolog B Mol Dev Evol*, **304**(1), 64–74.

Galvani, A. P. and Slatkin, M. (2003). Evaluating plague and smallpox as historical selective pressures for the CCR5-Delta 32 HIV-resistance allele. *Proc Natl Acad Sci U S A*, **100**(25), 15276–9.

Gardner, J. D., Tscharke, D. C., Reading, P. C., and Smith, G. L. (2001). Vaccinia virus semaphorin A39R is a 50-55 kDa secreted glycoprotein that affects the outcome of infection in a murine intradermal model. *J Gen Virol*, **82**(Pt 9), 2083–93.

Geddes, A. M. (2006). The history of smallpox. *Clin Dermatol*, **24**(3), 152–7.

Gershon, P. D., Kitching, R. P., Hammond, J. M., and Black, D. N. (1989). Poxvirus genetic recombination during natural virus transmission. *J Gen Virol*, **70**(Pt 2), 485–9.

Goebel, S. J., Johnson, G. P., Perkus, M. E., Davis, S. W., Winslow, J. P., and Paoletti, E. (1990). The complete DNA sequence of vaccinia virus. *Virology*, **179**(1), 247–66, 517–63.

Golini, F. and Kates, J. R. (1984). Transcriptional and translational analysis of a strongly expressed early region of the vaccinia virus genome. *J Virol*, **49**(2), 459–70.

Griffiths-Jones, S. (2006). miRBase: the microRNA sequence database. *Methods Mol Biol*, **342**, 129–38.

Grundhoff, A., Sullivan, C. S., and Ganem, D. (2006). A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses. *Rna*, **12**(5), 733–50.

Gubser, C. and Smith, G. L. (2002). The sequence of camelpox virus shows it is most closely related to variola virus, the cause of smallpox. *J Gen Virol*, **83**(Pt 4), 855–72.

Gubser, C., Hue, S., Kellam, P., and Smith, G. L. (2004). Poxvirus genomes: a phylogenetic analysis. *J Gen Virol*, **85**(Pt 1), 105–17.

Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, **52**(5), 696–704.

Hayakawa, T., Xu, J., and Hukuhara, T. (1996). Cloning and sequencing of the gene for an enhancing factor from pseudaletia separata entomopoxvirus. *Gene*, **177**(1-2), 269–70.

Henderson, D. A., Inglesby, T. V., Bartlett, J. G., Ascher, M. S., Eitzen, E., Jahrling, P. B., Hauer, J., Layton, M., McDade, J., Osterholm, M. T.,

et al. (1999). Smallpox as a biological weapon: medical and public health management. working group on civilian biodefense. *Jama*, **281**(22), 2127–37.

Herniou, E. A., Luque, T., Chen, X., Vlak, J. M., Winstanley, D., Cory, J. S., and O'Reilly, D. R. (2001). Use of whole genome sequence data to infer baculovirus phylogeny. *J Virol*, **75**(17), 8117–26.

Hertig, C., Coupar, B. E., Gould, A. R., and Boyle, D. B. (1997). Field and vaccine strains of fowlpox virus carry integrated sequences from the avian retrovirus, reticuloendotheliosis virus. *Virology*, **235**(2), 367–76.

Holmes, E. C. (2008). Evolutionary history and phylogeography of human viruses. *Annu Rev Microbiol*, **62**, 307–28.

Holzerlandt, R., Orengo, C., Kellam, P., and Alba, M. M. (2002). Identification of new herpesvirus gene homologs in the human genome. *Genome Res*, **12**(11), 1739–48.

Hu, F. Q., Smith, C. A., and Pickup, D. J. (1994). Cowpox virus contains two copies of an early gene encoding a soluble secreted form of the type II TNF receptor. *Virology*, **204**(1), 343–56.

Hua, Z., Lv, Q., Ye, W., Wong, C. K., Cai, G., Gu, D., Ji, Y., Zhao, C., Wang, J., Yang, B. B., *et al.* (2006). MiRNA-directed regulation of VEGF and other angiogenic factors under hypoxia. *PLoS ONE*, **1**(1), e116.

Huang, J., Huang, Q., Zhou, X., Shen, M. M., Yen, A., Yu, S. X., Dong, G., Qu, K., Huang, P., Anderson, E. M., *et al.* (2004). The poxvirus p28 virulence factor is an E3 ubiquitin ligase. *J Biol Chem*, **279**(52), 54110–6.

