Understanding the Evolutionary History of the Papillomaviruses

by

Seena D Shah

A thesis submitted for the degree of

Doctor of Philosophy

UCL

October 2012

I, Seena D Shah, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed _____ Date _____

Acknowledgments

The completion of this thesis has been the ultimate exercise in perserverance for me and I would not have reached this stage without the support of many people. I cannot express enough thanks to my supervisor, Richard Goldstein, for his patience, wisdom, advice throughout this whole process. I consider myself very lucky to have had such a mentor. I am also very fortunate to have been able to pick the brains of Mario dos Reis and Asif Tamuri and thank them for their valuable advice. I would also like to thank Willie Taylor, Jens Kleinjung and the rest of Mathematical Biology division at the NIMR for their support. The continual support and encouragement from my friends and family has helped me through several difficult periods. I thank my family for their patience and understanding. Bansi, Donna, Zoe and Laura have been my crutches and my cheerleaders; I could not have got through this without them.

Abstract

This thesis focuses on the evolutionary history of the papillomaviruses (PVs) using phylogenetic approaches. Two aspects have been examined: the first is the level of phylogenetic compatibility among PV genes and the second is determining the ancestral diversification mechanisms of the PVs in order to explain the origin of the observed associations with host species.

Bayesian phylogenetic analysis has been used to make evolutionary inferences. The existence of phylogenetic compatibility among genes was examined by estimating constrained and unconstrained phylogenies for pairs of PV genes. The Bayes' factor statistic derived from comparison of the constrained and unconstrained models indicated significant evidence against identical phylogenies between any of the 6 PV genes investigated and may indicate the existence of ancestral recombination events.

The formation of new host-virus associations can occur via a process of 'codivergence', where, following host speciation, the ancestral virus association is effectively inherited by the descendant host species; 'prior divergence' of the virus, which results in multiple virus associations with the host; and 'host transfer', in which the virus lineage is transferred between contemporaneous host species. To distinguish between these mechanisms of virus diversification, an approach based on temporal comparisons of host and virus divergence times was devised. Difficulties associated with the direct estimation of PV divergence times led to the incorporation of a biased sampling approach into Bayesian phylogenetic estimation. This allowed for viral divergence events to be biased in favour of codivergence but allowed sampling of times that violate this assumption and therefore indicate either prior divergence or host transfer. Statistical evaluation of the proportion of violations at each viral divergence identified significant evidence of prior divergence events behind many of the observed PV-host associations and one ancestral host transfer event.

Contents

Figures

Tables

Glossary

Capsid: a viral protein coat that forms around the viral genome to protect it from the environment outside of the host cell.

Codivergence: the process in which virus divergence occurs due to speciation of the host species it is associated with. Separation on different host species causes the two virus populations to evolve independently of each other.

Convergent evolution: the process by which the same trait or character is evolved in unrelated lineages. Lineages displaying traits/characters that have been acquired via convergent evolution may mistakenly be inferred to be closely related, and the similarity inferred to be inherited from a common ancestor.

Cospeciation: the process by which a vicariance event causing speciation of a host organism also causes speciation of the parasite species associated with it.

Cutaneous tissue: epithelial tissue made up of layers of keratinised stratified squamous cells.

Early genes: genes encoded in the 'early' region of the PV genome. These include the genes E1, E2, E4, E5, E6, E7, E8 and E9.

Epithelial hyperplasia: a proliferation of the epithelial cells; results in a wart or tumour

Fahrenholz' rule: strict cospeciation (codivergence) will result in parasite (virus) and host phylogenies with identical topologies.

Farris optimisation: an algorithm for calculating the minimum number of character changes along a proposed tree relating a set of data. Farris optimisation is employed when the characters are ordered and therefore changes are additive.

Fitch optimisation: similar to Farris optimisation but applied to data possessing unordered characters, e.g., the nucleotide states A, C, G, and T. For unordered characters all possible changes of character state have a cost of 1.

Homology: characters (or traits) that shared by two lineages that were inherited from their common ancestor

Homoplasy: characters (or traits) that are shared by two lineages but which were not inherited from their common ancestor.

Host transfer: the process in which a virus associated with one host species is able to establish infection on another host species. May also be referred to as host switch, or lateral/horizontal transmission.

Incomplete lineage sorting: the process by which following speciation of the host, the parasite or virus associates with only one of the descendant species.

Keratinocytes: stratified squamous cells that undergo a process of terminal differentiation in the epithelium that results in the loss of nuclei and filling of intracellular space by filaments of keratin. The keratinisation results in a tougher, waterproof tissue, which forms the skin and hair.

Late genes: genes encoded in the 'late' region of the PV genome. These include the genes L1 and L2, which express the capsid proteins.

Lesion: an abnormal growth of body tissue.

Monophyletic clade: a phylogenetic of taxa that unites all the descendents of a common ancestor.

Mucosal tissue: epithelial tissue that is moist and is made up of layers of non-keratinised stratified squamous cells. Mucosal tissue forms the lining of the mouth, the inner eye, and the ano-genital region.

Paraphyletic clade: a phylogenetic grouping of taxa that unites only some of the descendants from a common ancestor.

Phylogeny: a tree depicting the pattern and relative timing of the lineage splitting events that occurred among a group of species.

Polyphyletic clade: a phylogenetic grouping of taxa uniting species that do not share a recent common ancestor.

Polyploidy cell: a cell possessing more than two complete sets of chromosomes.

Post-speciation dispersal: the colonisation of different host species by the same parasite species.

Prior divergence: divergence of a viral lineage or parasite species in the absence of a host speciation event. Prior divergence allows the new lineage to exploit a different environment or resourse on the same host species.

Reassortment: the process in which two segmented virus genomes infecting the same cell exchange genome segments to create new reassortant strain.

Recombination: – the process in which a section of genetic material is exchanged between genomes. Recombination involves the breaking of DNA from one genome and insertion into another genome.

Stratified squamous cells: layers of flattened epithelial cells that form the tissue at anatomical locations subject to regular abrasion, e.g. the skin.

Tree topology: the lineage-splitting pattern, or branching pattern, that is observed in the phylogeny relating a group of species.

Abbreviations

aa amino acid

bp base pair

BF Bayes' factor

CI confidence interval

DBD DNA-binding domain

HME harmonic mean estimator

ILD incongruence length difference

LHT likelihood heterogeneity test

LRT likelihood ratio test

ML maximum likelihood

MP maximum parsimony

MRCA most recent common ancestor

mya millions of years ago

ORF open reading frame

ori origin of replication

PV papillomavirus

s.d. standard deviation

Chapter 1

Introduction

1.1 The Papillomaviruses (PVs)

The papillomaviruses (PVs) are small (approximately 55-60 nm in diameter), non-enveloped, double stranded DNA viruses that comprise the *Papillomaviridae* family. PV infection may cause lesions (epithelial hyperplasia) on mucosal and cutaneous tissue, referred to as warts, papillomas, and condylomas depending on the anatomical site. The lesions are generally benign and regress spontaneously; however, PVs may persist in the epithelial cells of their hosts for many years. Persistent infection has been identified as a key factor in the ability of a subset of PV types to cause infections that progress from benign, low-grade lesions to malignant tumors (Durst et al. 1983; Boshart et al. 1984; zur Hausen 1989; Ho et al. 1995; zur Hausen 2000; Campo 2002; Ferenczy and Franco 2002; Schiffman et al. 2005; Doorbar 2006). The oncogenic potential displayed by certain PV types has made the PVs medically important viruses and has resulted in increased research interest to understand the biology and pathology of these viruses.

The diversity demonstrated by the PVs also makes them interesting subjects for evolutionary study. In addition to differing histological preferences, site preferences and pathological severities among the PVs, different host species preferences are also observed. There is therefore a rich evolutionary history that is yet to be investigated among this family of viruses (Garcia-Vallve, Alonso and Bravo 2005; Bravo, de Sanjose and Gottschling 2011). For instance, substantial

PV diversity has been uncovered in humans and together these human-infecting lineages display the full complement of phenotypic variation that is observed among PVs (Ekstrom, Forslund and Dillner), e.g. some infect only cutaneous tissue at genital site, some infect only cutaneous tissue at non-genital sites, some infect only mucosal tissue at genital sites, other genotypes display dual tropism, etc. However, the molecular 'signatures' that correspond with these phenotypes (e.g., what defines a cutaneotrophic PV at the genotype level), are yet to be determined. Such a study will be of particular clinical benefit when applied to determine the molecular signatures for oncogenecity.

In this particular thesis, I attempt to investigate the mechanisms of PV diversification to different host species. Some day, a characterisation of the adaptive changes, occurring at the genotype level, that enable infection of a particular host species may be achieved but at present the focus is on determining the nature of the macroevolutionary processes (i.e. those occurring above the molecular level) by which PVs have diverged to new hosts (Gottschling et al. 2007b; Gottschling et al. 2011b).

1.2 Taxonomic Classification of the PVs

The PVs were initially assigned to the Papovaviridae family along with another group of tumour viruses, the polyomaviruses, based on morphological similarities, such as in capsid structure, between the two groups of viruses (Wildy 1971). However, the sequencing of PV and polyomavirus genomes revealed a lack of evidence for a homologous relationship, based on different genome organisations and protein sequence comparisons (Danos, Katinka and Yaniv 1982). Statistically significant sequence similarity was later identified between the large tumour-antigen of the simian virus 40 (a polyomavirus infecting monkeys) and the E1 protein of the PVs (Clertant and Seif 1984; Mansky, Batiza and Lambert 1997), both of which share helicase functionality. However, the lack of evidence to suggest homology between the genomes of the two groups of viruses effected their reclassification of to individual families by the International Committee on Taxonomy of Viruses (ICTV) (van Regenmortel et al. 2002).

Within the *Papillomaviridae* family, PV classification comprises of five taxonomic categories: “genus”, “species”, “type”, “subtype”, and “variant” (de Villiers et al. 2004).

Taxonomic classifications of newly isolated PVs are performed via sequence comparisons of the L1 gene (de Villiers et al. 2004). This is primarily due to the ease of amplifying this region of the genome as well as the high degree of sequence conservation observed in the L1 gene, but was also initially supported by phylogenetic evidence demonstrating that PV relationships deduced from the L1 gene region were congruent with the relationships determined from other genomic regions (Bernard et al. 1994; Myers et al. 1994; Chan et al. 1995). Subsequent phylogenetic analyses have, however, revealed differences in the evolutionary histories of PV genes from different genome regions (Garcia-Vallve, Alonso and Bravo 2005; Narechania et al. 2005; Bravo and Alonso 2007). A key observation among these differences is that early gene phylogenies show groupings consistent with the biological properties of the PV types (e.g. distinct clades of high-risk and low-risk mucosal PV types, grouping of genital cetacean PV types with genital primate PV types) whilst the late gene phylogenies do not. Consequently, there has been some suggestion (Bravo and Alonso 2007) that PV taxanomic classification based on conserved sequences from the E1 and E2 protein sequences would produce a taxonomic structure that grouped PV types by functional properties and hence, would be more appropriate than the L1 gene sequences. However, despite the phylogenetic inconsistencies of the L1 gene, the classification protocol has remained the same (Bernard et al. 2010); Table 1.1 outlines the criteria used to classify PV isolates based on L1 gene sequence similarities.

A PV isolate is declared a new type if its L1 gene shows more than 10% sequence divergence from its closest known PV type (de Villiers et al. 2004). For new PV types, the naming convention decided upon by PV researchers (Fauquet et al. 2005) (check Bernard et al. 2005) was to reference the scientific name of the infected host species, for example, PVs isolated from the common bottlenose dolphin (sp. *Tursiops truncatus*) are named Tursiops truncatus PVs (initialised as

TtPV). However, deviations from this naming system have resulted in some PV types being assigned the host species' common name. For example, the PV type isolated from the European elk (sp. *Alces alces*) was named 'EEPV' and 'CRPV' refers to the cottontail rabbit (sp. *Sylvilagus floridanus*) PV. Some PV names also incorporated the site of infection, for example, the oral PV types from dog (COPV), rabbit (ROPV) and hamster (HaOPV) species.

The different naming schemes that have been employed may produce replicate abbreviated forms, for instance, the PV type isolated from the European hedgehog (sp. *Erinaceus europaeus*) has been labelled EEPV1, which may be readily confused with EEPV from the European elk. Bernard et al. (2010) have corrected for these inconsistencies by renaming all PV types using only the scientific name of the host species, e.g. the European elk PV type (EEPV) is now known as the *Alces alces* PV type 1 (AaPV1) and the oral rabbit PV type (ROPV) is now known as the *Oryctolagus cuniculus* PV type 1 (OcPV1). Additional PV types isolated from the same host species are then numbered as type 2, 3, etc. Where replicate names are still possible under this scheme additional letters are used, e.g. PV types from the Western roe deer (sp. *Capreolus capreolus*) are abbreviated as CcaPV to avoid confusion with PV types isolated from the Loggerhead turtle (sp. *Caretta caretta*), which are abbreviated as CcPV. The only exceptions to this unified naming system occur with the PV types isolated from humans (sp. *Homo sapiens*), domestic cows (sp. *Bos taurus*) and domestic dogs (sp. *Canis familiaris*), each of which have retained their original abbreviated forms that reference only the genus of the host species, i.e., HPV, BPV, and CPV, respectively (Bernard et al. 2010).

The renaming of PV types by Bernard et al. (2010) coincided with the publication of the research carried out in this thesis (Shah, Doorbar and Goldstein 2010). To maintain consistency with the PV names used in the published paper, in the following text I refer to the animal PVs studied in this thesis by their originally assigned names. However, for the reader's reference, the revised names of those PV types will be given after, in parentheses, and are listed in Appendix

A1. PV types that have not been studied in this thesis will be referred to using the new names.

Taxonomic level	% Nucleotide identity (L1 gene)	E.g.
Genus	60-70	α
Species	71-89	α-9
Type	90-100	HPV16
Subtype	90-97	-
Variant	98-99	Tb-7

Table 1.1: Taxonomic levels within the PV family and the corresponding classification criteria as outlined by deVilliers (2004).

Intra-type diversity may be further classified into distinct subtypes when there is 2-10 % nucleotide difference with the L1 gene of the reference genome (i.e. first identified genome of that type), and distinct variants (Ong et al. 1993) when there is less than 2% sequence divergence from the reference genome. PVs isolated from bonobo (*Pan paniscus*, PcPV - now PpPV1) and the common chimpanzee (*Pan troglodytes*, CCPV1 - now PtPV1), which were classified as distinct types, possess sequence similarities indicative of a subtype relationship (de Villiers et al. 2004). Intensive sampling efforts of HPVs have so far identified little intra-type diversity at the subtype level, however, there has been substantial diversification at the variant level (Ho et al. 1993; Ong et al. 1993; Calleja-Macias et al. 2005; Chen et al. 2009). These diversifications may be associated with differences in biological behaviour, for instance, variants of the cervical cancer-associated types HPV16 and HPV18 differ in their ability to persist in epithelial cells and hence, their oncogenic potential (Villa et al. 2000; Burk et al. 2003; Sichero et al. 2007; Sichero, Simao Sobrinho and Villa 2012).

At higher taxonomic levels, PV types are grouped into species and PV species into genera. 18 PV genera were initially established (de Villiers et al. 2004) each

of which were designated a letter of the Greek alphabet (α-π). The continual discovery of new PV types has now extended the PV family to 32 distinct genera. To accommodate the additional genera into the existing nomenclature system, the Greek alphabet is re-used with the prefix 'dyo' (Bernard et al. 2010). However, some individual PV types still remain unclassified at the genus level, e.g. BPV7.

PV types are assigned to the same genus if they share 60-70% nucleotide identity. Nucleotide sequence identities between genomes of PV types from different genera are found to vary between 23-43 % (de Villiers et al. 2004). The genus classifications appear to unite PV types infecting closely related, if not the same, host species. For example, the α, β, γ, μ, and ν genera are all populated by PVs infecting species from the mammalian order Primates; the κ PVs have been isolated from different rabbit species from the order Lagomorpha; the δ, ε, and ξ PVs infect various ungulate hosts from the order Artiodactyla; the λ PVs infect species from the mammalian order Carnivora; the π genus currently consists of PV types infecting different species rodent species; and, the ο and υ PVs infect species from the mammalian order Cetacea.

Some genera may be further defined by biological and pathological properties beyond the observed host range. For instance, the δ and ε genera both comprise of artiodactyl PV types (including some isolated from bovine hosts) that cause fibropapillomas, which extend below the epithelial tissue of normal PV infection into the dermis (Nasir and Campo 2008). Bovine PV types from the ξ genus, however, cause only epithelial infection of cutaneous and mucosal tissue. Similarly, PV types from the β, γ, μ, and ν genera, which infect primates, all cause lesions in cutaneous tissue but, notably, not at genital sites. In contrast, the genus of α PVs, which also infect primates, is a mix of PV types specifically targeting mucosal and/or cutaneous epithelial cells at genital and non-genital sites.

Within each genus PV types are grouped into PV species (not to be confused with the host species), which are denoted by the genus name and a number, e.g., PV species from the α genus are named α-1, α-2, etc. The members of each PV

species are defined by 71-80% nucleotide sequence similarity in the L1 gene (de Villiers et al. 2004).

The viral species groupings are generally found to unite PV types with similar biological and pathological characteristics. The best example of this is observed in the α genus, which currently comprises of 15 different species (Figure 1.1, adapted from Narechania et al. 2005: Fig. 1). In the α genus, the PV species 1, 3, 10, 13, 14, and 15 all comprise of PV types that have been isolated from benign lesions of mucosal tissue at genital and/or oral sites. The α–10 species includes the PV types CCPV1 (now PtPV1) and PCPV1 (now PpPV1) from non-human primates (not included in Figure 1.1) - chimpanzee (Scinicariello et al. 1997, unpublished), and bonobo (Van Ranst et al. 1991), respectively, both of which were extracted from oral focal epithelial hyperplasias like their closest known relative – the human PV type HPV13.

The α species 5, 6, 7, 9 and 11 also contain PV types that specifically infect mucosal tissue; however, these types have the potential to cause malignant tumours and are therefore labelled as 'high-risk' (Munoz et al. 2003). The α-12 species (not included in Figure 1.1) currently comprises of only monkey PVs isolated from mucosal genital sites. The first α-12 PV type, RhPV1 (now MmPV1), was isolated from a metastatic penile squamous cell carcinoma (Kloster et al. 1988) and is observed to cluster with the high-risk HPV species in PV phylogenies. PV types from the α–2 species have been detected in skin warts at various anatomical sites whilst types from the α–4 and α-8 species have been isolated from benign lesions of both mucosal and cutaneous tissue and therefore display properties of dual tissue tropism.

As is demonstrated in the phylogeny of HPVs from the α genus (Figure 1.1), phylogenetic analysis of PV types reveals high support for monophyletic clustering at the PV species level. A similar observation is made for PV species of other genera (Bernard et al. 2010). Thus, phylogenetic groupings of the α-HPVs tend to correspond to similar biological and pathological properties. In particular, Figure 1.1 demonstrates high statistical support for the grouping of the high-risk HPV species (α species 5, 6, 7, 9 and 11) into a single clade, suggesting

a single lineage (internal branch 4 in Figure 1.1) for the origin of oncogenic potential among α HPV types.

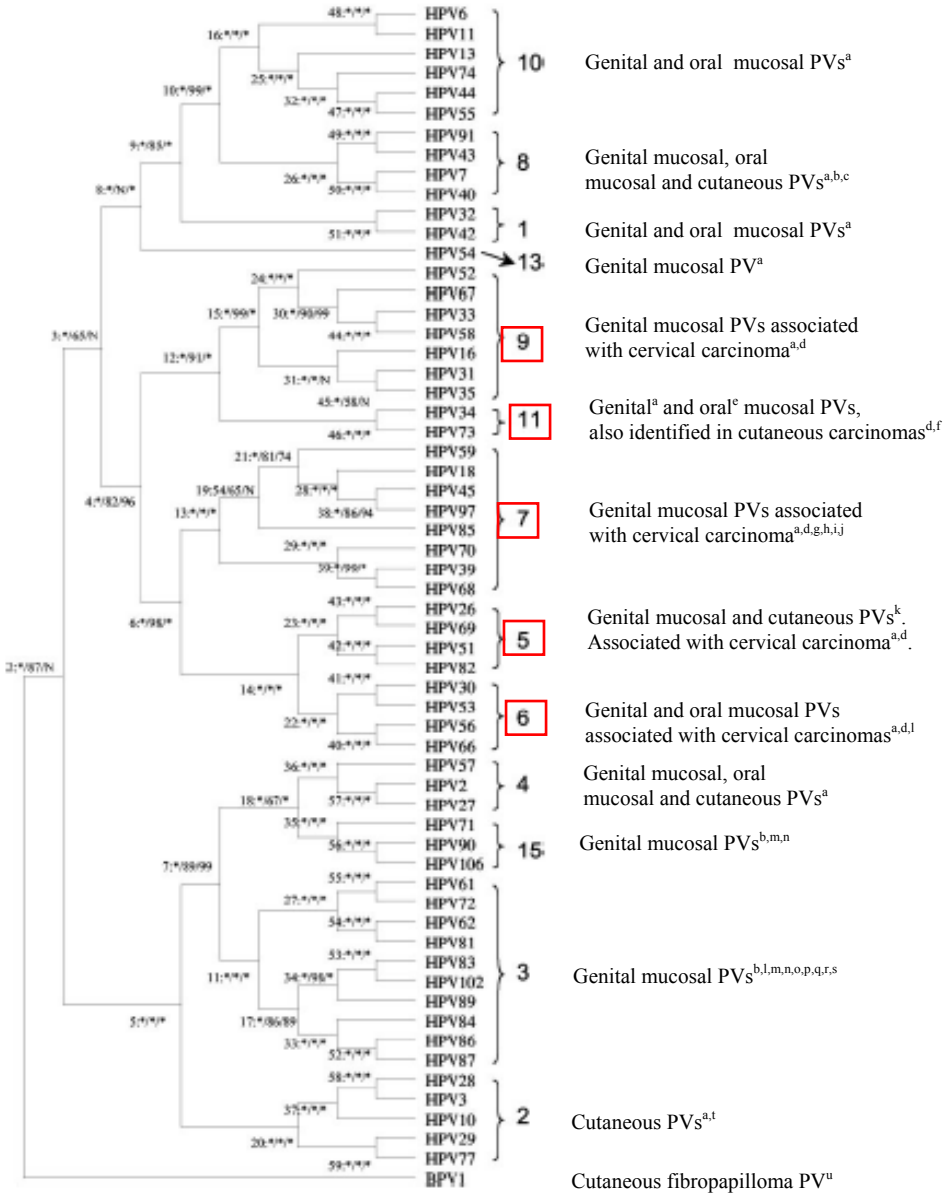


Figure 1.1: Phylogenetic tree depicting relationships among HPV types of the α genus. The tree was estimated from the concatenated protein and nucleotide sequences of the E6, E7, E1, E2, L2, and L1 ORFs using Bayesian phylogenetic methods, with the bovine PV type 1 (BPV1) functioning as an outgroup taxon. Branch labels indicate the node number followed by the support values from phylogenetic estimations using various methods in the following order: Bayesian clade credibility, maximum parsimony bootstrap percentage, and neighbour joining bootstrap percentage. Methods that show 100%

support for a branch are inidicated using an asterisk; disagreements between the Bayesian phylogenetic grouping and phylogenetic groupings from either of the other two methods are indicated by an 'N'. The HPV taxa, grouped by viral species classifications, constitute 14 of the 15 species identified within the α genus. 13 of these HPV-containing α species are represented in the tree. No HPV type is known for α PV species 12, which is currently populated by monkey PV types (RhPVs) only. Each α PV species consists of PV types with similar histological preferences and pathological outcomes. The PV species of high-risk PV types, which are associated with carcinomas, are highlighted in red. PV species descriptions were obtained from the following:

[a] (de Villiers 1989), [b] (Terai and Burk 2002), [c] (Greenspan et al. 1988), [d] (Munoz et al. 2003), [e](Volter et al. 1996), [f] (Kawashima et al. 1986), [g] (Chen et al. 2007a), [h] (Chow and Leong 1999), [i] (Forslund and Hansson 1996), [j] (Wu et al. 2009), [k] (Kino et al. 2000), [l] (Tachezy et al. 1994), [m] (Matsukura and Sugase 2001), [n] (Chen et al. 2007b), [o] (Fu et al. 2004), [p] (Brown et al. 1999), [q] (Terai and Burk 2001b), [r] (Terai and Burk 2001a), [s] (Menzo et al. 2001), [t] (Delius et al. 1998), and [u] (Chen et al. 1982). Adapted from Narechania et al. (2005: Fig. 1, scale not provided)

1.3 PV Biology

1.3.1 Genome Structure

PV genomes are unsegmented, circular structures, varying between 7000-9000 base pairs (bp) in length. They encode up to 9 genes, 5 of which are present in all PV types. A well-established feature of the PVs is that they display a relatively stable genome structure with a highly conserved genome organisation. The first PV genomes to be characterised were HPV1a (Danos, Katinka and Yaniv 1982) and BPV1 (Chen et al. 1982). These two PV types infect different host species (humans and cows, respectively) and are distantly related to each other but presented similar genome structures in which the relative positions of the 4 major ORFs – E1, E2, L1 and L2 – were highly conserved across the two genomes (Figure 1.2, reprinted from Chen et al. 1982: Fig. 5). The genomes of subsequent PV types have revealed similar organisations. In all PV genomes, the ORFs are transcribed from the same strand of DNA, and a non-coding upstream regulatory region (URR, aka the long control region (LCR)) is found at the 3' end of the genome (Figure 1.3, adapted from Doorbar 2006).

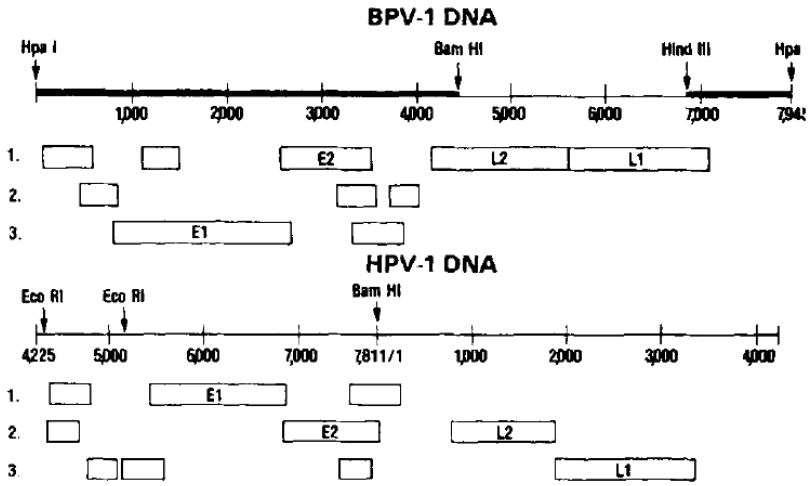


Figure 1.2: An alignment of the BPV1 and HPV1a genomes. The two genomes demonstrate similar size and position of the major ORFs E1, E2, L1 and L2. Differences are observed in the reading frames from which these ORFs are translated, however. The dark bar along the BPV1 genome indicates the region of the genome expressed in BPV1 transformed cells, thus the L1 and L2 gene products are not involved in cellular transformation. The identities of the smaller ORFs (< 500 bases) were not known in Chen et al. (1982) and therefore were not labelled. These ORFs correspond to the E6, E7, E4, E5, and E8 ORFs. Reprinted from Chen et al. (1982: Fig. 5).

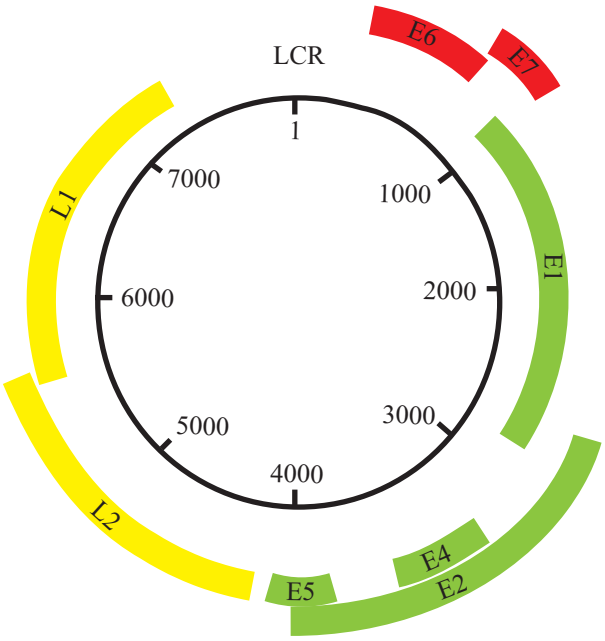


Figure 1.3: Schematic representation of the genome organisation of HPV16. The replication and structural genes E1, E2, E4, L1 and L2 are accompanied by the transforming genes E5, E6 and E7. The genes are transcribed from different reading frames on the same strand of DNA. The organisation of the ORFs is common across most PV genomes; however the transforming genes that are present may vary in different PV types. Some PV genomes may possess an E8 ORF in the E6-E7 region and the λ PVs possess a second NCR in the E2-L2 region. The extent of overlap between adjacent ORFs may also vary with PV types. Adapted from Doorbar (2006)

The URR varies in length from 360-700 bp among the PVs and contains the viral origin of replication (ori), as well as transcription factor binding sites and promoter elements (promoter sequences and late polyadenylation signal). The ori is approximately 60-80 bp in length and contains binding sites for two viral proteins – E1 and E2, which initiate viral replication and regulatory control upon ori binding. The genes in the coding portion of the PV genome are classified as either ‘early’ or ‘late’ genes, in reference to their time of expression during the virus life cycle (Danos, Katinka and Yaniv 1982).

The early genes E1, E2 and E4 encode proteins involved in replication and transcription, and are common to all PV genomes. The E4 ORF is contained within the E2 ORF and is translated from its second reading frame. The

remaining early genes – E5, E6 and E7 – manipulate cellular processes to promote viral replication in differentiated epithelial cells. E6 and E7 are found to be sufficient for cellular immortalisation (Munger et al. 1989a) and most PV genomes possess both these genes. However, an E6 ORF is not observed in the genomes of the ξ genus bovine PVs (Jackson et al. 1991; Hatama, Nobumoto and Kanno 2008; Hatama et al. 2011; Zhu et al. 2012) and two γ HPV types (HPV101 and 103) (Chen et al. 2007a), whilst all PV genomes sequenced from cetacean species lack an E7 ORF (Rehtanz et al. 2006; Van Bressem et al. 2007; Gottschling et al. 2011a; Robles-Sikisaka et al. 2012).

The PV genomes PePV1, FlPV1, and FcPV1, sequenced from avian species, all lack canonical E6-E7 ORFs (Terai, DeSalle and Burk 2002; Van Doorslaer et al. 2009) and instead possess a different ORF in the E6-E7 region. This ORF lacks significant sequence similarity to other E6 and E7 genes but possesses functionally important sequence motifs that are conserved among E7 proteins and is therefore also referred to as E7. It was proposed that the differences in the mammalian and avian genome organisations in the E6-E7 region may highlight a later acquisition of the transforming genes by mammalian PV genomes after diversification to the avian and mammalian hosts (Garcia-Vallve, Alonso and Bravo 2005). Recent characterisations of turtle (Herbst et al. 2009) and snake (Lange et al. 2012) PV genomes, however, reveal the presence of E6 and E7 ORFs. Snakes and turtles are more closely related to birds than to mammals, and therefore the presence of E6 ORFs in their PV genomes adds support to an alternative hypothesis of E6 gene loss by the avian PVs (Herbst et al. 2009).

The avian and reptilian PVs (Terai and Burk 2002; Herbst et al. 2009; Van Doorslaer et al. 2009; Lange et al. 2012) also lack an E5 ORF. Among mammals, the E5 ORF has been identified in the α, δ, ε (except BPV5), κ, and ξ genera. In genomes of the α, δ, ε and κ PVs, the E5 ORF is located in the genomic region between the E2 and L2 ORFs (the E2-L2 region), and may/may not overlap with one or both of these ORFs. Multiple E5 ORFs have been identified in the E2-L2 region of many PV types; however, phylogenetic analysis of the putative E5 proteins suggests that many of these multiple copies may be spurious translations

(Bravo and Alonso 2004). For the λ PVs, the absence of E5 ORFs occurs in spite of uncharacteristically large E2-L2 regions, over 1 kb long (Garcia-Vallve, Alonso and Bravo 2005).

The genomes of the ξ BPVs possess a small E2-L2 region; the E5 ORF was initially labelled as an E8 ORF (Narechania et al. 2004) as it is found to be located in place of the E6 ORF at the 5' end of the early region. In addition to an E5 ORF, the κ PVs also encode an E8 protein in the E6 region of the genome (Giri, Danos and Yaniv 1985; Christensen et al. 2000). The E8 protein appears E5-like in structure and function and may be more functionally active than the E5 ORF in κ PVs (Nonnenmacher et al. 2006).

The late genes, L1 and L2, of the PV geneome, are expressed in the final stages of a productive viral life cycle. L1 and L2 respectively express the major and minor capsid proteins, which are necessary for virus particle formation; their presence is therefore conserved in all PV genomes.

The order in which these genes occur in the PV genome is maintained among most PV genomes and only a few exceptions regarding the position of the transforming genes have been observed. Some curious additions are observed in specific PV genomes. The λ PVs (which appear exclusive to species from the order Carnivora) (Tachezy et al. 2002), the hedgehog-infecting EePV1 (Schulz et al. 2009), the horse-infecting EcPV2 and EcPV3 (Lange et al. 2011), and the snake MsPV1 (Lange et al. 2012) display one additional genomic element – a second non-coding region (NCR2) between the E2 and L2 ORFs. The NCR2 is found to be absent of the E2 binding sites and promoter elements found in the URR, and, aside from MsPV1, is found to be longer than the URR (e.g., 1172 nt vs. 472 nt for the length of the URR in EePV1). Further characterisation of the NCR2 remains to be performed.

The avian PePV1, FlPV1, and FcPV1 all possess an E9 ORF that is embedded within the E1 ORF (Van Doorslaer et al. 2009). The function of this protein has not been uncovered and it is not found to be homologous to any other PV proteins. In addition, the cetacean PVs that have been sequenced to date (except TtPV1) and the ξ bovine PVs (BPV3 and BPV4) display an additional ORF

embedded within the L1 ORF, which has been labelled L3 (Robles-Sikisaka et al. 2012); the function of this ORF is yet to be determined.

1.3.2 Protein Functions

The PVs rely on only a small set of proteins to carry out productive infection. Most of these proteins are therefore charged with multiple functions to regulate expression of cellular genes, manipulate cellular pathways, induce DNA synthesis and silence immune responses to viral presence in the host cells.

In the early region, the E1 protein has maintained a high level of sequence and structure conservation among the PVs (Longworth and Laimins 2004a). It plays a central role in initiating viral DNA replication and consequently is subject to strong functional constraints. The biological activities of E1 are largely achieved through a DNA-binding domain (DBD), approximately 150 amino acids (aa) long, which recognises E1 binding sites in the viral ori (Enemark et al. 2000), and a helicase domain, approximately 200 aa long, which forms dihexameric rings upon DNA-binding to separate the DNA strands for replication (Lin et al. 2002). Both the DBD and helicase domains are highly conserved among PVs whilst other regions of the E1 protein show less sequence conservation.

In addition to its role in DNA unwinding, the helicase domain was identified as an interaction partner of cellular DNA polymerase α-primase (Masterson et al. 1998), a necessary enzyme for DNA replication, and cyclinE/cyclin-dependent kinase (cdk)-2 complexes (Cueille et al. 1998), which are essential regulators of the cell cycle – this interaction may allow the virus to utilise cellular regulatory mechanisms in the regulation of its own replication. The less conserved N-terminal domain of E1 may also play a crucial role in replication: phosphorylation of various conserved sites within this domain by cdk complexes results in inactivation of its nuclear export signal and hence ensures nuclear retention of the viral protein for DNA replication (Deng et al. 2004).

The E2 protein comprises of two functional domains – a C-terminus DBD and an N-terminus transactivation domain. A necessary step in the initiation of viral replication is the formation of an $E1_2E2_2$ complex via interactions between the transactivation domain of E2 and the E1 helicase domain. This complex serves to increases the DNA-binding specificity of the E1 protein at the ori (Sedman and Stenlund 1995; Berg and Stenlund 1997; Stenlund 2003).. Upon loading of E1 onto the viral ori, cellular heat shock proteins effect the dissociation of E2 from the complex (Lin et al. 2002), terminating its role in replication initiation.

Of equal importance is the E2 protein's role as a viral transcription factor and its regulation of viral gene expression, which ensures viral infection is maintained at low copy numbers during the early stages of the viral life cycle (Steger and Corbach 1997). The E2 protein also appears to be responsible for ensuring newly replicated viral genomes survive basal cell division: prior to mitosis, the E2 DBD binds to the viral genome, whilst the E2 transactivation domain binds to the cellular chromosomal associated factor Brd4 (the bromodomain-containing protein 4) to ensure equal segregation and efficient transfer of the episomes into the nuclei of the daughter cells (You et al. 2004; Baxter et al. 2005). The highly functional transactivation domain and DBD of the E2 protein are separated by a stretch of sequence that is known as the E2 hinge region. The main function of the hinge region appears to be to provide the degree of structural flexibility required for the DBD and transactivation domains to perform their various functions and act independently of each other (Gauthier, Dillner and Yaniv 1991).

The E4 ORF lies within the hinge-encoding region of the E2 ORF. It lacks an initiation codon and is therefore translated from a spliced mRNA transcript containing the first 5 codons of the E1 ORF (Longworth and Laimins 2004a). There is little sequence conservation observed among the E4 proteins, which vary greatly in size from 50-331 aa. Although the E4 ORF is located in the early region of the genome, expression of E4 increases substantially in the latter stages of the virus life cycle. In fact, the E4 gene is found to be the most highly expressed PV gene during the productive life cycle (Longworth and Laimins 2004a).

Attempts to characterise E4 functions have identified a central role in facilitating cellular release by associating with and destabilising the cytokeratin networks (Doorbar et al. 1991) which confer structural integrity upon the cell. Interaction with keratin filaments is thought to occur via two conserved E1^E4 sequence segments: an N-terminus MADxxA motif, contributed by the E1 polypeptide, and an LLxLL leucine cluster, in the E4 polypeptide. Disintegration of the filaments may also cause secondary disruption of many cellular processes including signal transduction. The leucine cluster of the E4 protein has also been identified in interactions with mitochondria which are thought to induce apoptosis in terminally differentiated keratinocytes, thereby facilitating viral release from host cells (Raj et al. 2004).

Whilst the transforming proteins assume the task of preparing the differentiated cells for replication, E4 helps to maximise viral replication in these cells by preventing host genome replication during the viral-induced S (synthesis) phase of the cell cycle (Roberts et al. 2008). A process of progressive N-terminal cleavage occurs to generate smaller E4 polypeptices from the full length protein at various stages during the replicative phase. These polypeptides form multimer complexes which assist the full length protein in appropriating the host replication machinery for viral genome replication (Roberts et al. 2008). The various E4 species act to suppress host genome replication by inhibiting the binding of cellular replication licensing factors Mcm2 and Mcm7 to chromatin (Roberts et al. 2008) and inducing G2 cell-cycle arrest by preventing nuclear localisation of the cyclin B/Cdk1 complex which is necessary for cell-cycle progression to mitosis (Davy et al. 2002; Knight et al. 2004; Davy et al. 2005).

The transforming genes E6 and E7 both encode zinc finger proteins that interact with numerous cellular proteins to enforce a replicative state within differentiated epithelial cells. Both proteins share structural features and possess multiple domains of a Cys-x-x-Cys zinc-binding motif (typically, 4 motifs are found in E6 but only 2 in the E6 of reptilian PVs and 2 in E7, Barbosa, Lowy and Schiller 1989) and it has been proposed that the genes may have evolved from an ancient duplication of genetic sequence containing the zinc-binding motif (Cole

and Danos 1987). The specific cellular activities of the two proteins differ; E7 may be assumed to be the 'major' transforming protein as its functions are essential in causing cellular transformation and enabling DNA replication in the upper layers of the epithelium (Cheng et al. 1995; Flores et al. 2000), though the transforming functions of E6 complement those of E7. Functional analyses of the transforming proteins have largely been confined to those of the high-risk HPV types (in which they are referred to as 'oncoproteins') in an attempt to elucidate the particular interactions that contribute to the development of malignant states. Comparatively less information has been unearthed about the functions of the low-risk transforming proteins in productive infection.

The principal interaction of the E7 oncoprotein appears to be with the retinoblastoma tumour suppressor protein (pRb; (Munger et al. 1989b)). The E7 protein contains a pRb binding domain motif, Leu-x-Cys-x-Glu, which is conserved in most PVs. Interaction of E7 with pRB is likely to have numerous cellular implications, most of which remain to be studied; however an important consequence is activation of the E2F transcription factor, which is otherwise deactivated by pRb binding (Longworth and Laimins 2004a). By indirectly activating E2F, E7 increases transcription of genes necessary for DNA synthesis and therefore encourages a replicative state within the cell (Phelps et al. 1988). pRB interaction is not, however, guaranteed to induce cell transformation (Ciccolini et al. 1994; Schmitt et al. 1994) and the affinity of E7 for pRB varies among PV genomes (Munger et al. 1989b) with low-risk HPV E7 proteins showing a weaker pRb binding affinity than their high-risk homologs.

E7 is also able to interact with the retinoblastoma-like proteins, p107 and p130 (Phelps et al. 1988), via the pRB-binding motif; these proteins exhibit similar activities to pRB, including interaction with E2F transcription factors. E7 interaction with tumour suppressor proteins and increased expression of E2F enhances the activity of cyclin dependent kinase (cdk)/cyclin complexes which are regulators of cell cycle progression (Davies et al. 1993; Morozov et al. 1997). The pRB-binding motif is found to be absent in the δ and ε PVs associated with fibropapillomas, as well as in the γ HPVs and a few β HPVs though the HPVs are

able to transform the infected cell through pRB-independent mechanisms (Caldeira, de Villiers and Tommasino 2000).