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.

Hughes, A. L. and Friedman, R. (2000). Evolutionary diversification of protein-coding genes of hantaviruses. *Mol Biol Evol*, **17**(10), 1558–68.

Hughes, A. L. and Friedman, R. (2003). Genome-wide survey for genes horizontally transferred from cellular organisms to baculoviruses. *Mol Biol Evol*, **20**(6), 979–87.

Hughes, A. L. and Friedman, R. (2005). Poxvirus genome evolution by gene gain and loss. *Mol Phylogenet Evol*, **35**(1), 186–95.

Hulsen, T., de Vlieg, J., and Groenen, P. M. (2006). Phylopat: phylogenetic pattern analysis of eukaryotic genes. *BMC Bioinformatics*, **7**(398), 398.

Hulsen, T., Groenen, P. M., de Vlieg, J., and Alkema, W. (2008). Phylopat: an updated version of the phylogenetic pattern database contains gene neighborhood. *Nucleic Acids Res*, **2**, 2.

Hummel, S., Schmidt, D., Kremeyer, B., Herrmann, B., and Oppermann, M. (2005). Detection of the CCR5-Delta32 HIV resistance gene in bronze age skeletons. *Genes Immun*, **6**(4), 371–4.

Huynen, M. A. and Bork, P. (1998). Measuring genome evolution. *Proc Natl Acad Sci U S A*, **95**(11), 5849–56.

Inder, M. K., Ueda, N., Mercer, A. A., Fleming, S. B., and Wise, L. M. (2007). Bovine papular stomatitis virus encodes a functionally distinct VEGF that binds both VEGFR-1 and VEGFR-2. *J Gen Virol*, **88**(Pt 3), 781–91.

Iyer, L. M., Aravind, L., and Koonin, E. V. (2001). Common origin of four diverse families of large eukaryotic DNA viruses. *J Virol*, **75**(23), 11720–34.

Iyer, L. M., Balaji, S., Koonin, E. V., and Aravind, L. (2006). Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res*, **117**(1), 156–84.

Johnston, J. B. and McFadden, G. (2003). Poxvirus immunomodulatory strategies: current perspectives. *J Virol*, **77**(11), 6093–100.

Johnston, J. B., Barrett, J. W., Nazarian, S. H., Goodwin, M., Ricciuto, D., Wang, G., and McFadden, G. (2005). A poxvirus-encoded pyrin domain protein

interacts with ASC-1 to inhibit host inflammatory and apoptotic responses to infection. *Immunity*, **23**(6), 587–98.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, **8**(3), 275–82.

Katju, V. and Lynch, M. (2006). On the formation of novel genes by duplication in the caenorhabditis elegans genome. *Mol Biol Evol*, **23**(5), 1056–67.

Khayat, R., Tang, L., Larson, E. T., Lawrence, C. M., Young, M., and Johnson, J. E. (2005). Structure of an archaeal virus capsid protein reveals a common ancestry to eukaryotic and bacterial viruses. *Proc Natl Acad Sci U S A*, **102**(52), 18944–9.

Knowles, D. G. and McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome Res*, **2**, 2.

Koonin, E. V. and Dolja, V. V. (2006). Evolution of complexity in the viral world: the dawn of a new vision. *Virus Res*, **117**(1), 1–4.

Koonin, E. V. and Senkevich, T. G. (1992). Evolution of thymidine and thymidylate kinases: the possibility of independent capture of TK genes by different groups of viruses. *Virus Genes*, **6**(2), 187–96.

Koonin, E. V., Senkevich, T. G., and Dolja, V. V. (2006). The ancient virus world and evolution of cells. *Biol Direct*, **1**(29), 29.

Kotenko, S. V., Saccani, S., Izotova, L. S., Mirochnitchenko, O. V., and Pestka, S. (2000). Human cytomegalovirus harbors its own unique IL-10 homolog (cmvIL-10). *Proc Natl Acad Sci U S A*, **97**(4), 1695–700.

Krupovic, M. and Bamford, D. H. (2008). Virus evolution: how far does the double beta-barrel viral lineage extend? *Nat Rev Microbiol*, **6**(12), 941–8.

Kruse, N. and Weber, O. (2001). Selective induction of apoptosis in antigen-presenting cells in mice by parapoxvirus ovis. *J Virol*, **75**(10), 4699–704.

Kurilla, M. G., Swaminathan, S., Welsh, R. M., Kieff, E., and Brutkiewicz, R. R. (1993). Effects of virally expressed interleukin-10 on vaccinia virus infection in mice. *J Virol*, **67**(12), 7623–8.

Lawrence, J. G. and Ochman, H. (1998). Molecular archaeology of the escherichia coli genome. *Proc Natl Acad Sci U S A*, **95**(16), 9413–7.

Lawrence, J. G. and Ochman, H. (2002). Reconciling the many faces of lateral gene transfer. *Trends Microbiol*, **10**(1), 1–4.

Lee, H. J., Essani, K., and Smith, G. L. (2001). The genome sequence of yaba-like disease virus, a yatapoxvirus. *Virology*, **281**(2), 170–92.

Lee-Chen, G. J. and Niles, E. G. (1988). Transcription and translation mapping of the 13 genes in the vaccinia virus HindIII D fragment. *Virology*, **163**(1), 52–63.

Legrand, F. A., Verardi, P. H., Chan, K. S., Peng, Y., Jones, L. A., and Yilma, T. D. (2005). Vaccinia viruses with a serpin gene deletion and expressing IFN-gamma induce potent immune responses without detectable replication in vivo. *Proc Natl Acad Sci U S A*, **102**(8), 2940–5.

Li, S. C., Shiau, C. K., and Lin, W. C. (2008). Vir-Mir db: prediction of viral microRNA candidate hairpins. *Nucleic Acids Res*, **36**(Database issue), D184–9.

Li, Y., Carroll, D. S., Gardner, S. N., Walsh, M. C., Vitalis, E. A., and Damon, I. K. (2007). On the origin of smallpox: correlating variola phylogenics with historical smallpox records. *Proc Natl Acad Sci U S A*, **104**(40), 15787–92.

Likos, A. M., Sammons, S. A., Olson, V. A., Frace, A. M., Li, Y., Olsen-Rasmussen, M., Davidson, W., Galloway, R., Khristova, M. L., Reynolds, M. G., *et al.* (2005). A tale of two clades: monkeypox viruses. *J Gen Virol*, **86**(Pt 10), 2661–72.

Lin, Y. C., Li, J., Irwin, C. R., Jenkins, H., DeLange, L., and Evans, D. H. (2008). Vaccinia virus DNA ligase recruits cellular topoisomerase II to sites of viral replication and assembly. *J Virol*, **82**(12), 5922–32.

Liu, L., Xia, H., Wahlberg, N., Belak, S., and Baule, C. (2009). Phylogeny, classification and evolutionary insights into pestiviruses. *Virology*, **385**(2), 351–7.

Long, M., Betran, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nat Rev Genet*, **4**(11), 865–75.

Longini, I. M., J., Halloran, M. E., Nizam, A., Yang, Y., Xu, S., Burke, D. S., Cummings, D. A., and Epstein, J. M. (2007). Containing a large bioterrorist smallpox attack: a computer simulation approach. *Int J Infect Dis*, **11**(2), 98–108.

Lucas, A. and McFadden, G. (2004). Secreted immunomodulatory viral proteins as novel biotherapeutics. *J Immunol*, **173**(8), 4765–74.

Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, **290**(5494), 1151–5.

Macaulay, C., Upton, C., and McFadden, G. (1987). Tumorigenic poxviruses: transcriptional mapping of the terminal inverted repeats of shope fibroma virus. *Virology*, **158**(2), 381–93.

Macen, J. L., Upton, C., Nation, N., and McFadden, G. (1993). Serp1, a serine proteinase inhibitor encoded by myxoma virus, is a secreted glycoprotein that interferes with inflammation. *Virology*, **195**(2), 348–63.

Mackett, M. and Archard, L. C. (1979). Conservation and variation in orthopoxvirus genome structure. *J Gen Virol*, **45**(3), 683–701.

Mahalingam, S., Damon, I. K., and Lidbury, B. A. (2004). 25 years since the eradication of smallpox: why poxvirus research is still relevant. *Trends Immunol*, **25**(12), 636–9.

Martin, W. and Muller, M. (1998). The hydrogen hypothesis for the first eukaryote. *Nature*, **392**(6671), 37–41.

McCormack, G. P. and Clewley, J. P. (2002). The application of molecular phylogenetics to the analysis of viral genome diversity and evolution. *Rev Med Virol*, **12**(4), 221–38.