The E7 protein may also interact directly with the cyclin proteins to influence cell cycle progression (McIntyre, Ruesch and Laimins 1996) and is observed to prevent cellular growth inhibition by interactions with the cdk inhibitors, $p21^{CIP1}$, $p27^{KIP1}$ and $p15^{ink4A}$ (Pietenpol et al. 1990; Garcea and DiMaio 2007, p. 215). E7 is also observed to enhance E2F activity via associations with histone deacetylases (HDACs) (Phelps et al. 1992; Longworth and Laimins 2004b) since deacetylation of E2F serves to activate the transcription factor (Marks et al. 2001). E7-HDAC binding may serve an additional function in avoiding immune detection by preventing the expression of interferon regulatory factor 1 (IRF-1) (Park et al. 2000) which normally forms one of the first immune responses to detection of pathogen presence.

Like E7, the E6 protein is also involved in numerous cellular interactions to prevent cell cycle arrest and immune detection by the host immune system. Host cells may typically respond to the E7-pRB interaction by increasing expression of another tumour suppressor protein, p53 (Jones, Thompson and Munger 1997). p53 is responsible for either activating DNA repair proteins or inducing apoptosis, in response to DNA damage. E6 causes the inactivation of p53 by binding to the E6 associated protein (E6AP), a ubiquitin protein ligase, which is then able to engage p53 and mark it for ubiquitylation (Scheffner et al. 1990; Huibregtse, Scheffner and Howley 1991; Scheffner et al. 1993). The inactivation of p53 by the E6 ORF had been observed for all high-risk species of the α PVs and for a low risk PV type (HPV71); the activity has been linked to the presence of a non-basic residue at a site in close proximity to the E6AP binding region in the E6 ORF (Fu et al. 2010).

E6 also employs E6AP for the degradation of PDZ-domain containing proteins (Nakagawa and Huibregtse 2000; Favre-Bonvin et al. 2005), which serve as organising centres for complexes processes such as signal transduction, transcriptional regulation and receptor assembly (Garcea and DiMaio 2007, p. 203), thereby disrupting these processes. E6 plays an important role in cellular immortalisation by increasing expression of human telomerase reverse

transcriptase (hTERT), the catalytic subunit of telomerase (Klingelhutz, Foster and McDougall 1996). Telomere shortening is a mechanism employed by the cell to limit the number of cell divisions; E6-induced expression of hTERT results in elongation of telomere ends and thereby eliminates another key regulatory mechanism (Klingelhutz, Foster and McDougall 1996; Liu et al. 2009). hTERT expression is found to be a characteristic of oncogenic-E6 proteins and not of the low-risk E6 proteins (Van Doorslaer and Burk 2012).

E6-induced immune response suppression may take on a number of forms including binding to Interferon Regulatory Factor-3 (IRF-3) (Ronco et al. 1998), interacting with interleukin 18 (IL-18) cytokine to prevent activation of a cell-mediated immune response (Lee et al. 2001) and precluding antigen presentation by Langerhans cells via down-regulation of the E-cadherin molecules necessary for cellular adhesion (Matthews et al. 2003).

In infections with high-risk HPV types the E6 and E7 oncoproteins are observed to work together to impair the stability of the host genome and consequently increase the chances of malignant progression (Duensing et al. 2000). Expression of the E7 oncogene overstimulates centrosome synthesis causing cell division to proceed in the presence of multipolar mitotic spindles, which then prevents normal chromosome segregation and produces polyploid daughter cells (Heilman et al. 2009). These events render the host genome more susceptible to mutagenesis and are compounded by the actions of the E6 oncogene which inactivates various cellular proteins with functions in DNA repair, for instance, p53. Further interference by the oncoproteins with various cell-cycle checkpoints (Thompson et al. 1997; Thomas and Laimins 1998; Fan and Chen 2004) ensures cell cycle progression despite these unstable cellular conditions. Finally, it has also been proposed that the E7 oncoprotein facilitates integration of viral DNA into the host genome by causing DNA strand breaks; this is a key step in malignant progression of the infection (Duensing et al. 2000).

Additional transforming functions may be carried out by the E5 protein. Among HPVs, four distinct groups of E5 proteins labelled E5α, E5β, E5γ, and E5δ were defined in the α PV genus; a high level of evolutionary divergence,

approaching 80% aa difference, was observed between the E5 proteins from these different groups (Bravo and Alonso 2004). Despite low sequence conservation, the E5 ORFs from different PVs are characterised by similar structural and functional properties. All E5 proteins are small (42-83 aa), consisting of 1 or more hydrophobic transmembrane domains. Within the cell the protein is localised to the endosomal membranes and Golgi apparatus, and forms a dimer through interactions between its hydrophilic C-terminus domain (Surti et al. 1998).

Functional analyses suggest that the E5 protein is able to modulate cell-cycle progression via interactions with membrane proteins. BPV1 E5, which is the most studied of the E5 proteins, is found to stimulate platelet-derived growth factor (PDGF) beta-receptor tyrosine kinases thereby activating a signalling cascade that encourages mitosis of the cell (Lai, Henningson and DiMaio 2000). Excessive stimulation of receptor tyrosine kinases has been linked to the development of malignant tumours and may be a contributing factor in PV-induced cancers: E5 expression has been detected in bovine bladder cancers (Borzacchiello et al. 2003), whilst the E5 protein of the cervical cancer-causing HPV16 interacts with epidermal growth factor (EGF) receptor tyrosine kinases (Genther Williams et al. 2005). The E5 protein also helps induce DNA replication and mitosis by increasing expression of cyclin A for the formation of cyclinA-cdk2 complex important during the S phase of the cell cycle. E5 function has also been linked to the suppression of inter-cell communication by down-regulating the expression of the inter-cellular gap junction protein, connexion (Oelze et al. 1995). Prevention of cell signalling may allow E6- and E7-induced transformations to proceed without the risk of stimulating a defensive response from the surrounding uninfected cells.

The E5 protein also contributes to immune evasion during infection by down-regulating cell surface expression of major histocompatibility (MHC) antigens (Ashrafi et al. 2002; Marchetti et al. 2002; Longworth and Laimins 2004a; Doorbar 2006). Expermiental studies reveal that the E5 protein prevents acidification of the endosomes and Golgi apparatus, which then causes retention of MHC molecules in the Golgi (Marchetti et al. 2002). Both MHC class-I and class-II antigens are targeted to prevent the presentation of viral peptides to the

host immune system. Although E5 contributes to cell transformation, in human cervical cancers, the E5 gene is not found to be integrated into the host genome (Borzacchiello et al. 2003).

Of the two proteins expressed from the late region, L1 displays greater sequence conservation among PVs. The sole function of the L1 protein is to form a shell around the viral genome and protect it from the external environment once the virus particle is released from epithelial cells. The L1 polypeptide folds into an eight-stranded anti-parallel beta barrel structure known as the "jelly-roll" fold (Chen et al. 2000). Three large loop structures reside on one end of the beta-barrel and C-terminal alpha-helical domains project out from the other end (Garcea and Chen 2007). The L1 proteins self-assemble into pentavalent ring-shaped capsomers held together by multiple interactions between beta sheets and the loop structures (Chen et al. 2000). The C-terminal helical domains possess cysteine residues which permit the formation of inter-pentameric disulphide bonds; through these inter-helical bonds, 72 capsomer units are organised into a protein shell of icosahedral symmetry and ~55 nm in diameter (Chen et al. 2000; Modis, Trus and Harrison 2002). Sequence variability in the loop region has been associated with epitope function and can bind neutralising monoclonal antibodies that prevent cell binding and virion uncoating (Chen et al. 2000). The position of the C-terminal end of the L1 polypeptide, in the central region of the capsid shell, is thought to indicate a function in interactions with the encapsidated viral genome (Chen et al. 2000).

The L2 protein, a.k.a. the minor capsid protein, also forms part of the viral coat though its contribution is much smaller: estimates of the stoichiometry of L2 within the capsid vary from 12 molecules in total (Kawana et al. 1999; Modis, Trus and Harrison 2002) to 1 per capsomere (Doorbar 2006). In the formation of the capsid, the C-terminal portion of the L2 protein may form hydrophobic interactions with the central cavities of capsomeres (Finnen et al. 2003).

Functional studies of the L2 protein have identified two key roles for the L2 protein in the PV life cycle (Holmgren et al. 2005). The protein may influence the infectivity of the virus through its observed ability to bind to the cell surface and

facilitate viral entry independent of L1 (Kawana et al. 2001) and by nuclear localisation signals in the N- and C-termini of the protein that can efficiently transport the viral genome to the nucleus following virion uncoating (Roden et al. 2001). The second major function of L2 is to facilitate assembly of the virus particle. Although the L1 protein is capable of self-assembly into the capsid structure, L2 molecules, with the help of chaperone protein hsc70 (Florin et al. 2004), transport L1 capsomers into the nucleus for assembly at nuclear structures where L2 also associates with viral genome-bound E2 molecules to orchestrate genome encapsidation (Day et al. 1998; Okun et al. 2001). Surface-exposed portions of the L2 protein are also found to bind neutralising antibodies and may be utilised in PV vaccination development (Kawana et al. 1999; Gambhira et al. 2007).

1.3.3 Life Cycle

PV infection occurs within a specific type of epithelial cell, known as a squamous cell due to its scale-like appearance. The stratified squamous epithelium (Figure 1.4) consists of layers of closely packed squamous cells and serves as a protective barrier against the external environment (Madison 2003); the layered nature of stratified squamous epithelium offers greater protection in areas subject to regular abrasion. Stratified squamous epithelium make up the cutaneous tissue of the skin, lips, and part of the tongue, and the mucosal tissue of the cornea, mouth, oesophagus and the anogenital tract.

The bottom layer of epithelium is known as the basal layer; the basal cells are the only cells that undergo mitotic division in normal epithelium. Following each cell division, one daughter cell migrates up through the epithelium to replenish dead cells shed from the surface. As cells progress out of the basal layer and into the suprabasal layer, they exit the cell cycle and begin a process of terminal differentiation, which involves the loss of the nucleus thus precluding further cell division. The PV genome, which does not encode its own replication machinery, is entirely dependent on the production of replicative enzymes by the host cell to

ensure its propagation. It is therefore imperative that PV particles gain entry into the basal cells of the epithelium to ensure the chance of successful infection (Doorbar 2005; Lazarczyk et al. 2009); viral entry into the differentiated cells would be futile.

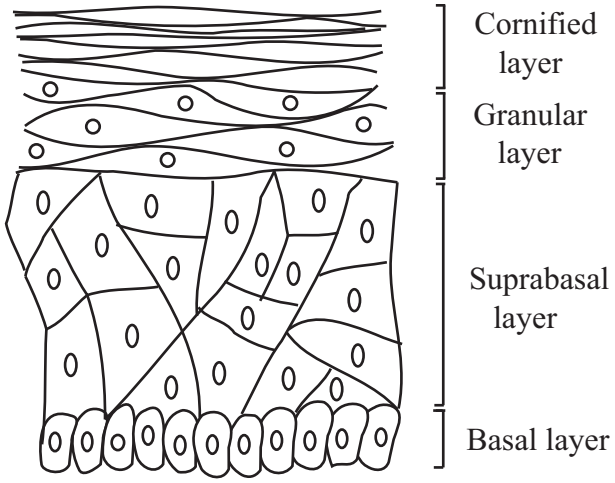


Figure 1.4: Cartoon representation of the cellular layers of the stratified squamous epithelium. The epithelial cells, called keratinocytes, undergo a process of terminal differentiation once they have progressed out of the basal layer. The differentiation process causes termination of the cell cycle, loss of nuclei and progressive cellular flattening towards the surface.

Entry into the basal cells can be gained in a number of ways. In many cases a micro-wound that punctures the epithelium down to the basal layer will provide a pathway for the virus. Hair follicles, also made of stratified squamous epithelium, are a rich supply of stem cells and are therefore also common entrance points for PVs (Doorbar 2005). Cervical infections can also arise via infection of columnar cells (Boxman et al. 2001) which exist below the stratified squamous epithelium of the cervix and eventually migrate into the basal layer of the cervical epithelium.

Characterisation of the process of cellular entry remains an on-going area of PV research (Horvath et al. 2010); however, putative viral receptors include the proteoglycan heparin sulphate (Giroglou et al. 2001; Shafti-Keramat et al. 2003), which acts as a cellular receptor for many viruses (e.g. Dengue virus and Herpes Simplex virus), and the α6 integrin (Evander et al. 1997; McMillan et al. 1999). Studies of HPV internalisation have identified the involvement of either clathrin-coated pits or caveolae (Bousarghin et al. 2003; Day, Lowy and Schiller 2003) although the process is found to be much slower than endocytosis of substrates and other viruses (Horvath et al. 2010). Following internalisation, the reducing environment of the cell causes cleavage of intercapsomere disulphide bonds, subsequent protease-driven cleavage of the carboxy-terminal helical arm of L1 releases the viral genome from the L1 capsomeres (Li et al. 1998). The viral genome is then transported into the nucleus for viral replication by the L2 capsid proteins (Doorbar 2006).

Inside the nucleus, the virus induces a brief period of replication in the basal layer in order to increase the number of infected cells (Doorbar 2006). This stage of the PV life cycle, referred to as 'genome maintenance', produces approximately 20-100 copies of the genome per cell (Longworth and Laimins 2004a), which exist as extra-chromosomal episomes. During genome maintenance, viral transcription proceeds from an early promoter located in the URR. The E1 and E2 proteins are expressed first to initiate replication but the E2 protein also acts as a regulator, restricting the amount of replication that occurs. This is achieved through a negative feedback mechanism involving adjacent binding sites in the URR for the E2 protein and cellular transcription factors necessary for activation of the early promoter. Since E2 binding site affinities vary, at low concentrations E2 engages with only two of its binding sites and permits activation of viral transcription from the early promoter. As E2 expression increases, E2 proteins bind to the remaining sites and simultaneously inhibit the binding of the cellular transcription factors, thereby repressing further viral transcription. The early promoter also initiates expression of low quantities of E6 and E7, whose contribution in the early stages of viral infection may be to eliminate cellular

checkpoints that block long-term retention of extra-chromosomal DNAs (Longworth and Laimins 2004a).

Following epithelial stem cell division, one daughter cell exits the basal layer and begins the process of terminal differentiation, involving termination of the cell cycle (Doorbar 2005). The cell cycle (Figure 1.5) is the sequence of events that occur in the process of cell division (mitosis) and is characterised by four distinct phases: G1, during which there is cellular growth; S, during which DNA replication occurs; G2, during which the cells prepare for division; and M, in which the chromosomes are separated and new daughter cells are formed. Each daughter cell enters the cell cycle in the G1 phase; in stratified squamous epithelium, the process of terminal differentiation requires exit from the cell cycle at the G1 phase. The PVs must induce cell cycle progression in cells beyond the epithelial basal layer to ensure propagation of their own genomes. As the cells progress out of the basal layer and begin differentiation, expression of the viral transforming genes is increased to cause G1/S phase transition (Doorbar 2005). S phase entry signals the expression and assembly of replication complexes for DNA synthesis and provides the viral genome with the necessary tools it lacks for its own replication. Changes in cellular factors and signalling in the differentiated cells have been associated with the activation of the late promoter located in the E7 ORF (Spink and Laimins 2005). This promoter increases expression of the viral replication genes (E1, E2, E4 and E5), and unlike the early promoter used for viral replication in the basal layer, is not regulated by the E2 protein, resulting in unrestricted genome replication for the production of new virus particles (Doorbar 2005).

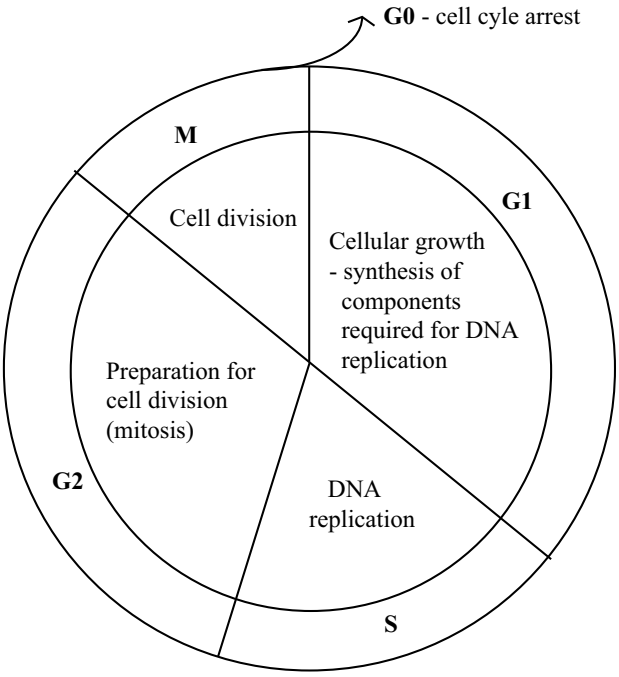


Figure 1.5: The 4 stages of the cell cycle.

The final stages of productive infection involve formation of the virus particle and its release from the epithelial lesion. Synthesis of the viral capsid proteins is restricted to the upper layers of the epithelium, in the highly differentiated epithelial cells, and may serve as a strategy for immune evasion (Doorbar 2006). L2 is expressed first and aids the assembly of the L1 proteins (Zhou et al. 1991). The high specificity DNA-binging capabilities of E2 may be utilised to aid encapsidation of the viral genome (Day et al. 1998; Zhao et al. 2000). The packaged viral particles are then ready to infect new cells. Viral release from the epithelium occurs naturally, with some help from the E4 protein, and unlike other viruses, is not self-instigated using lytic methods (Doorbar 2005). Released virus particles may either reinfect the same tissue or may separate from the infected individual. Virus transmission to a new individual typically occurs via direct epithelial contact e.g. from a mother to her newborn child; however, the virus particle is highly durable in external conditions (Roden, Lowy and Schiller 1997), and can therefore also be transmitted via indirect means, e.g. HPV2, which causes plantar warts on the soles of feet, commonly resides on the wet floors of communal showers and swimming pools.

Throughout the viral life cycle, the timing and level of expression of the various genes is crucial to ensuring a productive life cycle (Doorbar 2005). The different stages of PV infection are found to be finely tuned to the differentiated states of the epithelial cells. The expression of polycistronic mRNAs transcripts is common during the life cycle and transcripts containing the late genes, whose products are only active in terminally differentiated keratinocytes, have been observed in less differentiated cells (Stoler et al. 1989; Stoler et al. 1992). Alternative splicing mechanisms may be used to regulate the expression of different genes, particularly the oncogenes, in the different layers of the epithelium (Doorbar 2005; Tang et al. 2006; Zheng and Baker 2006; Mole, Milligan and Graham 2009; del Moral-Hernandez et al. 2010). However, the expression of PV genes has been shown to be altered by changing the particular codon specified for a degenerately encoded amino acid (Disbrow et al. 2003; Mossadegh et al. 2004; Muller 2005). It is therefore proposed that differences in codon usage preferences among individual PV genes correlate with differences in tRNA levels during cellular differentiation (Zhou et al. 1999; Zhao et al. 2005; Zhao and Chen 2011).

1.3.4 Pathogenicity

The epithelial lesions and hyperplasias symptomatic of PV infection are a consequence of viral-induced cell proliferation in the epithelium. In benign lesions, the cdk inhibitors act to reduce the amount of abnormal cell proliferation in the upper layers of the epithelium but in malignant lesions, increased activity of the oncogenes eliminates such cellular regulations (Doorbar 2006). Malignant transformation of PV-induced (cervical) lesions has been linked to the virus's ability to persist for years or even decades in the cells of the epithelial tissue (Bosch et al. 2002; Schiffman et al. 2005).

Persistent occupation of host cells affords the virus DNA the opportunity to transform the infected cells and to insert itself into the host's genomic material. The integration of the high-risk HPV16 and HPV18 genome into the host genome is found to be strongly correlated with the development of cervical cancer

(Hopman et al. 2004; Canadas et al. 2010). Most significantly, the insertion of the E6 and E7 genes enables continued expression of these genes thereby ensuring continued manipulation of cell-cycle regulation pathways by the virus.

Viral genome integration generally results in loss or disruption of the other viral genes, in particular the E2 gene. In productive infections the E2 protein plays an important role in regulating the expression of the transforming genes; its absence in oncogene-transformed cells results in uncontrolled expression of the integrated E6 and E7 genes. Inhibition of the functions of pRB by E7 results in continued DNA replication and cell division; errors sustained by the host's genomic material during each round of genome replication are allowed to accumulate due to the impairment of p53 by E6. When mutations supporting uncontrolled cell growth are incurred, PV infection can progress to an oncogenic stage.

The development of high grade cervical intra-epithelial neoplasias, which will progress to cancer, occurs over a period of several years and may be facilitated by the presence of other carcinogens, for example tobacco metabolites (Doorbar 2006). Genetic mutations caused by common carcinogens – UV radiation, pollutants, etc – in genes responsible for the regulation of cellular growth are maintained under the lack of cellular regulation induced by PV infection and hence are able to exert their harmful effects. Besides cervical cancer, genital HPVs have also been identified in cancers of the anus, vulva, vagina, penis, mouth and larynx (Hoory et al. 2008). HPVs have also been linked to cancers of the breast, lung, colon, rectum, prostate, oesophagus, head and neck (Chang et al. 1990; Cooper, Taylor and Govind 1995; Suzuk et al. 1996; Gillison 2004), and possibly of the bladder (Campo 2002), though not all of these cancers are limited to the stratified epithelium. HPV infection may play a causal role in some of the anogenital cancers but for the rest of these cancers, viral infection is assumed to play a secondary role to the more commonly known carcinogens, and can involve high-risk or low-risk HPV types (Munoz et al. 2006).

Skin cancers associated with the cutaneous HPVs, specifically the high-risk β HPV types 5 and 8, have been reported in patients suffering from epidermodysplasia verruciformis (EV), an inherited skin disease, as well as in

immunosuppressed individuals, such as transplant patients (Antonsson 2012). Among infections of non-human species there is also some evidence of PV oncogenicity: the cottontail rabbit PV type (CRPV) and a canine PV type (CPV2) can both cause cancer in their natural hosts (Giri, Danos and Yaniv 1985; Yuan et al. 2007). Carcinogenesis was also observed in experimental cross-species infections of BPV1 in hamsters (Robl and Olson 1968). In cows, natural PV infection results in only benign lesions but the development of squamous cell carcinomas remains a risk due to the presence in the wild of bracken fern which contains chemical mutagens and immuno-suppressants (Campo 1997). Ingestion of this plant has resulted in cases of bladder cancer in cows infected with BPV-1 or BPV-2 (Campo 2002) and cancer of the upper gastrointestinal (GI) tract in cows infected with BPV-4.

1.3.5 Immume Evasion Strategies

Viruses face constant selective pressure from their host's immune system and therefore must evolve strategies to negotiate through host defence mechanisms if they are to ensure successful infection and continuity of their lineage. The vertebrate hosts of PVs have evolved highly developed systems to deal with pathogen invasion and consequently the virus is observed to employ an array of preventative and defensive mechanisms to subdue stimulation of both innate and adaptive immune responses. These tactics allow the virus to delay immune clearance by their host.

Various characteristics of the PV life cycle appear to be optimised towards minimising immune detection. Firstly, the short-lived nature of epithelial cells is highly advantageous to viral survival as they are subject to less immune surveillance than other cells (Egelkrout and Galloway 2007). Likewise, PV infection remains localised within the epithelium and does not spread into the bloodstream where the detection of pathogenic presence can rapidly evoke an antibody response from the host (Schwarz and Leo 2008). The innate immune system is also alert to the presence of foreign DNA in the cytoplasm (Frazer

2009) and hence, we observe rapid import of the viral DNA into the nucleus upon viral uncoating. In the epithelial stem cells the viral proteins are expressed at low copy numbers thereby minimising the risk of initiating an immunogenic response; expression of the most immunogenic viral proteins – the capsid proteins – is delayed until the last stages of the PV life cycle, in the very upper layers of the epithelium (Frazer 2009). The natural process of cell death and desquamation of differentiated epithelial cells provides the new virus particles an inconspicuous release mechanism that avoids an inflammatory response and the risk of antigen presentation as is incurred following virus-enforced cytolytic release (Doorbar 2006; Schwarz and Leo 2008).

The PV proteins have also evolved mechanisms to obstruct immune reaction pathways of the innate response. In cervical neoplasias, the E6 and E7 proteins are found to cause multiple breakdowns in the immune response, including the prevention of intercellular signalling through inhibition of interleukin action (Lee et al. 2001) and of interferon-responsive gene expression (Ronco et al. 1998; Barnard and McMillan 1999; Chang and Laimins 2000; Park et al. 2000; Nees et al. 2001). It has also been suggested that the viral oncoproteins may prevent the Langerhans cells, which are stimulated by antigen binding, to initiate an adaptive immune response (Matthews et al. 2003; Zhang et al. 2003; Guess and McCance 2005) via cytotoxic T cells. Along with E6 and E7, the E5 protein protects the virus by preventing MHC-I and MHC-II antigen presentation at the cell surface (Ashrafi et al. 2002; Marchetti et al. 2002; Longworth and Laimins 2004a; Doorbar 2006). Given the extent of PV diversification observed, many undiscovered mechanisms are likely to be employed by the different PV types and their variants.

The ability to cause asymptomatic infection, where viral episomes are maintained in the basal layer without further progression of infection or any sign of clinical disease, provides a situation in which PV infection can be reactivated at a later stage when levels of immunosurveillance in the host decline. Latent PV infections can remain inactive for up to several years. During latent infection there is minimal gene expression, involving just E1 and E2 (Zhang et al. 1999),

which further reduces the risk of immune detection. Asymptomatic infections may be highly prevalent: PV DNA has been detected in the healthy skin of humans, including new-born babies (Forslund et al. 1999; Antonsson et al. 2000; Antonsson et al. 2003); non-human primates and ungulates (Antonsson and Hansson 2002; Ogawa et al. 2004); Australian animals (Antonsson and McMillan 2006) and in horses (Bogaert et al. 2008).

For most cases (~90%, Schwarz and Leo 2008) of PV infection the host's immune system is able to gain control over the virus: PV infections regress spontaneously with only a small percentage of infections becoming persistent. PV infections can be cleared by the immune system in less than a year (Hopfl et al. 2000). In humans, it is observed that a successfully resolved PV infection protects against future infections by the same PV type (Frazer 2009). However, when the host's immune system is incapable of defending against viral infection, PV infection can spread around the body more easily and epithelial lesions are found to become more prevalent as well as increasing in severity. Reduced immuno-competency affects patients suffering from Epidermodysplasia verruciformis (EV), who often develop skin cancer due to activation of latent PV infection; transplant patients (Halpert et al. 1986; Petry et al. 1994); aging individuals; patients infected with the immunodeficiency virus (Frisch, Biggar and Goedert 2000) and species facing extinction (Sundberg et al. 2000; Rector et al. 2007).

The last decade has seen the emergence of the first two vaccines for protection against HPV infection (Koutsky et al. 2002; Harper et al. 2004; Villa et al. 2005). Both vaccines are prophylactic vaccines that introduce innocuous virus-like particles (VLPs), consisting of an L1-derived protein coat that is absent of genomic material, into the host to stimulate production of neutralising antibodies. The antibodies produced are specific towards the viral epitopes encountered and therefore the vaccines do not offer general protection against all HPV infections; however, priority has been given to immunisation against PV types causing cervical cancer. *Cervarix*^{TM}, developed by GlaxoSmithKline, consists of VLPs derived from the L1 protein of HPV16 and HPV18, whilst *Gardasil*^{TM}, developed by Merck, consists of VLPs for the HPV types 6, 11, 16,

and 18. Clinical studies have found the HPV vaccines to be effective in protecting against both new and persistent infections, and the production of high levels of the neutralising antibodies was maintained for several years after vaccination (Harper et al. 2004; Harper et al. 2006; Mao et al. 2006; Villa et al. 2006).

1.4 PV Evolution

1.4.1 Rate of PV Evolution

Estimated rates of the evolution of PV types show that, contrary to the common perception of viruses as rapidly evolving pathogens, PV evolution appears to proceed slowly. Sequence comparisons of two bovine PV type 1 variants sequenced almost 30 years apart, and from different continents, exhibited less than 0.1 % nucleotide differences across a 4807 bp long sequence from the early and late region of the genome (Ahola et al. 1983). In humans, analysis of HPV-16 and HPV-18 variants obtained from different ethnic groups revealed a high degree of similarity among the variants, with only 5% sequence divergence in the most variable genomic regions (Ong et al. 1993). This is in stark contrast to HIV or influenza A genomes which can achieve 1% nucleotide differences within the space of a year (Gibbs, Calisher and Garcia-Arenal 1995). Whilst these RNA viruses evolve at a rate of ~ 10^{-3} nucleotide substitutions/site/year (Duffy, Shackelton and Holmes 2008), estimates of the evolutionary rate of PVs lie on the order of 10^{-8} nucleotide substitutions/site/year.

The amount of evolution that has occurred between molecular sequences can be estimated using models of sequence change. If it is possible to specify the time over which the estimated changes occurred (i.e., calibrate the timescale of evolution) then the rate at which the sequences have been evolving can also be estimated. Various estimates of the evolutionary rate of PVs have been obtained in this manner. However, the slow-evolving nature of PV genomes, as demonstrated by the BPV1 variants, means that little change occurs over measurable timescales. Consequently, evolutionary rates have been estimated

from PV types infecting different host species, using the time at which the host species diverged, to represent the time at which the corresponding PV types diverged. Various estimates have been reported, each one being derived from a different subset of PV types.

An evolutionary rate of 3.3×10^{-8} nucleotide substitutions/site/year was estimated from the E6 gene of PV types isolated from chimpanzee (PtPV1) and bonobo (PpPV1) (Van Ranst et al. 1995). A second evolutionary rate of 3.6×10^{-8} nucleotide substitutions/site/year was estimated from the E6 gene of these PVs and their closest relative – the human PV type 13 (Van Ranst et al. 1995). Tachezy et al. (2002) obtained lower average rate estimates of $7.3\text{-}9.6 \times 10^{-9}$ nucleotide substitutions/site/year from from the E6, E1 and L1 genes of PV types isolated from the domestic cat (FdPV1) and domestic dog (CPV1) (Tachezy et al. 2002). A re-estimation of the evolutionary rate following the detection of novel PV types (PcPV1, LrPV1, PlpPV1, UuPV1) from different Felidae species (puma, bobcat, Asiatic lion, and snow leopard), provided rates ranging from $1.76\text{-}2.69 \times 10^{-8}$ nucleotide substitutions/site/year across the different genomic regions and an overall rate of 1.95×10^{-8} nucleotide substitutions/site/year for the coding region of the feline PVs (Rector et al. 2007).

Average evolutionary rates ranging from $0.9\text{-}2.2 \times 10^{-8}$ nucleotide subs/site/year were estimated for the E1, E2, L1and L2 genes of turtle (CmPV1, CcPV1) and avian (FcPV1, and PePV1) PV types (Herbst et al. 2009). These rates appear to agree with estimates from the feline PV dataset of Rector et al. (2007), which ranged from $1.76\text{-}2.13 \times 10^{-8}$ nucleotide subs/site/year among the E1, E2, L1and L2 genes; however, substantial overlap in the confidence intervals for the estimated rates is only observed for the L1 ORF. Moreover, calculation of the divergence time of the turtle-avian PV split, using the evolutionary rates estimated from the feline PVs, produced estimates of approximately 60 My, which is over 3 times more recent than fossile estimates for the time of divergence of the corresponding hosts (Herbst et al. 2009). Three possible explanations exist for the inconsistency. The estimates for the evolutionary rates of the feline PVs, or the reptilian-avian PVs, or both, may be grossly incorrect. Alternatively, the different estimates obtained for different PV datasets may

indicate evidence against a constant rate of evolution, which was assumed in all cases, across different PV lineages. The third explanation is that the PVs have not codiverged with their hosts and therefore use of the host's divergence times to calibrate the timescale of PV evolution is erroneous.

Despite the lack of consensus among the estimated rates, the values estimate suggest that the PVs evolve at a rate more comparable to that of their hosts, which is estimated to be on the order of 10^{-9} nucleotide substitutions/site/year (Miyamoto et al. 1988; Makalowski and Boguski 1998). Slow evolutionary rates are characteristic among other DNA viruses and are largely attributed to the fact that viral genome replication is performed by the host's own DNA polymerases, which possess proof-reading and error-correcting mechanisms to ensure high-fidelity in replication (Shadan and Villarreal 1993).

1.4.2 PV-Host Associations

The Papillomaviridae have diverged to infect a diverse set of host species from the mammalian, avian and reptilian classes of vertebrates (Bernard et al. 2010). However, it is not known how the observed host range was acquired by this family of viruses. Individual PV types demonstrate high specificity for the host they were isolated from and the slow evolutionary rates suggest against the ability to 'jump' between host species with ease. However, closely related PV types have been detected on distantly related hosts (Chan et al. 1992a; Myers et al. 1996b), therefore raising the question, "How have PV associations been formed on different host species?" To answer this question we must consider the different processes that enable virus diversification to new hosts.

1.4.2.1 Similarities to Parasite-Host Associations

Viruses are non-living entities that possess genetic material but which lack a cellular structure and therefore do not possess the machinery to replicate their genomes and create progeny virus particles. They must infect the cells of living organisms, where they can use the host cell's replication machinery to create new copies of their own genetic material. During infection virus proteins must interact

with numerous cellular molecules and will therefore adapt to the cellular environment to ensure successful infection. Host cell infection will also place them at risk from the host's immune system and therefore they are under constant selective pressure from the host to maintain the association. Thus, viruses can develop a high degree of specialisation for a particular host species.

The relationship between a virus and its host is therefore similar to that between a parasite and its host. Parasites are living organisms (uni- or multi-cellular) which associate with another organism (the host) for nutritional gain. Viruses and parasites are both smaller than their hosts and reproduce at faster rates. As with viruses, in parasites it also appears that there is selection for specialisation. For the parasite, the benefits of specialisation can be 'optimal foraging' (i.e. continual exploitation of the most rewarding resource) and a more efficient use of the resources that gives it an advantage over invading competitors (Futuyma and Moreno 1988). Many parasites therefore demonstrate narrow host ranges and are highly specialised to their hosts (Janz, Nyblom and Nylin 2001). Since both viruses and parasites both benefit by maintaining a specialised association with their hosts, these associations may evolve in similar ways.

The coevolutionary dynamics of parasite-host associations have long been a subject of interest (Klassen 1992); five key processes are considered (Page 2003). The commonly assumed mechanism acting on parasite-host associations is cospeciation. Cospeciation describes the process by which an associated host and parasite assemblage speciate together as the vicariance event (geographical isolation) causing host speciation affects the associated parasite species. Cospeciation therefore produces two "new" parasite-host associations among the descendant species. In virus-host associations, the virus residing within host cells is not directly affected by the vicariance event; however speciation of the host will cause separation of the virus population between the descendant hosts. Independent evolution of the separated populations will then result in divergence of the ancestral virus lineage (i.e., 'codivergence' of host and virus).

A different outcome following host speciation is that the parasite species remains associated with only one of the descendant species of the host. This process is often referred to as 'missing the boat' or 'incomplete lineage sorting'.

Alternatively, cospeciation occurs but one of the new parasite-host associations is terminated due to extinction of the parasite species. Extinction of the host species will, of course, result in extinction of the associated parasite species and therefore, as neither species remains, this is not a process of concern when studying the evolutionary history of observed parasite-host associations. I will use the collective term 'sorting events' to refer to processes resulting in the absence of a virus association on a host (i.e., missing the boat and virus extinction).

As the parasite genome replicates, it may acquire random mutations which enable the organism to exploit a new resource or environment thereby resulting in speciation of the parasite independently of the host. When the descendent parasite species evolve to utilise different resources on the same host species, multiple parasite species will be observed to infect a particular host species. This process is commonly referred to in the literature as 'parasite duplication'. If the mutations allow the parasite to utilise the resources on a different co-existing host species, then colonisation of this host will result in the formation of a new parasite-host association. This mechanism of parasite diversification is referred to as a 'host switch' or 'host transfer' and may result in closely related parasite species infecting distantly related host species. If the colonising parasite species encounters competition from a pre-existing parasite species, which is utilising the same resources, competition for those resources will result in the eventual extinction of one of the parasite species from the host.

Similarly, mutations occurring in a virus lineage may cause it to diverge independently of the host. The divergent lineages may then exploit new cellular environments on the same host (referred to as 'prior divergence') or exploit cellular environments on new hosts ('host transfer').

1.4.2.2 Fahrenholz' Rule for Codiverging Parasite-Host Associations

Comparisons of the phylogenies of associated hosts and parasites can provide indications of the processes that have produced the observed associations. The field of host-parasite cophylogenetic analysis rests on one key rule. Fahrenholz

(1913) made the general observation that closely related host species were infected by closely related parasite species and distantly related host species were infected by distantly related parasite species (Klassen 1992). Thus, the degree of divergence among parasites species tended to match that of their associated hosts. These observations, along with the assumption (of the time) that codivergence was the only process by which parasites could speciate, led Eichler (Eichler 1942) to propose Fahrenholz' rule, which states that strict host-parasite codivergence results in identical phylogenetic relationships for hosts and their associated parasites (Klassen 1992). Fahrenholz' rule is therefore employed in methods of cophylogenetic analyses (Brooks 1981; Brooks 1990; Page 1994b; Page 1994a; Charleston 1998; Ronquist 2002) to test for evidence of parasite-host codivergence: identical phylogenies indicate strict codivergence, incongruent phylogenies indicate some that degree of parasite diversification must be explained via non-codiverging mechanisms.

1.4.2.3 Do PV-Host phylogenies obey Fahrenholz' rule?

Assuming that the coevolutionary dynamics affecting virus-host associations are similar to those described for parasite-host associations, Fahrenholz' rule may be applied to examine the evidence for codivergence of PV types with their hosts. Initial PV phylogenies generated from small data sets demonstrated large evolutionary distances between PV types from distantly related hosts and the evolution of primate and non-primate PV types along separate branches (Chan et al. 1992a). These observations were thought to support the PV-host codivergence hypothesis (Bernard 1994; Van Ranst et al. 1995; Rector et al. 2007). Further supported was obtained from phylogenies demonstrating that diversification within the globally prominent HPV types 16 and 18 has followed the biogeographical patterns of their human hosts, therefore indicating codivergence within a host (Chan et al. 1992b; Ho et al. 1993; Ong et al. 1993; Bernard 1994).

However, complete concordance between estimated PV phylogenies and the host phylogeny, as per Fahrenholz's rule, was not observed. In particular, Chan et al. (1992a) found the bovine PV type BPV4 to be more closely related to some

human PV types than to other bovine PV types (BPV1 and BPV2). A further discrepancy is observed with the non-human primate PV types, RPV (now MmPV1) and CgPV1, which, rather than branching off at the base of a clade of HPV types in accordance with the host's relationships, assume positions nested within clades of human PV types (Fig. 4 and 5, Chan et al. 1992a).

As the PV database has increased over the last two decades new types have been detected in previously unknown hosts as well as in known hosts; phylogenetic analysis of these larger data sets (Chan et al. 1997a; Chan et al. 1997b; Garcia-Vallve, Alonso and Bravo 2005; Bravo and Alonso 2007; Gottschling et al. 2007a; Gottschling et al. 2007b; Gottschling et al. 2011a; Gottschling et al. 2011b; Lange et al. 2012; Robles-Sikisaka et al. 2012) have uncovered multiple discordances in the pattern of host and virus divergence events (Figure 1.6, reprinted from Gottschling et al. (2007b: Fig. 2)), which clearly cannot be reconciled with a strictly codiverging mechanism of PV diversification among vertebrates.

Conflicting branching patterns between PV and host phylogenies have been rationalised by the proposal of host transfer events. Several authors have highlighted the polyphyletic arrangement of various non-human primate PVs among large clades of α and β HPVs as an indicator of PV host transfer events (Chan et al. 1992a; Myers et al. 1994; Myers et al. 1996a; Chan et al. 1997b; Gottschling et al. 2007a). It was once thought that HPV7, prevalent in skin warts on the hands of butchers, may represent PV transfer from other animals to humans (Orth et al. 1981); however, estimated phylogenies places HPV7 among other HPV types and no related PV lineage has been detected in these animals to support this assumption.

Only two PV types, the bovine PV types 1 and 2, have been isolated from more than one host species (Otten et al. 1993; Bloch, Breen and Spradbrow 1994; Antonsson and Hansson 2002; Chambers et al. 2003; Bogaert et al. 2008). Unlike normal PV infection of the epithelium these zoonotic BPVs produce non-productive fibroblastic tumours, or sarcoids, in horses and donkeys. Among rabbits, the cottontail rabbit papillomavirus produces productive infections in its

natural host (the cottontail rabbit), but produces only poorly productive infections that can progress to cancers in domestic rabbits (Rous and Beard 1935).

The evidence seems to suggest that PV types are unable to complete a productive life cycle in non-native hosts. The difficulty of successful host transfer of PVs is not surprising given the slow evolutionary rate, and thus long adaptation times, of DNA viruses. Specific obstacles may be presented by the highly adapted nature of molecular interactions between the virus and host regulatory proteins (Shadan and Villarreal 1993). Whilst the lack of physical evidence suggests against recent host transfer events, we cannot discount the possibility that ancestral host transfer events contribute to the observed phylogenetic incongruities.