McCraith, S., Holtzman, T., Moss, B., and Fields, S. (2000). Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc Natl Acad Sci U S A*, **97**(9), 4879–84.

McFadden, G. (2005). Poxvirus tropism. *Nat Rev Microbiol*, **3**(3), 201–13.

McGeoch, D. J. (1990). Protein sequence comparisons show that the 'pseudoproteases' encoded by poxviruses and certain retroviruses belong to the deoxyuridine triphosphatase family. *Nucleic Acids Res*, **18**(14), 4105–10.

McGeoch, D. J. and Davison, A. J. (1999). The descent of human herpesvirus 8. *Semin Cancer Biol*, **9**(3), 201–9.

McLysaght, A., Baldi, P. F., and Gaut, B. S. (2003). Extensive gene gain associated with adaptive evolution of poxviruses. *Proc Natl Acad Sci U S A*, **100**(26), 15655–60.

Mercer, A. A., Lyttle, D. J., Whelan, E. M., Fleming, S. B., and Sullivan, J. T. (1995). The establishment of a genetic map of orf virus reveals a pattern of genomic organization that is highly conserved among divergent poxviruses. *Virology*, **212**(2), 698–704.

Mercer, J. and Helenius, A. (2008). Vaccinia virus uses macropinocytosis and apoptotic mimicry to enter host cells. *Science*, **320**(5875), 531–5.

Merchlinsky, M. (1990). Mutational analysis of the resolution sequence of vaccinia virus DNA: essential sequence consists of two separate AT-rich regions highly conserved among poxviruses. *J Virol*, **64**(10), 5029–35.

Mockett, B., Binns, M. M., Boursnell, M. E., and Skinner, M. A. (1992). Comparison of the locations of homologous fowlpox and vaccinia virus genes reveals major genome reorganization. *J Gen Virol*, **73 (Pt 10)**(Pt 10), 2661–8.

Montague, M. G. and Hutchison, C. A. (2000). Gene content phylogeny of herpesviruses. *Proc Natl Acad Sci U S A*, **97**(10), 5334–9.

Morales, M., Ramirez, M. A., Cano, M. J., Parraga, M., Castilla, J., Perez-Ordoyo, L. I., Torres, J. M., and Barcena, J. (2009). Genome comparison of a nonpathogenic myxoma virus field strain with its ancestor, the virulent Lausanne strain. *J Virol*, **83**(5), 2397–403.

Moreira, D. and Brochier-Armanet, C. (2008). Giant viruses, giant chimeras: the multiple evolutionary histories of mimivirus genes. *BMC Evol Biol*, **8**(12), 12.

Moroziewicz, D. and Kaufman, H. L. (2005). Gene therapy with poxvirus vectors. *Curr Opin Mol Ther*, **7**(4), 317–25.

Moss, B. (1990). Regulation of orthopoxvirus gene expression. *Curr Top Microbiol Immunol*, **163**, 41–70.

Moss, B. (2001). Poxviridae: The viruses and their replication. In D. Knipe and P. Howley, editors, *Fields Virology*. Lippincott Williams and Wilkins, Philadelphia, 4 edition.

Moss, B., Shisler, J. L., Xiang, Y., and Senkevich, T. G. (2000). Immune-defense molecules of molluscum contagiosum virus, a human poxvirus. *Trends Microbiol*, **8**(10), 473–7.

Moyer, R. W. and Graves, R. L. (1981). The mechanism of cytoplasmic orthopoxvirus DNA replication. *Cell*, **27**(2 Pt 1), 391–401.

Murphy, P. M. (1993). Molecular mimicry and the generation of host defense protein diversity. *Cell*, **72**(6), 823–6.

Nerenberg, B. T., Taylor, J., Bartee, E., Gouveia, K., Barry, M., and Fruh, K. (2005). The poxviral RING protein p28 is a ubiquitin ligase that targets ubiquitin to viral replication factories. *J Virol*, **79**(1), 597–601.

Nichols, R. J., Stanitsa, E., Unger, B., and Traktman, P. (2008). The vaccinia virus gene I2L encodes a membrane protein with an essential role in virion entry. *J Virol*, **82**(20), 10247–61.

Notebaart, R. A., Huynen, M. A., Teusink, B., Siezen, R. J., and Snel, B. (2005). Correlation between sequence conservation and the genomic context after gene duplication. *Nucleic Acids Res*, **33**(19), 6164–71.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**(1), 205–17.

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**(6784), 299–304.

Odom, M. R., Hendrickson, R. C., and Lefkowitz, E. J. (2009). Poxvirus protein evolution: Family-wide assessment of possible horizontal gene transfer events. *Virus Res*, **20**, 20.

Pan, X., Zhang, B., San Francisco, M., and Cobb, G. P. (2007). Characterizing viral microRNAs and its application on identifying new microRNAs in viruses. *J Cell Physiol*, **211**(1), 10–8.

Parrish, S., Resch, W., and Moss, B. (2007). Vaccinia virus D10 protein has mRNA decapping activity, providing a mechanism for control of host and viral gene expression. *Proc Natl Acad Sci U S A*, **104**(7), 2139–44.

Pead, P. J. (2003). Benjamin Jesty: new light in the dawn of vaccination. *Lancet*, **362**(9401), 2104–9.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, **85**(8), 2444–8.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, **96**(8), 4285–8.

Perriere, G. and Gouy, M. (1996). WWW-query: an on-line retrieval system for biological sequence banks. *Biochimie*, **78**(5), 364–9.

Pfeffer, S., Zavolan, M., Grasser, F. A., Chien, M., Russo, J. J., Ju, J., John, B., Enright, A. J., Marks, D., Sander, C., *et al.* (2004). Identification of virus-encoded microRNAs. *Science*, **304**(5671), 734–6.

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grasser, F. A., van Dyk, L. F., Ho, C. K., Shuman, S., Chien, M., *et al.* (2005). Identification of microRNAs of the herpesvirus family. *Nat Methods*, **2**(4), 269–76.

Phillips, M. J., Delsuc, F., and Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol*, **21**(7), 1455–8.

Pickup, D. J., Ink, B. S., Parsons, B. L., Hu, W., and Joklik, W. K. (1984). Spontaneous deletions and duplications of sequences in the genome of cowpox virus. *Proc Natl Acad Sci U S A*, **81**(21), 6817–21.

Pickup, D. J., Ink, B. S., Hu, W., Ray, C. A., and Joklik, W. K. (1986). Hemorrhage in lesions caused by cowpox virus is induced by a viral protein that is related to plasma protein inhibitors of serine proteases. *Proc Natl Acad Sci U S A*, **83**(20), 7698–702.

Piskurek, O. and Okada, N. (2007). Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals. *Proc Natl Acad Sci U S A*, **104**(29), 12046–51.

Podell, S. and Gaasterland, T. (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol*, **8**(2), R16.

Pourbohloul, B., Meyers, L. A., Skowronski, D. M., Krajden, M., Patrick, D. M., and Brunham, R. C. (2005). Modeling control strategies of respiratory pathogens. *Emerg Infect Dis*, **11**(8), 1249–56.

Prangishvili, D., Stedman, K., and Zillig, W. (2001). Viruses of the extremely thermophilic archaeon sulfolobus. *Trends Microbiol*, **9**(1), 39–43.

Prichard, M. N., Keith, K. A., Johnson, M. P., Harden, E. A., McBrayer, A., Luo, M., Qiu, S., Chattopadhyay, D., Fan, X., Torrence, P. F., *et al.* (2007). Selective phosphorylation of antiviral drugs by vaccinia virus thymidine kinase. *Antimicrob Agents Chemother*, **51**(5), 1795–803.

Rajagopal, I., Ahn, B. Y., Moss, B., and Mathews, C. K. (1995). Roles of vaccinia virus ribonucleotide reductase and glutaredoxin in DNA precursor biosynthesis. *J Biol Chem*, **270**(46), 27415–8.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., and Claverie, J. M. (2004). The 1.2-megabase genome sequence of mimivirus. *Science*, **306**(5700), 1344–50.

Ray, C. A., Black, R. A., Kronheim, S. R., Greenstreet, T. A., Sleath, P. R., Salvesen, G. S., and Pickup, D. J. (1992). Viral inhibition of inflammation: cowpox virus encodes an inhibitor of the interleukin-1 beta converting enzyme. *Cell*, **69**(4), 597–604.

Reeves, P. M., Bommarius, B., Lebeis, S., McNulty, S., Christensen, J., Swimm, A., Chahroudi, A., Chavan, R., Feinberg, M. B., Veach, D., *et al.* (2005). Disabling poxvirus pathogenesis by inhibition of Abl-family tyrosine kinases. *Nat Med*, **11**(7), 731–9.