A salient feature of the PV database is that some hosts appear more than once. For instance, the phylogeny presented in Figure 1.6 contains 18 PV types isolated from *Homo sapiens* (human), 7 PV types from *Bos Taurus* (bovine), and 2 PV types from *Canis familiaris* (dog), *Ovis aries* (sheep), and *Sylvilagus floridanus* (cottontail rabbit) host species. Multiple associations with a particular host species can be formed when an associated virus lineage diverges independently of the host (i.e., prior divergence); however, if this was the case, all PV lineages associated with a particular host species would cluster together in a monophyletic clade. The fact that the PV types from human, bovine and dog hosts each fail to form a monophyletic clade suggests that if any prior divergence occurred, it was in an ancestral host species and not the observed host species.

1.4.2.4 Elucidating the History of PV-Host Association Mechanisms

The phylogenetic incongruities observed between the PVs and their hosts conflict with Fahrenholz' rule for strictly codivergence. Instead, the incongruities are thought to encode an amalgamation of codivergence, within-host adaptive radiation (prior divergence), host transfer and sorting events (Gottschling et al. 2007b). No attempt has been made, however, to decipher the evolutionary history of PV-host associations. In Chapter 4, I describe various methods of

cophylogenetic analysis that have been developed to determine the evolutionary causes behind observed host-parasite phylogenetic incongruities. I will describe the limitations of these methods that prevent their application to the PV-host phylogenies and present a new approach, based around temporal comparison of host and virus lineage splitting events, to characterise the processes behind viral divergence events and resolve virus-host phylogenetic incongruities. In the absence of known or estimable viral divergence times, I apply a biased sampling approach to divergence time estimation in Bayesian phylogenetic methods. The distributions of sampled times for various viral divergences are used to evaluate the support for codivergence, host transfer and prior divergence along the PV phylogeny. The results indicate that the observed PV-host phylogenetic incongruities can be largely explained by substantial prior divergence of PV lineages in the ancestors of extant hosts. An ancestral host transfer event is also inferred.

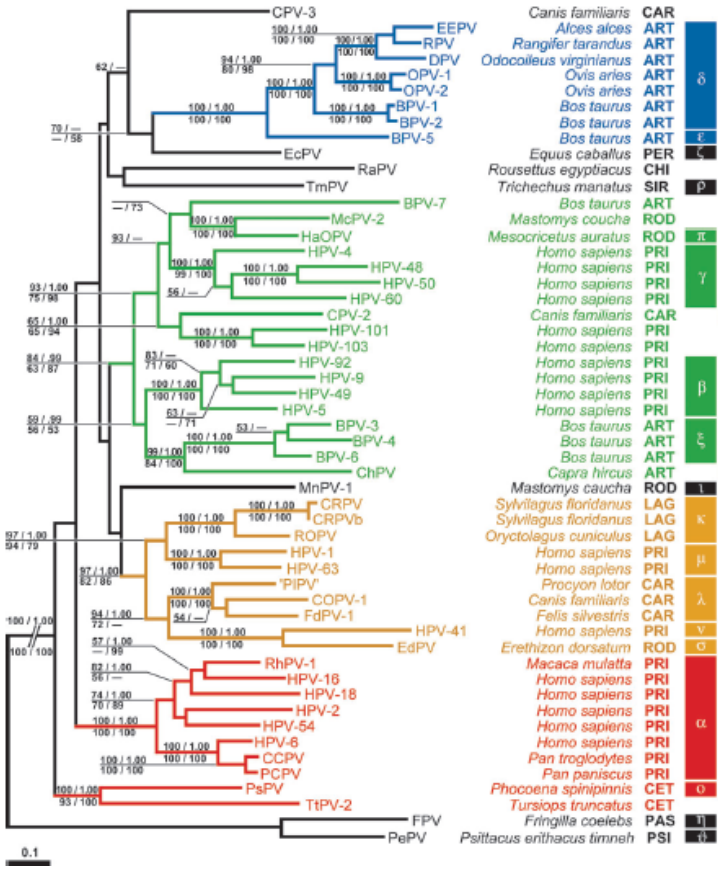


Figure 1.6: A maximum likelihood PV tree generated from combined E1-E2-L1 aa sequences. Multiple PV types infecting *Homo sapiens*, *Bos taurus* and *Canis familiaris* fail to form monophyletic clades suggesting that the path of PV lineage evolution has deviated from that of its host on multiple occassions. Reprinted from Gottschling et al. (2007b: Fig. 2).

1.4.3 Phylogenetic Incongruity of PV Genes

Characterisation of the mechanisms of PV diversification requires estimation of the phylogenetic tree relating PV types. To ensure that this characterisation was based on an accurate estimate of the PV evolutionary history, I first investigated the degree of phylogenetic compatibility among the PV genes. Phylogenetic compatibility is a necessary consideration as the evolutionary trajectory of individual genes may often deviate from that of the species at various points during its evolutionary history. Many bacterial and viral genomes have encountered lateral gene transfer, recombination or reassortment

events during their evolution (Lefeuvre et al. 2009; Simon-Loriere and Holmes 2011; Koonin and Wolf 2012; Nelson et al. 2012). These events provide additional avenues of diversity and are important in ensuring the survival of the species; however, they also produce genomes with discordant evolutionary histories. For instance, when a recombination event occurs, the evolutionary history of the inserted portion may be different to that of the non-recombined genomic region (Figure 1.7). In such cases not only would it be incorrect to make evolutionary inferences from a single phylogeny but, in some cases, the discordant phylogenetic signals may even cause the estimation method to settle on a tree topology that fails to represent any of the true evolutionary histories along the sequence (Posada and Crandall 2002).

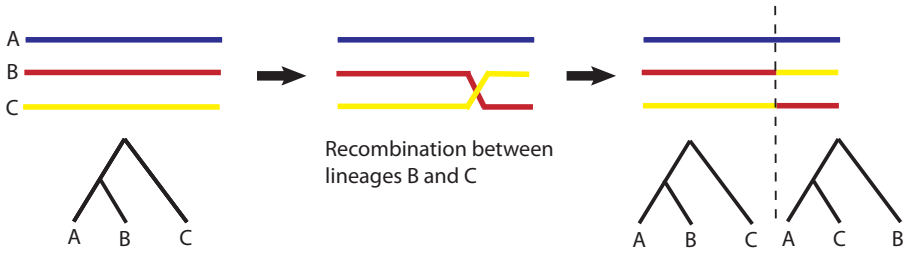


Figure 1.7. An illustration of the effect of a recombination event on phylogenetic estimation. The phylogenies of the sequences are different before and after the recombination event (breakpoint) therefore phylogenetic estimation from the entire sequence length will produce a phylogeny that is incorrect for some if not all of the sites.

1.4.3.1 Observed Phylogenetic Incongruities

PV genomes, being unsegmented structures, are not capable of reassortment; however, recombination and lateral gene transfer events could have influenced the evolutionary history of these viruses. Despite a lack of physical evidence to support the occurrence of recombination or gene transfer, various phylogenetic estimates have revealed possible discordances between the genes of certain PV genomes. In particular, within the α genus, the high-risk PVs present different

phylogenetic arrangements for the early proteins and the late proteins (Bravo and Alonso 2004; Garcia-Vallve, Alonso and Bravo 2005; Narechania et al. 2005). All early genes appear to support the monophyletic grouping of the high-risk PVs in a sister clade to the low-risk alpha PVs; however, phylogenetic estimations from late genes present a polyphyletic arrangement of the high-risk PVs (Figure 1.8, adapted from Narechania et al. 2005: Fig. 2). Phylogenetic analysis of multiple PV genera (Garcia-Vallve, Alonso and Bravo 2005) further revealed differences in the phylogenetic arrangements of the regulatory proteins, the structural proteins, and the transforming proteins; in general the PVs show consistent phylogenetic grouping of PV types by genus across all genes but differ in the respective placement of genus clades and of PV types within genus clades.

1.4.3.2 Recombination Detection in PV sequences

The observation of conflicting tree topologies for different PV genes has prompted studies searching for statistically significant phylogenetic conflicts which may indicate recombination events among PV sequences. Varsani et al. (2006) applied a suite of recombination detection programs to identify significant recombination signal in the genomes of 105 human and non-human PV types. Their analysis uncovered 7 potential recombination signals, 4 of which were supported by topological incongruities when the respective genetic regions were reanalysed. The L2 gene was identified again: 4 different recombination signals were detected, one of which involved the high-risk α PVs – the α-5, α-6, and α-7 species – that are phylogenetically separated from the remaining high-risk α PVs in late gene phylogenies. The parent sequences of the recombinant were identified as distant relatives of the α-10 PV type HPV3 and the α-13 PV type HPV54. The PV types from the α-5, α-6, and α-7 species may therefore be descendants of a recombinant type.

A second potential recombination event in the L2 gene involving α PV sequences was identified for HPV42 from the α-1 species with an α-5 PV type as one of the parent sequences. The L2 genes of the γ-HPVs were also found to be descendants of a recombinant sequence that comprised some of the L2 gene of an

ancestral β-PV type. Another recombinant sequence identified was PsPV1; once again recombination was located in the L2 gene and was thought to involve a relative of a β HPV sequence as one of the parent sequences, the nature of the other parent sequence could not be determined. Two recombinant signals were located in the L1 gene: one involved recombination within the α PV genus and the other within the β PV genus. The remaining putative recombination event is located in the E1 gene of ν-HPV41; the recombinant region within this PV type is thought to contain sequence from an ancestor of COPV and another unknown PV.

In a separate analysis of just the α HPVs (Angulo and Carvajal-Rodriguez 2007; Carvajal-Rodriguez 2008), the coalescent composite likelihood method, which utilises models of evolutionary change (McVean et al 2002; Carvajal - Rodriguez et al 2006), was used to detect recombination signals in the E6, E7, L1 and L2 genes. The α HPVs were analysed in 4 distinct groups: significant evidence for recombination was found in the E6 and L2 genes of HPVs from all species of the high-risk group; in the L1 and L2 genes of a group of low-risk mucosal HPVs which are phylogenetically closely related (α-1, α-8 and α-10); in the L2 gene of another group of low-risk mucosal HPVs (α-3 and α-15); and in the E7 gene of a group of HPV16 variants. Specific details on the locations and particular sequences from which the signals were the strongest were not obtainable from the analysis. These results support those of Varsani et al (2006) in the identification of the L2 gene as a recombination hotspot for PVs, however, this may be an artifact from increased sequence divergence in the L2 gene and it is interesting to see that significant recombination signal was not detected in the more closely related group of HPV16 variants.

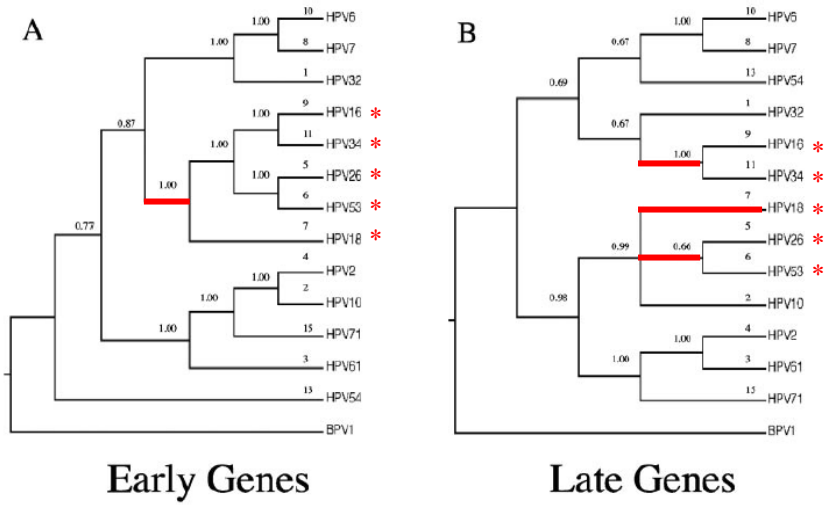


Figure 1.8. Phyogeny of the early genes and the late genes of the α HPVs. A single HPV type is used to represent each PV species within the α genus. Species 5, 6, 7, 9, and 11 comprise of the 'high-risk' (*) genital PV types whilst species 1, 8 and 10 comprise of the 'low-risk' genital PV types. The phylogenetic arrangement of the high-risk PVs differs in the two trees (branches leading to clades of high-risk PVs are highlighted in red): the high-risk PV types form a monophyletic clade in the early gene tree but are split into two (possibly three) distinct clades in the late gene tree. Adapted from Narechania et al. (2005: Fig. 2)

1.4.3.3 Testing Phylogenetic Incongruence

The observed phylogenetic incongruities and detection of multiple recombination signal among PV sequences, suggests against combined phylogenetic analysis of PV genes without first testing whether the observed differences are statistically significant. In Chapter 3, I describe some tests that can be performed to evaluate the significance of phylogenetic differences among data partitions and discuss the results of previous incongruence tests of PV sequences. These results are specific to the analysed data sets, which differ from the data set analysed in this thesis and therefore can not be applied to my data set. I have performed an independent analysis of phylogenetic compatibility among PV genes using Bayesian phylogenetic methods and report findings of significant phylogenetic incongruence between the E1, E2, E6, E7, L1 and L2 PV genes.

The results presented in Chapters 3 and 4 have been published in Shah, Doorbar and Goldstein (2010).

Chapter 2

Phylogenetic Analysis using Bayesian Methods

This chapter provides an introduction to the theory and procedure behind Bayesian methods of phylogenetic analysis, which I have used to infer evolutionary events from PV genetic sequences. The evolutionary history of a virus family is best determined from molecular sequence data, i.e., genetic or protein sequences, as other biological data, such as morphological or serological characteristics, may be less discriminative and convey less information on evolutionary rates and speciation times.

2.1 Multiple Sequence Alignment

Molecular phylogenetic analysis is based on the implicit assumption that the sequences being analysed are homologs, i.e., they are all descendants of a common ancestral sequence. A phylogenetic estimate is a proposal of the evolutionary tree relating the homologs to each other. The root of the tree represents the common ancestor of the analysed sequences, which assume positions at the ends of the terminal branches, or tips, of the evolutionary tree. The relationships among the sequences are indicated through the lineage branching patterns (tree topology) and the extent of evolutionary change along each lineage (if estimated) is indicated through the branch lengths.

Branching points (‘nodes’) along the tree represent the divergence of an ancestral lineage to form two daughter lineages. However, estimated phylogenies may postulate mutifurcating nodes (also referred to as ‘polytomies’), where a lineage diverges into more than two daughter lineages. Polytomies may either

represent a series of divergence events happened in close succession such that the exact order of lineage splitting events cannot be determined or the inability of the inference method to resolve the phylogenetic placement of a subset of sequences.

For a set of observed homologous genetic sequences, the nucleotide base $x \in \{\text{T/U(DNA/RNA)}, \text{C}, \text{A}, \text{G}\}$ observed at each position, or site, in each sequence has evolved from a base in the genetic sequence of the LCA. Thus, individual sites across sequences will be related to each other through a common ancestral base. The exceptions to this are those sites which have been inserted into individual sequences at some point during the evolutionary process - these sites will not be shared by all sequences in the data set. Deletion events causing the removal from a particular sequence will also reduce the number of sites shared by all sequences.

In order to infer the evolutionary relationships among the sequences, it is first necessary to identify site homologies across the sequences. This is achieved using alignment algorithms which generate an $n \times m$ alignment matrix where each row corresponds to one of the n sequences and each column corresponds to a homologous site. A scoring matrix indicating the cost of aligning the different nucleotide bases against one another is used to determine the optimal alignment of sites across the sequences.

The alignment algorithm may introduce gaps into some sequences to achieve a full alignment of all sites. Alignment columns possessing the gapped ('-') characters may indicate insertions of the non-gapped sites in the respective sequences or deletions of sites from the corresponding gapped sequences. Alternatively, they may indicate highly divergent sites where it is difficult to obtain a favourable alignment of the observed bases. The patterns of insertion and deletion events can contribute valuable information for phylogenetic reconstruction (Lloyd and Calder 1991); however, the difficulty in accurately inferring these events from the alignment, along with the difficulty of modelling the evolutionary processes of insertions and deletions, means that gapped sites in the aligned data matrix are sometimes excluded from the phylogenetic estimation process. Thus, evolutionary trees are generated by considering only the mutational changes that have occurred among molecular sequences.

2.2 An Overview of Non-Bayesian Methods of Phylogenetic Analysis

2.2.1 Distance Matrix Methods

Phylogenetic estimation from aligned sequence data can be performed using a variety of methods. The distance matrix approach differs from other methods of sequence-based phylogenetic estimation in that evolutionary relationships are not inferred directly from the observed sequences, but from estimated evolutionary distances quantifying the expected amount of character change between pairs of sequences. A crude estimate of the pairwise sequence distances can be obtained from the proportion of non-identical sites between sequences. However, since only the most recent character replacements at each site are detectable from the observed sequences, this method of distance estimation will underestimate the amount of change between sequences when multiple substitution events have occurred per site. The use of probabilistic models of sequence evolution (described below) allows the raw estimates of sequence distances to be corrected for the possibility of 'hidden' substitutions.

Phylogenetic estimation from the distance matrix then requires a search through all possible tree topologies relating the m sequences. For each topology, a least squares (LS) method is used to ensure that the branch lengths along this tree provide the closest fit to the estimated evolutionary distances between all pairs of sequences (Cavalli-Sforza and Edwards 1967):

$$S = \sum_{i<j} \left(d_{ij} - \delta_{ij} \right)^2 , \tag{2.1}$$

where d_{ij} is the additive distance between sequences i and j, obtained by summing the branch lengths along the shortest path from i to j, and δ_{ij} is the evolutionary distance between i and j, estimated using a model of evolutionary change. For a given distance matrix, the optimal phylogeny can be specified as the one that

provides the best fit to the estimated distances, i.e. the one with the smallest value of the least squares measure, *S*. Alternatively, a minimum evolution (ME) criterion (Edwards and Cavalli-Sforza 1963; Kidd and Sgaramella-Zonta 1971; Rzhetsky and Nei 1993) can be specified to search for the phylogeny conveying the least amount of evolutionary change among the sequences, measured by the sum of all branch lengths.

For large data sets, a search for either the LS or ME phylogeny is encumbered by the number of tree topologies that need to be evaluated. For a data set with *m* taxa, there are $\frac{(2m-5)!}{2^{m-3}(m-3)!}$ possible unrooted, labelled tree topologies that must be considered. Thus, for a data set consisting of 10 sequences, a search for the optimal phylogeny requires evaluation of the selected optimality criterion (e.g. the smallest tree length) for each of the 2,027,025 possible unrooted tree topologies. To facilitate the analysis of larger data sets, phylogenetic methods employ heuristic methods that optimise the phylogenetic estimate without searching through the entire tree space.

In distance matrix methods, clustering algorithms are employed to approximate the best topology. Two commonly used algorithms are the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Sokal and Sneath 1963), which is applied to estimate a rooted phylogeny under the LS method, and Neighbour Joining (NJ) (Saitou and Nei 1987), which enables estimation of an unrooted phylogeny under the ME criterion. The UPGMA algorithm constructs the optimal tree from the distance matrix in a step-wise manner. At each stage the closest related pair of taxa/clades are joined together in a clade on the tree. The selected pair are removed from the distance matrix and replaced by the newly formed clade. Distances from this clade to the remaining taxa/clades are recalculated by taking the unweighted arithmetic mean of the distances from each of the original components in the newly joined cluster. The NJ method adopts a different approach in which a resolved bifurcating topology (i.e., one in which each internal node has only 3 connecting branches) is derived from an unresolved star tree topology (i.e., one possessing a single internal node

that is connected to all taxa). At each stage, the NJ algorithm pairs up the taxa or clades that produce the greatest reduction in the tree length.

Distance matrix methods provide a fast method of phylogenetic estimation as the transformation of the aligned sequence matrix into a distance matrix provides a substantial reduction in the amount of data that has to be analysed. Simulation studies (Saitou and Nei 1987) have found the NJ method to be highly efficient at estimating the true tree topology for closely related sequences separated by small evolutionary distances. This is because the number of unobserved substitutions is small and therefore evolutionary distances can be estimated more accurately. However, when sequences are separated by larger evolutionary distances, the sampling error in the estimated distances may be great, affecting the accuracy of phylogenetic reconstruction. Developments to the NJ method, such as BIONJ (Gascuel 1997) and Weighbor (Bruno, Socci and Halpern 2000) are found to improve the accuracy of phylogenetic estimates under these conditions.

2.2.2 The Maximum Parsimony method

Other methods of phylogenetic estimation attempt to derive evolutionary relationships directly from the character changes observed at each homologous site, with the correct phylogeny assumed to be the one that provides the best fit to the observed evolutionary patterns among the sequences. The maximum parsimony (MP) method of phylogenetic estimation defines the best fitting tree as the one postulating the least amount of evolutionary change along the tree (Camin and Sokal 1965). The approach is similar in concept to the ME criterion utilised in distance matrix methods, however, the evolutionary models implied in MP are much simpler than the probabilistic models employed in distance methods and generally disregard the possibility of multiple substitutions per site.

For a given tree topology, a dynamic programming algorithm is used to determine the minimum amount of evolutionary change required along each branch, starting from the tips, for which sequence data exists, and progressing down the tree towards the root. Each site is assumed to evolve independently

allowing the changes at each site to be considered individually. Along each branch, the minimum amount of evolution can either be evaluated in terms of the number of character changes (Fitch 1971) or, by employing a cost matrix to weight the different changes, in terms of the cost of character changes (Sankoff 1975). The sum of either the number or cost of changes along each branch provides the measure of evolutionary change at each site (the site length); the sum of the site lengths provides the total measure of evolutionary change along the tree (the tree length). The MP phylogenetic estimate is the phylogeny with the smallest tree length.

Not all sites of the sequences are utilised by MP methods. Sites which show no change of character or with only single representatives of characters are considered uninformative in parsimony methods as they are equally supported by all tree topologies. Thus, the MP tree is not derived from all available data. Further underestimation, by MP, of the amount of evolutionary change occurring along a tree comes from the assumption that changes are equally likely along each branch of the tree. Branches representing higher evolutionary rates or longer evolutionary periods will encounter more changes of state than branches representing lower evolutionary rates or shorter evolutionary periods but these differences are not modelled by the MP method.

In some instances, MP methods have been identified as statistically inconsistent methods of phylogenetic estimation (Felsenstein 1978), meaning that even an infinite amount of data will not guarantee estimation of the correct tree. Specifically, a 'long branch attraction' problem has been identified for phylogenetic reconstruction of certain topological structures under MP. For sequences simulated along phylogenies with long branches separated by short internal branches, the greater number of changes occurring along the long branches means that it is possible for the terminal sequences to display identical site characters purely by chance (convergent/parallel evolution) despite having evolved from different ancestors. In MP estimation, the homoplasious (i.e., parallel) changes along the long branches are misinterpreted as homologous states inherited from their common ancestor, causing the long branches to be erroneously clustered together (Felsenstein 1978; Hendy and Penny 1989; Huelsenbeck and Hillis 1993). The influence of LBA can be reduced by including

more sequences closely related to those possessing long branches, thus removing the long branches from the tree. However, the identification of such sequences may be non-trivial.

2.2.3 The Maximum Likelihood Method

In maximum likelihood (ML) phylogenetic estimation, the optimal tree is defined as the one with the highest likelihood, that is, "the highest probability of evolving the observed data" (Felsenstein 1981). The probability of the observed sequences given a particular phylogeny is computed using the probabilities of character change specified by a model of sequence evolution. The evolutionary models employed to calculate the likelihood are also utilised in Bayesian phylogenetic estimation and are described in more detail below. For each topology, the set of parameter values constituting the evolutionary model (i.e., substitution model parameters and branch lengths) that maximise the likelihood of the tree is determined. ML phylogenetic estimation can proceed under simple evolutionary models consisting of very few parameters or under models of increasing degree of complexity. Complex models allow for a better representation of the evolutionary process; however, given a finite amount of data, estimation of optimal parameter values will be more difficult than for a simpler evolutionary model consisting of fewer parameters. ML phylogenetic estimation may therefore be difficult for more descriptive evolutionary models. However, ML phylogenetic estimates are found to be robust to violations of the specified evolutionary model (Yang, Goldman and Friday 1994) which is beneficial since even complex evolutionary models are likely to be simplifications of the real process. In addition, when used with a model that suitably reflects the complexity of the data set, the ML method is found to be statistically consistent (Yang 1994b).

2.2.4 Heuristic Methods for MP and ML Phylogenetic Estimation

As with distance methods, for large data sets, both MP and ML phylogenetic estimations utilise heuristic methods to obtain an optimal phylogeny. In contrast to the clustering algorithms of distance methods, MP and ML perform tree rearrangements to explore different topologies in tree space. Starting from an initial tree topology, various perturbation methods can be used to generate new topologies. These methods include nearest neighbour interchange (NNI), tree bisection and reconnection (TBR), and subtree prune and regraft (SPR). The NNI algorithm randomly selects an internal (i.e., non-terminal) branch and proposes a change to one of the two neighbouring tree topologies derived from alternative arrangements of the four subtrees protruding from the selected branch. The NNI operator can also be extended to allow the swapping of any two randomly selected subtrees (known as a subtree swap operation). The TBR algorithm randomly selects an internal branch for removal from the tree; a new topology is derived by randomly selecting a branch from the resulting subtrees and reconnecting the two subtrees at the selected branches. The SPR algorithm is similar to TBR in that the tree is bisected at an internal branch; however the subtree that is pruned from the main portion of the tree is reattached via the same node to a randomly selected branch in the remainder of the tree, thus maintaining the order of branching events in that subtree.

Traditionally, a hill-climbing approach was used to search for the optimal phylogeny, i.e., perturbations that improve the fit of the tree to the data are accepted until no further improvements are achieved. However, as the number of possible topologies increases, it becomes more likely that the sequence of rearrangements that must be performed to reach the optimal topology consists of rearrangements to topologies that reduce the fit to the data (Maddison 1991). Thus, hill-climbing algorithms may fail to reach the global optimum when tree space is large. A number of alternative search algorithms (e.g. genetic algorithms (Lewis 1998; Goloboff 1999; Brauer et al. 2002), simulated annealing (Salter and Pearl 2001; Barker 2004), and divide and conquer algorithms (Goloboff 1999)), which allow exploration through sub-optimal regions of tree space are steadily being implemented into phylogenetic software packages (Giribet 2007).

2.2.5 Confidence Measures for Estimated Phylogenies

To assess the level of confidence that can be attached to estimated phylogenies Felsenstein (1985) proposed application of a non-parametric bootstrapping approach (Efron 1979). In the absence of additional data, bootstrapping provides a means of generating artificial data sets from the same distribution as the observed data, from which the variability of parameter estimates can be evaluated. For phylogenetic estimations, we wish to generate data sets representing the same underlying evolutionary process as the observed sequences. The artificial data sets, of the same length as the analysed sequences, are obtained by repeated sampling of sites, with replacement, from the original data-matrix. This makes the assumptions that each site evolves independently and that each site in the alignment is observed with the same frequency at which it is observed in the population of site patterns generated according to the underlying phylogeny. Hence, resampling from the alignment is equivalent to sampling from the distribution of site patterns produced under the tree (Felsenstein 1985). Each artificial data set represents a bootstrap sample which can then be analysed using the same phylogenetic reconstruction method (distance matrix, MP, or ML) as the original data to produce a bootstrap phylogeny.

Topological uncertainties in the phylogenetic estimate from the real sequences are assessed by determining what proportion of the bootstrap trees display the same phylogenetic relationships. Thus each clade in the original tree estimate is assigned a bootstrap support value, which indicates the proportion of times the clade is observed in repeat samples. However, inferences of phylogenetic confidence from a non-parametric bootstrapping approach are more commonly made from a ‘majority-rule consensus tree’. This tree depicts the set of clades that are observed in the majority (i.e., at least 50%, although a higher threshold may be used if greater confidence is desired) of bootstrap trees. Taxa which fail to show a preferred clustering pattern in the majority of trees are placed into a polyphyletic clade at the base of the tree to indicate the uncertain nature of their phylogenetic positions. Clades with high bootstrap proportions may be assumed to be well supported by the majority of sites in the real

sequences and inspire confidence in the phylogenetic estimate, whilst low bootstrap proportions indicate regions where the phylogenetic signal may be highly variable among the sites.

The determination of bootstrap support values for a given phylogeny requires substantial additional computation as phylogenetic estimation must be performed on each bootstrap data set. If the assumptions of the evolutionary model or phylogenetic method used are too simplistic for the sequences being studied, incorrect phylogenies will be estimated for both the original data set and the bootstrap data sets. The concern here is in obtaining incorrect bootstrap phylogenies consistent with the estimated phylogeny so as to produce high bootstrap support values for incorrect phylogenetic groupings. The bootstrap method therefore provides an indication of the level of precision of the original phylogenetic estimate rather than its accuracy. Various simulation studies have found bootstrap support values to provide conservative estimates of the confidence in the estimated topology (Hillis and Bull 1993; Alfaro, Zoller and Lutzoni 2003; Huelsenbeck and Rannala 2004). Hiilis and Bull (1993) found that bootstrap support values > 70% corresponding with clade probabilities > 95% and therefore branches with lower bootstrap support values should be inferred as uncertain.

2.3 Bayesian Phylogenetic Analysis

2.3.1 The Bayesian Statistical Framework

Bayesian methods of phylogenetic analysis differ from the above methods by directly enabling determination of the uncertainty of all estimands, i.e. the model parameters and the phylogeny. In Bayesian statistics, inferences about any hypothesis or parameter, θ, are made from the posterior probability distribution of θ, which is the conditional probability distribution of θ given the observed data X (in phylogenetics, X may be the aligned matrix of molecular sequences). Thus, rather than searching for the optimal estimate, the Bayesian approach allows

determination of the conditional probability distribution for a given parameter. The posterior probability density of θ, $f(\theta \mid X)$, is given by Bayes' theorem (Bayes 1763) as

$$f(\theta \mid X) = \frac{f(\theta) f(X \mid \theta)}{f(X)}. \tag{2.2}$$

$f(\theta)$ represents the prior (unconditional probability) distribution of θ and is intended to represent our knowledge of θ gained independently of the data. $f(X \mid \theta)$ represents the likelihood function of θ, i.e., the probability of the observed data given θ. The product of $f(X \mid \theta)$ and $f(\theta)$ represents the joint distribution of θ and $X, f(X, \theta)$. Thus, the posterior probability density $f(\theta \mid X)$ is obtained by dividing the joint density $f(X, \theta)$ by the marginal probability of the data, $f(X)$. $f(X)$ acts as a normalising constant to ensure that the posterior density over all θ integrates to 1:

$$f(X) = \int f(\theta) f(X \mid \theta) \, \mathrm{d}\theta. \tag{2.3}$$

Bayes' theorem demonstrates how posterior inferences of θ from the observed data are made by updating our prior knowledge of θ with the corresponding information contained in the data.

Bayesian phylogenetic analysis is performed using the same probabilistic models of evolution as applied in distance matrix and ML phylogenetic methods. These models can accommodate substantial heterogeneity in the evolutionary process and may therefore involve a large number of parameters. Whilst ML methods may encounter difficulties in parameter estimation for complex models, the Bayesian approach provides a better framework for the analysis of data under multi-parameter models as it does not require parameter optimisation. When there is more than one unknown quantity in the model, Bayes' theorem provides the joint posterior distribution of these parameters. For example, Bayesian phylogenetic analysis may comprise of the following parameters: the tree

topology τ, the branch lengths t and various model parameters represented by the vector θ . The joint posterior distribution of these parameters is

$$f(\tau, t, \theta \mid X) = \frac{f(\tau, t, \theta) f(X \mid \tau, t, \theta)}{\int f(\tau, t, \theta) f(X \mid \tau, t, \theta) \, \mathrm{d}\tau \, \mathrm{d}t \, \mathrm{d}\theta} \tag{2.4}$$

The posterior distribution for any single parameter can be obtained from the joint distribution by integrating over all other parameters. For example, the posterior distribution of tree topologies τ can be obtained from $f(\tau, t, \theta \mid X)$ by integrating over all branch lengths t and model parameters θ :

$$f(\tau \mid X) = \int f(\tau, t, \theta \mid X) \, \mathrm{d}t \, \mathrm{d}\theta \tag{2.5}$$

By integrating over parameters that are not of interest in further inference (so-called 'nuisance' parameters), their influence on the parameters of interest is ameliorated. Thus, inferences made from the posterior distribution of topologies are less likely to be influenced by sampling errors affecting the model parameters and branch lengths. The posterior probability of a particular tree topology then tells us the probability that that topology is true conditional on the observed data and the other parameters. The posterior distribution of topologies can be analysed to determine posterior probabilities for specific groupings of taxa, without requiring generation of new data sets as in the bootstrap method.

2.3.2 Computing Bayesian Posterior Probabilities

To perform a Bayesian phylogenetic analysis, Bayes' theorem states that we require prior probabilities for any unknown parameters and a method to compute the likelihood.

2.3.2.1 Computing the Likelihood of an Evolutionary Hypothesis

The likelihood of a tree tells us the probability that the observed data evolved along that tree. Calculation of the likelihood requires a means of calculating the

probability of evolutionary events – nucleotide changes – along the tree. Various probabilistic models can be applied in phylogenetic analysis of molecular sequences to calculate the probability, $p_{ij}(t)$, of change from state i to state j during a period of time of length t.

2.3.2.1.1 Models of Nucleotide Substitution

To model substitution events at the nucleotide level it is assumed that these events occur randomly in time and form a Markov chain such that the probability of nucleotide change from state i to state j, is dependent only on state i, the current state, and not on the history of previous substitutions at that site. Under these assumptions, the substitutional process can be described as a continuous-time Markov process in which the states of the Markov chain are defined by the four nucleotides. For ease of computation, it is also assumed that each site in a sequence evolves independently of other sites, and can therefore be analysed individually. The Markov chain generated by the evolutionary model therefore provides a probabilistic description of the sequence of nucleotide replacements at individual sites.

Various nucleotide substitution models have been proposed; each model is derived from a substitution-rate matrix (**Q**) of instantaneous rates of change between states (i.e., nucleotides). The instantaneous rate of change from any nucleotide i to nucleotide j, q_{ij}, is determined by the equilibrium frequency π_j of nucleotide j and the exchange rate r_{ij} between i and j. The substitution-rate matrix for one such model, the general time-reversible (GTR) (Tavare 1986) model, is

$$\mathbf{Q}^{\mathrm{GTR}} = \{q_{ij}\} = \begin{bmatrix} \cdot & \pi_{\mathrm{C}} a & \pi_{\mathrm{A}} b & \pi_{\mathrm{G}} c \\ \pi_{\mathrm{T}} a & \cdot & \pi_{\mathrm{A}} d & \pi_{\mathrm{G}} e \\ \pi_{\mathrm{T}} b & \pi_{\mathrm{C}} d & \cdot & \pi_{\mathrm{G}} f \\ \pi_{\mathrm{T}} c & \pi_{\mathrm{C}} e & \pi_{\mathrm{A}} f & \cdot \end{bmatrix} \tag{2.6}$$

The GTR model allows the equilibrium frequencies to differ among nucleotides, subject to the constraint that all nucleotide frequencies sum to 1, and allows exchange rates to differ among nucleotide pairs. It does, however, impose the

restriction that substitution events are time-reversible, meaning that the amount of flow from nucleotide state *i* to *j*, $\pi_i q_{ij}$, is equivalent to the amount of flow from *j* to *i*:

$$\pi_i q_{ij} = \pi_j q_{ji}, \qquad \text{for all } i \neq j. \tag{2.7}$$

Under this condition, $r_{ij} = r_{ji}$, and therefore $a = r_{TC} = r_{CT}$, $b = r_{TA} = r_{AT}$, $c = r_{TG} = r_{GT}$, $d = r_{CA} = r_{AC}$, $e = r_{CG} = r_{GC}$, and $f = r_{AG} = r_{GA}$. The total substitution rate for a nucleotide *i* in the **Q** matrix is given by $\sum_{j:j\neq i} q_{ij}$. Each row in the **Q** matrix must sum to 0 and therefore the diagonal elements, which represent the rates of leaving each state, are assigned rates that ensure this condition is met:

$$q_{ii} = -\sum_{j:j\neq i} q_{ij}. \tag{2.8}$$

Other time-reversible substitution models are derived by applying additional restrictions on the GTR model. For instance, the HKY85 model (Hasegawa, Yano and Kishino 1984; Hasegawa, Kishino and Yano 1985) constrains the rate parameters by categorising nucleotide changes as either transitions or transversions. Thus, $r_{TC} = r_{AG} = \alpha$, the transition rate parameter, and $r_{TA} = r_{TG} = r_{CA} = r_{CG} = \beta$, the transversion rate parameter. This model still allows for some heterogeneity in nucleotide exchange rates but reduces the number of free parameters from 9 in the GTR model to 5. The most constrained 4-state model is the JC69 model (Jukes and Cantor 1969) in which there is a single substitution rate for all substitution events and the nucleotide frequencies are all equal.

2.3.2.1.2 Obtaining Probabilities of Nucleotide Change

Following specification of a substitution rate matrix, the probabilities of nucleotide substitutions over a period of time *t* are obtained by taking the matrix exponential of the product of the rate matrix and *t*:

$$\mathbf{P}(t) = \{p_{ij}(t)\} = \mathrm{e}^{\mathbf{Q}t} \tag{2.9}$$

Each element $p_{ij}(t)$ in the transition-probability matrix represents the probability that given the current nucleotide state is i, it will be j a time t later. The substitution rates of the **Q** matrix can be converted into relative rates such that the average substitution rate is 1 per unit time. Since the probability of a nucleotide substitution is dependent on the product of the substitution rate and the time elapsed, this scaling makes the time t equal to the evolutionary distance (measured in units of expected number of substitutions per site). Thus, given a nucleotide substitution model, the probability of nucleotide change along any branch of length t substitutions per site is provided by the transition-probability matrix $P(t)$. To obtain the elements $p_{ij}(t)$, the matrix exponential of $\mathbf{Q}t$ is calculated through diagonalisation of the matrix $\mathbf{Q}t$:

$$\mathrm{e}^{\mathbf{Q}t} = \mathbf{U}\ \mathrm{diag}\{\exp(\lambda_1 t), \exp(\lambda_2 t), \exp(\lambda_3 t), \exp(\lambda_4 t)\}\mathbf{U}^{-1} \tag{2.10}$$

where (λ_1, λ_2, λ_3, λ_4) constitute the eigenvalues of **Q**, whilst the columns of **U** form the right eigenvectors of **Q** and the rows of $\mathbf{U}^{-1}$ form the left eigenvectors of **Q** (Yang 2006: p. 12). For example, the transition-probability matrix under the HKY85 model is

$\mathbf{P}(t) =$

$$\begin{bmatrix}
\pi_T + \frac{\pi_T\pi_R}{\pi_Y}\mathrm{e}^{\lambda_2 t} + \frac{\pi_C}{\pi_Y}\mathrm{e}^{\lambda_4 t} & \pi_C + \frac{\pi_C\pi_R}{\pi_Y}\mathrm{e}^{\lambda_2 t} - \frac{\pi_C}{\pi_Y}\mathrm{e}^{\lambda_4 t} & \pi_A(1-\mathrm{e}^{\lambda_2 t}) & \pi_G(1-\mathrm{e}^{\lambda_2 t}) \\
\pi_T + \frac{\pi_T\pi_R}{\pi_Y}\mathrm{e}^{\lambda_2 t} - \frac{\pi_T}{\pi_Y}\mathrm{e}^{\lambda_4 t} & \pi_C + \frac{\pi_C\pi_R}{\pi_Y}\mathrm{e}^{\lambda_2 t} + \frac{\pi_T}{\pi_Y}\mathrm{e}^{\lambda_4 t} & \pi_A(1-\mathrm{e}^{\lambda_2 t}) & \pi_G(1-\mathrm{e}^{\lambda_2 t}) \\
\pi_T(1-\mathrm{e}^{\lambda_2 t}) & \pi_C(1-\mathrm{e}^{\lambda_2 t}) & \pi_A + \frac{\pi_A\pi_Y}{\pi_R}\mathrm{e}^{\lambda_2 t} + \frac{\pi_G}{\pi_R}\mathrm{e}^{\lambda_3 t} & \pi_G + \frac{\pi_G\pi_Y}{\pi_R}\mathrm{e}^{\lambda_2 t} - \frac{\pi_G}{\pi_R}\mathrm{e}^{\lambda_3 t} \\
\pi_T(1-\mathrm{e}^{\lambda_2 t}) & \pi_C(1-\mathrm{e}^{\lambda_2 t}) & \pi_A + \frac{\pi_A\pi_Y}{\pi_R}\mathrm{e}^{\lambda_2 t} - \frac{\pi_A}{\pi_R}\mathrm{e}^{\lambda_3 t} & \pi_G + \frac{\pi_G\pi_Y}{\pi_R}\mathrm{e}^{\lambda_2 t} + \frac{\pi_A}{\pi_R}\mathrm{e}^{\lambda_3 t}
\end{bmatrix} \tag{2.11}$$

with $\pi_R = \pi_A + \pi_G$, $\pi_Y = \pi_C + \pi_T$, $\lambda_2 = -\beta$, $\lambda_3 = -(\pi_R\alpha + \pi_Y\beta)$, and $\lambda_4 = -(\pi_Y\alpha + \pi_R\beta)$ (reprinted from Yang (2006: equation 1.20)).