Rice, P., Longden, I., and Bleasby, A. (2000). Emboss: the european molecular biology open software suite. *Trends Genet*, **16**(6), 276–7.

Roberts, K. L. and Smith, G. L. (2008). Vaccinia virus morphogenesis and dissemination. *Trends Microbiol*, **11**, 11.

Rocha, E. P., Cornet, E., and Michel, B. (2005). Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genet*, **1**(2), e15.

Rothenburg, S., Seo, E. J., Gibbs, J. S., Dever, T. E., and Dittmar, K. (2008). Rapid evolution of protein kinase PKR alters sensitivity to viral inhibitors. *Nat Struct Mol Biol*, **30**, 30.

Rubins, K. H., Hensley, L. E., Bell, G. W., Wang, C., Lefkowitz, E. J., Brown, P. O., and Relman, D. A. (2008). Comparative analysis of viral gene expression programs during poxvirus infection: a transcriptional map of the vaccinia and monkeypox genomes. *PLoS ONE*, **3**(7), e2628.

Ruby, J. G., Jan, C. H., and Bartel, D. P. (2007). Intronic microRNA precursors that bypass Drosha processing. *Nature*, **448**(7149), 83–6.

Sabeti, P. C., Walsh, E., Schaffner, S. F., Varilly, P., Fry, B., Hutcheson, H. B., Cullen, M., Mikkelsen, T. S., Roy, J., Patterson, N., *et al.* (2005). The case for selection at CCR5-Delta32. *PLoS Biol*, **3**(11), e378.

Sarnow, P., Jopling, C. L., Norman, K. L., Schutz, S., and Wehner, K. A. (2006). MicroRNAs: expression, avoidance and subversion by vertebrate viruses. *Nat Rev Microbiol*, **4**(9), 651–9.

Seet, B. T., McCaughan, C. A., Handel, T. M., Mercer, A., Brunetti, C., McFadden, G., and Fleming, S. B. (2003). Analysis of an orf virus chemokine-binding protein: Shifting ligand specificities among a family of poxvirus viroceptors. *Proc Natl Acad Sci U S A*, **100**(25), 15137–42.

Sekiguchi, J. and Shuman, S. (1997). Ligation of RNA-containing duplexes by vaccinia DNA ligase. *Biochemistry*, **36**(29), 9073–9.

Senkevich, T. G., Bugert, J. J., Sisler, J. R., Koonin, E. V., Darai, G., and Moss, B. (1996). Genome sequence of a human tumorigenic poxvirus: prediction of specific host response-evasion genes. *Science*, **273**(5276), 813–6.

Senkevich, T. G., Ojeda, S., Townsley, A., Nelson, G. E., and Moss, B. (2005). Poxvirus multiprotein entry-fusion complex. *Proc Natl Acad Sci U S A*, **102**(51), 18572–7.

Shackelton, L. A. and Holmes, E. C. (2004). The evolution of large DNA viruses: combining genomic information of viruses and their hosts. *Trends Microbiol*, **12**(10), 458–65.

Shchelkunov, S. N., Blinov, V. M., and Sandakhchiev, L. S. (1993). Ankyrin-like proteins of variola and vaccinia viruses. *FEBS Lett*, **319**(1-2), 163–5.

Shisler, J. L., Senkevich, T. G., Berry, M. J., and Moss, B. (1998). Ultraviolet-induced cell death blocked by a selenoprotein from a human dermatotropic poxvirus. *Science*, **279**(5347), 102–5.

Shuman, S. and Moss, B. (1988). Factor-dependent transcription termination by vaccinia virus RNA polymerase. evidence that the cis-acting termination signal is in nascent RNA. *J Biol Chem*, **263**(13), 6220–5.

Slabaugh, M., Roseman, N., Davis, R., and Mathews, C. (1988). Vaccinia virus-encoded ribonucleotide reductase: sequence conservation of the gene for the small subunit and its amplification in hydroxyurea-resistant mutants. *J Virol*, **62**(2), 519–27.

Smith, G. L., Howard, S. T., and Chan, Y. S. (1989). Vaccinia virus encodes a family of genes with homology to serine proteinase inhibitors. *J Gen Virol*, **70**(Pt 9), 2333–43.

Smith, S. K., Olson, V. A., Karem, K. L., Jordan, R., Hruby, D. E., and Damon, I. K. (2009). In vitro efficacy of ST246 against smallpox and monkeypox. *Antimicrob Agents Chemother*, **53**(3), 1007–12.

Snel, B., Bork, P., and Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nat Genet*, **21**(1), 108–10.

Sonnberg, S., Seet, B. T., Pawson, T., Fleming, S. B., and Mercer, A. A. (2008). Poxvirus ankyrin repeat proteins are a unique class of F-box proteins that associate with cellular SCF1 ubiquitin ligase complexes. *Proc Natl Acad Sci U S A*, **30**, 30.

Spyropoulos, D. D., Roberts, B. E., Panicali, D. L., and Cohen, L. K. (1988). Delineation of the viral products of recombination in vaccinia virus-infected cells. *J Virol*, **62**(3), 1046–54.

Srinivasan, V., Schnitzlein, W. M., and Tripathy, D. N. (2001). Fowlpox virus encodes a novel DNA repair enzyme, CPD-photolyase, that restores infectivity of UV light-damaged virus. *J Virol*, **75**(4), 1681–8.

Stack, J., Haga, I. R., Schroder, M., Bartlett, N. W., Maloney, G., Reading, P. C., Fitzgerald, K. A., Smith, G. L., and Bowie, A. G. (2005). Vaccinia virus protein a46r targets multiple toll-like-interleukin-1 receptor adaptors and contributes to virulence. *J Exp Med*, **201**(6), 1007–18.

Stern-Ginossar, N., Saleh, N., Goldberg, M. D., Prichard, M., Wolf, D. G., and Mandelboim, O. (2009). Analysis of human cytomegalovirus encoded-microRNAs activity during infection. *J Virol*, **5**, 5.

Su, J. R. (2004). Emerging viral infections. *Clin Lab Med*, **24**(3), 773–95, vii.

Sullivan, C. S., Grundhoff, A. T., Tevethia, S., Pipas, J. M., and Ganem, D. (2005). SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. *Nature*, **435**(7042), 682–6.

Susko, E., Leigh, J., Doolittle, W. F., and Bapteste, E. (2006). Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the gamma-proteobacteria. *Mol Biol Evol*, **23**(5), 1019–30.

Takemura, M. (2001). Poxviruses and the origin of the eukaryotic nucleus. *J Mol Evol*, **52**(5), 419–25.

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*, **10**(3), 512–26.

Thiel, T., Whiteman, N. K., Tirape, A., Baquero, M. I., Cedeno, V., Walsh, T., Uzcategui, G. J., and Parker, P. G. (2005). Characterization of canarypox-like viruses infecting endemic birds in the galapagos islands. *J Wildl Dis*, **41**(2), 342–53.

Thomassen, G. O., Rosok, O., and Rognes, T. (2006). Computational prediction of microRNAs encoded in viral and other genomes. *J Biomed Biotechnol*, **2006**(4), 95270.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**(22), 4673–80.

Tian, X., Pascal, G., Fouchecourt, S., Pontarotti, P., and Monget, P. (2009). Gene birth, death, and divergence: The different scenarios of reproduction-related gene evolution. *Biol Reprod*, **7**, 7.

Tsirigos, A. and Rigoutsos, I. (2005). A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. *Nucleic Acids Res*, **33**(12), 3699–707.

Tulman, E. R., Delhon, G., Afonso, C. L., Lu, Z., Zsak, L., Sandybaev, N. T., Kerembekova, U. Z., Zaitsev, V. L., Kutish, G. F., and Rock, D. L. (2006). Genome of horsepox virus. *J Virol*, **80**(18), 9244–58.

Turner, P. C. and Moyer, R. W. (1990). The molecular pathogenesis of poxviruses. *Curr Top Microbiol Immunol*, **163**, 125–51.

Turner, P. C. and Moyer, R. W. (2008). The vaccinia virus fusion inhibitor proteins SPI-3 (K2) and HA (A56) expressed by infected cells reduce the entry of superinfecting virus. *Virology*, **380**(2), 226–33.

Uetz, P., Dong, Y. A., Zeretzke, C., Atzler, C., Baiker, A., Berger, B., Rajagopala, S. V., Roupelieva, M., Rose, D., Fossum, E., *et al.* (2006). Herpesviral protein networks and their interaction with the human proteome. *Science*, **311**(5758), 239–42.