Using this matrix, the probability of evolutionary changes along a tree, and hence the likelihood of the tree can be computed. For example, for a simple two taxon tree relating observed sequences A and B to their ancestral sequence V, the likelihood of the tree T relating them defined by a topology τ and branch lengths **t**, and the specified substitution model is given by

$$P(\mathrm{A},\mathrm{B} \mid T,\boldsymbol{\theta}) = \prod_{s=1}^{n} \left(\sum_{x_{\mathrm{V}}^{s} \in \{\mathrm{T},\mathrm{C},\mathrm{A},\mathrm{G}\}} \pi_{x_{\mathrm{V}}^{s}} p_{x_{\mathrm{V}}^{s} x_{\mathrm{A}}^{s}}(t_{\mathrm{A}},\boldsymbol{\theta}) p_{x_{\mathrm{V}}^{s} x_{\mathrm{B}}^{s}}(t_{\mathrm{B}},\boldsymbol{\theta}) \right), \qquad (2.12)$$

where $\boldsymbol{\theta}$ is a vector consisting of the parameters of the specified substitution model, n is the number of sites, and x_{V}^{s}, x_{A}^{s}, and x_{B}^{s} are the nucleotide states at site s in sequences V, A, and B, respectively. The assumption of independent evolution at each site allows us to consider the probability of nucleotide changes at each site independently. In addition, it is assumed that each lineage (branch) evolves independently. For each site s, the probability that the observed states x_{A}^{s} and x_{B}^{s} evolved from x_{V}^{s} along branch lengths t_{A} and t_{B}, respectively, is obtained by taking the product of the probability $\pi_{x_{\mathrm{V}}^{s}}$ of the state in the ancestral sequence and the transition probabilities $p_{x_{\mathrm{V}}^{s} x_{\mathrm{A}}^{s}}(t_{\mathrm{A}},\boldsymbol{\theta})$ and $p_{x_{\mathrm{V}}^{s} x_{\mathrm{B}}^{s}}(t_{\mathrm{B}},\boldsymbol{\theta})$ along each branch. $\pi_{x_{\mathrm{V}}^{s}}$ corresponds to the nucleotide frequencies specified in the **Q** matrix of the chosen evolutionary model. Since the ancestral sequence is not observed, the identity of x_{V}^{s} at each site s is unknown and so, to account for this uncertainty, the likelihood of the tree T is computed by summing over all possible states at each site in the ancestral sequence.

In modelling substitution events as a reversible process, the probability that sequences A and B evolved along T can actually be calculated without reference to the ancestral species. This is because the direction of evolution is irrelevant under a reversible model:

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t) \qquad (2.13)$$

Thus, in equation 2.12,

$$\pi_{x_{\mathrm{V}}^s} p_{x_{\mathrm{V}}^s x_{\mathrm{A}}^s}(t_{\mathrm{A}}, \boldsymbol{\theta}) = \pi_{x_{\mathrm{A}}^s} p_{x_{\mathrm{A}}^s x_{\mathrm{V}}^s}(t_{\mathrm{A}}, \boldsymbol{\theta}). \tag{2.14}$$

Applying this equality, equation 2.12 becomes

$$P(\mathrm{A}, \mathrm{B} \mid T, \boldsymbol{\theta}) = \prod_{s=1}^{n} \left(\sum_{x_{\mathrm{V}}^s \in \{\mathrm{T}, \mathrm{C}, \mathrm{A}, \mathrm{G}\}} \pi_{x_{\mathrm{A}}^s} p_{x_{\mathrm{A}}^s x_{\mathrm{V}}^s}(t_{\mathrm{A}}, \boldsymbol{\theta}) p_{x_{\mathrm{V}}^s x_{\mathrm{B}}^s}(t_{\mathrm{B}}, \boldsymbol{\theta}) \right) \tag{2.15}$$

The probability that sequences A and B evolved along T can therefore be obtained by calculating the probability that sequence B evolved from sequence A via sequence V. Under the reversible model of evolution, this is also equivalent to the probability that sequence A evolved from sequence B via sequence V. The Chapman-Kolmogorov equation for transition probabilities under a Markov process allows the calculation of this probability to be simplified even further as it states that

$$\sum_{\mathrm{V}} \left(P(\mathrm{A}, \mathrm{V} \mid t_{\mathrm{A}}) \times P(\mathrm{B}, \mathrm{V} \mid t_{\mathrm{B}}) \right) = P(\mathrm{A}, \mathrm{B} \mid t_{\mathrm{A}} + t_{\mathrm{B}}) \tag{2.16}$$

Thus,

$$P(\mathrm{A}, \mathrm{B} \mid T, \boldsymbol{\theta}) = \prod_{s=1}^{n} \left(\pi_{x_{\mathrm{A}}^s} p_{x_{\mathrm{A}}^s x_{\mathrm{B}}^s}(t_{\mathrm{A}} + t_{\mathrm{B}}, \boldsymbol{\theta}) \right) \tag{2.17}$$

and the transition probabilities calculated using time-reversible Markov models therefore allow us to account for the unobserved substitutions (e.g. from A to V and V to B) that may have occurred in the evolutionary period between sequences A and B.

For larger trees the likelihood is computed in the same manner as demonstrated for the two-taxon tree: taking the product of the frequency of the state at the root node and the transition probabilities along the branches on the tree. However, for a tree with m taxa, there will be m-1 internal nodes; if the sequences at all of these nodes are unknown then, for each site, there will be 4^{m-1} possible configurations of nucleotide states along the tree and hence the likelihood calculation would have to sum over each of the possible

configurations for each site. Felsenstein (1981) presented an efficient 'pruning' algorithm which substantially reduces the amount of computation required to calculate the likelihood of a tree. This algorithm computes the conditional probability of nodes along the tree at each site *s*. For any node *v* of the tree, the conditional probability $L_{x_v^s}$ is the probability of the states observed at site *s* in the tips descendant from node *v*, given that node *v* has state x_v^s :

$$L_{x_v^s} = \left(\sum_{x_y^s \in \{T, C, A, G\}} p_{x_v^s x_y^s}(t_y, \theta) L_{x_y^s} \right) \left(\sum_{x_z^s \in \{T, C, A, G\}} p_{x_v^s x_z^s}(t_z, \theta) L_{x_z^s} \right) \qquad (2.18)$$

where *y* and *z* are the nodes descendant from node *v*. If *y* (or *z*) is a tip node, $L_{x_y^s}$ is 1 when x_y^s is the observed state and 0 when it is not. If *y* (or *z*) is not a tip node, its conditional probability is computed in the same manner as above. Thus, we can start at the tips and progress down the tree, computing the conditional probability for each internal node, which is then used to compute the conditional probability for its parent node, and so on, until the conditional probability $L_{x_0^s}$ of the root node *v*=0 is obtained. The likelihood of the tree is then

$$L = \prod_{s=1}^{n} \left(\sum_{x_0^s \in \{T, C, A, G\}} \pi_{x_0^s} L_{x_0^s} \right). \qquad (2.19)$$

2.3.2.1.3 Among-Site Rate Variation

The evolutionary model presented above considers all sites to be evolving at the same rate however this assumption may often be violated in reality (Fitch and Margolish 1967; Wakeley 1993; Excoffier and Yang 1999). Within protein-coding genes, some sites may be more functionally constrained than others if the codons they comprise code for structurally and functionally important amino acid residues of the protein. Thus varying selective constraints across sites can result

in differential propensities for substitutions along a gene. The degeneracy of the genetic code itself confers a general pattern of differing constraints on each codon position: the four-fold degenerate third position of most codons undergo more substitution events than the first and second codon positions at which a nucleotide change is likely to result in translation of a different amino acid (Bofkin and Goldman 2007). Substitution rates can also vary among genes and may therefore need to be accounted for if multiple gene sequences are being combined in phylogenetic analysis.

For a simple treatment of rate variation across sites, we can perform a visual inspection of a sequence alignment to identify invariant sites from mutating sites and apply a discrete two-rate class model in which the invariant sites evolve at a rate of zero and the mutating sites evolve at a constant, non-zero rate that may be fixed at a value such that the mean evolutionary rate across sites is 1 (Hasegawa, Kishino and Yano 1985; Palumbi 1989). Additional rate classes can be applied to further distinguish among varying rates of evolution at mutating sites. A model with K rate classes comprises of K parameters specifying the probability p_k of each rate class and K parameters specifying the rate r_k of each class. As with the invariant sites model, the rates r_k are assigned such that the mean rate across sites is 1 i.e., $\sum_k p_k r_k = 1$. The rate r_k assigned to a particular site determines how much the nucleotide exchange rates are increased or decreased for that site and therefore the transition-probability matrix for a site belonging to rate class k is

$$\mathbf{P}(r_k t) = \{p_{ij}(r_k t)\} = \mathrm{e}^{\mathbf{Q} r_k t} \tag{2.20}$$

A three-rate class model can be used to account for different rates of evolution at each codon position: all first-position sites are assumed to evolve at rate r_1, all second-position sites are assumed to evolve at rate r_2, and all third-position sites are assumed to evolve at rate r_3. When *a priori* knowledge of the distribution of sites among the discrete rate classes is unavailable, the likelihood calculation must average over the k rate classes for each site s :

$$\mathrm{L} = \prod_{s} \left(\sum_{k} p_k \times f(x^{(s)} \mid \tau, \mathbf{t}, \boldsymbol{\theta}, r = r_k) \right) \tag{2.21}$$

As this increases the time taken to compute the likelihood by a factor of k, relative to a model with no rate variation among sites, use of the discrete rates model with more than three rate classes is not recommended (Yang 1996).

For each site, the discrete rates model effectively draws the evolutionary rate from one of the K specified classes; a more realistic approach to modelling the rate variation would be to assume a continuous distribution of rates across sites. A variety of distributions have been used for this purpose but the application of a gamma distribution (Nei and Gojobori 1986; Jin and Nei 1990; Yang 1993) has become the standard in phylogenetic analysis. The gamma distribution is parameterised by a shape parameter α and a scale parameter β:

$$g(r; \alpha, \beta) = \frac{\beta^{\alpha} \mathrm{e}^{-\beta r} r^{\alpha-1}}{\Gamma(\alpha)}, \; r > 0 \tag{2.22}$$

Under a gamma distributed model of rate variation, a mean rate of 1 across sites is achieved by setting β equal to (Yang 1993).

In contrast to the discrete rates model which consists of 2K-1 rate parameters (the K rates and K-1 free parameters for the rate-class frequencies), the continuous gamma distribution of rate variation offers a more detailed model of rate variation that is specified by only one parameter: α. Figure 2.1 illustrates the gamma distribution under various values of α. With $\beta = \alpha$, the variance of the gamma distribution is $1/\alpha$ and therefore small values of α (<1) model substantial rate variation among sites with a large proportion of sites evolving at a low rate and a smaller proportion of sites evolving at very high rates. As α increases, the variance in rates decreases such that the majority of sites evolve at similar rates; as $\alpha \to \infty$, the gamma-distributed model approaches a constant rate of evolution across all sites.

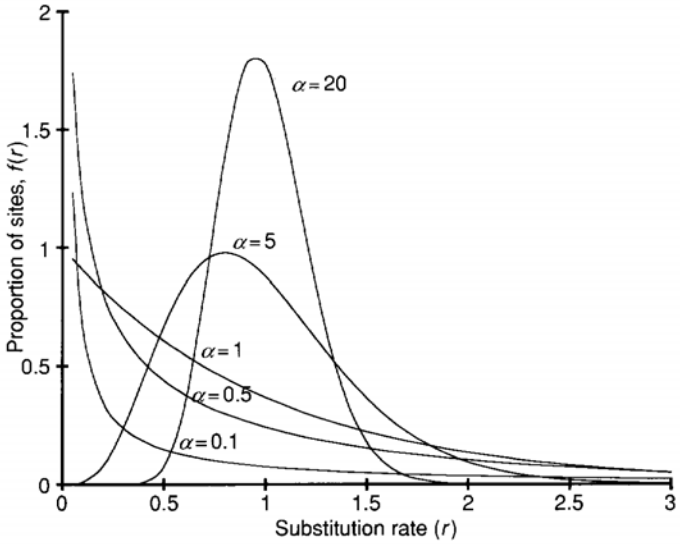


Figure 2.1: Probability density function of the gamma distribution at different values of the shape parameter α. The gamma distribution is used to model rate variation among sites. Reprinted from Yang (1996: Fig. 1).

Given α, the probability of the data at site s is obtained by integrating over the resulting gamma distribution of rates r:

$$f(x^s \mid \tau, \mathbf{t}, \boldsymbol{\theta}, \alpha) = \int_0^\infty g(r) f(x^s \mid r, \mathbf{t}, \boldsymbol{\theta}) dr \tag{2.23}$$

Evaluation of this integral within a reasonable timescale becomes impossible as the size of the data set (number of sequences) increases. For moderate to large-sized data sets, a discretised approach (Yang 1994a), in which the gamma distribution is sectioned into K distinct rate categories of equal density, is used to approximate the continuous gamma model of rate variation. Under a discrete-gamma model, the probability at each site is derived as

$$f(x^s \mid \tau, \mathbf{t}, \boldsymbol{\theta}, \alpha) = \sum_{k=1}^{K} \frac{1}{K} f(x^s \mid r = r_k, \mathbf{t}, \boldsymbol{\theta}) \tag{2.24}$$

A substantial reduction in computational time is obtained by summing over the rates r_K of the K rate categories rather than integrating over the entire gamma distribution. The rate r_k of the k^{th} rate category of the gamma distribution is represented by either the mean or the median rate of that category. Since the rates

r_k are determined by the gamma distribution, the only parameter in the discrete-gamma model is α, making the discrete-gamma approach more appealing than the discrete-rates model for incorporating rate variation among sites. The larger K is, the closer the approximation to the continuous distribution; for most data sets that demonstrate evidence of rate variation, at least 4 rate categories are required to provide a suitable approximation of the continuous distribution (Yang 1994a).

2.3.2.1.4 Rate Variation Across Lineages

When modelling the evolutionary process along a tree it is also necessary to consider the possibility of temporal rate variation. The molecular clock hypothesis (Zuckerkandl and Pauling 1965) states that the rate of evolution remains constant in time; however, as lineages diverge from one another, various differences may arise in the factors that govern the rate of molecular change (e.g. population size, generation time) and consequently molecular evolution may proceed at different rates in divergent species (Thorne, Kishino and Painter 1998). Whilst the molecular clock hypothesis may hold for closely related sequences, for trees involving distantly related species, the evolutionary rate is unlikely to be the same across all lineages. The false assumption of a molecular clock, under which all sequences are assumed to have undergone the same amount of change since their common ancestor, can therefore affect the estimation of tree topologies and node divergence times along the tree.

2.3.2.1.4.1 Testing the Molecular Clock Hypothesis

A common method of evaluating the validity of the molecular clock assumption for a set of sequences is to perform a likelihood-ratio test (LRT) (Felsenstein 1988). An LRT provides a means of evaluating the support, shown by the data, for a particular hypothesis (the null hypothesis) in a likelihood framework. The maximum likelihood of the null hypothesis, given the data, is compared to the maximum likelihood of an alternative hypothesis, given the data. To test the molecular clock assumption, the null hypothesis specifies an

evolutionary model in which all branches of the phylogeny evolve at the same rate. An alternative hypothesis to evolution under a molecular clock is the independent rates hypothesis, where each branch is assigned its own rate. The molecular clock hypothesis represents a special case of the independent rates hypothesis when all branch rates have the same value. The nested nature of the two hypotheses means that the likelihood under the alternative hypothesis cannot be worse than the likelihood under the null hypothesis; thus, the LRT determines whether the increased likelihood under the independent rates hypothesis is statistically significant to justify rejection of the molecular clock hypothesis.

The likelihood-ratio test statistic, $2\Delta\ell$, which is twice the difference of the log likelihoods under the two hypothesis, is approximately distributed according to a χ^2 distribution. The statistical significance of the LRT statistic can be determined by comparison against the χ^2 distribution with degrees of freedom equal to the difference in the number of free parameters between the two models being tested. In the ML framework, phylogenetic estimation of an m taxon tree requires estimation of the model parameters, the topology and the $2m$-3 branch lengths. Under the molecular clock hypothesis, each taxon will have undergone the same amount of evolutionary change from the root and therefore all tip sequences must be equidistant from the root. This constraint reduces the number of branch length parameters requiring estimation under the molecular clock to the m-1 internal node heights. A likelihood-ratio test of the molecular clock hypothesis therefore has $(2m\text{-}3)\text{-}(m\text{-}1) = m\text{-}2$ degrees of freedom.

2.3.2.1.4.2 Incorporating Rate Variation Along a Tree

When there is significant evidence against a constant rate of evolution among lineages, the amount of evolutionary change v_l along each branch l can not be assumed to be proportional to the time duration t_l of each branch but must be determined by accounting for the evolutionary rate of the branch:

$$v_l = r_l t_l \,, \tag{2.25}$$

and therefore the transition-probability matrix for each branch is:

$$\mathbf{P}(v_l) = e^{\mathbf{Q}r_l t_l} \tag{2.26}$$

A number of approaches have been developed to incorporate rate variation in phylogenetic analysis. In ML methods, local clocks (Hasegawa, Kishino and Yano 1989; Rambaut and Bromham 1998) are utilised to partition the tree into distinct regions, each of which evolves at a constant rate. A local clock model allows rate variation along the tree whilst also accounting for the possibility that closely related lineages are likely to evolve at a similar rate; it is therefore more economical than the independent rates model. The non-parametric rate smoothing (NPRS) (Sanderson 1997) and penalised likelihood (PL) (Sanderson 2002) methods specify independent rates for each branch; NPRS attempts to minimise the rate variation among branches by minimising the sum of squared differences in rates between adjacent branches whilst the PL approach specifies an additional parameter λ which determines how much deviations from a molecular clock model penalise the likelihood. In Bayesian methods of phylogenetic analysis, a variable rates model across the tree is used to specify the prior probability distribution of the evolutionary rate. Various stochastic models of rate variation have been proposed, some of which are described later in the chapter.

2.3.2.2 Prior Probabilities

The second component required for calculation of posterior probabilities is the specification of prior probabilities for all model parameters. In a phylogenetic analysis this will entail specifying prior distributions for the tree topology, branch lengths, and parameters of the evolutionary model. In the Bayesian framework, the prior distribution of a parameter θ is intended to represent our prior beliefs about the parameter and provides a means of incorporating knowledge of uncertainty about θ into the inference process. For the phylogenetic parameters, however, there is often little information available to guide us in specifying the most appropriate prior distribution for a data set. As a consequence, many parameters are assigned 'vague' prior distributions, designed to provide an

unbiased distribution over a range of values that is large enough to encompass the true (unknown) values. The application of vague priors in Bayesian analysis is, however, quite controversial: whilst vague priors are not entirely uninformative, the diffuse nature of the specified distribution does not provide any further probabilistic distinction between values within the accepted range thus limiting our ability to fully exploit the power of the Bayesian approach. Attention must also be paid to ensure that a vague prior specified for one parameter does not induce a biased prior on a related parameter. Such a scenario was illustrated by Felsenstein (2004: p. 302) using a uniform (0,5) prior on branch length *t*. Under the JC69 model of sequence evolution, t is related to the probability of nucleotide change at a site by:

$$p = \frac{3}{4} - \frac{3}{4}\exp\{-4t/3\} \qquad (2.27)$$

As this relation is not linear, a uniform prior on *t* does not translate into a uniform prior on *p* - instead the prior on *p* assumes an exponential form. The more diffuse the prior applied to *t* is, the more extreme is the effect on the prior distribution of *p*. Thus, vague, unbiased prior distributions may exert some influence on the estimated posterior distribution. However, if we are using a suitable evolutionary model and the data is informative about the parameters of that model, then the likelihood will dominate the posterior distribution and the prior distribution should have little influence on the conclusions drawn from the resulting posterior (Yang 2006).

2.3.2.2.1 Prior Distributions for Model Parameters

For parameters of the substitution model and additional models of rate heterogeneity, the exact nature of the prior applied is not a major concern as these parameters are generally found to have sharp likelihood profiles and vary little across trees (Yang, Goldman and Friday 1994). Thus, provided the corresponding prior distributions for model parameters assign a non-zero probability to regions of parameter space with high likelihood, they are not found to have a substantial

influence on the posterior. For nucleotide frequencies and substitution rate parameters, Dirichlet priors are commonly applied. When a discrete-gamma model of rate variation across sites is specified, the prior distribution of the parameter α is typically assigned a uniform distribution. An additional parameter *pInv*, modelling the proportion of invariant sites, may be specified as part of the variable rates model and is also assigned a uniform (0,1) prior. If the evolutionary model accounts for gamma-distributed rate variation across sites, then there is no need to specify *pInv*, as the gamma distribution will also account for invariable sites. Prior distributions on other phylogenetic parameters, namely the tree shape and branch lengths, require more consideration as they are capable of exerting greater influence over the resulting posterior distribution.

2.3.2.2.2 Prior Distribution for Tree Topologies

The main goal of a phylogenetic analysis is often to determine how the sequences in the data set are related to one other. This information is obtained from the topology of the tree which depicts the branching patterns of extant and inferred ancestral lineages. The prior probabilities of all possible tree topologies for the *m* sequences of the data set are therefore required for Bayesian phylogenetic estimation. Previous phylogenetic or cladistic studies may provide us with some prior information on the expected phylogeny of the group of organisms under study; however, the probability distribution over the entire tree space is not easily obtained. The set of possible topologies (the 'tree space') expands with the size of the data set and therefore it is more common to represent our ignorance of the prior distribution over tree space by assigning equal probability to all topologies. For a data set consisting of *m* taxa, this will assign a prior probability of $\frac{2^{m-3}(m-3)!}{(2m-5)!}$ to every possible unrooted tree topology. In some cases, the overall topology may be unknown but knowledge of subgroup relationships within the data set may exist. This information can be incorporated into the prior distribution by specifying monphyletic constraints on groups of taxa. These constraints indicate that the corresponding taxa all cluster together in

the phylogeny with probability of 1 causing the resulting prior on topologies to be non-uniform.

2.3.2.2.2 Prior Distributions for Node Times

2.3.2.2.3.1 Priors Generated from an Evolutionary Model

The models of nucleotide substitution can be specified such that the mean evolutionary rate is 1.0 substitution per site, causing the branch lengths to represent the amount of time elapsed between divergence events. Thus, the prior distribution of branch lengths will be specified by the prior distribution of node times. This distribution is typically obtained by modelling the underlying branching process. The Yule pure-birth process (Yule 1925) models lineage speciation events along the tree with a birth rate parameter λ. For each lineage, the probability of a speciation event occuring in the infinitesimal amount of time dt is λdt. The Yule pure birth process evolves lineages along a tree until m lineages are obtained, from which the times of the m-1 speciation events are obtained relative to the time of the root (t_1=1). Each tree generated under this model describes an 'unlabelled history', which comprises of the topology and the order of speciation events but lacks assignments of taxa to the external branches. For each unlabelled history τ generated under the Yule model, the probability density of node times **t** (conditional on a time t_1=1 for the root node and t_m=0 for the final m lineages) is

$$f(\tau, \mathbf{t} \mid \lambda, m, t_1) = \frac{\lambda^{m-2} \exp\{-\lambda \sum_{i=2}^{m-1} t_i\}}{(m-1)(1-\mathrm{e}^{-\lambda})^{m-2}} \qquad (2.28)$$

(Edwards 1970). For each unlabelled history, there are $m!/2^{m-1}$ ways of labelling the external branches to generate distinct labelled histories, and therefore, under the Yule model, the prior probability density of node times for a particular labelled history τ with m lineages at time t=0 is

$$f(\tau, \mathbf{t} \mid \lambda, m, t_1) = \frac{2^{m-1}}{m!} \times \frac{\lambda^{m-2} \exp\{-\lambda \sum_{i=2}^{m-1} t_i\}}{(m-1)(1-\mathrm{e}^{-\lambda})^{m-2}} \tag{2.29}$$

The probability density of node times over all $(m\text{-}1)!m!/2^{m-1}$ labelled histories is then

$$f(\mathbf{t} \mid \lambda, m, t_1) = (m-2)! \times \frac{\lambda^{m-2} \exp\{-\lambda \sum_{i=2}^{m-1} t_i\}}{(1-\mathrm{e}^{-\lambda})^{m-2}}. \tag{2.30}$$

Further detail can be added to the model by accounting for extinction of lineages during the evolutionary process, via a death rate μ, and the fact that phylogenetic reconstruction is usually performed using a sample of lineages from the complete set of extant taxa, via a sampling fraction ρ (Yang and Rannala 1997). When μ=0 and ρ=1, the birth-death process corresponds to a pure birth process and the probability density of node times is that given in equation 2.31.

The pure birth and birth-death models both produce a uniform prior over labelled histories (Edwards 1970; Yang and Rannala 1997); this is not equivalent to a uniform distribution on topologies, however. When m>3, the tree space viewed in terms of labelled histories is larger than that viewed in terms of topologies as multiple labelled histories may be derived from a single tree topology. This can be illustrated for tree topologies with m=4 taxa (Figure 2.2) for which there are 15 possible topologies and 18 possible labelled histories. The 12 asymmetrically-shaped topologies produce 1 labelled history each whilst the 3 symmetrically-shaped topologies each accommodate 2 distinct labelled histories. A uniform distribution on topologies which assigns 1/15 probability to each topology therefore gives greater weighting to the labelled histories associated with asymmetric topologies but less probability to labelled histories associated with symmetric topologies than would a uniform distribution on labelled histories and would therefore inadvertently introduce bias in the relative order of speciation events. Thus, the uniform prior on tree topologies is only suitable when knowledge of the branching pattern, and not speciation times, is desired.

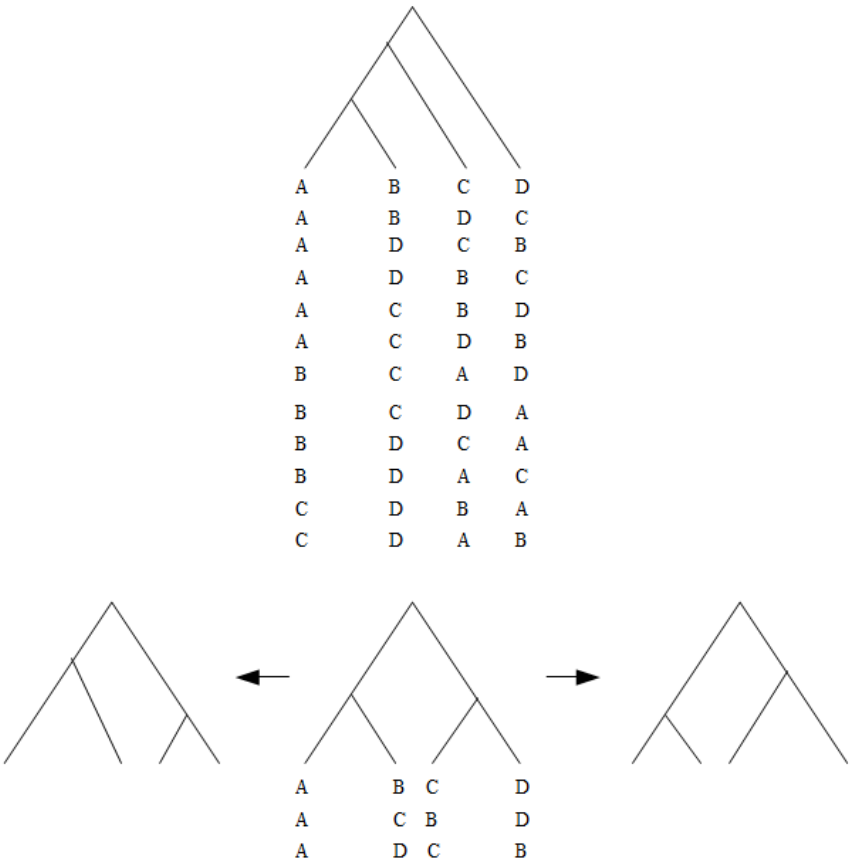


Figure 2.2: The possible topologies and labelled histories for a four-taxon tree. The 15 possible topologies consist of two types of branching patterns: pectinate (top) and symmetrical (bottom). There is only one way of ordering speciation events in the pectinate trees resulting in only one possible labelled history per topology. For the symmetrical topologies, however, the branching events of the two distinct clades occur at different times with respect to one another, resulting in two possible labelled histories per topology. A four-taxon tree therefore has 15 possible rooted topologies but 18 possible labelled histories.

2.3.2.2.3.2 Fossil calibrations

Independent information on the ages of ancestral species, such as that estimated from fossil specimens, can also be employed in phylogenetic analysis to date divergence events along the tree. When the sequences have evolved in a clock-like manner, the specification of at least one fossil date on a phylogenetic tree is sufficient to calibrate the tree to the actual timescale of evolutionary events and estimate the evolutionary rate. Molecular divergence time estimates derived using fossil calibrations may in turn serve as calibration information for clades that have limited representation in the fossil record.

In Bayesian phylogenetic methods, the use of calibration dates is one of the few examples where prior probabilities can be derived from previously obtained information. The calibration priors for the respective nodes are typically specified using statistical distributions, rather than a point mass on each node, to account for errors and uncertainties in the prior estimates (Yang and Rannala 2006). When a fixed tree topology is assumed, the joint distribution of the calibration densities will reflect the influence of the topological constraints. For instance, if overlapping divergence time densities are applied on a pair of adjacent nodes, the prior density on the descendant node will be reduced for times coinciding or preceding those of the parent node and similarly for the density on times of the parent node that overlap with those of the child node.

The joint distribution of the calibration densities must then be incorporated with the uncalibrated prior distribution on all node times obtained using the Yule model, for instance. In BEAST (Drummond and Rambaut 2007), the Bayesian phylogenetic package utilised in this thesis, this is achieved by multiplying the calibration densities with the Yule prior on divergence times to obtain the marginal calibration density. However, the resulting prior density on calibrated nodes can differ from the specified distribution and it is therefore important to compare the nature of the combined prior against the posterior distribution of divergence times at each node to determine the extent of the influence of the prior on the posterior estimates.

2.3.2.2.4 Prior Distributions for the Rate of Molecular Evolution

The evolutionary tree can also be calibrated by specifying a prior distribution on the evolutionary rate. Under a molecular clock, the evolutionary rate can either be fixed to a known value or a statistical distribution can be applied to account for uncertainties in the estimated rate. For sequences demonstrating significant support against a constant rate of evolution, the prior distribution must account for variations in the evolutionary rate. Within the Bayesian framework of phylogenetic analysis, various models have been proposed to obtain a prior density for variable rates across lineages.

The Bayesian inference program MultiDivTime (Thorne and Kishino 2002) models the degree of autocorrelation, represented by the autocorrelation parameter v, in rates between adjacent lineages (Thorne, Kishino and Painter 1998; Kishino, Thorne and Bruno 2001). When v is small, the evolutionary rate will be similar among closely related lineages. When v is large, changes of rate between adjacent nodes will be uncorrelated. The amount of evolutionary change along a branch is then determined by the product of the time duration of the branch and the mean of the rates at the parent and descendant nodes of that branch. To implement this variable-rates model in a Bayesian framework, prior distributions need to be specified for the rate at each node and for the autocorrelation parameter.

A very different approach to rate variation is implemented in MrBayes (Huelsenbeck and Ronquist 2001), where changes in the evolutionary rate along a tree are uncorrelated and are allowed to occur at any point in time along a branch (Huelsenbeck, Larget and Swofford 2000). Given an initial rate or rate distribution at the root node, changes in the rate along the tree are modelled as a Poisson process. At each point i of rate change, a gamma-distributed rate multiplier r, multiplies the rate m prior to i to give a new rate m' after i. The amount of change, v_l, encountered along a branch l is obtained by integrating the rates along l:

$$v_l = \int_{t_{\sigma(l)}}^{t_l} r_l(u)\,\mathrm{d}u\,, \tag{2.31}$$

where $t_{\sigma(l)}$ and t_l are the node ages at each end of branch l. Prior distributions are required for the initial rate m at the root node, the Poisson process parameter λ, which represents the frequency of rate change events, and the rate multiplier r.

Like MrBayes, BEAST (Drummond and Rambaut 2007) also employs uncorrelated models of rate variation; however, changes of rate occur only once per branch. For each branch, the prior distribution of rates can take the form of either an exponential or a log-normal distribution (Drummond et al. 2006). Thus, for branch l the probability density for the branch rate r_l is

$$f(r_l;\lambda) = \lambda \mathrm{e}^{-\lambda r_l} \tag{2.32}$$

under an exponential model, and

$$f(r_l;\mu,\sigma^2) = \frac{1}{r_l\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\ln r_l - \mu)^2}{2\sigma^2}\right) \tag{2.33}$$

under a log-normal model. Depending on the model used, prior densities will be required for the hyperparameters λ or μ and σ. The prior distribution of branch rates is the product of the densities for each branch.

A comparison of several different models of rate variation in Bayesian analysis revealed a high degree of concordance among divergence time estimates under the different models, indicating that the nature of the model is largely inconsequential as long as it allows changes in the evolutionary rate along a tree (Aris-Brosou and Yang 2002).

2.3.2.2.5 Influence of phylogenetic priors

The difficulty of determining suitable priors for phylogenetic parameters means that it is tempting to use the default settings of the selected Bayesian phylogenetic inference program. However, the use of an informative but incorrect prior can have significant consequences on the posterior distribution obtained from the analysis and therefore the specification of prior densities deserves careful attention. A cautionary tale is provided by the effect of the MrBayes default branch length prior on branch length estimates (Brown et al. 2010; Marshall 2010; Rannala, Zhu and Yang 2012).

The default branch length prior applied in MrBayes is an exponential distribution with a mean of 0.1 substitutions/site. This distribution places substantial prior density on long branch lengths, thus implies a prior model in which substantial evolutionary change has occurred in each lineage. For instance, ML and Bayesian phylogenetic estimates derived from various genes of the frog genus *Acris* (Gamble et al. 2008: Fig 5 and 6, respectively) are observed to differ in scale by over an order in magnitude. Repeated analysis of this data set by Brown et al. (2010) under the same conditions as the initial analysis produced a ML tree length of 0.64 substitutions/site whilst the 95% credible interval (CI) of tree lengths from the Bayesian analysis was (0.81, 1.10). The effect of the prior was re-established whe reanalysis of data sets previously analysed under the default branch length prior in MrBayes with a more restrictive branch length prior lead to a shortening of the estimated tree length in all cases (Brown et al. 2010; Marshall 2010).

To remedy the branch length bias in Mr Bayes, Rannala et al. (2012) propose the use of a different prior. A compound Dirichlet prior, which applies a diffuse prior on the tree length (the sum of all the branch lengths in the tree) and a Dirichlet prior on the lengths of all branches in the tree, improves the estimation of realistic branch lengths by MrBayes by regulating branch lengths through the tree length prior (Rannala, Zhu and Yang 2012).

Longer branch length priors can impact the MCMC analysis through various means. An important point to note in analyses run in MrBayes is that the MCMC chain is initialised with a tree with large branch lengths (all branch lengths are

assigned a starting value of 0.1 substitutions per site). Thus, the chain starts off in a region of parameter space which is likely to possess a low likelihood for most real data sets. The application of branch length priors that place substantial weight in this region of parameter space then makes it difficult for the chain to move towards regions of higher likelihood and shorter branch lengths. Difficulties in MCMC mixing and convergence may also ensue due to multiple local peaks in the posterior that arise from the different prior and likelihood distributions (Rannala, Zhu and Yang 2012).

2.3.2.3 Computing the Marginal Probability of the Data

The specification of an evolutionary model from which probabilities of nucleotide change can be derived and prior densities for all parameters allows the posterior probabilities of phylogenies to be derived using Bayes' theorem (equation 2.2). However, calculation of the posterior probabilities requires calculation of the normalising constant $f(\mathbf{X})$. This is the marginal probability of the data and therefore requires integration over all of parameter space, i.e., over all possible tree topologies and, for each topology, integration over all branch lengths and parameter values. As the number of taxa studied grows, the number of tree topologies that must be evaluated to calculate $f(\mathbf{X})$ grows exponentially e.g., for a data set comprising of 6 taxa, 105 unrooted topologies must be considered but the addition of just one taxon expands the tree space to 945 unrooted topologies. In addition, the increase in the number of branches with each additional taxon, increases the dimensionality of the integral that must be evaluated for each topology. Thus, for all but the smallest data sets ($m \leq 5$), evaluation of $f(\mathbf{X})$ is computationally unfeasible within a reasonable timescale.

2.3.3 MCMC Simulation of the Posterior Distribution

2.3.3.1 The MCMC Algorithm

For larger data sets, Bayesian phylogenetic estimation is achieved using Markov chain Monte Carlo (MCMC) sampling algorithms, which provide a means of sampling from the posterior distribution of interest without requiring calculation of $f(\mathbf{X})$. Given the observed data, an evolutionary model and associated prior probabilities, MCMC algorithms allow us to sample a Markov chain whose stationary distribution is the joint posterior distribution of our parameters and therefore inferences from the posterior distribution can be made from the sampled chain. Bayesian phylogenetic analysis using an MCMC algorithm proceeds as follows:

1. Each state in the Markov chain is defined by a particular phylogeny (τ, $\mathbf{v}$) and set of model parameter values ($\boldsymbol{\theta}$). The initial state ($\mathbf{k}_i=\{\tau, \mathbf{v}, \boldsymbol{\theta}\}$) of the Markov chain can be either specified using the results of a previous analysis of the data or obtained by randomly sampling from the prior distributions of each parameter. The likelihood $f(\mathbf{X} \mid \mathbf{k}_i)$ and joint prior probability $f(\mathbf{k}_i)$ is calculated for the initial state.
2. In the next iteration of the algorithm, the Markov chain samples new values for τ, $\mathbf{v}$, and $\boldsymbol{\theta}$. New values are obtained using proposal mechanisms specified for each parameter. The proposed values $\mathbf{k}^*= \{\tau^*, \mathbf{v}^*, \boldsymbol{\theta}^* \}$ are either accepted or rejected depending on the acceptance ratio, α:

$$\alpha = \frac{f(\mathbf{k}^* \mid \mathbf{X}) q(\mathbf{k}_i \mid \mathbf{k}^*)}{f(\mathbf{k}_i \mid \mathbf{X}) q(\mathbf{k}^* \mid \mathbf{k}_i)} \tag{2.34}$$

where $f(\mathbf{k} \mid \mathbf{X})$ is the joint posterior probability of the parameter values in the state and $q(\cdot|\cdot)$ is the proposal density, i.e., the probability of the proposed changes in states. In evaluating the ratio of the posterior

probabilities of each state, the marginal probability of the data, $f(\mathbf{X})$, which forms the denominator in Bayes' theorem and remains constant over all states, cancels out and therefore calculation of α simply requires computation of the ratios of the prior densities, the likelihood, and the proposal densities under the proposed and current states:

$$\alpha = \frac{f(\mathbf{k}^*)f(\mathbf{X}\,|\,\mathbf{k}^*)q(\mathbf{k}_i\,|\,\mathbf{k}^*)}{f(\mathbf{k}_i)f(\mathbf{X}\,|\,\mathbf{k}_i)q(\mathbf{k}^*\,|\,\mathbf{k}_i)} \qquad (2.35)$$

If $\alpha \geq 1$, the proposed values are accepted and $\mathbf{k}_{i+1} = \mathbf{k}^*$. If $\alpha < 1$, the proposals are accepted with probability α, i.e., a random number, r, is chosen from a uniform distribution (U(0,1)) and if $\alpha > r$ the proposals are accepted, otherwise they are rejected and $\mathbf{k}_{i+1} = \mathbf{k}_i$.

The likelihood and joint prior probability for state $\mathbf{k}_{i+1}$ is calculated.

3. Step 2 is repeated for a large number of iterations (typically $>10^5$, depending on the complexity of the model and the size of the data set) to allow sufficient sampling from the posterior distribution. In each iteration, the acceptance ratio specifies the probability of the proposed changes to phylogenetic and model parameters being accepted.

2.3.3.2 MCMC Proposal Mechanisms

The proposal mechanisms used to move between states in parameter space form an essential component of MCMC algorithms. A simulation will typically begin at a random point in parameter space and it is the proposal densities and evaluation of the acceptance ratio at each stage of the simulation that guides the chain of sampled states to converge on the posterior distribution. Three necessary conditions to ensure the posterior distribution is achieved are that the proposal mechanisms must allow random sampling, they must produce a Markov chain that is aperiodic and must produce a chain which allows all states to be reached from any other state, i.e., proposal densities between any two states *i* and *j* must be non-zero. Under the Metropolis algorithm (Metropolis et al. 1953) of MCMC

sampling all proposal densities are symmetric, i.e., the proposal densities between states i and j are the same in either direction, thus the proposal ratio is not a component of the acceptance ratio. Other MCMC algorithms e.g. the Metropolis-Hastings method, correct for asymmetric proposal densities via the proposal ratio (a.k.a. the Hastings ratio) (Hastings 1970). The proposal ratio computes the ratio of proposal densities in the reverse direction, thus correcting for biases in the proposal densities. For instance, say the change of state from i to j is twice as likely as a change in the reverse direction, the ratio of $q(i|j)/q(j|i)$ will only allow these proposals to be accepted with probability 0.5. This correction ensures that the sampled Markov chain (and hence inferences of the posterior distribution) is not affected by proposal biases.

Different proposal mechanisms, or operators, are employed to assist the Markov chain move through parameter space. The type of operator applied depends on the nature of the parameter. To propose new values for numerical parameters either a sliding-window mechanism or a scale factor mechanism is used. Scale factor operators generate a random multiplier to either decrease or increase the current value of a parameter. Under a sliding-window mechanism, a new parameter state is selected from a distribution centred on the current parameter state. The size of the window, a pre-specified constant, determines the potential size of the jump in parameter space made by the Markov chain. Sliding-window proposals can be made by specifying either a uniform or normal distribution, with the window size determined by the width or variance, respectively. Large jumps can facilitate greater exploration of the parameter space; however, if most of the posterior density of a parameter is concentrated in a small region of parameter space then many of the proposed changes will be rejected and the chain will spend a number of iterations stuck in the same state. Conversely, small jumps restrict the chain to only small movements in parameter space and will therefore require a much larger number of iterations to effectively sample the posterior distribution. Thus the window size can greatly affect the extent of mixing in the Markov chain.

For the proposal of new tree topologies, operators based on the topological rearrangement algorithms commonly applied in heuristic tree searching methods, i.e., NNI, TBR, and SPR, can be used. The TBR algorithm takes larger jumps in tree space than SPR, which in turn is able to take larger jumps in tree space than NNI. Local changes to branch lengths within a tree can be performed by sliding a randomly selected node a distance up or down a path running from the root to one of the nodes' descendant tips. If the node traverses either the parent or one or more descendant nodes lying on that path, a topological change will also occur.