Upton, C. and McFadden, G. (1986). Identification and nucleotide sequence of the thymidine kinase gene of shope fibroma virus. *J Virol*, **60**(3), 920–7.

Upton, C., Macen, J. L., Maranchuk, R. A., DeLange, A. M., and McFadden, G. (1988). Tumorigenic poxviruses: fine analysis of the recombination junctions in malignant rabbit fibroma virus, a recombinant between shope fibroma virus and myxoma virus. *Virology*, **166**(1), 229–39.

Upton, C., Slack, S., Hunter, A. L., Ehlers, A., and Roper, R. L. (2003). Poxvirus orthologous clusters: toward defining the minimum essential poxvirus genome. *J Virol*, **77**(13), 7590–600.

van Buuren, N., Couturier, B., Xiong, Y., and Barry, M. (2008). Ectromelia virus encodes a novel family of F-box proteins that interact with the Scf complex. *J Virol*, **6**, 6.

van Dongen, S. (2000). *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht.

VanSlyke, J. K. and Hruby, D. E. (1990). Posttranslational modification of vaccinia virus proteins. *Curr Top Microbiol Immunol*, **163**, 185–206.

Vieira, P., de Waal-Malefyt, R., Dang, M. N., Johnson, K. E., Kastelein, R., Fiorentino, D. F., deVries, J. E., Roncarolo, M. G., Mosmann, T. R., and Moore, K. W. (1991). Isolation and expression of human cytokine synthesis inhibitory factor cDNA clones: homology to Epstein-Barr virus open reading frame BCRFI. *Proc Natl Acad Sci U S A*, **88**(4), 1172–6.

Walzer, T., Galibert, L., and De Smedt, T. (2005). Poxvirus semaphorin A39R inhibits phagocytosis by dendritic cells and neutrophils. *Eur J Immunol*, **35**(2), 391–8.

Wang, N., Baldi, P. F., and Gaut, B. S. (2007). Phylogenetic analysis, genome evolution and the rate of gene gain in the herpesviridae. *Mol Phylogenet Evol*, **43**(3), 1066–75.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**(6684), 440–2.

Westphal, D., Ledgerwood, E. C., Hibma, M. H., Fleming, S. B., Whelan, E. M., and Mercer, A. A. (2007). A novel Bcl-2-like inhibitor of apoptosis is encoded by the parapoxvirus orf virus. *J Virol*, **81**(13), 7178–88.

WHO (2008a). Smallpox factsheet.

WHO (2008b). WHO recommendations concerning the distribution, handling and synthesis of variola virus DNA, may 2008. *Wkly Epidemiol Rec*, **83**(44), 393–5.

Xing, K., Deng, R., Wang, J., Feng, J., Huang, M., and Wang, X. (2006). Genome-based phylogeny of poxvirus. *Intervirology*, **49**(4), 207–14.

Xu, R. H., Cohen, M., Tang, Y., Lazear, E., Whitbeck, J. C., Eisenberg, R. J., Cohen, G. H., and Sigal, L. J. (2008). The orthopoxvirus type I IFN binding protein is essential for virulence and an effective target for vaccination. *J Exp Med*, **205**(4), 981–92.

Yanai, I., Wolf, Y. I., and Koonin, E. V. (2002). Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biol*, **3**(5), research0024.

Yao, X. D. and Evans, D. H. (2001). Effects of DNA structure and homology length on vaccinia virus recombination. *J Virol*, **75**(15), 6923–32.

Zhang, L., Villa, N. Y., Rahman, M. M., Smallwood, S., Shattuck, D., Neff, C., Dufford, M., Lanchbury, J. S., Labaer, J., and McFadden, G. (2009a). Analysis of vaccinia virus-host protein-protein interactions: validations of yeast two hybrid screenings. *J Proteome Res*, **28**, 28.

Zhang, L., Villa, N. Y., and McFadden, G. (2009b). Interplay between poxviruses and the cellular ubiquitin/ubiquitin-like pathways. *FEBS Lett*, **583**(4), 607–14.

Zhou, Q., Snipas, S., Orth, K., Muzio, M., Dixit, V. M., and Salvesen, G. S. (1997). Target protease specificity of the viral serpin CrmA. Analysis of five caspases. *J Biol Chem*, **272**(12), 7797–800.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S., and Wang, W. (2008). On the origin of new genes in drosophila. *Genome Res*, **18**(9), 1446–55.