For trees free of the molecular clock constraint, under which all tips must be equidistant from the root, Larget and Simon (1999) proposed a combined branch length and topological operator for making local tree changes. For any randomly selected internal branch, two of its four adjacent branches (one from either side) are randomly selected and the lengths of these three selected adjacent branches are altered by a randomly selected multiplier between 0 and 1. Following local branch length modification, one of the two nodes joined to the middle branch is selected for translocation (along with its subtending subtree) to a new position along the path described by the three branches. Topological changes arise if the size of the translocation is greater than the branch length. In an MCMC simulation, each iteration may consist of multiple topological moves, and several topological operators may be employed to ensure adequate sampling of tree space is achieved.

2.3.3.3 Determining Convergence of an MCMC Simulation

The MCMC algorithms produce a Markov chain of sampled parameter states in which each state is dependent only on its preceding state. At the start of the simulation the chain will randomly sample states from parameter space, however, as the simulation progresses the acceptance ratio will cause the chain to converge to a stationary distribution. Convergence can be monitored by observing the simulated sequence of states for each parameter. In the early stages of the simulation, known as the 'burn-in period' the chain may sample from widely different areas of parameter space. However, when the chain has converged it will appear localised to a particular region of parameter space and will

predominantly sample states from within that region. If the algorithm has been appropriately set up and run for a sufficient number of iterations, this stationary distribution will correspond to the target distribution, i.e., joint posterior distribution of the parameters.

In Bayesian phylogenetics, the nature of this distribution is not known due to the difficulties encountered in deriving the marginal probability of the data $f(\mathbf{X})$. Thus, it is not possible to determine if the simulated chain has converged on the true posterior distribution. For large, multi-parameter models, the posterior distribution is likely to be multi-modal and it is therefore possible for a chain to converge on a distribution that is localised in one high density region of parameter space rather than the complete posterior distribution. When the true posterior distribution is unknown, a comparison of the stationary distributions of multiple, independent Markov chains generated from different starting points provides the best means of assessing whether the target distribution has been achieved. One measure used to statistically determine the degree of consistency among independent chains is the potential scale reduction statistic $\hat{R}$, which estimates "the factor by which the scale of the current distribution for [an estimand] x might be reduced if the simulations were continued in the limit $n\rightarrow\infty$" (Gelman and Rubin 1992). This is achieved by comparing the variance, τ^2, of the target distribution represented by all the chains with the variance, W, of the distributions represented by each individual chain (i.e., the within-chain variance):

$$\hat{R} = \sqrt{\frac{\tau^2}{W}} \tag{2.36}$$

For m chains, each of length n (following removal of the burn-in states), τ^2 measures the variance of estimand x in the mn sampled states and is determined from the weighted average of the within-chain variance, W, and the between chain variance, B (Yang 2006: p. 173):

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{x}_{.j} - \bar{x}_{..})^2 , \tag{2.37}$$

where $\bar{x}_{.j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, and $\bar{x}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{x}_{.j}$

$$W = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{.j})^2\right) \tag{2.38}$$

$$\tau^2 = \frac{n-1}{n}W + \frac{1}{n}B \tag{2.39}$$

When chains have converged on the true posterior distribution, the within-chain variance will be similar to the variance of the target distribution and therefore $\hat{R} \approx 1$. If the chains have not converged on the posterior distribution, the variance of the target distribution (i.e., all sampled states) will be greater than the within-chain variance and therefore $\hat{R} > 1$.

If there are indications that the chain has not converged on the posterior distribution, either the simulation can be run for longer or the proposal mechanisms can be modified to improve mixing among regions of parameter space. For more complex distributions, a Metropolis-coupled MCMC (MC^3) algorithm (Geyer 1991) can be run in which heated chains are run alongside the standard ('cold') chain. The MCMC algorithms for the heated chains are run in the same manner as the cold chain however the target densities for the heated chains are modified so as to flatten the posterior distribution and thereby facilitate mixing between different regions of parameter space. In each iteration an additional proposal mechanism is employed to propose a swap of states between a hot chain *j* and the cold chain *i*, enabling movement to a state that may have been rejected in the cold simulation. Inferences are then made from the posterior distribution sampled by the cold chain.

2.3.4 Deriving Phylogenetic Inferences from the Posterior Distribution

For chains demonstrating convergence on a single distribution it is expected that all states after the burn-in period are sampled in proportion to their posterior probabilities and therefore provide an approximation of the joint posterior distribution. The posterior distribution of any particular parameter, for instance, θ_1 from $\boldsymbol{\theta} = \{ \theta_1, \theta_2, \ldots, \theta_n \}$, can be obtained from the joint posterior distribution by integrating over the remaining parameters:

$$f(\theta_1 \mid \mathbf{X}) = \int f(\tau, \mathbf{v}, \boldsymbol{\theta} \mid \mathbf{X}) \, \mathrm{d}\tau \, \mathrm{d}\mathbf{v} \, \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_n \qquad (2.40)$$

The joint posterior distribution is represented by the distribution of sampled states from the MCMC simulation, and therefore the marginal distribution for θ_1 is readily obtained from the Markov chain by summing over the remaining parameters. Inferences regarding are then based on analysis of the resulting posterior distribution which specifies the probability distribution of θ_1 given the other parameters and the observed data.

To make phylogenetic inferences from the posterior distribution, one can either extract the tree topology with the highest posterior probability, i.e., the maximum *a posteriori* (MAP) tree, or obtain a majority-rule consensus tree that summarises the entire sample of tree topologies in the chain. In both the MAP and consensus trees, branch lengths represent the mean values from the posterior distribution and each internal (i.e., non-terminal) branch is associated with a posterior probability that represents the proportion of sampled states in which the partitioning of taxa conferred by that branch is observed. The posterior probabilities for each branch are referred to as posterior clade probabilities and tell us the probability that the clade formed by the group of taxa subtending from that branch is true given the data and the evolutionary model. The posterior clade probabilities therefore provide an indication of the accuracy of phylogenetic groupings that the bootstrap support values, determined when using other methods of phylogenetic estimation, are unable to provide.

A concern with Bayesian methods of phylogenetic inference is that they may produce inflated posterior clade probabilities that can lead to false conclusions of certainty in the estimate topology (Yang and Rannala 2012). A simulation-based example of this is observed in the 'star-tree paradox' (Lewis, Holder and Holsinger 2005; Yang and Rannala 2005; Yang 2007) where Bayesian phylogenetic analysis converges on a single optimal bifurcating topology for data simulated along a four-taxon star-tree (i.e., a tree possessing only one internal node), even as the amount of data increases to infinity. In this example, the inflated posterior probabilities were attributed to the specification of an inappropriate internal branch length prior which biased the posterior distribution towards trees with a long internal branch (Yang 2007). This example again illustrates the influence of inappropriate priors on a Bayesian analysis. However, the prior distribution is not the only factor affecting the posterior.

Simulation studies (Buckley 2002; Huelsenbeck and Rannala 2004; Lemmon and Moriarty 2004) invegstigating the reliability of clade posterior probabilities highlight the effect of the evolutionary model on posterior probabilities. These studies all found that an under-specificied evolutionary model, i.e., one that accounts for less heterogeneity among evolutionary processes, results in over-estimated posterior probabilities on the resulting phylogenetic estimates. Over-specified models had less impact on the posterior probabilities, i.e., the posterior probability was close to the probability that the tree was correct (given the evolutionary model). The authors therefore recommend the complex evolutionary models (without over-parameterising) for Bayesian phylogenetic analysis. However, as even the most complex evolutionary models available fail to account for all aspects of the molecular evolutionary process, the influence of the evolutionary model on the posterior distribution for a real data set is uncertain.

Variability in the posterior distribution of an estimand is typically presented using credible intervals. A 95% credible interval indicates the range of estimated values (or trees) which has 0.95 probability of containing the true value. As the posterior probability of any single tree is the probability that it is the true tree (given the observed data, the evolutionary model, and the prior probabilities), the 95% credible interval of trees can be constructed by taking the smallest set of

trees producing a total posterior probability of 0.95. However, unlike numerical parameters, the credible interval of phylogenies can be difficult to represent. If the interval consists of a few trees, the variations can be viewed individually but for large trees, where the number of possible topologies is large, the credible interval may also comprise of too many topologies and therefore uncertainties in an estimated topology will be inferred from the posterior clade probabilities.

2.4 Summary

The Bayesian approach provides a more flexible method of phylogenetic analysis than distance-based, MP and ML methods. Complex evolutionary models can be incorporated with ease enabling a more accurate consideration of the process of molecular evolution, the estimation process is enhanced through the use of prior information regarding parameters, and uncertainties in parameter estimates are accounted for in the resulting posterior probability distribution. However, to obtain meaningful results from Bayesian phylogenetic analysis it is necessary incorporate this complexity wisely. In particular, evolutionary models and prior distributions for the model parameters should be investigated for their effect on the posterior distribution before inferences are made. Ensuring convergence and efficient mixing of the MCMC chains is also vital in determining that the posterior distribution has been sampled.

Chapter 3

Evaluating Phylogenetic Incongruence Among PV Genes

3.1 Introduction

This chapter focuses on determining the extent of phylogenetic compatibility among the PV genes in my data set. Topological differences among phylogenies derived from different PV genes have been reported (Bravo and Alonso 2004; Garcia-Vallve, Alonso and Bravo 2005; Narechania et al. 2005; Gottschling et al. 2007b) but it is important to determine whether the observed disparities are simply due to sampling errors, in which case a combined analysis of all genes together should provide a more accurate estimate of the evolutionary history of the taxa, or if the differences reveal real differences in evolutionary histories. A statistical evaluation of the observed phylogenetic differences among PV genes is therefore required before any further evolutionary analyses can be performed.

3.1.1 Hypotheses for evaluating phylogenetic incongruence

When different data partitions from the same set of taxa present phylogenies with conflicting topologies then, assuming confidence in the phylogenetic estimation method, the topological incongruities among the partitions can be evaluated from two perspectives. The differences can be evaluated under the null hypothesis that the partitions are phylogenetically independent (Lapointe and Legendre 1990; Lapointe and Legendre 1992; Miyamoto and Fitch 1995; Campbell, Legendre and Lapointe 2011), thereby testing for significant

correlations in the phylogenies derived from each partition. Alternatively, one may propose that the observed differences are due to sampling errors, i.e., errors in phylogenetic estimation that result from the use of an incomplete set of data. By evaluating the null hypothesis that the data partitions share a single phylogenetic structure (Rodrigo et al. 1993; Farris et al. 1994; Huelsenbeck and Bull 1996), significant support for phylogenetic incongruence among the partitions can be identified.

3.1.1.1 Testing for phylogenetic congruence

The null hypothesis of phylogenetic independence among partitions of a group of taxa is evaluated by measuring the degree of topological similarity among the trees estimated from each partition. This implies that we can be confident in the estimated phylogenies, which may not always be correct. Topological similarity can be measured using a tree metric such as the partition distance (Robinson and Foulds 1981; Penny and Hendy 1985), which calculates the number of partitions (i.e. splits of taxa) that are not shared by all trees. Measures of the degree of concordance among distance matrices derived from the estimated phylogenies have also been used in tests of phylogenetic congruence (Legendre and Lapointe 2004; Campbell, Legendre and Lapointe 2011).

The value of the chosen congruence-test statistic estimated from the real data is then compared to a distribution of such values obtained for data constructed under the null hypothesis of phylogenetic independence (the null distribution). Thus, if the test statistic falls in the upper tail of the null distribution (the top 5% is a commonly-used threshold), the null hypothesis is rejected. A rejection of the null hypothesis of phylogenetic independence signifies that, despite observed differences, the trees are more similar than expected by chance. The observed differences can be attributed to random error and further phylogenetic analysis of the data can be performed using a 'total-evidence' (Kluge 1989) approach (i.e., all partitions can be analysed together).

3.1.1.2 Testing for phylogenetic incongruence

Examples of statistical tests examining the null hypothesis of phylogenetic congruence include the incongruence length difference (ILD) test (Farris et al. 1994) and the likelihood heterogeneity test (LHT, Huelsenbeck and Bull 1996). These tests evaluate whether the individual partition phylogenies provide a significantly better 'fit' to the data than a single ('total evidence') phylogeny estimated when the partitions are analysed together. A rejection of the null hypothesis indicates significant evidence against a shared phylogeny for all partitions and therefore warns against the combination of data partitions in a total evidence phylogeny. A failure to reject the null hypothesis suggests that the topological differences observed among individual partition phylogenies can be attributed to sampling error.

The main topic of consideration in thesis is the characterisation of the processes causing phylogenetic incongruities between the *Papillomaviridae* family and their vertebrate hosts. This characterisation will be compromised if it is based on a total-evidence PV phylogeny estimated from genes that have conflicting evolutionary histories. Thus, I have chosen to evaluate the null hypothesis of phylogenetic congruence among the genes rather than evaluate the hypothesis of phylogenetic independence. Only genes that do not demonstrate significant evidence against the null hypothesis of phylogenetic congruence will be combined in further phylogenetic analysis. This follows the 'conditional data combination' approach (Bull et al. 1993; De Queiroz 1993; Miyamoto and Fitch 1995; Huelsenbeck, Bull and Cunningham 1996).

3.1.2 Tests of phylogenetic incongruence

3.1.2.1 The Incongruence Length Difference Test

The ILD test (Farris et al. 1994) tests the null hypothesis that distinct data partitions, e.g., X and Y, are phylogenetically compatible in a parsimony framework by evaluating the difference between length of the total-evidence phylogeny (L_{X+Y}) and the total length of the phylogenies estimated for each individual partition ($L_X + L_Y$):

$$\delta_{XY} = L_{X+Y} - (L_X + L_Y) \tag{3.1}$$

When distinct partitions for a set of taxa have different evolutionary histories, the length of the estimated total-evidence MP tree will be much greater than the sum of tree lengths for the individual partitions as the method will likely struggle to find a total-evidence tree topology that fits all the partitions equally well. As a result, for some sites the estimated total-evidence tree will propose a greater number of character changes than the optimal tree for those sites and hence, the more phylogenetic discordance there is between partitions, the greater the value of δ will be.

The significance of the calculated ILD for the observed sequences is ascertained by non-parametric bootstrapping of the observed data partitions i.e. randomly repartitioning sites in the total-evidence matrix into new partitions of the same size as X and Y, and performing the ILD test on the repartitioned data. This is repeated at least 100 times to provide a distribution of permuted δ values against which δ_{XY} for the real data can be evaluated. If δ_{XY} falls in the top 5% of the distribution then X and Y are taken to demonstrate significant phylogenetic incongruence.

As the ILD test is based on MP phylogenetic estimation, it will suffer the limitations of this method, most notably the difficulty in distinguishing between homology and homoplasy. In not accounting for the possibility of multiple substitution events, MP methods bias against homoplasious changes and

therefore may underestimate the amount of evolution that has occurred. In addition, simulation studies (Dolphin et al. 2000; Barker and Lutzoni 2002; Darlu and Lecointre 2002) have shown that when the amount of homoplasy is different among the partitions under investigation, the performance of the ILD test is affected in two ways. First, as the amount of evolutionary change apportioned to homoplasy is minimised in parsimony estimation, the tree lengths of partitions with more homoplasious characters will be underestimated, resulting in a larger ILD value. Second, when bootstrapping sites from the total data matrix, the homoplasious character sites will be spread out among the bootstrapped partitions resulting in larger tree lengths for the individual perturbed partitions and hence, a smaller ILD for most replicates. Thus, the ILD test is associated with a high type-I error rate resulting in the false rejection of the null hypothesis of congruence among partitions.

3.1.2.2 The Likelihood Heterogeneity Test

The LHT (Huelsenbeck and Bull 1996) examine phylogenetic compatibility among data partitions in a likelihood framework; the test statistic in the LHT is the difference in the ML of the total-evidence topology and the total MLs of each tree estimated for the individual partitions:

$$\delta = LL_{X+Y} - (LL_X + LL_Y) \qquad (3.2)$$

Unlike MP, where the fit of the data to the tree is measured by the number of estimated substitutions, in ML methods, the log-likelihood is a function of the parameters specified in the evolutionary model. The parameter values may therefore also influence the analysis of phylogenetic compatibility among data partitions. A plausible scenario is that different data partitions all support the same tree topology but have different evolutionary rates. ML methods allow us to decouple topological incongruence from "process" incongruence by optimising parameter values for individual partitions even under the topological constraints.

The significance of the estimated δ (δ_{obs}) can be determined by estimating the null distribution of δ i.e., the distribution of δ under the null hypothesis of phylogenetic congruence between partitions, using parametric bootstrapping (aka Markov/Monte Carlo simulation, Goldman 1993). Whereas non-parametric bootstrapping generates new data sets by resampling from the observed data matrix (as observed in the ILD test); parametric bootstrapping uses the model parameters to generate new data. So the maximum likelihood evolutionary model (including the tree topology) estimated for the combined data partitions are used to simulate the evolution of new sequences under the null model. This allows us to determine the extent of stochastic variation in δ estimates when the data partitions are phylogenetically congruent and assess whether δ_{obs} falls within this range, thereby implying phylogenetic compatibility of the data partitions, or outside this range, thereby implying phylogenetic incongruence of the partitions.

3.1.3 Previous studies of phylogenetic incongruence among PV genes

To date, there have been two studies that have explicitly tested phylogenetic incongruity among PV genes; these were each performed on different data sets and using different methods. Narechania et al. (2005) assessed the significance of the topological incongruities observed between trees inferred for the early genes and the late genes of the α HPVs using a localised ILD (LILD) test (Thornton and DeSalle 2000). Whereas the ILD test allows the identification of significantly incongruent partitions, the localised ILD test identifies significant incongruence at specific phylogenetic nodes of a given partition. In contrast, Gottschling et al. (2007b) analysed a more diverse set of PV types, comprising 18 different PV genera. Significant phylogenetic incongruence among the E1, E2, L1 and L2 genes of this data set was examined by performing the ILD test in a pairwise manner (implemented as the partition homogeneity test (PHT) in the PAUP suite of phylogenetic software (Swofford 1998)).

3.1.3.1 Phylogenetic incongruence among genes of the α HPVs

The LILD test employed by Narechania et al. (2005) to study phylogenetic incongruence among the α HPVs evaluates the tree length difference between an MP gene tree estimated under the constraint of a particular node and the MP gene tree obtained in the absence of any such constraint. The nodes used to constrain individual gene trees are taken in turn from a tree topology presumed to relate the taxa. This test therefore aims to identify tree nodes causing significant incongruence between a gene trees and the overall phylogeny. It is therefore useful for identifying sequences to remove from the data set when there is specific interest in performing a combined phylogenetic analysis of multiple genes.

In the absence of a PV topology derived from independent data, Narechania and colleagues estimated the total-evidence phylogeny from a concatenated data set of the E1, E2, E6, E7, L1, and L2 genes and proteins. The significance of the tree-length difference obtained for each total-evidence node-gene pair was determined by evaluation against a null distribution of tree length differences obtained by non-parametric bootstrapping of the concatenated data matrix to allow identification of nodes in the total evidence phylogeny which are significantly incongruent with single gene phylogenies.

The total-evidence phylogeny, estimated using Bayesian phylogenetic methods, possessed well-supported (p>0.99) monophyletic clades of PV types grouped within the same PV species classification, and monophyletic grouping of the high-risk PV species 9, 11, 7, 5, and 6. The parent node of this high-risk clade in these trees is referred to by the authors as the "oncogenic node". The LILDs obtained at the oncogenic node for both the L1 and L2 genes was found to be statistically significant ($P \leq 0.01$), strongly suggesting that the existence of a single oncogenic node in the evolutionary history of the α HPVs is not supported by either of the late genes.

In total, 12 significant LILDs were found; these corresponded to various node-gene pairs, with all but two of these significant incongruities identified in the late genes. Four of the 10 significant LILDs in the late genes were associated with nodes in the high-risk clade, whilst the remainder were located within a

clade comprised of low risk mucosal and cutaneous PV species (α–4, α–15, α–3, and α–2). A significant LILD in the E7 gene was observed at the basal node for the clade of the PV species 4, 15, 3, and 2. Significant LILDs in the early genes were observed at the basal node for the clade of the PV species 4, 15, 3, and 2 in the E7 gene and within the high-risk clade of either the E1 or E6 gene (Narechania et al. (2005) indicate significant LILD at node 12 in the E1 gene in Fig. 5 but report significant LILD at node 6 in the E6 gene in the text).

A potential source of error in the analysis of Narechania et al. (2005), is the use of the total-evidence phylogeny to represent the PV phylogeny since the authors are making the *a priori* assumption that the genes share the same evolutionary history. This phylogeny is then used to examine incongruities with each gene tree. If genuine phylogenetic incongruence does exist amongst the genes, the conflicting phylogenetic signals may produce a total-evidence tree topology that fails to reflect any of the gene histories; if this is the case the LILD test will be investigating incongruence at nodes which never occurred in the evolutionary history of the PVs.

In this particular study however, the results may still be of some significance as, for the α PVs at least, it is observed that there is substantial resemblance between the topologies of the total-evidence tree and the early gene phylogenies, particularly with respect to the arrangement of the high-risk HPVs. This is likely a consequence of the greater proportion of sites from the early genes E1, E2, E6, and E7 in the total-evidence matrix. Based on the early gene topology of the total-evidence tree and the findings that the majority (10/12) of significant incongruities with the total-evidence tree occurred for the late genes, their results certainly highlight significant incongruities between the evolutionary histories of the late genes and the early genes of mucosal α HPVs. Additional pairwise similarity scans of the α HPVs revealed a distinction in similarities to high-risk and low-risk α PVs for the E6 and L2 ORFs, which may suggest that the early gene-late gene phylogenetic incongruities of the high-risk α PVs may be driven by changes in these two ORFs.

3.1.3.2 Phylogenetic incongruence among the genes of multi-genera PVs

Gottschling et al. (2007b) applied the PHT (i.e., the ILD test performed on pairs of genes only) on the E1, E2, L1 and L2 genes of 53 PV types from 18 different genera to determine which PV genes may be combined in phylogenetic analysis. Significant phylogenetic heterogeneity ($P \leq 0.001$) was determined for the E1-L2, E1-L1, and L1-L2 paired gene (first and second codon positions only) partitions. Analysis of the respective protein sequences, however, found significant incongruence only in partition pairs involving the L2 protein and of these three partition pairs only the E1-L2 partition had $P \leq 0.001$. Once again, L2 was identified as a source of phylogenetic incongruence among the PV genes though only 35% of the sites from the full L2 alignment were used in the test (phylogenies for the other genes were derived from more than 75% of sites from the original alignment).

Differences in the results obtained for the nucleotide and amino acid sequences may indicate the effect of differences in the amount of data analysed using each data type. Tests of phylogenetic congruity performed on the nucleotide sequences using the first and second positions of each codon examined twice the number of sites than were available in the amino acid sequences. However, the amino acid sequences provide more character states (20 amino acids vs. 4 nucleotides) and may therefore provide greater phylogenetic resolution, despite the availability of less sites.

PV gene trees generated for the data set studied revealed certain taxa (HPV-16, HPV-1, HPV-63, and PlPV) which assume different phylogenetic position in the individual gene trees; removal of these taxa was found to remove any source of significant incongruence among the PV genes, including L2. This may either indicate the possibility of recombination involving each of the removed sequences or simply that insufficient sampling has resulted in an inability to resolve the phylogenetic positions of these taxa.

The overall finding of this study of phylogenetic incongruence among PV genes was that the E1, E2, and L1 proteins can be combined in phylogenetic analyses. However, as the tests of phylogenetic heterogeneity were performed on a reduced data set in which PV-genus species possessing multiple HPV types

were each represented by a single HPV type (e.g. HPV type 6 was used to represent HPV types from the α-10 PV species), the conclusions made in this study may not hold for expanded data sets of PV sequences. For instance, the results obtained by Narechania et al. (2005) suggest that the inclusion of more high-risk and low-risk PV types from the α HPVs would render the L1 aa tree topology significantly incongruent with the topologies of the E1 and E2 aa trees. Thus, when performing phylogenetic analysis of a PV data set, it is necessary to first evaluate the evidence for phylogenetic incongruence among the genes or proteins of the specific data set under analysis, if this has not been performed previously.

I chose to investigate phylogenetic incongruence among PV genes using a test based on the methods of the LHT (Huelsenbeck and Bull 1996) but implemented in a Bayesian framework, as suggested by Nylander et al. (2004). This approach was also used by Stevenson et al. (2007) to evaluate observed phylogenetic incongruities among genes in members of the spirochete species *Leptospira interrogans*. Bayesian methods are preferable to an ML approach as they can account for any phylogenetic and model uncertainties. This allows us to ensure that the tests of phylogenetic incongruence are not influenced by a single incorrect phylogeny.

3.2 Method

3.2.1 The PV data set

Amino acid and nucleotide sequences of the PV genes E1, E2, E6, E7, L1 and L2 were obtained from Genbank (Benson et al. 2005). The data set consisted of the nucleotide sequences of 108 PV types from 14 different genera (Appendix A.1). The PV types RaPV1, MnPV1, TmPV1, EcPV1, EcPV2, CPV3, and ChPV1 were initially included in the data set but demonstrated variable phylogenetic positions and thus have been omitted from further analyses. The

genes E1, E2, L1 and L2 are present in all 108 PVs, while 6 PV types (PePV1, PsPV1, TtPV2, BPV3, BPV9, and BPV10) lack either an E6 or E7 ORF and therefore incongruence tests involving the transforming genes were performed using a data set of 102 PV sequences. The protein sequences were aligned individually using Muscle (Edgar 2004). Nucleotide alignments were then constructed from the amino acid alignments using Pal2nal (Suyama, Torrents and Bork 2006). Gapped positions in the resulting nucleotide alignments were removed resulting in alignment lengths of 1389, 480, 300, 192, 1266, and 681 sites for the E1, E2, E6, E7, L1, and L2 genes, respectively.

3.2.2 Testing the Molecular Clock

Likelihood ratio tests were performed first to evaluate support for a constant rate of evolution in each gene. Tests of the molecular clock hypothesis were performed by estimating ML phylogenies for each gene under models of clock-like and non clock-like evolution using the PAML phylogenetic analysis software (Yang 1997). In each case, a HKY + Γ(4) (Hasegawa, Yano and Kishino 1984; Yang 1994a) model was specified to model nucleotide changes. The significance of the likelihood ratio test statistic, which is twice the difference of the likelihoods under each model, was determined by comparison against a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters between the clock-like and non clock-like models.

3.2.3 A Bayesian test of phylogenetic incongruence

I analysed the sequence data using the BEAST software (Drummond and Rambaut 2007) for Bayesian phylogenetic estimation. For each analysis, I specified the HKY + Γ(5) + Inv evolutionary model, with each codon position partitioned and branch rates selected from a relaxed clock log-normal distribution (Drummond et al. 2006). The HKY model (Hasegawa, Yano and Kishino 1984) is a fairly general model which is commonly used in nucleotide sequence

analyses. It provides a distinction between transition and transversion substitutions with different parameters for each type of event and unequal nucleotide frequencies which is more representative of the sequences to be studied. The HKY model is used over the more generalized GTR model as the difference of 4 parameters between the two models benefits us with a substantial reduction in computational time, given our large data set.

To account for variations in the evolutionary rate at different sites I modeled rates across sites using a gamma distribution with 5 rate categories (Γ(5)). In addition, a parameter *pInv* relating to the proportion of invariant sites was specified. The inclusion of a separate parameter for invariant sites is in fact unnecessary as the gamma distribution accounts for such sites and therefore there is slight overparameterisation in my model. Future phylogenetic analyses should ensure against over-parameterisation of the evolutionary model as the inclusion of too many correlated parameters can affect the convergence of the MCMC algorithm and unduly increase the influence of the prior distribution on the posterior (Rannala 2002).

The HKY + Γ(5) + Inv model was applied to each codon position to account for the different selective pressures that generally act on each position. I used a relaxed clock model following rejection of the molecular clock assumption. A Yule model of speciation was specified for the tree prior. In each analysis, the initial tree was generated randomly.

To determine if any of the six genes shared the same evolutionary history and could be combined in further phylogenetic analysis, I investigated the phylogenies of the genes in pairs. By employing a Bayesian approach, the evidence for incongruence between any two genes could be determined in the absence of confident phylogenetic trees for either. For each gene pair I ran two separate MCMC chains, each sampling over two separate phylogenetic trees – one for each gene.

I applied the methods of Huelsenbeck and Bull (1996) to determine whether constraining the sampling process to only consider pairs of trees with identical topologies would produce a significantly worse fit to the observed data,

quantified as change in the total log likelihood, indicating differences in tree topologies and evidence for incongruence. For each gene pair, the first MCMC chain sampled topologies which were constrained to be the same for both genes whilst the second MCMC chain samples independent topologies for each gene. In both chains, evolutionary parameters were constrained across gene partitions but branch lengths were allowed to vary for each gene tree. The possibility of phylogenetic incompatibility due to process incongruence (i.e. different mutational processes) was also investigated by modifying the test above to remove the constraint of identical evolutionary parameter estimates for each gene in both the constrained and unconstrained chains. The analysis was repeated again to ensure convergence of the sampled distributions.

Phylogenetic incongruence among PV genes may be attributed to convergent evolution at the amino acid level or recombination. To determine whether convergent evolution provides a plausible explanation, I performed the incongruence tests using only the third codon positions. The redundancy of the genetic code means that nucleotide substitutions at the third codon position are incapable of changing the amino acid coded for and therefore selective pressures driving convergent evolution will not act at third codon positions. Tests of phylogenetic compatibility at the third codon sites were not performed on the transforming genes due to the shorter alignments and higher evolutionary rates of these genes.

For each paired-gene run, the MCMC algorithm was run for 30,000,000 generations with sampling of states every 1,000 generations. In each chain, the first 5,000 sampled states were discarded as the burn-in period of the algorithm (i.e., the time taken for the chain to reach equilibrium), leaving 25,000 states for analysis. Each state in the Markov chain is sampled dependent on the previous state and therefore there will be some degree of correlation among states in the chain. The effective sample size (ESS) for a parameter indicates the number of independent points that have been sampled. It is calculated by dividing the number of post-burn-in sampled states by the auto-correlation time (the average minimum number of states between two uncorrelated sample points). For each

chain I used the Tracer software distributed with BEAST to determine the ESSs of sampled parameters. Tracer flags ESS values less than 200 as the chain may not contain enough independent samples to provide a sufficient representation of the posterior distribution. For all chains, sampled parameters of the evolutionary model had high ESS values (the lowest ESS was 1991.59), whilst the likelihood, prior and posterior had ESS values ranging from 329.895 to 2886.523. Thus, each chain contained a sufficient number of independent samples to suggest a good amount of mixing and sampling from the posterior distribution.

To ensure convergence of each chain on the posterior distribution, MCMC runs were repeated, starting from a different, randomly obtained initial staring point. The sampled distributions for all parameters, likelihood and posterior distributions can be compared to see if similar distributions are obtained across multiple, independent runs. I calculated the PSRF statistic (equations 2.37 – 2.40) for each MCMC component. The distribution of calculated PSRF values for all components, across all chains run, had a mean of 1.02 (s.d. = 0.027), indicating generally good agreement of sampled distributions between independent MCMC chains run under the same set of constraints for each gene partition. Convergence of all chains on the same distribution may therefore be inferred.

Each MCMC chain samples over many different topologies, to make topological comparisons the maximum *a posteriori* (MAP) tree for each gene was obtained from the phylogenies sampled for individual gene MCMC runs performed using the same evolutionary model as specified above. The posterior probability associated with each branch in the MAP tree tells us the probability that the grouping of taxa observed subtending from that branch (i.e., the subtending clade of taxa) is correct, given the data and the evolutionary model. Clade posterior probabilities greater than 0.9 (p>0.9) are taken to provide significant support for the estimated clade. Clades associated with lower posterior probabilities indicate uncertain phylogenetic relationships.

In Bayesian statistics, different hypotheses or models are compared through the Bayes factor (Jeffreys 1935). The Bayes factor (BF_{10}) comparing model M_1

against M_0 is derived from Bayes' theorem as the ratio of the marginal likelihoods under the two models:

$$BF_{10} = \frac{p(X \mid M_1)}{p(X \mid M_0)} \tag{3.3}$$

For a model M with parameters $\boldsymbol{\theta}$,

$$p(X \mid M) = \int p(X \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \tag{3.4}$$

BF_{10} tells us how much the data favour M_1 over the null hypothesis (M_0). Kass and Raftery (1995) provided guidelines for the interpretation of Bayes factors (Table 3.1) Converting the BF to a logarithmic scale by taking twice the natural logarithm of BF_{10}, they suggested that positive support for H_1 (and against H_0) may be inferred when $2\ln BF_{10} >= 2$, with values greater than 10 indicating very strong support for the alternative hypothesis.

$2\ln(BF_{10})$	Evidence against H_0
0-2	Not worth more than a bare mention
2-6	Positive
6-10	Strong
>10	Very strong

Table 3.1: Interpretation of Bayes factor values determined for the comparison of two distinct models or hypotheses. Reprinted from Kass and Raftery (1995).

Bayesian phylogenetic analysis is performed using MCMC sampling methods, which enable estimation of the posterior distribution without performing the difficult calculation of the marginal likelihood. Thus, the BF for different phylogenetic models is not easily obtained. Newton and Raftery (1994) applied the Monte Carlo method for approximating integrals to the evaluation of the marginal likelihood for a model M (as defined in equation 3.4). They found

that given a sample from the posterior distribution, the marginal likelihood is estimated by the harmonic mean of the sample likelihoods:

$$f(X \mid \mathrm{M_i}) = \left[\frac{1}{N} \sum_{j=1}^{N} \frac{1}{f(X \mid \tau_j, \nu_j, \theta_j)} \right]^{-1} \tag{3.5}$$

The BF can therefore be estimated from the ratio of the harmonic means of the likelihoods of the MCMC chains generated under each model. The harmonic mean estimator (HME) possesses the important property of statistically consistent; however, it is associated with problems of infinite variance and a tendency to over-estimate the marginal likelihood (Newton and Raftery 1994), which is a consequence of under-sampling of points from regions of low likelihood in a finite sample.

A proposed alternative to the HME applies the thermodynamic integration (TI) method of statistical physics to the estimation of the marginal likelihoods under/for each hypothesis (Lartillot and Philippe 2006). In the model-switching application of TI, the sampling process is used to integrate along a continuous path connecting two models defined on the same parameter space. The integral of this path provides the BF for the two models and is approximated by sampling at discrete points along the path, with a Markov chain being run at each point. Lartillot and Phillipe found TI to provide a more reliable estimate of the BF than the HME; however, the computational demands of the method limit its applicability to large datasets.

More recently, a stepping-stone (Fan et al. 2011; Xie et al. 2011) method, which applies importance sampling to the path sampling approach of TI, has been proposed for approximating marginal likelihoods. In the estimation of marginal likelihoods, the SS method is found to provide a similar degree of accuracy to TI; however, it is a less computational intensive method that TI as it does not require sampling from the posterior distribution (Fan et al. 2011). Thus, the SS method may be deemed the preferred method of ML estimation and should be investigated for future testing of hypotheses of phylogenetic incongruence among partitions.

In order to estimate the BF of chains run on the PV genes, I rely on the observation by Lartillot and Philippe (2006) that marginal likelihoods estimated using the HME and TI were similar when the models tested were of similar dimensions. In analyzing phylogenetic incongruence among PV genes, the two models tested consist of the same parameters and the only difference between them is the constraint placed on the tree topology; thus I am making the assumption that the HME will perform as well as TI here.

For each gene pairing, I calculated the marginal log likelihoods, estimated by calculating the log of the harmonic mean of the likelihoods at each state using Equation 3.5, for both the topologically constrained and the unconstrained MCMC chains.

$$\text{Marginal log likelihood} = \ln(n / \sum_{i=1}^{n} \exp(LL_i)) \tag{3.6}$$

To account for the uncertainty in BF estimates, I determined the 95% confidence interval (CI) by resampling 1000 times from the posterior distributions of the unlinked and linked chains and calculating the BF for each bootstrapped sample. Resampling was done in accordance with the method used by Suchard et al. (2003) in which blocks of states, rather than individual states, are sampled from the MCMC chain so as to preserve the correlated nature of consecutive states during the MCMC simulation. The auto-correlation time (lag), i.e. the minimum length between uncorrelated states in an MCMC chain was obtained from the sampled chains using the Tracer application in BEAST. The bootstrap method provides only an approximation of the error in the BF since we are not actually sampling new chains from the posterior distributions; however, the computational resources required to determine the error using the latter method are too great to make it worthwhile.

3.3 Results

3.3.1 Testing the molecular clock assumption

Table 3.2 shows the results of the likelihood ratio test performed to evaluate support for the molecular clock hypothesis in the E1, E2, E6, E7, L1, and L2 genes. For each gene we find significant support (P<0.001) against the null hypothesis of a constant evolutionary rate and hence reject the assumption of a molecular clock.

	$\ln L(H_0=MC)$	$\ln L(H_1=NC)$	$2\Delta \ln L$	Df	χ^2 P-value
E1	-94528.66	-93563.04	1931.24	106	<0.001
E2	-40356.20	-39931.92	848.56	106	<0.001
E6	-23581.55	-23215.96	731.18	100	<0.001
E7	-14177.63	-13987.23	380.8	100	<0.001
L1	-81862.18	-81162.88	1398.6	106	<0.001
L2	-52814.43	-52335.73	957.4	106	<0.001

Table 3.2: Results of likelihood ratio tests performed on each gene to evaluate support for a constant rate of evolution. MC indicates the null hypothesis of a molecular clock and NC indicates the alternative hypothesis of non clock-like evolution. df indicates the degrees of freedom in each test.

3.3.2 Bayesian tests of phylogenetic incongruence

In all fifteen pairings of the E1, E2, E6, E7, L1, and L2 genes, higher log likelihoods were observed for chains run with independent topologies for each gene than when both genes were constrained to the same topology at each state in the chain; this is true whether the evolutionary parameters are constrained or not (Appendix A.2 and A.3), suggesting that differences in evolutionary history contribute to the differences in likelihoods. The Bayes factor was used to

determine whether the differences between topologically constrained and unconstrained chains for each gene pairing were significant. All gene pairings produced values of $2\ln BF > 20$ (Tables 3.3 and 3.4) thus demonstrating significant support for topological incongruence among the PV genes.

I performed two runs of the constrained and unconstrained topology chains for each gene and generated BF estimates for each run so as to determine the consistency of the estimates. For most gene pairs the BFs from the separate runs are similar and where larger differences are observed there is overlap in the associated 95% credible intervals (CIs). The E7-E2 pairing provides the only instance of non-overlapping credible intervals for the BF from separate runs: (19.51, 30.44) vs. (36.14, 57.33). The range of the 95% CI is observed to be about 20 log units on average. For all gene pairs, the CIs for estimated BFs point to significant evidence for independent gene topologies.

For MCMC chains generated with independent evolutionary parameters for each gene, the greatest values are observed when an early gene is paired with a late gene, with the E1-L2 pairing giving BF values of 263.8456 (254.6650, 271.8332) and 264.5274 (255.7516, 274.4575), the E1-L1 pairing giving BF values of 200.1557 (191.1177, 203.5548) and 203.5124 (195.0476, 208.3344), and the E2-L2 pairing giving BF values of 109.7836 (99.2848, 130.0872) and 111.7983 (109.3592, 122.8943). However, the L1-L2 pairing also demonstrated substantial evidence of phylogenetic incongruence with BF values of 103.1562 (97.0436, 105.9685) and 106.1399 (93.3945, 121.0006). Similar results are obtained from chains generated under a heterogeneous evolutionary model across gene partitions.

The results of the incongruence test performed on third codon sites of the core genes are shown in Table 3.5. For each gene pairing, the MCMC chain generated for unlinked topologies had a higher log likelihood than the MCMC chain generated for linked topologies and in all cases the BF values indicate significant phylogenetic incongruence at the third codon positions of the genes.

	$2\ln BF_{UC}(1^{st}$ chain)	$2\ln BF_{UC}(2^{nd}$ chain)
Linked EP		
E1-E2	111.58 (85.22, 136.48)	120.45 (108.58, 128.69)
E1-L1	400.32 (382.24, 407.10)	407.02 (390.10, 416.67)
E1-L2	527.70 (509.34, 543.66)	529.05 (511.50, 548.92)
E2-L1	155.00 (146.78, 166.10)	139.09 (127.17, 172.31)
E2-L2	219.56 (198.56, 260.18)	223.60 (218.72, 245.79)
L1-L2	206.32 (194.09, 211.94)	212.28 (186.79, 242)
Unlinked EP		
E1-E2	136.16 (118.50, 151.29)	97.80 (85.18, 124.56)
E1-L1	378.92 (360.80, 389.70)	388.68 (361.29, 397.96)
E1-L2	503.55 (490.69,511.39)	509.55 (486.36, 520.97)
E2-L1	153.69 (133.00, 159.47)	154.04 (116.83, 161.38)
E2-L2	275.77 (250.87, 281.63)	234.01 (213.68, 272.90)
L1-L2	213.17 (191.92, 235.10)	199.39 (177.82, 215.04)

Table 3.3: Results of phylogenetic incongruence test of the core PV genes: calculated Bayes factors for paired genes with independent (unconstrained) tree topologies against paired genes with the same (constrained) tree topology. Values in parenthesis indicate the 95% CIs for the Bayes Factor estimates.

	$2\ln BF_{UC}$(1st chain)	$2\ln BF_{UC}$(2nd chain)
Linked EP		
E6-E1	113.21 (105.22, 118.59)	98.38 (90.25, 111.65)
E6-E2	134.04 (117.10, 141.14)	115.25 (99.86, 136.40)
E6-E7	79.57 (73.40, 97.29)	80.61 (73.33, 101.05)
E6-L1	192.35 (184.29, 213.20)	157.29 (141.84,218.75)
E6-L2	216.84 (210.67, 232.61)	218.39 (198.72, 243.42)
E7-E1	55.31 (37.13, 70.90)	60.13 (48.09, 69.90)
E7-E2	69.22 (49.00, 77.21)	31.53 (25.92, 43.60)
E7-L1	103.46 (96.94, 112.16)	137.58 (102.49, 145.85)
E7-L2	106.18 (91.42, 133.68)	90.24 (84.44, 133.54)
Unlinked EP		
E6-E1	91.08 (74.19, 117.71)	111.59 (80.64, 117.71)
E6-E2	117.82 (108.71, 124.44)	143.46 (112.31, 151.53)
E6-E7	88.39 (73.28, 107.90)	92.70 (83.17, 109.02)
E6-L1	234.42 (218.23, 249.71)	219.30 (208.60, 235.34)
E6-L2	214.83 (209.03, 223.91)	227.80 (203.84, 239.83)
E7-E1	45.97 (37.30, 70.26)	54.56 (44.34, 88.76)
E7-E2	24.03 (19.51, 30.44)	51.78 (36.14, 57.33)
E7-L1	102.59 (86.33, 127.21)	82.24 (75.26, 112.96)
E7-L2	120.28 (114.35, 125.67)	163.85 (113.70, 182.10)

Table 3.4: Results of phylogenetic incongruence test of the PV oncogenes: calculated Bayes factors for paired genes with independent tree topologies against paired genes with the same tree topology. Values in parenthesis indicate the 95% CIs for the Bayes Factor estimates.

	$2\ln BF_{UC}$ (1st chain)	$2\ln BF_{UC}$ (2nd chain)
Linked EP		
E1-E2	88.12 (68.34,114.34)	103.63 (87.26,108.89)
E1-L1	103.63 (130.79, 150.85)	139.14 (126.20, 164.95)
E1-L2	164.61 (156.17, 170.79)	151.88 (141.62, 167.91)
E2-L1	115.62 (92.99, 140.50)	112.57 (109.26,118.73)
E2-L2	125.63 (95.34, 147.39	113.58 (100.80, 141.97)
L1-L2	68.64 (59.23, 85.91)	61.06 (51.71, 90.88)
Unlinked EP		
E1-E2	90.74 (80.63, 96.43)	100.50 (52.14, 127.46)
E1-L1	147.79 (126.65, 159.44)	143.22 (136.61, 164.66)
E1-L2	142.11 (127.67, 164.59)	156.91 (150.04, 167.13)
E2-L1	119.45 (104.57, 130.46)	87.62 (73.60, 123.80)
E2-L2	100.92 (86.20, 142.55)	138.22 (129.32,147.46)
L1-L2	78.32 (73.58, 94.78)	94.79 (80.41, 105.41)

Table 3.5: Results of phylogenetic incongruence tests of the core PV genes (third codon sites only): calculated Bayes factors for paired genes with independent (unconstrained) tree topologies against paired genes with the same (constrained) tree topology. Values in parenthesis indicate the 95% CIs for the Bayes factor estimates.

3.3.3 Estimated phylogenetic differences among PV genes

The maximum *a posteriori* (MAP) PV trees obtained from the sampled phylogenies for each gene are provided in Appendix A.6-11. No two genes produce identical MAP tree topologies; however, some similarities do exist. Overall, the estimated gene trees display high posterior probabilities ($p<0.9$) for genus-based groupings of PV types. In the E1 and L1 gene trees, monophyletic clades of taxa representing each genus are observed with posterior probabilities greater than 0.99. A similar observation is made in the other gene trees, with a few exceptions. Uncertainties are observed in the grouping of the E2-α genus ($p=0.57$), the L2-λ genus ($p=0.71$), the E6-λ genus ($p=0.78$), and the E6-ε genus ($p=0.55$). In addition, E6 sequences from the κ genus fail to cluster together and are instead observed to be distantly related. In the E7 tree, the κ PVs and the γ PVs each form paraphyletic clades. All 6 gene trees also place the δ and ε PVs together in a monophyletic clade of fibropapilloma-causing PVs ($p>0.99$). The ν and σ PV lineages, which were isolated from cutaneous papillomas in human and porcupine species, respectively, cluster together with a posterior probability of 1.0 in the trees of 5 PV genes but with less certainty in the E7 gene tree ($p=0.8$).

Topological differences between the gene trees are observed in the relative ordering of the genus clades. However, the proportion of inter-genus branches (i.e., branches joining together PV types from different genera) in each gene tree with $p<0.9$ is as follows: 0.15 (E1), 0.57 (E2), 0.75 (E6), 0.93 (E7), 0.36 (L1), and 0.71 (L2), thus indicating substantial topological uncertainty deeper in the trees of the E2, E6, E7, and L2 genes. Further comparison of the gene topologies is therefore restricted to the E1 and L1 gene trees.

Figure 3.1 shows a splits network generated by combining the MAP topologies for the E1 and L1 genes in SplitsTree. A split is defined as the partition of taxa obtained following removal of any branch in the tree. SplitsTree obtains all the splits for the E1 and L1 MAP trees and creates a network consisting of edges for each split observed in the two trees. Regions of the gene

trees which are congruent are represented by single edges in the network and are 'tree-like' in appearance; however, if two taxa (or sets of taxa) are connected to each other in different ways in the two gene trees this is represented in the network by a set of parallel edges or 'reticulations'. Such regions in the network therefore display where incongruities between the evolutionary histories of the two genes lie.

The network shows several incongruent regions, the majority of which are located at the base of the network, but incongruent regions are also observed at the base of the clade of α PVs. Direct comparison of the E1 and L1 gene MAP tree topologies reveals several key differences. The ν+σ PV clade occupies a basal position in the E1 gene tree and is excluded from the clade formed by all other mammalian PVs, besides the δ+ε PV clade, with p=1.0. However, in the L1 gene tree the ν+σ PV clade clusters within the μ+κ PV clade (p=0.94). Different positions are also observed for the PsPV1+TtPV2 clade, which associates with the α PVs (p=1.0) in the E1 gene tree, but with the ξ PVs (p=0.98) in the L1 gene tree.

Another notable difference in the two gene trees concerns the arrangement of the high-risk species α-5, α-6, α-7, α-9, α-11, and α-12. In the E1 gene tree, these high-risk species PV types cluster together (p=1.0); however, in the L1 gene tree the high-risk PV types are split: PV species α-9, α-11, and α-12 cluster with the low-risk PV species α-8 and α-10 (p=0.91), whilst PV species α-5, α-6, and α-7 form a distinct clade (p=0.97).

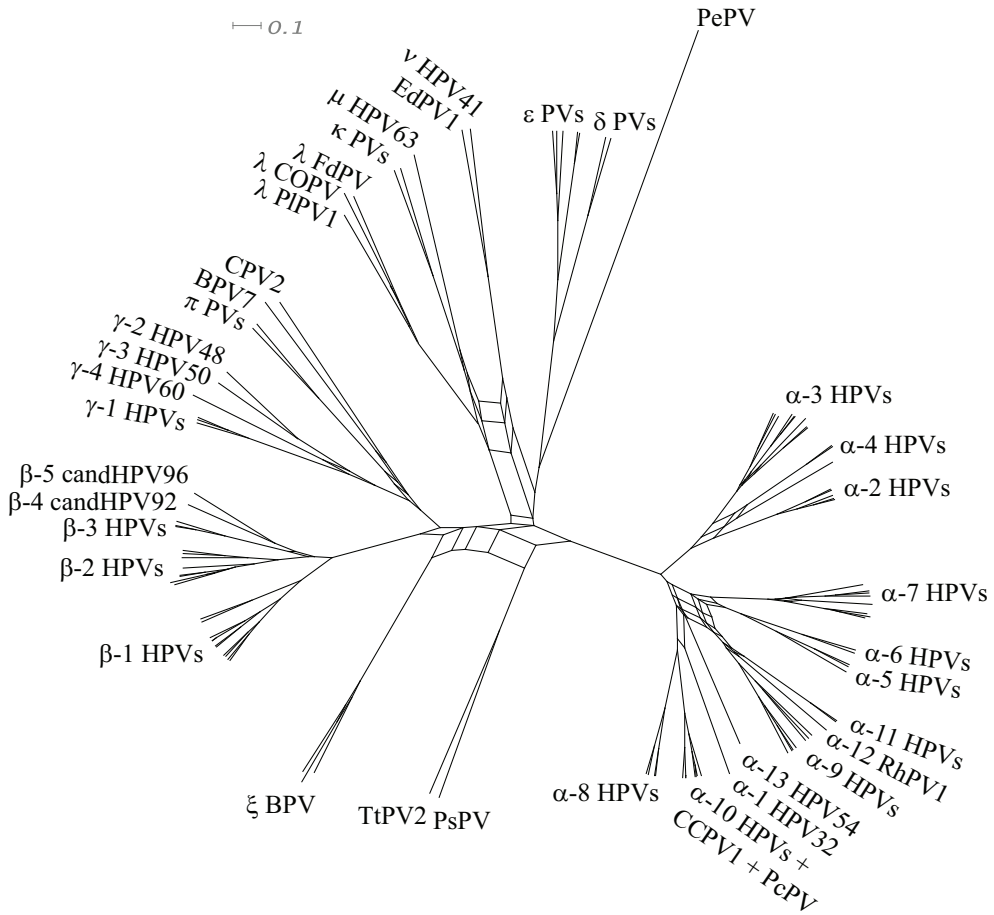


Figure 3.1: A splits network generated from the E1 and L1 MAP phylogenies using SplitsTree. Sets of parallel edges in the network indicate locations of topological incongruence between the trees.

3.4 Discussion

A Bayesian phylogenetic approach has been used to examine phylogenetic compatibility among the genes of a divergent set of PV sequences. In contrast to previous studies of phylogenetic incongruity among the PV genes (Narechania et al. 2005; Gottschling et al. 2007b), the tests employed here are not conditioned on the assumption that the individual gene tree topologies can be determined with absolute confidence or, that a total-evidence phylogeny constructed from all genes represents the "true" phylogeny. Instead, the MCMC methods employed in Bayesian phylogenetic analysis allow us to integrate over uncertainties in the

specific topology of each gene so that the results are not biased by incorrect topologies.

The statistical support for phylogenetic incongruence was determined by estimation of the Bayes factor for linked vs. unlinked topologies for paired genes. The calculated Bayes factors are much larger than the values suggested by Kass and Raftery (1995) as indicators of significant evidence for one hypothesis over another. However, it has been noted that large errors may be associated with the harmonic mean estimator used to estimate the marginal likelihoods of MCMC chains under each phylogenetic hypothesis. I have attempted to determine the uncertainty in the estimation by calculating individual Bayes factors for repeated chains and determining the 95% credible interval of the Bayes factor for each chain by bootstrapping the sampled likelihoods. Accounting for these errors, the tests performed detected significant phylogenetic incongruence in all pairings of the E1, E2, E6, E7, L1 and L2 genes. Significant phylogenetic incongruence between the genes remained when accounting for differences in the evolutionary process acting on each gene. The overwhelming conclusion from the tests performed is therefore that no two PV genes (for the set of PV types studied) can be combined in phylogenetic analysis.

Topological incongruities were observed at multiple branches between the gene phylogenies and involved rearrangements between multiple sets of taxa. The MAP trees derived for the E1 and L1 genes, which are the most conserved of the PV genes, both demonstrated high posterior clade probabilities (p>0.9) along their trees and therefore produce strongly supported conflicting phylogenies. The uncertainties observed in phylogenetic estimates for the E2, E6, E7, and L2 genes may be due to shorter alignments that provided fewer sites possessing strong phylogenetic signal. Entropy measurements performed on PV genes from a diverse set of PV types have identified the presence of more than double the number of highly variable (Shannon entropy > 1.6) sites in the E2, E6, E7, and L2 genes than the E1 and L1 genes (Batista et al. 2011). The inference of accurate phylogenies from these genes may therefore be difficult for diverse sets of taxa. In such cases we might have expected that gene pairings of the E2, E6,

and E7 genes with the E1 gene, and of L2 with L1, would have aided phylogenetic estimation of the shorter genes, resulting in higher likelihoods for the constrained chains than the unconstrained chains. This was not observed, however, and we may assume from this that, although the phylogenetic signal in each of the four genes is poorer than of the core genes, it is still strong enough to demonstrate phylogenetic incongruence among the genes.

The inference of different evolutionary histories for each gene is an intriguing finding. The E1 and E2 genes express proteins which perform regulatory functions during the viral life cycle. The E1 and E2 proteins even interact with each other to initiate viral genome replication. The E6 and E7 genes both manipulate cellular pathways to ensure a replicative state is maintained in differentiated epithelial cells. The L1 and L2 genes are expressed in the latter stages of the viral life cycle; their protein products make up the viral capsid. Given the overlapping functions of E1 and E2, E6 and E7, and L1 and L2, it would be expected that in each of these pairs, the evolutionary histories of the genes would be highly similar and this assumption has been made previously for a similar data set of PV types to that studied here (Garcia-Vallve, Alonso and Bravo 2005). This illustrates the importance of testing phylogenetic congruence of PV genes before making further inferences from the estimated phylogenies.

Some interesting patterns are observed from the conflicting topologies derived for the E1 and L1 genes. Three topological rearrangements, concerning the positions of PsPV1+TtPV2, the σEdPV1+ νHPV41, and the α−5+α−6+α−7 HPVs, are observed with high support (p>0.9) between the two gene trees. In the E1 gene tree, the cetacean genital PsPV1+TtPV2 cluster with the genital PV containing clade of primate α PVs (p=1.0), the porcupine σEdPV1 + human νHPV41 are distantly related to all other PV types, and the high-risk α−5+α−6+α−7 HPVs form a monophyletic clade with other high-risk α PVs (p=1.0). However, in the L1 gene tree the cetacean genital PsPV1+TtPV2 cluster with the bovine non-genital ξ−PVs (p=0.98), the porcupine σEdPV1 + human νHPV41 cluster with the human μHPV63 and lagomorph κ PVs (p=0.94), and the high-risk α−5+α−6+α−7 HPVs form a paraphyletic 'high-risk' clade within the α PV clade (p=0.97).

Thus, it appears that phylogenetic groupings within E1 gene tree reflect the biological and pathological characteristics of the lineages, whereas groupings in the L1 gene tree reflect similar host preferences. Since the E1 gene has a key role in replication, this may suggest against PV host-specificity being governed by replication factors within the cell as has been postulated for the polyomaviruses (Schneider et al. 1994). Instead, host-specificity may be exclusively governed by the specificity of the capsid proteins (L1 and L2) for the host cell surface receptors allowing virion attachment and entry into the cell (Webby, Hoffmann and Webster 2004).

Narechania et al. (2005) proposed that phylogenetic incongruities between the early genes and the late genes of the high-risk α HPVs may be a consequence of convergent evolution, in either the early genes or late genes, arising due to development under similar evolutionary pressure. The patterns observed above may agree with such a hypothesis. For example, convergent evolution may have occurred among the early genes of the cetacean PVs and the α primate PVs due to similar environments presented by genital tissue. The incongruence tests performed here on the core genes using only the third codon sites, which are immune to the influence of convergent evolution, revealed significant phylogenetic incongruence at these sites, which strongly suggests against the hypothesis of convergent evolution. However, the phylogenies estimated for each gene did not agree with estimates from the full gene and displayed greater topological uncertainty. Thus, we cannot be sure that the significant BF values at the third codon positions indicate genuine phylogenetic differences or if they are due to random phylogenetic signal. A more detailed examination of the possibility of convergent evolution among PV sequences is required.

Strongly supported conflicting phylogenies for different genes from the same set of taxa can also suggest the possibility of recombination events. The E1-L1 splits network (Figure 3.1) generated using SplitsTree (Huson and Bryant 2006) shows several incongruent regions, the majority of which are located at the base of the network and therefore may suggest multiple ancestral recombination events. Despite a lack of physical evidence for recombination among PVs,

several recombination detection studies have reported findings of significant recombination signal in PV sequences (Varsani et al. 2006; Angulo and Carvajal-Rodriguez 2007; Carvajal-Rodriguez 2008). These studies identified recombination signal in the L2 gene of multiple PV types, which may explain the greater topological uncertainty observed in the L2 gene phylogeny. The L2 gene, encodes a capsid protein, which interacts with host cell receptors and is immunogenic. Recombination in this region may therefore provide a means of generating more genetic diversity to infect new hosts and/or evade host immune mechanisms. The phylogenetic incongruities observed between the E1 and L1 genes, may therefore indicate recombination in the L1 gene, which, like L2, encodes a protein that forms a part of the viral capsid and is also immunogenic.

The number of cetacean PV types has increased since this analysis was performed, with two classified PV genera (omnikron and upsilon) comprising of PV types infecting multiple cetacean species and 1 cetacean PV type infecting *Phocoena phocoena* (PphPV3) currently without genus classification (Gottschling et al. 2011a; Robles-Sikisaka et al. 2012). Phylogenetic estimations with these new types show well-supported conflicting phylogenetic arrangements of the cetacean PVs in early gene and late gene trees, with only PphPV3 (isolated from the harbour porpoise – *Phocoena phocoena*) demonstrating a consistent phylogenetic placement (with the α PVs) across the genomic regions (Gottschling et al. 2011a; Robles-Sikisaka et al. 2012). The fixed position of PphPV3 among the incongruent gene phylogenies suggests a scenario of recombination among similar hosts: the ancestral relative of PphPV3 recombined with another ancestral cetacean PV lineage to produce a lineage with early genes from the PphPV3 ancestor and late genes from the unknown cetacean PV ancestor (Gottschling et al. 2011a). The biological plausibility of this scenario suggests that further studies of recombination among PVs should focus on the cetacean PVs in addition to the high-risk HPV types.

Extracting evidence of recombination from sequence data is a non-trivial exercise since the ability to detect recombination rests largely on obtaining the correct alignment of the sequences; however, the occurrence of a recombination event itself and any subsequent mutational events in the recombinant region may adversely affect the ability to derive the correct sequence alignment. In addition,

a recombination breakpoint is inferred wherever statistically significant phylogenetic conflicts are estimated between adjacent segments of sequence, thus the process of phylogenetic estimation must also be accurate. Bayesian methods of recombination detection (Husmeier and McGuire 2003; Minin et al. 2005; Martins Lde, Leal and Kishino 2008; Bloomquist, Dorman and Suchard 2009; Webb, Hancock and Holmes 2009) allow statistically significant changes in tree topology along a sequence to be identified whilst accounting for model uncertainties and may therefore provided a less biased approach to investigating recombination among PV sequences.

Genuine topological differences observed in the phylogenies estimated from different genomic regions of the same set of taxa imply the influence of convergent evolution and/or recombination on the evolution of the sequences. However, we cannot discount the possibility that the conflicting phylogenies are a result of errors in the phylogenetic estimation process. These systematic errors can arise if an inappropriate method or evolutionary model is used to analyse molecular sequences. For instance, MP methods of phylogenetic estimation have been shown to be inconsistent (will not converge on the right tree even with infinite data) when the rates of evolution vary considerably among the branches of a tree (Felsenstein 1978). Simulations studies (Huelsenbeck and Rannala 2004; Lemmon and Moriarty 2004; Grievink et al. 2010) have shown the effects of underspecification of the evolutionary model on accurate phylogenetic estimation by ML and Bayesian methods. If the evolutionary model used fails to provide a good representation of the evolutionary process that produced the observed sequences, phylogenetic estimations for one or all of the partitions are liable to be incorrect. This may result in the observation of phylogenetic incongruence where none exists. More importantly, the incongruities may have high statistical support, as is observed in the PV gene trees, and thus falsely suggest genuine differences in the evolutionary histories.

Although the mathematical models used are not able to capture all aspects of the evolutionary process, the recommendation is that one should account for as much heterogeneity in the evolutionary process as is possible (Lio and Goldman 1998; Huelsenbeck and Rannala 2004; Lemmon and Moriarty 2004). A

parameter-rich evolutionary model can, however, slow down the phylogenetic estimation process and therefore various hypothesis tests can be performed to determine which components of the evolutionary model have the greatest impact on phylogenetic estimation (Goldman 1993; Huelsenbeck and Rannala 1997).

In the present analysis, I used a fairly general model of evolution that represents a great deal of variation in the evolutionary process. I partitioned each gene by codon position; for each codon position an evolutionary model that accounted for differences in rates of transitions and transversions between bases, unequal nucleotide frequencies, rate variation among sites and rate variation across lineages (following rejection of a molecular clock hypothesis) was specified. However, I did not investigate the appropriateness of the priors that were applied to the model parameters. For all parameters the default prior settings of BEAST were used and in future analysis it would be sensible to first the impact of each prior on the joint prior and the posterior distribution.

Phylogenetic estimation may also be affected by unequal taxon sampling (Heath et al. 2008). The correct phylogenetic placement of taxa that are only distantly related to other taxa is difficult to achieve as the larger evolutionary distances make it difficult to infer the exact amount of evolution between these taxa. Thus, one should attempt to have a balanced data set by either removing taxa with no close relations or including more closely related sequences.

The PV data set analysed is quite imbalanced as it includes many closely related primate PV types and a smaller proportion (23/108) of less-closely related non-primate PV types. The fact that most of the incongruent regions indicated in the E1-L1 splits network occur at the base of the gene trees, could suggest a difficulty in resolving deeper divergences. However, in order to perform an analysis of virus-host association mechanisms (see Chapter 4), the inclusion of PV types from a wide variety of hosts was necessary. Some reduction in the number of primate PV types included could have been possible, although a reduction down to only single representative types for PV species may allow for topological uncertainties within genus clades. Efforts to increase knowledge of PV diversity among vertebrates have resulted in genome sequences from more non-human species. However, these types appear to have increased the number of PV genera (Bernard et al 2010) rather than sampling more diversity within a

particular genus and thus, the identification of more distantly related PV types may further increase the problem of accurate phylogenetic inference for the complete PV family.

The direct consequence of this analysis and the observation that incongruent nodes are located at various locations in the gene trees is that a combined phylogenetic analysis of PV genes will only be possible with small, closely related data sets of PVs. Evidence of this is apparent in the combined E1-E2-L1 protein topologies estimated for a data set of diverse PV types by Gottschling et al (2007), which places the PsPV1 and TtPV2 clade of Cetacean PVs as an outgroup to all other mammalian PVs, whilst the individual gene phylogenies estimated in this analysis indicate a close relationship with the α PVs in the early genes and a close relationship with the ξ BPVs in the late genes.

Chapter 4

Analysis of PV-Host Phylogenetic Incongruence Using a Biased Sampling Approach

4.1 Introduction

Phylogenetic analysis of PV types infecting different hosts has revealed a branching pattern that is not consistent with the evolutionary history of their hosts (Gottschling et al. 2007b). In studies of parasite-host systems, congruent phylogenetic patterns are taken to be highly indicative of host-linked parasite evolution via a cospeciating mechanism, as per Fahrenholz's rule. Observed disparities between the host and parasite phylogenies imply that some parasite diversifications occurred independently of host speciation events. Host-independent evolutionary changes in parasite lineages can occur via parasite duplication on the same host species (prior divergence) or transfer and successful colonisation on new host species (host transfer). Inferring these events from the observed host-parasite associations and the respective evolutionary histories is the task of cophylogenetic analyses. Different methods have been developed, each of which examines the coevolutionary history of parasite-host associations to different extents.

Cophylogenetic methods may be applied to characterise the mechanisms by which observed virus-host associations were formed as these associations are affected by mechanisms similar to those affecting parasite-host assocations. However, the application of these methods to study PV-host associations is difficult due to various short-comings of the methods and the complex nature of the discrepancies observed between phylogenies of the PVs and their hosts. I will describe some of the different

approaches and the issues preventing their use before presenting the approach taken to examine phylogenetic incongruities between the PVs and their mammalian hosts.

4.1.1 Characterisation of virus-host phylogenetic incongruities

4.1.1.1 Commonly used cophylogenetic methods of host-parasite analysis

4.1.1.1.1 Brooks' Parsimony Analysis (BPA)

BPA (Brooks 1981; Brooks 1990) is one of the oldest programs in the field of cophylogenetic analyses and, whilst it is a common tool in studies of cladistic biogeography, the initial development of the method was targeted towards phylogenetic analysis of host-parasite systems. The method was developed as a solution to Hennig's "parasitological method" which proposed the utilisation of parasites "as markers of evolutionary relationships among hosts" (Brooks 1981) but lacked the means to resolve phylogenetic incongruities between parasites and their hosts.

In the BPA approach, parasite taxa are evaluated as character states of the host and their distributions are used to reconstruct host relationships in the same manner as morphological and molecular data. Homology among host taxa may be assumed if they are found to share the same or closely related parasite associations. To distinguish between homology and homoplasy i.e. convergent/parallel evolution of host-parasite associations, the method requires data from multiple parasite species infecting the same group of host species. The parasite phylogenies (either known or estimated) are converted to parasite-host cladograms in which parasite taxa labels are replaced with the names of the associated host species. Under strict cospeciation, the individual parasite-host cladograms should depict identical relationships of the host species, which can then be inferred as representing the relationships among the host species.

Brooks used parsimony analysis to resolve topological differences among parasite-host cladograms. The relationships observed in each parasite-host cladogram are converted into binary representation and collected in an additive binary matrix

from which the most parsimonious host cladogram i.e., the one that is supported by the majority of parasite phylogenies is obtained using the Wagner algorithm or Hennigian argumentation. Multiple occurrences of a host species in a parasite-host cladogram, which arise when multiple parasite lineages are associated with the same host, are dealt with by introducing dummy host species to distinguish among the multiple lineages. The absence of a host species in one or more parasite-host cladograms indicates that no parasite association has been detected but may exist and is therefore coded as missing ('?') data in the additive binary matrix.

Differences between the BPA-estimated host cladogram and individual parasite phylogenies allow inferences to be made regarding the extent of host tracking by the parasites and the likely causes of disparate branching patterns. In BPA, congruent phylogenetic patterns are attributed to associations that arose via descent i.e., due to cospeciation, but also parasite duplications or lack of parasite speciation following host speciation. Multiple occurrences of a host species in the binary matrix will produce instances of homoplasy in the estimated host cladogram. The BPA method assumes that the underlying events causing homoplasy produce distinct phylogenetic patterns and can therefore be readily characterised. When homoplasy occurs within a single, monophyletic clade, it is found to be strongly indicative of parasite duplication. Homoplasious events across different clades, however, suggest that the multiple parasite lineages do not share a recent common ancestor and so the observed associations must be due to host transfer events.

The observed parasite associations are examined to infer the exact mechanism of host transfer: a ‘post-speciation dispersal’, which produces host range expansion such that more than one host species is associated with the same parasite lineage, is inferred when homoplasy involves the same parasite lineage; alternatively, “speciation by host-switching”, is inferred for homoplasious characters involving different parasite lineages from the same parasite group.

Missing parasite associations of hosts are also characterised using the simplest explanation: if the majority of parasite groups demonstrate associations with a host species, then absent parasite taxa from this host are interpreted as extinction events, and if most parasite groups lack any association with this host, then the situation is

assumed to involve a lineage sorting event i.e., the host was never infected with these parasites (also referred to as 'primitive absence' in BPA terminology).

Although BPA was not conceived for the direct comparison of host and parasite phylogenies, the method has been investigated for such purposes (Siddall 1996; Dowling 2002; Siddall and Perkins 2003). Referred to as ‘Type II BPA’, parasite characters are mapped onto the host tree, with the most parsimonious reconstruction sought via Farris optimisation.

As a method for reconstructing ancestral host-parasite association mechanisms, BPA presents several difficulties. The various characterisations described above are made by *a posteriori* interpretation of the estimated host cladogram. This means that there is no way to assess the degree of confidence associated with the inferred reconstruction. In addition, characterisation of ancestral events will become more difficult and uncertain as the number of taxa and the degree phylogenetic discordance increases. The BPA method will also be sensitive to uncertainties in parasite phylogenies and alternative topologies will require individual analyses.

Examinations of the BPA method have also revealed problems of ‘ghost lineages’ in the solutions (Page 1990a; Dowling 2002). The ghost lineages are additional instances of ancestral parasite lineages which appear due to the postulation of host switches or extinction events of descendent lineages. For any host lineage upon which the placement (removal) of a parasite taxon leads to the inference of a host transfer (extinction) event, the ancestral lineages of that host will also incur host transfer (extinction) events, resulting in an overestimation of the number of host switching and extinction events.

4.1.1.1.2 TreeFitter

Unlike BPA, the TreeFitter approach (Ronquist and Nylin 1990; Ronquist 1995) is specifically designed for the determination of ancestral host-parasite association mechanisms. An evolutionary model consisting of successive specialisation (i.e., association by descent) and host switching (association by colonisation) events is applied to fit the parasite tree into the host tree. The most parsimonious fit is

determined by the designated event-costs. Successive specialisation events are assigned a uniform cost of 1 whilst the relative switching cost - generally greater than 1 - may be determined by the user.

The tree-fitting process proceeds first with the calculation of a cost matrix which details the minimum cost of transfers between all pairs of host tree branches. Movement between branches proceed in one of three ways: from a parent branch to a descendant branch, from a descendant branch to a parent branch and laterally between branches separated by at least one branch. The first of these transfer events represents a cospeciation event, which is considered a special case of successive specialisation, and so all such transfers are assigned a cost of 1. Transfers from a descendent species to its parent species are assigned infinite cost thus ensuring that they are never postulated. All other transfer events are either defined by a single host switch event or by a combination of host switch and successive specialisation events, depending on the relative switching cost applied. The calculated cost matrix is then used to determine the most parsimonious host state at each node in the parasite phylogeny via post-order traversal i.e., starting from the tips, where the host state is known, and progressing down the tree to the root node.

To prevent back transfers occurring across the tree, a 'segment coding' technique is employed whereby each branch is partitioned at the speciation times of other branches. This ensures that only host transfer events between co-existing segments of branches are permitted, which forms an important consideration when evaluating transfers between distantly related branches. Temporal considerations are easily incorporated when the relative speciation times are known, i.e., if the host and parasite trees each conform to a constant rate of evolutionary change, but when this is not the case the method has to evaluate against all possible sequences of host speciation times – a process only feasible for very small sets of host taxa.

Determination of the most appropriate host switching cost may present a problem since there is often little information available to quantify the likelihood of host switching events. The optimal host switching weight should be one that neither precludes nor overestimates host switching events; Ronquist (1995) suggested examining a range of weights for a particular host-parasite system and using the weight that provides the greatest reduction in the number of successive specialisation

(tracking) events postulated since reconstructions biased against host switching events will have to postulate an unnecessarily large number of tracking events to explain the observed associations. Other concerns stem from the one-host-per-parasite requirement of the method: this restricts the reconstruction potential of the method since all host-switching parasite lineages must terminate their association with the pre-existing (source), host and duplication events cannot be invoked.

4.1.1.1.3 Reconciliation methods: TreeMap and Jungles

Reconciliation methods have become the standard approach to resolving host and parasite phylogenetic incongruities. Their origins stem from the separate, but conceptually similar, field of gene tree-species tree incongruence. More so than with parasites, genes are expected to track their hosts with complete fidelity yet the predicted evolutionary histories often differ due to additional evolutionary events acting on the genes such as gene duplications, gene losses, and lateral gene transfers. Goodman et al. (1979) suggested that the estimation of gene trees required a consideration of the various gene-specific events in addition to mutational changes at the sequence level. They proposed reconciling the incongruent gene trees estimated from sequence analysis with the species trees by postulating either a gene duplication event, which results in paralogous gene lineages within a species, or a gene loss event, which may comprise of either gene deletion or gene inactivation/reactivation processes, at each incongruent cladogenetic event in the tree. The most parsimonious explanation of the gene tree, i.e., the reconciliation requiring postulation of the least number of gene events, is sought.

Page (1994a; 1994b) adapted the reconciliation method for the host-parasite problem; the method is implemented in the software package TreeMap (Page 1995). In phylogenetic terms, gene duplication events are analogous to adaptive radiation of a parasite within a host species, as both events produce multiple gene/parasite lineages per species/host, and gene losses are analogous to extinction or sorting events of parasite lineages, as the associate lineage is rendered absent from the host lineage. Following the methods of Goodman and colleagues in gene tree-species tree reconciliation, TreeMap invokes cospeciation, duplication and sorting events on the

parasite tree to explain the observed incongruities with the host tree. Given the host and parasite phylogenies and the existing associations between terminal host and parasite taxa, TreeMap performs a post-order traversal on each tree such that each internal node is assigned the union of the host sets of its descendant nodes. Each ancestral parasite node is therefore labelled by the group of host species parasitised by their descendants. A reconciliation of the host and parasite phylogenies is achieved by matching host sets of the parasite nodes to the nearest equivalent host set of nodes on the host tree. A cospeciation event is assigned when mapped host and parasite nodes possess identical sets of descendant host species'. Duplication events are assigned whenever a parasite node and its ancestral node both map to the same node on the host tree. Sorting events are inferred from the termination of host tracking by parasite lineages. By mapping to similar host sets, TreeMap aims to derive a reconciliation that maximises cospeciation events along the parasite tree.

The mapping process employed by TreeMap does not readily allow for the detection of host transfer methods. Figure 4.1 illustrates the problem with a simple example. A host transfer event of the parasite on the ancestral species of extant hosts A and B to host C is mistaken as an ancestral duplication event based on the mapping obtained between host sets. The method therefore lacks the means to distinguish a host transfer from an ancestral duplication followed by lineage sorting. This ambiguity increases the number of potential solutions. Initial developments of the reconciliation method avoided consideration of host transfer events in an effort to reduce the complexity of the problem. This was found to have an adverse effect on the reconciliation process: TreeMap reconciliations failed to produce the maximum number of cospeciation events possible due to the postulation of additional duplication-lineage sorting events required to resolve the incongruent branching patterns. Page (1994a) therefore suggested invoking host transfer events that permitted an increase in the overall degree of cospeciation observed between the host and parasite trees. Under this scheme, parasite lineages postulated as host transfer events had to be removed from the parasite tree to allow mapping of the remaining parasite nodes to the host tree. Since the removal of colonising lineages also resulted in the removal of their descendant lineages, and therefore precluded further characterisation of subsequent events, Page (1994b) proposed an alternative solution

that incorporated host transfer events into the mapping process. For each host transfer event considered, the host set assigned to the colonising parasite lineage is removed from its ancestral lineage so as to inform the mapping algorithm that the association between the transferred parasite and its colonised host is not an ancestral one and thereby prevent a mapping of the ancestral parasite node to the ancestral host node. The combination of duplication, host transfer and sorting events then required to resolve the incongruities between the host and parasite trees is simply that which provides the most number of cospeciation events.

TreeMap has since benefitted from the development of another approach, Jungles (Charleston 1998), which is now implemented in TreeMap 2. Jungles analyses the same information as TreeMap i.e., a host tree, parasite tree and the associations between the terminal taxa, to construct a network of optimal solutions (an example is provided in Figure 4.2, reprinted from Charleston (1998: Fig. 7)) for the evolutionary history of associations between the host and parasite groups. In this network each vertex describes a potential association of a parasite node with the host tree; this association may either be with a node or an edge in the host tree. Direct mappings of parasite tree nodes to host tree nodes are indicative of cospeciation events. Duplication and host switching events are both assumed to occur in the absence of a host speciation event and so are inferred whenever parasite tree nodes map to host tree branches. The arcs of the jungle link together vertices (i.e., associations) of the ancestral parasite species with those of their descendant species thus maintaining the correct order of parasite speciation events in the jungle. The properties of each arc provide a distinction between a duplication event and a host switch event: if the host (h_a) associated with the ancestral parasite species is itself an ancestor of the host (h_b) associated with the descendant parasite species, then a duplication event will be inferred; if h_a is not an ancestor of h_b, a host switching event is inferred. Sorting events are inferred whenever the path between h_a and h_b traverses additional host species.

To derive the optimal reconstruction from this network of potential associations, event costs are assigned and the lowest scoring reconstruction is extracted via a dynamic programming algorithm. The most parsimonious solution, i.e., the one with

the maximum number of cospeciation events, is achieved by assigning a negative cost to cospecation events and positive costs to non-cospeciation events. The default costs do not differentiate between duplication, lineage sorting and host transfer, though host transfer events may be assigned a higher cost – if they are believed to occur less frequently than duplications and sorting events. The optimal reconstruction of ancestral host and parasite associations determined by Jungles is therefore highly dependent on the event costs assigned.

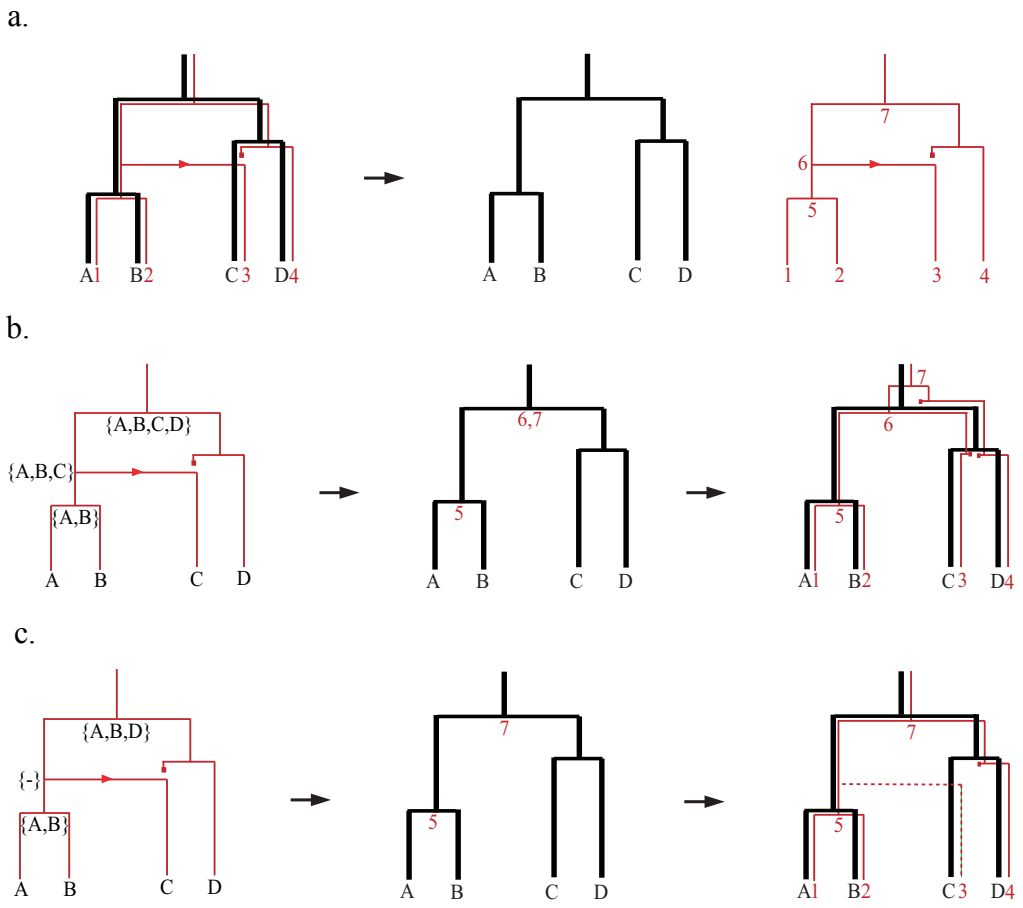
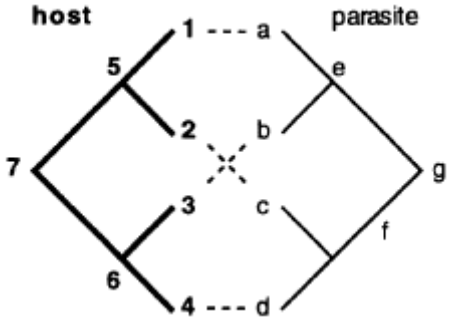


Figure 4.1: The difficulties of inferring host transfer events in TreeMap. a) A four-taxon host tree (black) overlaid with the phylogeny of the associated parasite group, which fails to track the host tree entirely due to a host transfer event from the ancestor of hosts A and B to host C. The individual host and parasite phylogenies are shown on the right with the ancestral parasite nodes numbered. b) Using the TreeMap method, a post-order tree traversal is performed to assign host sets to the internal nodes of the parasite tree. The mapping dictates that nodes 6 and 7 of the parasite tree both map to the root node of the host tree, the observed phylogenetic incongruity is therefore interpreted as being due to a duplication event (at node 7) followed by 3 lineage sorting events. The possibility of a host transfer event cannot be considered under this scheme. c) Using Page's modification to allow consideration of host-switching lineages, we now explicitly consider parasite taxon 3 to have been acquired by host C via a host transfer event, therefore host C is removed from all ancestral nodes of parasite 3 and parasite node 6 is "undefined". The reconciliation produced from this mapping then presents the true sequence of events.

As with TreeFitter, determining the best costing scheme for a host-parasite system presents substantial difficulty for the user since this information cannot be empirically derived. Under the default event costs, where the non-cospeciating events are weighted equally, the host switch vs. duplication-lineage sorting problem will arise at all conflicting branching points. As a result, it is highly likely that multiple equally parsimonious solutions will be obtained, which may even differ to the extent that the set of cospeciating nodes proposed in each solution do not overlap. To obtain a single solution one may have to investigate a range of event costs: any reconstruction found to be optimal under a number of costing schemes should be favoured as the most probable explanation of the incongruent phylogenies. The greater the degree of topological incongruence between a parasite tree and its host's tree, the more combinations of events that serve as plausible explanations, making it increasingly difficult for these methods to derive a unique solution for ancestral association mechanisms without additional information to discern between alternative solutions.

a.



b.

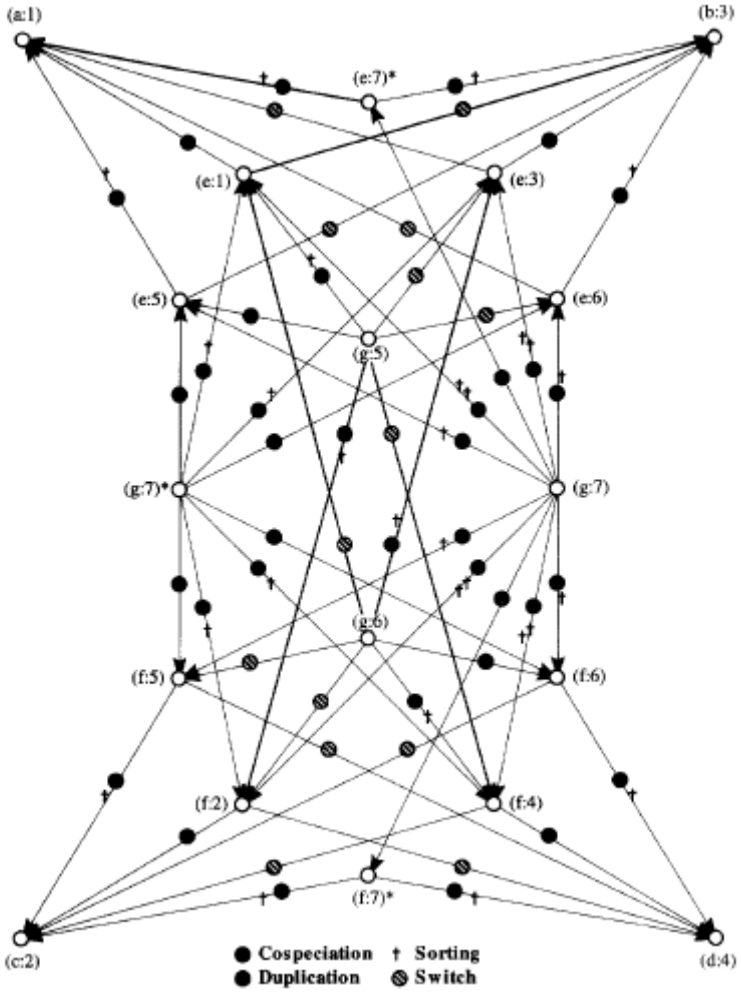


Figure 4.2: Analysing host-parasite associations using Jungles. a) A tanglegram depicting the respective cladograms of associated host and parasite taxa and the associations (dotted lines) between the terminal taxa. b) A jungle constructed from the above tanglegram, displaying all potentially optimal reconciliations of parasite nodes with the host tree. Reprinted from Charleston (1998: Fig. 7).

4.1.1.2 Statistical methods in host-parasite cophylogenetic analysis

The methods described above are commonly used to predict ancestral host-parasite association mechanisms; however, they all lack a sound statistical basis upon which the predicted events may be evaluated. A statistical reconstruction estimating the probability of reconstructions would be more preferable. The maximum likelihood and Bayesian methods applied in phylogenetics may be extended to cophylogenetic analyses to provide statistical interpretations of host-parasite phylogenetic incongruities. A few developments have already been made in this area.

4.1.1.2.1 Statistical analysis of evolutionary distances

Correlations in evolutionary distances of associated host and parasite taxa may provide an indication of the extent of cospeciation: if there is complete or substantial cospeciation among the two groups then we would expect evolutionary distances between taxa in the parasite tree to be proportional to those of the associated taxa in the host tree. Two different tests of cospeciation have been proposed based on the analysis of evolutionary distances. To identify cases of significant cospeciation, both methods test the null hypothesis that the host and parasite species are randomly associated and therefore do not share a coevolutionary history. Only statistically significant correlation values between host and parasite evolutionary distances will cause rejection of the null hypothesis and provide significant evidence in favour of cospeciation.

4.1.1.2.1.1 The Mantel test

The Mantel test (Mantel 1967) examines the evidence for host-parasite cospeciation by evaluating the extent of correlation between the evolutionary distances of extant taxa. Hommola et al. (2009) modified the Mantel test to allow multiple associations among host and parasite species. Host and parasite distance matrices, D^H and D^P, indicating the distance (either phylogenetic or observed) between all pairs of terminal taxa in the host and parasite trees are constructed. The

associations between terminal host and parasite species are then considered in pairs: for each host-parasite interaction, the distance between the corresponding hosts is recorded in one vector and the distance between the corresponding parasites is recorded in another vector. The test then calculates Pearson's correlation coefficient, r^{obs}, for the host and parasite distance vectors. Under the null hypothesis of independent evolution, no significant correlation is expected between the distances. Significance is determined by permuting the terminal taxa on the host and parasite phylogenies and repeating the analysis using the observed host-parasite associations to obtain a distribution of correlation coefficients for the randomised data. If r^{obs} is greater than this distribution at the determined α level, this provides evidence of significant correlation in the host and parasite distances and the null hypothesis is accordingly rejected in favour of cospeciation. The method provides a simple test of cospeciation and does not offer further characterisation of phylogenetic incongruities.

4.1.1.2.1.2 ParaFit

A slightly-more elaborate analysis of evolutionary distances is provided by ParaFit (Legendre, Desdevises and Bazin 2002). ParaFit assesses the extent of host-parasite cospeciation by investigating the 'fourth-corner problem' (Legendre 1997). Given a matrix **A** detailing the presence/absence of parasite associations with a group of host species, the fourth-corner problem refers to the estimation of the relationship between specific characteristics of the associated hosts and parasites and determination of whether the estimated parameters of the relationship are indicative of a non-random association. Patristic distance matrices calculated from the parasite and host phylogenies are converted to principal coordinate matrices (**B** and **C**, respectively) from which ParaFit calculates the fourth-corner matrix, **D**:

$$\mathbf{D} = \mathbf{CA'B} \tag{4.1}$$

The elements of **D** (d_{ij}) therefore consist of the cross products of principal coordinates of associated host and parasite taxa and are evaluated for evidence of cospeciation by deriving the test statistic ParaFitGlobal:

$$\text{ParaFitGlobal} = \sum (d_{ij})^2 \tag{4.2}$$

The significance of the ParaFitGlobal is determined by randomising each row in A, i.e., randomising the associations, computing D for the randomised associations and

hence obtaining the distribution of ParaFitGlobal under the null hypothesis of random association of host and parasite species. Rejection of the null hypothesis indicates significant evidence for cospeciation of the parasites and their hosts. ParaFit includes a further test to identify which terminal associations in particular display significant evidence of cospeciation. This is achieved by removing a terminal association (k) and determining whether the difference between the new value of the test statistic ParaFitGlobal(k) is significantly worse than that of ParaFitGlobal. If the removed association was a consequence of cospeciation, then it is expected that its removal will decrease the value of ParaFitGlobal. In this manner ParaFit provides a means of identifying which terminal associations are due to cospeciation and which are not, thus allowing incongruent host and parasite phylogenies to be pruned down to the cospeciating lineages. Although ParaFit is one of the few methods accommodating multiple associations between host and parasite species, the method can only evaluate terminal associations and therefore further examination of ancestral associations is not possible.

4.1.1.2.2 Likelihood ratio tests of cospeciation

A series of likelihood ratio tests were proposed for evaluating the evidence for cospeciation between incongruent host and parasite trees (Huelsenbeck, Rannala and Yang 1997). Applying the methods of Huelsenbeck and Bull (1996) to examine topological congruence among data partitions, these tests determine the statistical significance of the difference in maximum likelihood when various phylogenetic parameters are optimised over the host and parasite sequences together and when they are independently optimised for each data set. The proposed tests systematically examine the null hypothesis of cospeciation, and hence, congruence, between host and parasite phylogenies at the topological, temporal, and substitution rate levels.

Under the null hypothesis of cospeciation, in which parasite speciation events occurred only in response to speciation events of the associated host, the estimated host and parasite phylogenies should be topologically identical. Observed incongruities may be due to stochastic error or rare occurrences of non-cospeciating

events. The likelihood ratio test of topological congruence therefore examines whether the observed phylogenetic incongruities are significant or whether they appear more likely to be artefacts of the estimation process. Under the null hypothesis, the data sets are analysed together with the tree topology constrained to be identical across both data sets whilst all other parameters are free to vary. Under the alternative hypothesis of no cospeciation, i.e., random association of hosts and parasites, the data sets are analysed independently with the topological constraint lifted. The ratio of the maximum likelihood under each hypothesis is obtained and its significance is determined by parametric bootstrapping under the null hypothesis. Significant support against the null hypothesis indicates that genuine phylogenetic incongruities exist between the host and parasite groups, and that the non-cospeciating events may obscure detection of any coevolutionary history that exists between the species. If, however, the observed phylogenetic differences do not provide enough evidence to reject the null hypothesis, the possibility of cospeciation exists and can be further tested.

Assuming identical branching patterns in the host and parasite phylogenies, one can then examine the degree of concordance between the times of corresponding branching points, i.e., speciation times. In the absence of actual speciation times, if both groups display support for a constant rate of evolution among their respective data sets, the amount of evolution observed in each lineage will be proportional to the amount of time between speciation events. The branch lengths can therefore be evaluated for temporal congruence of host and parasite speciation events. Under the null hypothesis, the topologies and corresponding branch lengths are constrained to be identical between the host and parasite trees. Under the alternative hypothesis, the constraint of identical branch lengths is relaxed; however, the topological constraint remains since topological identity is assumed. Significance of the likelihood ratio test of node times may be determined by parametric bootstrapping, as above, or more simply by comparing twice the value of the test statistic against a χ^2 distribution with n-2 degrees of freedom, where n is the number of taxa (this is the difference in the number of nodes for which speciation times are estimated in the unconstrained and constrained analyses).

Rejection of the null hypothesis of identical node times suggests a scenario in which the host and parasite phylogenies have identical topologies but different speciation times. Such a scenario cannot be reconciled with a strictly cospeciating mechanism of parasite evolution and indicates the influence of additional parasite diversification mechanisms on parasite evolution. The relationship between relative speciation times of the host and parasite nodes will indicate the likely nature of the mechanism. If significant support against temporal congruence is not found, cospeciation of hosts and parasites may be inferred.

The final test proposed evaluates the evidence for identical evolutionary rates given identical topologies and branch lengths. Tests of identical evolutionary rates will be more applicable when host and parasite taxa are represented by the same family of genes or proteins. A failure to reject the null hypothesis of identical evolutionary rates, suggests substantial synchrony in the coevolutionary history of the host-parasite system; however, rejection of the null hypothesis in this test does not refute the former conclusion of cospeciation.

These likelihood ratio tests are applicable when investigating associations limited by one-host-per-parasite, but cannot readily accommodate multiple associations between species. There are several solutions proposed to deal with this: one can increase the data set by adding replicate sequences for taxa involved in multiple associations, although this may affect tests of node times and evolutionary rates; eliminate the complexity by removing taxa involved in multiple associations; or reduce the multiple associations in a stepwise manner to identify non-cospeciating taxa (Huelsenbeck, Rannala and Yang 1997). Another approach may be to reduce multiple associations to a one-to-one correspondence and evaluate the multiple associations independently. None of these methods are ideal, however, as they involve modifications of the real data set and thus, a loss of information.

In addition to the likelihood ratio tests above, Huelsenbeck and colleagues applied Bayesian phylogenetic methods to determine the posterior probability of identical topologies between gophers and their louse parasites. The method involves individual estimation of the host and parasite posterior distributions and summation of the products of posterior probabilities for all pairs of identical topologies sampled. The

advantage of this method over the likelihood ratio tests is that the Bayesian approach accounts for all possible identical topologies not just the most optimal topology. Inferences of topological congruity therefore average over uncertainties in the exact tree topology relating the sequences. In theory, one could then apply the constraint of topological congruence in a subsequent Bayesian analysis of the host and parasite sequences and extract the posterior probability of identical branch lengths from the posterior distributions; however, the large number of rooted trees possible for $n>5$ sequences means that unless knowledge of evolutionary rates is available, the probability of observing identical trees with identical branch lengths is likely to be very small, even if cospeciation has occurred.

4.1.1.2.3 Bayesian estimation of host switching

Host transfer events have generally presented difficulties in cophylogenetic analyses; however, Huelsenbeck, Rannala and Larget (2000) developed a method that specifically evaluated host switching events along the incongruent phylogenies of gophers and their louse parasites. They introduced a host switching rate parameter, λ, into Bayesian phylogenetic analysis of the host and parasite sequences, **H** and **P**. The host switching events are modelled using a Poisson process on the host tree; the number of host switching events, the source branches of each switch, the target branches of each switch, and the times of the host switching events proposed at each state in the MCMC chain are recorded in a vector, **e**. Proposed host switches will cause temporal and topological discordances between the host and parasite trees (unless a host switch occurs between sister taxa in which case topological differences will not be observed). Estimation of the parasite phylogeny under a host switching model is therefore linked to that of the host phylogeny such that in the absence of any proposed host switching events the topologies and speciation times of both trees are identical.

The joint posterior probability density of λ and the host and parasite substitution model parameters, θ_H and θ_P, is then

$$f(\lambda,\boldsymbol{\theta}_{\mathrm{H}},\boldsymbol{\theta}_{\mathrm{P}} \mid \mathbf{H},\mathbf{P}) = \frac{\ell(\lambda,\boldsymbol{\theta}_{\mathrm{H}},\boldsymbol{\theta}_{\mathrm{P}})f(\lambda,\boldsymbol{\theta}_{\mathrm{H}},\boldsymbol{\theta}_{\mathrm{P}})}{f(\mathbf{H},\mathbf{P})} \tag{4.3}$$

where,
$$f(\mathbf{H},\mathbf{P}) = \int \ell(\lambda,\boldsymbol{\theta}_{\mathrm{H}},\boldsymbol{\theta}_{\mathrm{P}})\,\mathrm{d}F(\lambda,\boldsymbol{\theta}_{\mathrm{H}},\boldsymbol{\theta}_{\mathrm{P}}) \tag{4.4}$$

and the likelihood function of λ, $\boldsymbol{\theta}_{\mathrm{H}}$, and $\boldsymbol{\theta}_{\mathrm{P}}$ integrates over host switching events, host speciation times and host phylogenies thus preventing uncertainties in these parameters from influencing the likelihood:

$$\ell(\lambda,\boldsymbol{\theta}_{\mathrm{H}},\boldsymbol{\theta}_{\mathrm{P}}) = \int f(\mathbf{H} \mid \tau_{\mathrm{H}},\mathbf{t}_{\mathrm{H}},\boldsymbol{\theta}_{\mathrm{H}})f(\mathbf{P} \mid \tau_{\mathrm{H}},\mathbf{t}_{\mathrm{H}},\boldsymbol{\theta}_{\mathrm{P}},\mathbf{e})\mathrm{d}F(\mathbf{e} \mid \tau_{\mathrm{H}},\mathbf{t}_{\mathrm{H}},\lambda)\mathrm{d}F(\tau_{\mathrm{H}},\mathbf{t}_{\mathrm{H}}) \tag{4.5}$$

The posterior distribution of **e** enables identification of host lineages associated with high posterior probabilities of host switching events and the times of these events, which can be used to examine specific scenarios of host switching in the parasite phylogeny. The advantage of utilising sequence data in the estimation of host switches is that the method is not restricted to predicting host switches at locations of topological incongruence only. This is observed in the gopher-louse system analysed, where there is a high posterior probability of a host switch between two taxa (*G. b. majusculus* and *G. b. halli*) displaying a phylogenetic pattern which is concordant with the associated host relationships (Huelsenbeck, Rannala and Larget 2000: Fig. 13).

The use of Bayesian MCMC methods allows estimation of a host switching rate and host transfer events of parasite lineages without conditioning these events on one particular tree topology and is therefore highly advantageous for the analysis of virus systems. The evolutionary model can be extended to model specificities such as evolutionary distance-dependent host switching probabilities. The computational time will, however, be much greater than for TreeFitter or TreeMap and this may present an obstacle in analysis of large data sets. An additional limitation is that, in its current implementation, the method is only applicable when there is a one-to-one correspondence of host and parasite taxa. This is because the stochastic nature by which host switching events are modelled may produce host switches within a host if

duplicate/dummy taxa are allowed. Under this assumption it is also difficult to incorporate duplication and extinction/lineage sorting events, which would be required for a more complete statistical evaluation of the mechanisms affecting ancestral lineages of associated hosts and parasites.

4.1.2 A suitable approach for analysis of the PVs

The one-host-one-parasite requirement that is common to most of the above methods presents a dilemma for the analysis of PV-host phylogenetic incongruence since some hosts are infected by multiple PV types. The paraphyletic arrangement of PV types infecting the same host means that it is difficult to select one representative PV type per host species. In many methods the constraint can be circumvented by introducing replicate host sequences; however, the paraphyletic arrangement of PV types infecting the same host means that the introduction of replicate host sequences does little to reduce the complexity of the problem. In humans alone there are over 100 different PV types; these HPV types comprise 5 distinct PV genera, which are distantly related to each other. Thus, to elucidate the mechanisms by which PV associations were formed with humans, at least 5 different HPV types (representing each genus clade) must be considered. Likewise, using representative BPVs would demand a minimum inclusion of 4 different BPVs as the lack of monophyly among BPVs means that the number of host-virus associations cannot be reduced any further without excluding significant diversification events (i.e., of viruses on distinct host species) from the analysis.

4.1.2.1 Previous studies of PV-host phylogenetic incongruence

The general acceptance, among PV researchers, of a strictly codiverging mechanism of PV diversification to new host species means that there is a lack of studies applying cophylogenetic methods to resolve incongruities between phylogenies of the PVs and their hosts. Prior to the publication of Shah, Doorbar and Goldstein (2010) just one case existed: Jackson (2005) analysed a small dataset of 17

PV types from 16 mammalian host species (Figure 4.3, reprinted from Jackson 2005: Fig. 9b) with Jungles. The polyphyletic lineages of human and bovine PVs were reduced to one type per host in order to facilitate detection of codivergence events (although two PV types infecting *Colobus guereza* – CgPV1 and CgPV2 – were retained).

The reconciliation analysis performed both with and without host transfer events, produced a set of potentially optimal solutions with 26 and 24 maximum codivergence events, respectively, both of which were found to be statistically significant numbers. The most parsimonious solution under each model was not derived on account of the fact that the event costs required could not be accurately assigned. The proposed codivergence events in these solutions included the chimp-bonobo, human-chimp, human-monkey, cat-dog, human-rabbit, and cervidae (deer-elk) PV divergences. The correlation in genetic distances for codiverging host and parasite taxa (Figure 4.4, reprinted from Jackson 2005: Fig. 12) is not as convincing, however, with $r^2 = 0.596$ (although error bars are not shown). This may indicate a conflict in the support for cospeciation between the branching patterns and corresponding speciation times, as only the topological structure was examined for the reconciliation.

4.1.2.2 Topological comparisons lack discriminative power

The reconciliation methods, TreeMap and Jungles, currently provide the only method of cophylogenetic analysis that models codivergence (cospeciation), host transfer, prior divergence (parasite duplication) and lineage sorting events and have therefore become a popular tool to resolve phylogenetic incongruities between hosts and their associates (i.e., the parasite or virus). Howver, the statistical power of tree reconciliation methods is affected by a bias towards codivergence/cospeciation and, consequently, these methods will struggle to identify trees displaying 'false congruence' (Jackson 1999).

A key requirement of tree-fitting/reconciliation methods is "if the parasite and host trees are identical (isomorphic) when the labels of the parasite terminals are exchanged with the labels of their associated hosts, then the method must produce

only one optimal solution, fitting the elements in the parasite tree to the corresponding elements in the host tree" (Ronquist 2002, p.28). In other words, when complete topological congruence is observed, the reconciliation should consist of only codiverging events. However, whilst Fahrenholz' rule states that strict codivergence produces congruent topologies, converse is not always true. Instances of false congruence (also referred to as 'pseudo-cospeciation', Hafner and Nadler 1988) arise when congruent branching patterns, which are the result of non-cospeciating (non-codiverging) events acting on the respective lineages, are mistakenly attributed to cospeciation (codivergence).

A simple example of how false congruence can arise for two sister host taxa (A and B) parasitized by sister parasite lineages (1 and 2, respectively) is presented in Figure 4.5. When the parasite clade is overlayed on the host tree (Figure 4.5a), the identical topologies suggest that the parasite lineages have codiverged with their associated hosts. However, the observed associations and phylogenetic relationships can also be explained either by invoking a duplication event followed by two sorting events (Figure 4.5b) or by invoking a host transfer event (Figure 4.5c). The observed associations may in fact stem from a more complicated scenario involving a combination of these events but only the last event to occur would be inferable.

Simulation studies (Charleston and Robertson 2002; De Vienne, Giraud and Shykoff 2007) have generated instances of false congruence between host and parasite phylogenies by a mechanism of preferential host switching i.e., between closely related hosts. Reconciliation methods applied to characterise diversification mechanisms along the simulated parasite phylogenies falsely postulate a significant number of cospeciation events. This illustrates the difficulties faced by any method that attempts to characterise the mechanisms behind incongruent host and parasite trees by only examining the respective topological structures. Considering the fact that more than one possible explanation exist for congruent branching patterns, the number of possible solutions for incongruent topologies must be even greater. Discerning between these solutions requires knowledge of the likelihood of each event; however, this is rarely known.

Figure 4.3: A PV-host tanglegram generated using Jungles. A representative data set was used, excluding multiple bovine and human PV types each of which span more than one clade and therefore contribute significantly to phylogenetic incongruities between the PVs and their hosts. Reprinted from Jackson (2005: Fig. 9b).

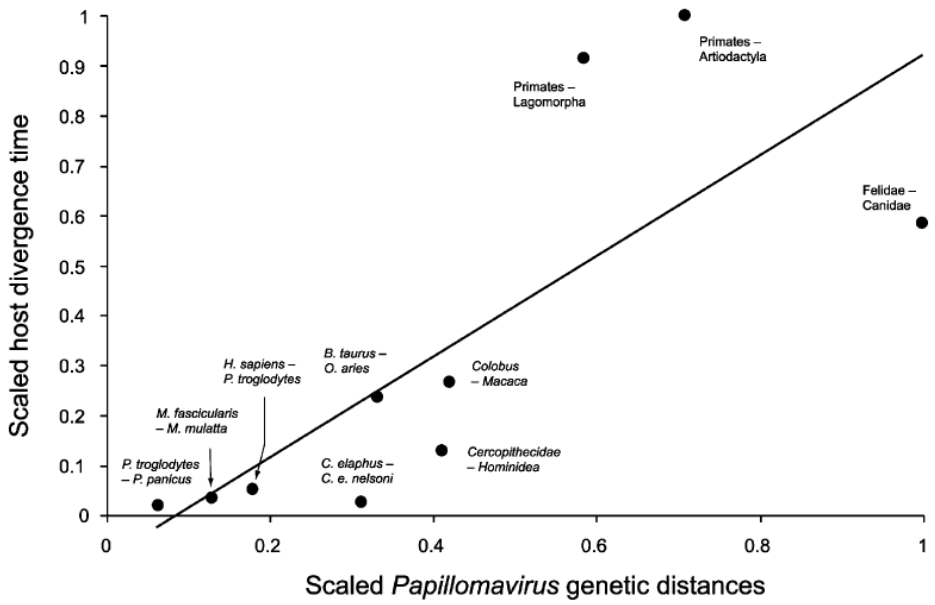


Figure 4.4: The correlation between genetic distances of Jungles-predicted cospeciating host and PV nodes; r^2 = 0.596. The number of data points appears too few to permit confident inference of a linear correlation between host and virus speciation times, whilst the absence of a molecular clock of evolution among the diverse PV lineages renders the use of PV genetic distances invalid. Reprinted from Jackson (2005: Fig. 12).

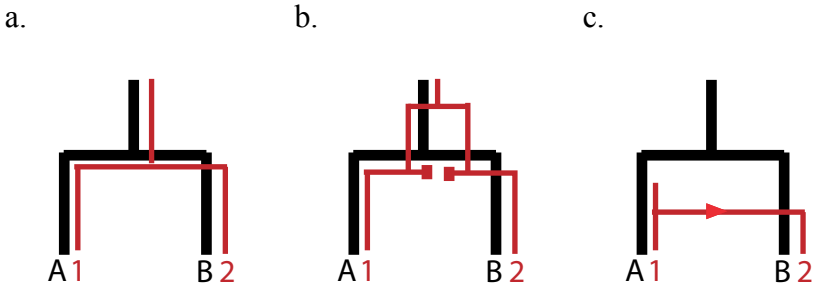


Figure 4.5: Three different explanations for topological congruence between sister host (A and B) and parasite (1 and 2) lineages. a) cospeciation of the host and parasite lineages; b) a duplication/adaptive radiation of the ancestral parasite lineage associated with the ancestral host species, followed by two sorting events during host speciation resulting in each descendent host inheriting only one of the two parasite lineages; or c) transfer of the parasite lineage associated with one host to another, in this case unoccupied, host species.

4.1.2.3 Utilisation of divergence times

Charleston (2002) presented an example of false congruence among the phylogenies of the simian lentiviruses and their primate hosts by utilising the times of host and virus divergence. Topological congruities previously attributed to codivergence of the primates and the lentiviruses (e.g., Beer et al. 1999) were contradicted by vast disparities (on the order of millions of years) between the corresponding divergence times, which should be similar under a scenario of codivergence. Page (1990b) evaluated temporal congruence between host and parasite speciation events to resolve ambiguities from a reconciliation analysis of the gopher and louse data set and, unsurprisingly, found that the temporal information contradicted the evolutionary scenarios produced by the reconciliation. Speciation times have also been utilised in the methods of Huelsenbeck, Rannala, and Yang (1997) and Huelsenbeck, Rannala, and Larget (2000) to evaluate evidence for codivergence and host transfer, respectively.

A comparison of the relative times of branching events is more informative than the branching patterns as the various events that cause diversification of parasite lineages each occur at different times relative to speciation of the associated hosts. This is evident in Figure 4.5: in a) parasite duplication on a host species occurs *prior* to the divergence event that produced the extant parasitized hosts, in b) parasite diversification via host transfer occurs *after* the speciation event of the extant parasitized hosts but for c) cospeciating lineages, the host and parasite speciation events occur within a similar time frame (it would be naïve to assume that host and parasite speciate at exactly the same time). The distinctive host-parasite temporal relationships characterising each event offer an alternative approach to resolve host-parasite phylogenetic incongruities.

In the study of virus data sets, however, utilisation of temporal data is difficult as viruses do not leave fossils from which the ancestral divergence times may be estimated. Viral divergence times may still be estimated from evolutionary distances if the evolutionary rate is known. Various estimates of the PV evolutionary rate have been derived from different subsets of PVs (Van Ranst et al. 1995; Tachezy et al.

2002; Rector et al. 2007; Herbst et al. 2009) but the individual estimates fail to converge on the same rate and, more importantly, the estimates were obtained assuming correspondence between viral and host divergence times, i.e., making the assumption of codivergence, which is the hypothesis that we are interested in testing. In addition, I found significant evidence against a constant rate of evolution among our data set of heterogeneous PV types, thus prohibiting examination of virus-host genetic distances in place of actual divergence times (as was performed by Jackson 2005).

To characterise the events that influenced PV diversification and the formation of new host associations, I chose to develop a method in which the temporal relationships between corresponding host-virus divergence events could be evaluated. In the absence of known estimates of PV divergence times and a constant rate of evolution among PV lineages, I took advantage of Bayesian methods to estimate the posterior distribution of PV divergence times at each node. Besides providing a statistical analysis, the Bayesian phylogenetic approach is advantageous in that it allows sampling of different phylogenies, thus accommodating topological uncertainties in the virus phylogeny. If the Bayesian estimation is performed correctly, each PV divergence, i.e. internal node, will be sampled according to its probability of being correct. The posterior probability density of divergence times at nodes with high posterior probabilities can then be compared to the corresponding host divergence times to infer the likely diversification mechanism of the viral lineages and thereby elucidate the likely mechanisms behind the observed phylogenetic incongruities. The Bayesian phylogenetic method implemented in the BEAST software was used to perform the analysis.

4.2 Method

4.2.1 Sampling of viral divergence times

My initial approach was to use the sampling algorithms incorporated in Bayesian phylogenetic methods to investigate how often sampled divergence times for the papillomavirus sequences corresponded to the known host divergence times. In the absence of an accurate evolutionary rate for the viruses, however, the viral divergence time is equally likely to be at any time in the past; the probability that the estimated viral divergence times correspond to the rather narrow interval of the host speciation time is extremely remote, as is the probability that the viral divergence time occurred after the origin of life or within the lifetime of the universe. Thus, the estimation of divergence times in the PV tree required calibration information either in the form of an evolutionary rate or by specifying times for some of the divergence events. However, for the PVs there was no reliable evolutionary rate estimate available and it would be illogical to apply fixed constraints to node times that imply assumptions about cospeciation that presuppose the relationships that we are interested in investigating.

The solution arrived at to deal with this problem was to apply a biased sampling approach, based on the importance sampling techniques used in stochastic simulations, to the sampling of divergence times. Importance sampling provides a way of guiding random sampling algorithms to reduce the variance of the sampled points such that meaningful inferences about the process under investigation can be made from the resulting distributions. The aim is to "avoid taking sample points where the value of the function is negligible and to concentrate the sample points where the value of a function is large" (Borcherds 2000). For instance, we can be sure that speciation events of the PV sequences involved in the analysis did not occur in the last few years, and nor did they occur prior to the existence of eukaryotes, so it makes sense to discourage the chain from sampling at these extremities.

For the sampling of PV divergence times, it is assumed that codivergence of the virus with its host is common, and therefore MCMC sampling of viral divergence times is biased in favor of large number of codivergence events. However, an

important feature of this approach is that the assumption is made *without* fixing any specific viral divergences to be codiverging only (as is the case when using known divergence times to calibrate a tree). Thus, sampled viral divergence times may or may not correspond to those of the associated host. The overall assumption of codivergence allows the identification of PV nodes which show significant evidence against this assumption. This is achieved by imposing a penalty term in the log-likelihood calculation for each node where codivergence is violated.

Violations of codivergence occur when the sampled divergence time for a node does not coincide with the speciation time of the corresponding host. When host and virus divergence times coincide, no penalty is imposed on the likelihood. When a sampled virus divergence time does not match that of the corresponding host, a penalty will be imposed on the log likelihood of the tree thus discouraging substantial sampling of times that disagree with those of the host. Significant violations of codivergence at a specific node will only be observed when adherence to codivergence has a more severe effect than that imposed by the violation penalty on the overall likelihood. This approach should result in enhanced sampling of trees and timings where codivergence is common, but avoids the imposition of any fixed constraints.

Importance sampling techniques enable the random sampling process to focus predominantly on areas of higher density; however, it is necessary to account for the resulting bias in the sampled points by proportionally downweighting any sample points which we biased for and proportionally upweighting sample points we biased against (Borcherds 2000). Unfortunately, in the sampling of PV divergence times, this would result in a situation similar to that encountered prior to the imposition of the biases; the calculation, appropriately corrected, would again be dominated by the vast space of possible trees where cospeciation occurs at some random time in the past. The resulting posterior probabilities would be too small to permit inference of ancestral events.

An alternative approach is to consider violations of codivergence at individual nodes as a measure of the evidence against codivergence, given the overall bias towards cospeciation. At each node, the posterior density of times sampled before the

host speciation times will represent the magnitude of violations of codivergence in favour of prior divergence, and the posterior density of times sampled after the host speciation times represents the magnitude of violations of codivergence in favour of host transfer. The observed violations will represent a conservative estimate due to the general assumption of codivergence The magnitude of these violations can then be translated into statistical significance through a parametric bootstrapping (Monte Carlo simulation) procedure.

Parametric bootstrapping allows us to assess the probability that a similar or greater degree of violation would be observed if cospeciation had in fact occurred at that node. This is achieved by constructing synthetic data modelled on a papillomavirus phylogeny in which speciation times of all of the nodes under investigation have been adjusted to conform to cospeciation. For nodes which show significant violations of cospeciation, the nature of the mechanism will be revealed by the timing of viral divergence relative to that of host divergence – significant violations prior to host speciation indicate prior divergence, whilst violations after host speciation indicate host transfer.

The biased distribution applied to each node takes the form of a uniform distribution within the bounds of the host speciation range, outside these bounds it is flat but assumes a non-zero value (Figure 4.6). The same penalty is applied to divergence times on either side of the host speciation range as I am not making any further assumptions about the relative likelihood of host transfer over prior divergence and vice versa. A flat distribution for all times outside the host-speciation range means we are not considering time-dependent effects on the probability of non-cospeciating events, i.e., that a host transfer may become more unlikely with increasing time after the host speciation.

The analysis was performed using the BEAST program for Bayesian phylogenetic analysis as the BEAST algorithm accommodates the sampling of different topologies during a run and the estimation of divergence times under a variable rates model across lineages, both of which are required in the analysis of the PVs. To implement the log likelihood penalty in BEAST, I modified the getLogPriorComponent method of the class dr/inference/prior/UniformParameterPrior.java such that the likelihood value returned when the sampled parameter is outside the bounds of the uniform

distribution is equal to the natural logarithm of the penalty value. To allow sampling outside of the host speciation range I modified the isWithinBounds method of dr/inference/model/Parameter.java to return TRUE even when the sampled time lay outside the bounds of the uniform distribution, i.e., the host speciation range.

To determine a suitable penalty for the data set, i.e. one that permits sampling outside the host speciation times but still restricts the sampling of times to the desired time scale, I experimented with values within the range $0 < x < 1$. Log likelihood penalties of ln(0.5) and ln(0.1) were found to be too weak to restrict the sampling of times to within the host speciation range whilst the ln(0.0005) penalty was found to be too stringent with the result that independent chains failed to converge over the number of states sampled. The intermediary penalties ln(0.05) and ln(0.005) did allow sampling within the time scales of interest, however, and I observed convergence of multiple chains, making these penalties ideal for this analysis.

Owing to the phylogenetic incongruities observed between the E1 and L1 genes of the PVs, I analysed each gene independently. The biased sampling of divergence times cannot be performed at all internal nodes of the PV gene trees as the inclusion of multiple human and bovine PV types, means that the trees contain clades where all PV divergences appear to have involved the same host species. For such nodes there is therefore no corresponding host speciation event that has occurred and therefore no host speciation time which can be utilised to bias the sampling of the viral divergence times. Biased sampling, using host speciation times, was performed specifically at PV nodes which formed the most recent common ancestor (MRCA) to PV lineages from different hosts and for which the corresponding host divergence times were available. By a process of visual inspection of the E1 and L1 gene MAP phylogenies obtained in Chapter 3, I identified 19 nodes (excluding the root node), in each gene tree, which met the first criterion. These nodes are highlighted in Figure 4.7. *A priori* knowledge of the viral phylogeny (or the posterior distribution of trees) is necessary in order to identify nodes at which biased distributions can be applied, but some degree of topological uncertainty in the phylogeny can be accommodated as we restrict bias sampling to those nodes which can be confidently identified by their high posterior probabilities.

Of the 19 nodes identified in the E1 and L1 gene MAP trees, posterior probabilities greater than 0.90 were observed at 17 and 15 of these nodes, respectively. In each gene tree, the lowest node posterior probability was observed at the human-monkey PV split for which posterior probabilities of 0.6 were observed in both gene trees. The low posterior probability associated indicates that there is uncertainty in the phylogenetic placement of the monkey PV type RhPV1 (now MmPV1). However, an examination of multiple independent MCMC chains revealed consistency in the MAP topology and posterior probabilities of this clade. Thus, I proceeded with the biased sampling at this node in spite of its lower posterior probability.

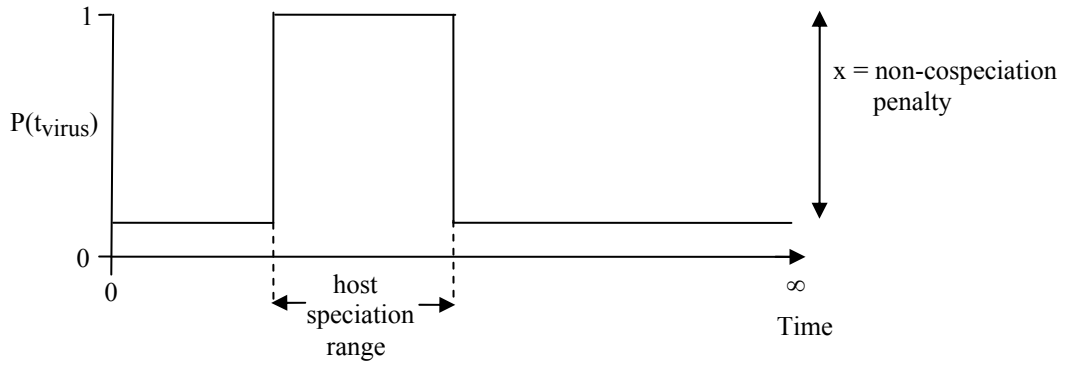


Figure 4.6. Profile of the biased distribution applied to test cospeciation at viral nodes. The distribution is biased towards cospeciation and therefore all virus divergence times sampled from within the range of the host speciation times is assigned a probability of 1. All times sampled outside this range are assigned a lower but non-zero probability thus penalising but not prohibiting sampling of non-cospeciating times. Log likelihood penalties of ln(0.005) (x = 0.995) and ln(0.05) (x = 0.95) were applied in turn.

For each node selected for biased sampling, I specified the corresponding set of terminal taxa and assigned the modified "uniform distribution prior" on the node age;

the upper and lower bounds of the distribution were dictated by the corresponding host speciation time range obtained from molecular estimates along a mammalian supertree (Bininda-Emonds et al. 2007). The data set analysed by Bininda-Emonds et al. (2007) comprised more mammalian species (a total of 4,510) than previous studies that have attempted to date the mammalian (or vertebrate) tree (Kumar and Hedges 1998; Springer et al. 2003). This tree also includes all host species' included in our data set, allowing us to investigate diversification mechanisms at the maximum number of inter-host nodes. However, it is noted that there are errors in this analysis which may have affected the estimated times.

In the estimation of the mammalian tree topology, Bininda-Emonds and colleagues used a supertree approach in which the full tree, comprising all species in the data set, is derived by combining 'source' trees generated from subsets of the full data set. The set of source trees are represented in a matrix, from which the supertree is estimated by parsimony analysis. The supertree approach deconstructs the task of estimating phylogenies for large data sets into smaller, more manageable chunks. However, the fact that the supertree is not obtained from direct analysis of the entire data set as a whole, means that not all phylogenetic relationships among the taxa are analysed and consequently, the estimated phylogeny may be an inaccurate representation of the true relationships.

The methods of Bininda-Emonds and colleagues have been further criticised by more recent attempts to date the mammalian tree (Meredith et al. 2011; Dos Reis et al. 2012). These criticisms highlight potential errors in the source trees chosen for construction of the supertree. In estimating node ages along the tree, they point out that the authors do not appropriately account for lineage rate variation, for uncertainty in fossil calibrations and for uncertainty in branch lengths. Newer estimates of mammalian divergences, obtained from simultaneous analysis of the entire data set in phylogenetic estimation and use of Bayesian MCMC methods to estimate divergence times along the tree, propose younger ages (closer to fossil estimations) for intra-ordinal divergences within the placental mammalian clades (e.g. within the Lagomorpha, Primates, Carnivora, Artiodactyla, etc.) compared to those estimated by Bininda-Emonds and colleagues. Future studies requiring molecular divergence dates for the mammalian hosts should look to the recent estimates (e.g., Dos Reis et al. 2012) for a more accurate temporal comparison of virus-host divergence.

The high posterior probabilities observed at the majority of the E1 and L1 gene tree nodes (Appendix A6 and A10) meant that the nature of the MRCA of the corresponding host species could be guessed with a high degree of certainty and therefore simplified the assignment of biased distributions at these nodes. For example, the E1 and L1 MAP phylogenies both contain a clade of δ PVs, which consists of PV types infecting species from the Cervidae (deer, elk and roe deer) and Bovidae families. In both gene trees, the δ cervid PVs cluster together (p=1.0) and the δ Bovidae PVs cluster together (p=1.0). Thus, I used the estimated times for the Cervidae-Bovidae host divergence to bias the divergence times sampled at the node joining the δ Cervidae PV clade and the δ Bovidae PV clade together.

Where topological arrangements presented a more complicated scenario, I applied multiple biases covering the different combinations possible. In the E1 gene tree for example, the human γ PVs, the murid π PVs, the canine PV type 2 (CPV2), and the bovine PV type 7 (BPV7) all cluster together with high posterior probabilities (p=1.00); however, the relationships within this clade are incongruent with those of the corresponding hosts: we would expect PVs infecting the Euarchontoglires (murid and human) and the Laurasiatheria (canine and bovine) to each cluster together and the corresponding host speciation times could be used to bias the murid-human, canine-bovine and Euarchontoglire-Laurasiatheria PV divergence times. This presents a tricky situation for the application of host speciation times. However, the fact that the analysis is not based on a fixed topology provides some flexibility in the specification of biased distributions. I therefore considered the 6 possible pairings of the 4 host lineages within this polyphyletic clade (e.g., dog-bovine, dog-murid, dog-human) and assigned individual biased distributions to the divergence times of each pairing. I also considered the 3 different pairings of the cat, dog, and raccoon λ PVs due to the differences between the two trees in the topology of this clade. The host speciation times applied to bias the divergence times of PV nodes highlighted in Figure 4.7 are shown in Table 4.1.

This modified BEAST analysis was performed on both the E1 and L1 genes from the PV dataset, with the same model specifications as before and the biased prior distributions on the ancestral node ages of the specified subsets of PV taxa. The parrot

PV (PePV), which shows the greatest evolutionary distance to all mammalian PVs, was specified as the outgroup, and an additional bias corresponding to the mammal-avian divergence was specified at the root. For each of these nodes, the MCMC chains were examined to determine the proportion of the sampled states in which the node age agreed with the associated host speciation time, the proportion in which the node age pre-dates host speciation (in agreement with prior-divergence) and the proportion in which the node ages post-date host speciation (in agreement with host transfer). Each BEAST analysis was run for 30,000,000 generations with states sampled every 1,000 generations. For each gene, I ran three chains, to ensure convergence of the chains. Convergence was determined by calculation of the PSRF statistic. Sampled components of all chains had PSRF values close to 1.00; the average PSRF was 0.99 (s.d. = 0.015). In addition, the ESS values of sampled parameters in all chains were greater than 500, indicating sufficient number of independent states for inferences to be made from the sampled chains.

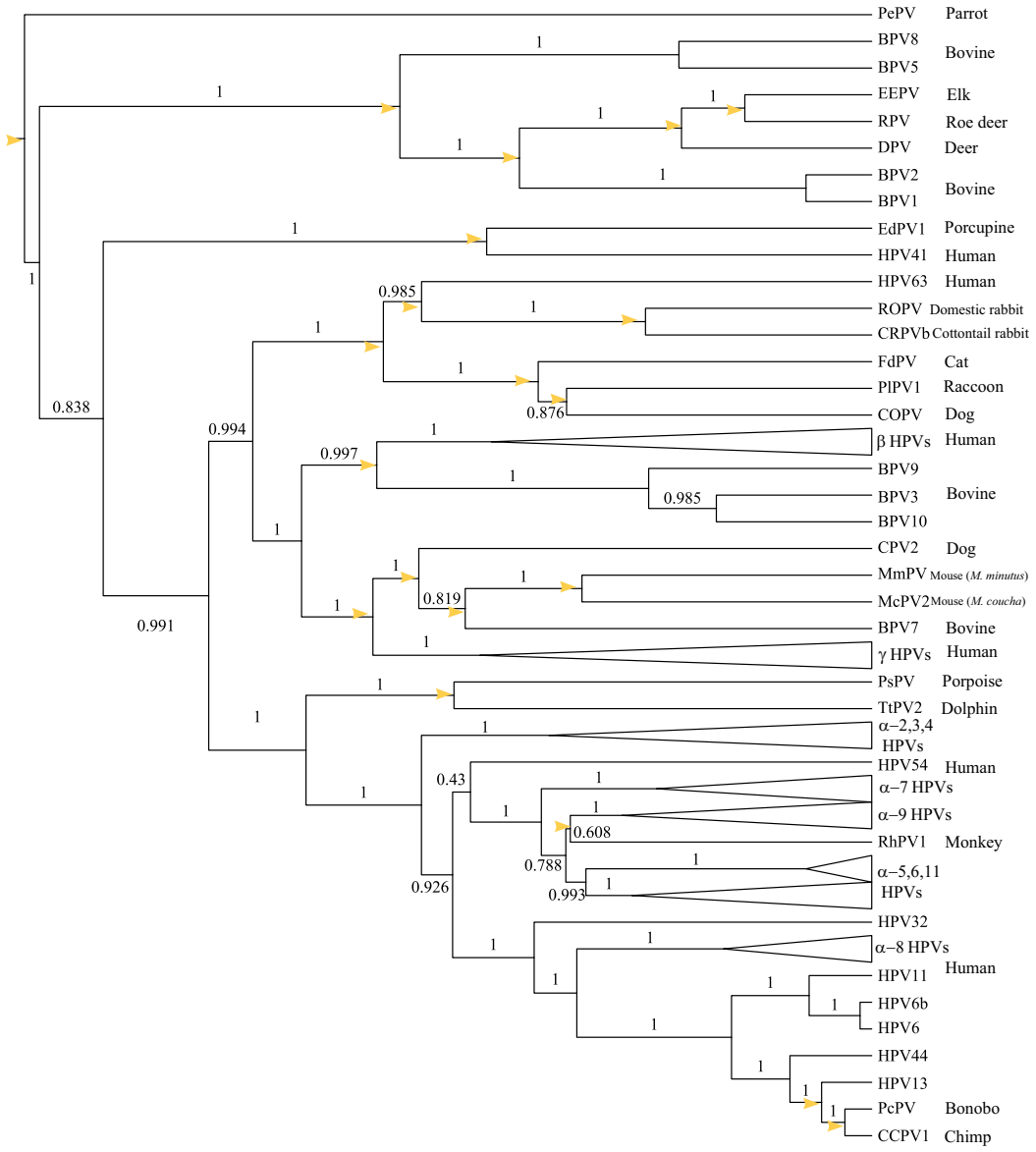


Figure 4.7a: Nodes selected for biased sampling of divergence times in the E1 gene tree. Divergence time distributions, biased in favour of codivergence, were applied to the highlighted nodes to investigate the support for a non-codiverging mechanism.

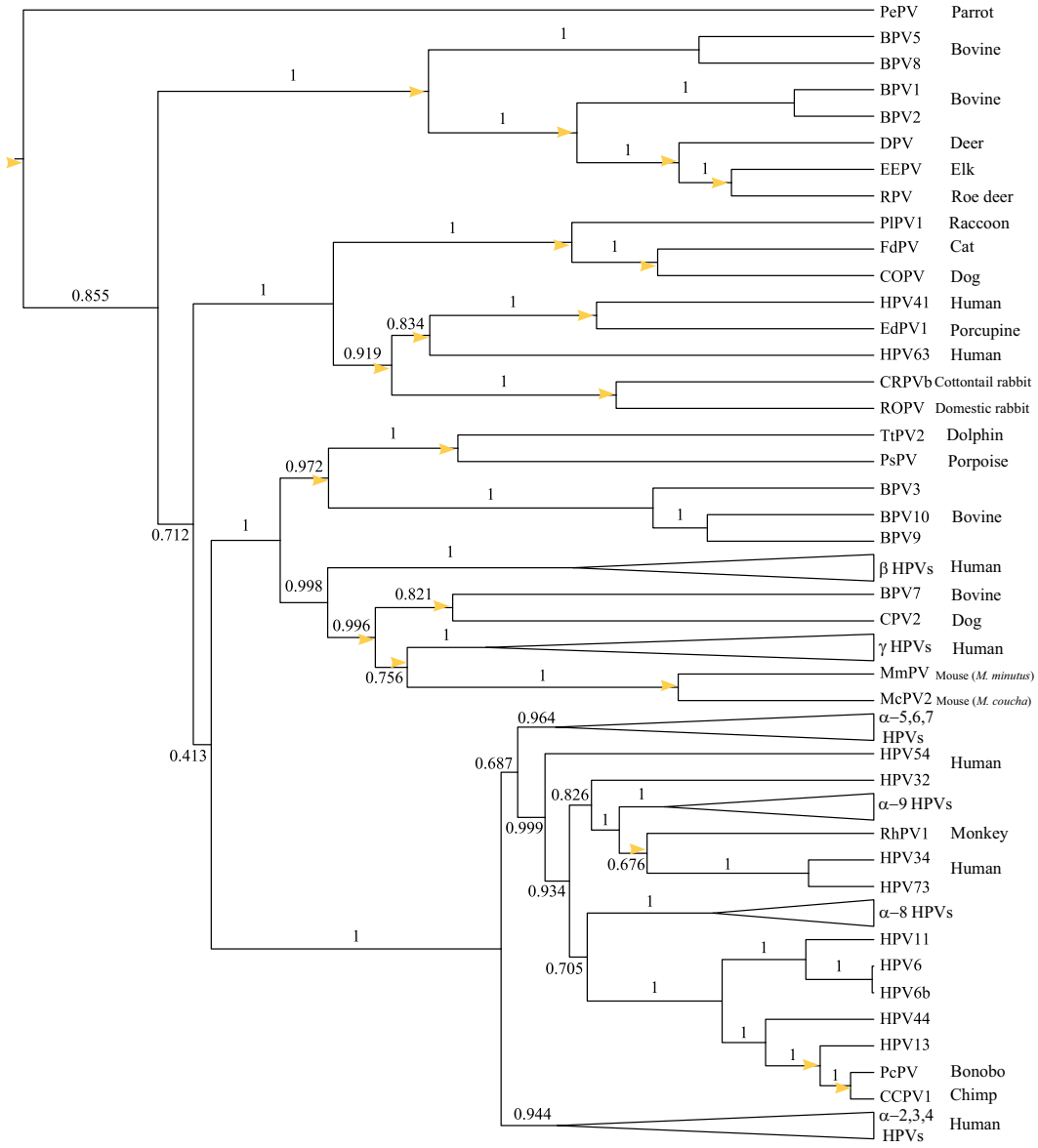


Figure 4.7b: Nodes selected for biased sampling of divergence times in the L1 gene tree. Divergence time distributions, biased in favour of codivergence, were applied to the highlighted nodes to investigate the support for a non-codiverging mechanism.

Super-order	Order	Host divergence	Host speciation time (mya)
Euarchontoglires	Primates	Chimp-Bonobo	3.4-5.5
		Human-(Chimp/Bonobo)	7.8-9.5
		Human-Monkey	19.0-31.8
	Glires	Muridae: M. coucha-M. minutus	25.9-31.9
		Cotton tail rabbit-European (Domestic) rabbit	19.0-31.8
	Primates-Glires[a]		90.0-93.8
Laurasiatheria	Carnivora	Cat-(Dog/Raccoon)	59.8-67.1
		Dog-Raccoon	53.2-59.8
	Cetacea	Dolphin-Porpoise	27.8-32.1
	Cervidae	Deer-(Elk/Roe deer)	14.7-18.5
		Elk-Roe deer	14.0-16.8
	Ruminantia	Cervidae-Bovine	24.3-31.9
	Cetartiodactyla[b]	Cetacea-Ruminantia	59.2-63.9
	Carnivora-Cetartiodactyla		83-85.8
Euarchontoglires-Laurasiatheria[b]			92.9-98.4

Table 4.1: Host speciation times (estimated by Bininda-Emonds et al. (2007)) used to sample PV divergence times .

[a] The Primates-Glires speciation time was applied to PV splits of human-rabbit, human-porcupine, human (γ)-Muridae.

[b] The Cetartiodactyla speciation time was applied to the divergence of the Cetacean PVs and the ξ BPVs in the L1 tree only.

[c] The Euarchontoglires-Laurasiatheria speciation time was applied to PV splits of human-rabbit-Carnivora, human (γ)-bovine, human (γ)-dog, Muridae-bovine, Muridae-dog, human (γ)-Muridae-bovine-dog, Primate-Cetacea (E1 only), human (β)-bovine (E1 only).

4.2.2 Monte Carlo simulation under the null hypothesis

In order to calculate p-values for the violations of cospeciation observed at each node of this biased BEAST analysis, PV E1 and L1 gene data sets were simulated under a model of cospeciation (at the nodes of interest). Cospeciation was conferred on the MAP trees from the above BEAST analysis by specifying times randomly sampled from the corresponding host speciation times, assuming a uniform distribution, and re-estimating the times of the remaining internal nodes using r8s (Sanderson 2003) and the non-parametric rate smoothing (NPRS) algorithm, which allows for rate heterogeneity between branches. I repeated the process using different sets of sampled times to produce ten trees with different divergence times of the internal nodes. To convert the branch lengths from units of time to units of distance we sampled rates for each branch from the distribution of branch rates obtained in the above BEAST analysis. Sequences were simulated along the resulting trees using Evolver from the PAML package (Yang 1997; Yang 2007). Each codon position was simulated separately using the mean values of substitution parameters κ and α obtained from the partitioned BEAST analysis. Ten data sets were simulated for each tree, resulting in 100 simulated dataset in total. The biased BEAST analysis was then performed on each simulated dataset using the same settings and evolutionary model as was applied in the analysis of the real data.

4.3 Results

4.3.1 PV-Host Tree Incongruence

The PV data set analysed consisted of 107 mammalian PVs from 18 different species and 1 avian PV; Figure 4.8 shows tanglegrams (constructed by hand) of the host tree and the PV MAP trees derived from independent Bayesian analysis of the E1 and L1 genes. The MAP trees for each PV gene possessed high (p>0.9) posterior probabilities at the majority (97/106 and 93/106, respectively) of internal branches and therefore we are confident about most of the topological associations of PV taxa

in the gene trees. The tanglegrams show that when each PV gene tree is compared against the host tree, the host and virus topologies are far from congruent. Several clades in the virus tree show associations across the host tree. The human, bovine and canine host species are each infected by multiple PV types which fail to cluster together in one clade. Instead, the PV types of each host species are observed to be distantly related to other types infecting the same host. Of the three non-human primate PVs included in the analysis, the chimpanzee and bonobo PV types (CCPV and PcPV1 – now labelled, PtPV1 and PpPV1, respectively) are nested within the clade of the low-risk α HPVs while rhesus monkey PV (RhPV1 - now labelled MmPV1) is nested within the clade of the high-risk α HPVs.

The lack of monophyly among some PVs at the hosts' species level continues at the order and superorder levels. The dataset contains PV types isolated from the mammalian orders of Rodentia (murid and porcupine), Primates (human, chimpanzee, bonobo, monkey), Lagomorpha (rabbit), Carnivora (cat, dog, raccoon), Cetacea (porpoise and dolphin) and Artiodactyla (bovine, elk, deer, roe deer). Rodentia, Primates and Lagomorpha fall under the superorder Euarchontoglires, whilst the remaining orders fall under the superorder Laurasiatheria. Among the Rodentia, the murid PVs and the porcupine PV are in different parts of the tree: the σ porcupine EdPV1 clusters with ν HPV41 whilst the π murid McPV2 and MmPV (now MnPV1) cluster with the γ HPVs. The PV trees do not show an early divergence of sequences from Euarchontoglires and Laurasiatheria but instead we see Euarchontoglire-derived PVs clustering with Laurasiatheria-derived PVs in several well-supported clades in both the E1 and L1 gene trees.

The phylogenetic incongruities between the E1 and L1 gene trees not only reveal differences in the associations between PV types but also differences in the grouping of PVs from different host species. The cetacean PVs cluster with Primate α PVs in the E1 tree but with the ξ BPVs in the L1 tree. The ν HPV41-σ EdPV1 clade occupies different position in the two trees, and although this clade associates with human and rabbit PVs in the L1 tree, the Glire PVs (rabbits, porcupine) do not cluster together.

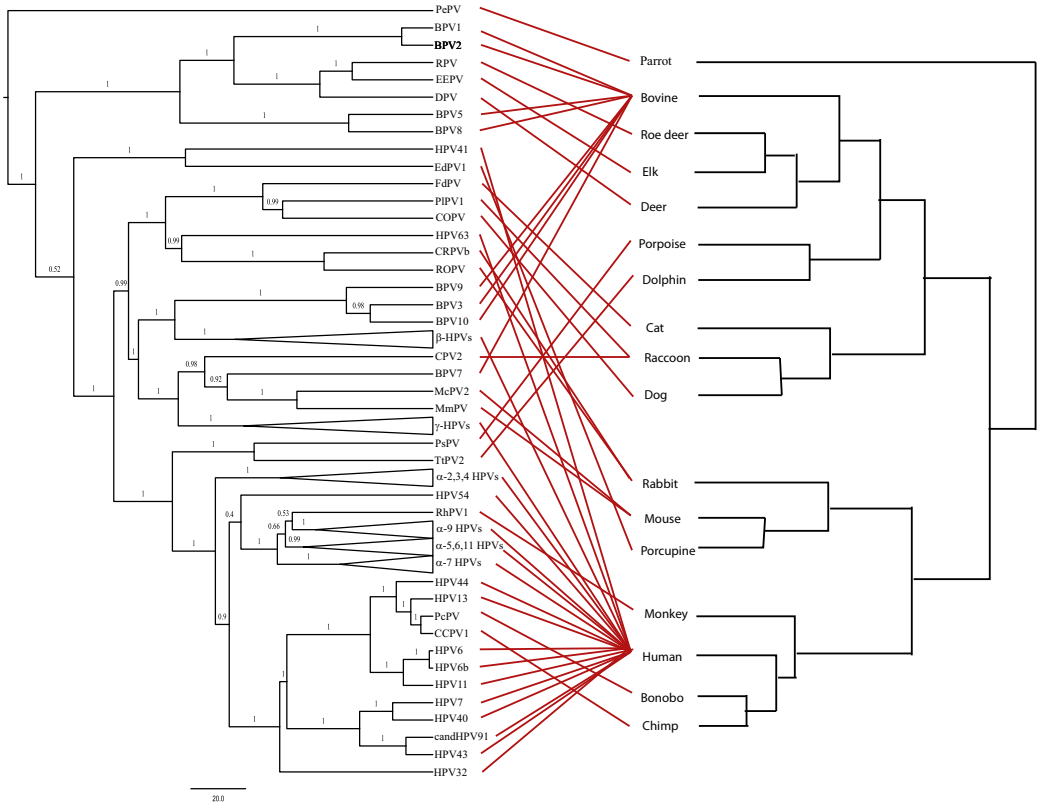


Figure 4.8a: PV-host tanglegrams based on the E1 gene MAP tree of PVs. Terminal associations between associated host and virus taxa are indicated by the red lines. Despite topological differences between the E1 and L1 gene trees, neither gene tree shows complete concordance with the phylogeny of the associated hosts. PV types infecting humans, in particular, span most of the tree and are interspersed by PV associations with other hosts.

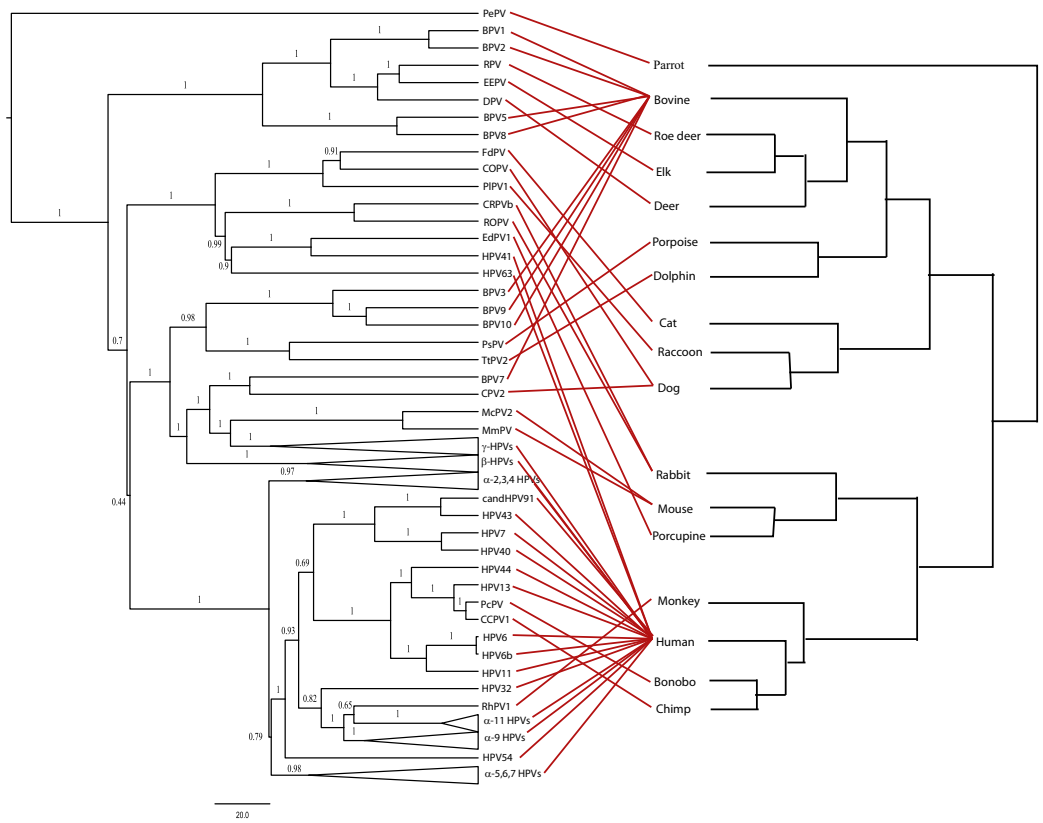


Figure 4.8b: PV-host tanglegrams based on the L1 gene MAP tree of PVs. Terminal associations between associated host and virus taxa are indicated by the red lines. Despite topological differences between the E1 and L1 gene trees, neither gene tree shows complete concordance with the phylogeny of the associated hosts. PV types infecting humans, in particular, span most of the tree and are interspersed by PV associations with other hosts.

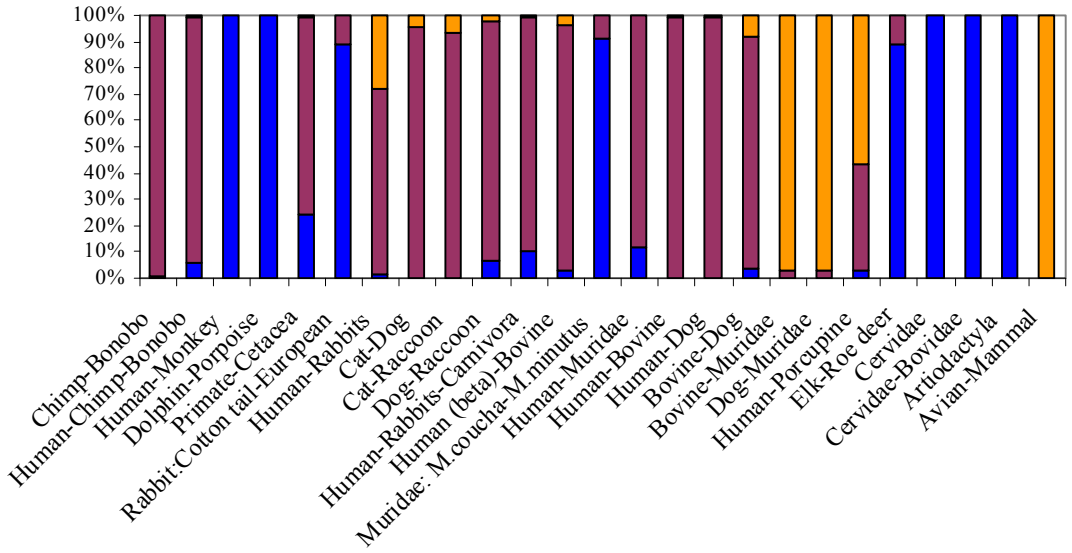
4.3.2 Biased sampling of divergence times

At each node biased for the sampling of divergence times, I observed consistency in the distribution of sampled times (Appendix B.1 and B.2) obtained in multiple independent chains run under the different likelihood penalties. I further compared the distribution times sampled at each node against the prior distribution of times (Appendix B.3 and B.4) to ensure that the results were not dominated by the prior. The density plots shown in Appendix B.3 and B.4 reveal some degree of prior density associated with viral divergence times outside the host speciation times; however, at each node, the prior distribution of times is different to the sampled distribution thus demonstrating the influence of the data on the sampling of divergence times.

The sampled times are categorised as representing codivergence if they fall within the biased range of the host speciation times, prior divergence if they pre-date the host speciation range, and later divergence if they post-date the host speciation range. The amount of sampling observed from each category differs among the nodes analysed (Figure 4.9 and 4.10). Almost all the nodes show some degree of sampling outside the host speciation range. The only exceptions to this were observed using the stricter ln(0.005) likelihood penalty for violations of cospeciation at the γ HPV-CPV2, γ HPV-BPV7, π murid PV-CPV2, and π murid PV-BPV7 divergences of the L1 gene. These four lineages cluster together to form a polyphyletic clade with uncertain topology (p=0.82 for the branch joining CPV2 and BPV7, and p=0.76 for the branch joining the γ HPVs and the π murid PVs); however, all of the times sampled for these four divergences fell within the host speciation range and produced high posterior probabilities (p>0.9) for the CPV2-BPV7 and γ HPV-π murid PV groupings.

Several nodes showed only a small amount (<10 %) of sampling from outside the host speciation range, e.g. at divergences of the chimp-bonobo, human-(chimp-bonobo), cat-dog, dog-raccoon, cat-raccoon, γ human-BPV7 and γ human-CPV2 divergences, among others. For some of these nodes, the minor violations were found to be statistically significant but only under the stricter likelihood penalty – this may be due to inefficient mixing (see below).

a.



b.

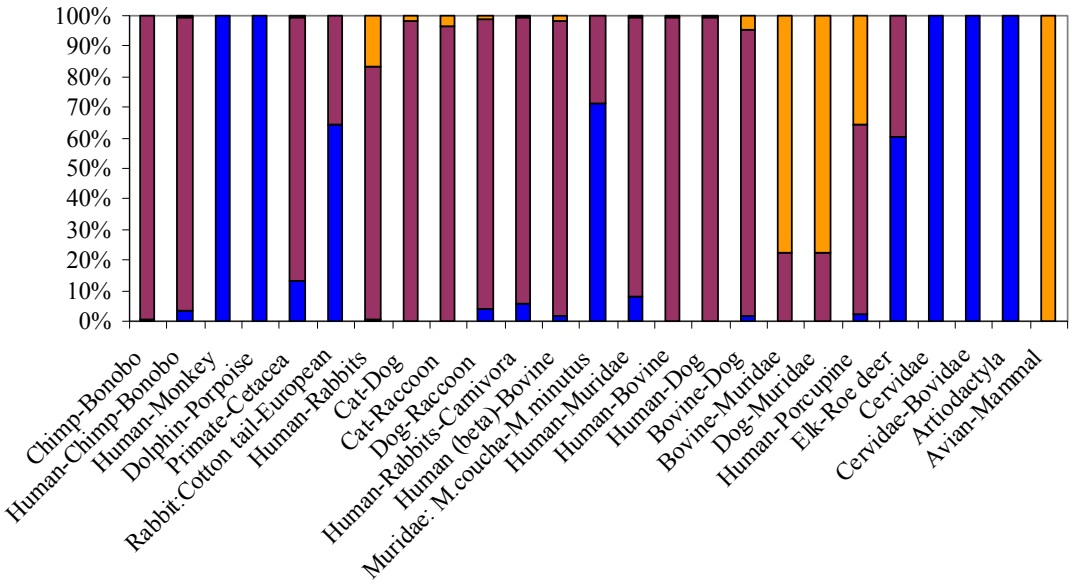
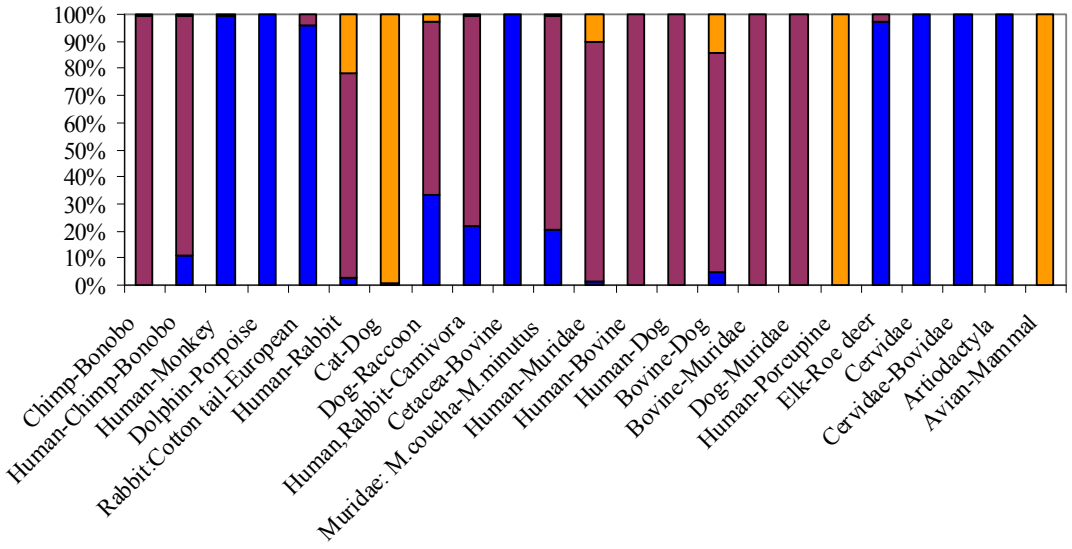


Figure 4.9: Proportion of sampling of viral divergence times reflecting codivergence (red), prior divergence (blue) and later divergence (orange) for nodes in the E1 gene tree: a. results obtained under the ln(0.05) likelihood penalty and b. results obtained under ln(0.005) likelihood penalty.

a.



b.

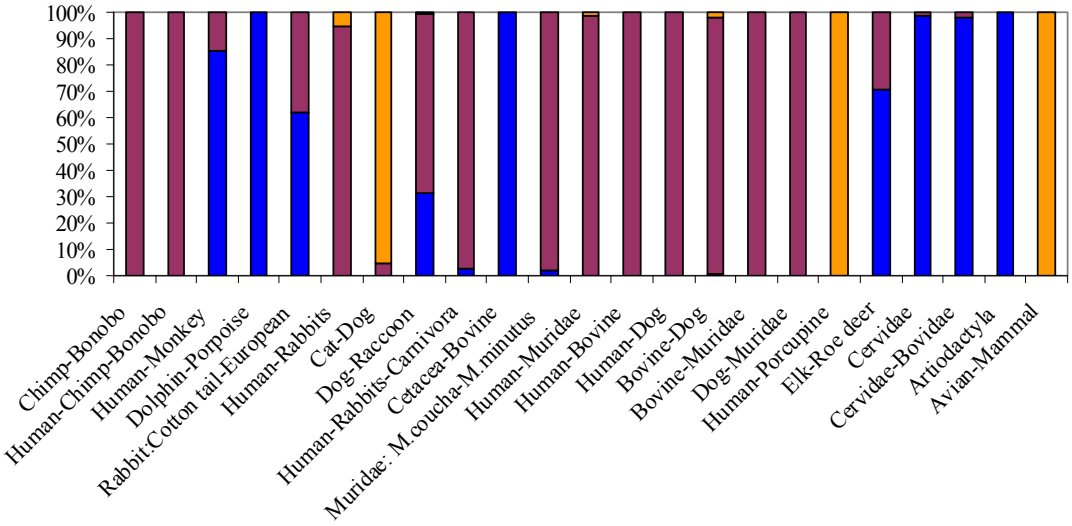


Figure 4.10: Proportion of sampling of viral divergence times reflecting codivergence (red), prior divergence (blue) and later divergence (orange) for nodes in the L1 gene tree: a. results obtained under the ln(0.05) likelihood penalty and b. results obtained under ln(0.005) likelihood penalty.

For both the E1 and L1 genes, divergences of the α human-monkey, cetacean, Cervidae, Cervidae-Bovidae, and δ-ε artiodactyla PV lineages all show 100% (or close to) violation of codivergence in favour of prior divergence and the mammalian-avian PV divergence for both genes showed 100% sampling of times indicative of a host transfer event. In addition, the L1 gene shows near 100% violation of codivergence in favour of host transfer times at the ν human-σ porcupine and the λ cat-dog PV divergences, whilst a 97% violation of codivergence times in favour of host transfer is observed at the π muridae PV-BPV7 and π muridae PV-CPV2 divergences of the E1 gene (under the weaker penalty only). For the remaining nodes there is a mix of sampling from within and outside of the host speciation range but violations are largely restricted to either prior divergence or later divergence – there are no nodes for which a substantial proportion of the chain sampled times from both sides of the host speciation range.

A comparison of the sampled times under the different biases applied reveals how the extent of sampling outside the host speciation range is affected by the size of the bias. For both the E1 and L1 genes, with the stronger bias (greater likelihood penalty) there is less sampling outside of the biased times at divergences of the κ rabbit, μ human-κ rabbit, π muridae and δ elk-roe deer PVs than is observed with the weaker bias (lower penalty). This effect is also evident at the ν human-σ porcupine divergence of the E1 gene and the α human-monkey divergence of the L1 gene. For other nodes, the posterior distribution of sampled times remains the same under both penalties.

In order to make inferences about the nature of the diversification mechanism occurring at each node the statistical support for the observed violations of cospeciation was determined using parametric bootstrapping under a model of cospeciation. The posterior probabilities of codivergence, prior divergence and host transfer for the 100 E1 and L1 simulated datasets are summarised in Figure 4.11. Posterior probabilities of cospeciation at PV divergences of the simulated data sets are generally higher than the posterior probabilities of host transfer or prior divergence. For the E1 simulations, exceptions occur at the ν human-σ porcupine, π muridae-CPV2 and π muridae–BPV7 divergences, for which the posterior probability of host

transfer events is sometimes greater than that of codivergence. The BPV7-CPV2 divergence appears to favour host transfer over codivergence in some of the L1 simulations.

The distribution of posterior probabilities of prior divergence and host transfer for the data sets simulated under a model of codivergence allow us to evaluate the statistical significance of violations observed for the real data set. Tables 4.2-4.5 show the statistical support for prior divergence or host transfer at selected nodes in the E1 and L1 gene trees. Most nodes do not reject the codivergence, suggesting that the data are consistent with our assumption of the generality of this process of viral divergence.

Statistically significant support for prior divergence at the ancestral PV nodes of α human-monkey, dolphin-porpoise, κ domestic-cottontail rabbits, δ elk-roe deer, δ Cervinae (deer)-Capreolinae (elk, roe deer), and δ Cervidae-Bovidae types was observed for both genes. As the branching patterns at these nodes are congruent with those of the corresponding host species, the temporal analysis has allowed identification of non-cospeciating mechanisms at nodes where topological methods would most likely have assumed cospeciation. As seen in Figure 4.12, there is generally good agreement between the timing of these prior divergence events in both gene trees, arguing against recombination at these points. In addition to significant violations at these nodes, there is strong support for prior divergence of the E1 genes of the Muridae (harvest and multimammate mouse) PVs, whereas for the L1 genes, the divergence times sampled for this node largely agree with the host speciation times. For the L1 genes, prior divergence at the ancestral node of the cetacean PV-ξ BPV was also found to be statistically significant; these two groups of PVs do not share an immediate common ancestor in the E1 gene tree.

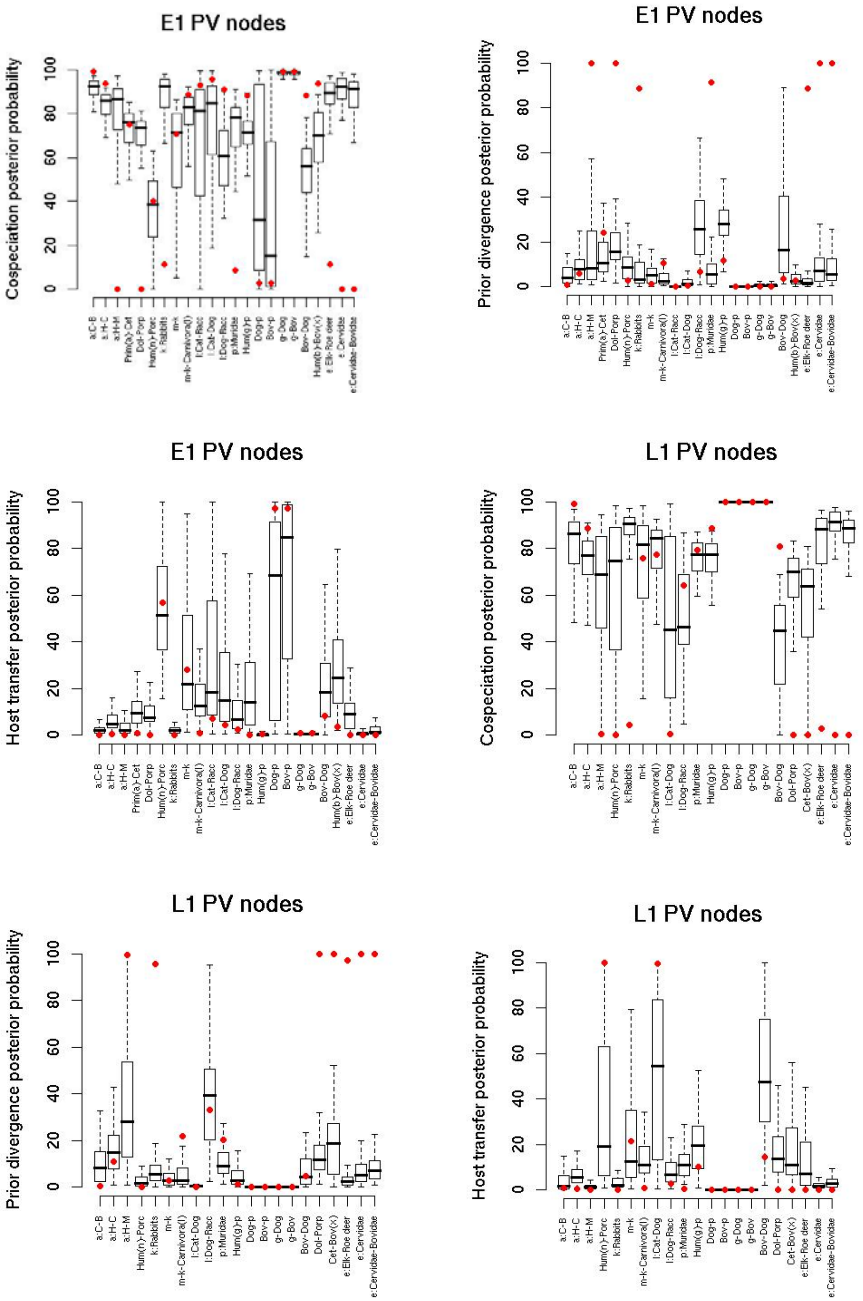


Figure 4.11: The "posterior probabilities" of codivergence, prior divergence and host transfer for 100 simulated data sets of the E1 and L1 genes. Red points indicate corresponding posterior probability for real data set.

Results obtained with the stronger codivergence bias of ln(0.005) were in general similar, as shown in Tables 4.3 and 4.5. All nodes for which codivergence was rejected with the weaker bias produced similar results with the stronger bias, with the exception of the prior divergence of the Cottontail and European rabbit divergence of the E1 gene, which was strongly supported with the weaker bias ($P < 0.01$) but not as strongly supported with the stronger bias ($P < 0.06$). A number of nodes seemed to reject codivergence with the stronger bias based on minimal posterior probabilities, for example the cat-raccoon, γ human-BPV7 and γ human-CPV2 divergences of the E1 gene and the μ human-κ rabbit-λ Carnivora divergence of both the E1 and L1 genes. The violations of codivergence at these divergences corresponded to phylogenetic trees that were far from the MAP tree. When analyzing the simulated datasets using the higher bias, the MCMC chain does not appear to sample these topologies. It is therefore possible that there is inadequate mixing of the MCMC sampling procedure at this higher bias.

Based on the events inferred and the corresponding sampled divergence times, the PV gene phylogenies may be redrawn, to scale, on top of the host phylogeny (from which the host speciation times used in this analysis were obtained) to illustrate the evolutionary trajectory of PV lineages (Figure 4.13). There are uncertainties in the inference, but we can characterize the overall picture.

There was a wide diversification of PVs among mammals starting from around 150 mya. Starting with an early divergence of the δ-ε Artiodactyla PV lineage from that of the other mammalian PVs, by the time of the Euarchontoglires-Laurasiatheria divergence approximately 96 mya both genes had well-defined α, β, δ-ε Artiodactyla, ξ Bovine, Cetacean, and λ (Carnivora excluding CPV2) lineages. In addition, the L1 gene had diverged into two lineages ancestral to the γ human-π muridae PV and BPV7-CPV2 types, while the E1 appears to demonstrate a divergence into lineages ancestral to the γ HPV and mouse-BPV7-CPV2 lineages. The μ human-ν human-σ porcupine-κ rabbit clade present at this time in L1 was divided into μ human-κ rabbit PVs and ν human-σ porcupine clades in E1.

The L1 gene exhibits codivergence of the μ HPV-ν HPV-EdPV1 and κ rabbit PV lineages in the ancestral Euarchontoglires species, followed by a divergence of μ human and ν human-σ porcupine PV lineages and a host transfer event between humans (ν HPV41) and porcupine (σ EdPV1). The E1 gene follows a different trajectory, with the ν human-σ porcupine lineage diverging from other of the PV lineages quite early; the μ human-κ rabbit PV lineage diverges from the λ clade sometime later, but still prior to the split between Euarchontoglires and Laurasiatheria.

E1 gene PV divergence (ln(0.05))	Prior divergence	Codivergence	Later divergence
Chimp-Bonobo	0.60	99.22	0.18
Human-Chimp-Bonobo	5.66	93.77	0.57
Human-Monkey	100**	0	0
Cetacea: Dolphin-Porpoise	100**	0	0
Primate-Cetacea	24.30	74.98	0.72
Rabbit: Cotton tail-European	88.86**	11.14	0
Human-Rabbit	1.27	70.72	28.00
Cat-Dog	0.25	95.48	4.27
Cat-Raccoon	0.03	93.03	6.94
Dog-Raccoon	6.60	90.91	2.48
Human-Rabbit-Carnivora	10.38*	88.73	0.89
Human (beta)-Bovine	2.92	93.69	3.39
Muridae: *M.coucha-M.minutus*	91.31**	8.69	0
Human-Muridae	11.65	88.15	0.20
Human-Bovine	0.05	99.18	0.77
Human-Dog	0.05	99.18	0.77
Bovine-Dog	3.42	88.23	8.35
Bovine-Muridae	0	2.63	97.37
Dog-Muridae	0	2.79	97.21
Human-Porcupine	2.81	40.21	56.98
Elk-Roe deer	88.69**	11.30	0.01
Cervidae : Deer-Elk-Roe deer	100**	0	0
Cervidae-Bovidae	100**	0	0

Table 4.2: Observed distribution of diversification mechanisms at PV divergences of the E1 gene from the biased sampling analyses run with likelihood penalties of ln(0.05) for sampled times that violate the corresponding host-speciation times.

* indicates P-values < 0.05 obtained from the biased sampling analysis of simulated data generated under a model of codivergence at each node. ** indicates P-values < 0.01 from this analysis.

E1 gene PV divergence (ln(0.005))	Prior divergence	Codivergence	Later divergence
Chimp-Bonobo	0.35	99.56	0.09
Human-Chimp-Bonobo	3.21	96.49	0.30
Human-Monkey	100**	0	0
Cetacea: Dolphin-Porpoise	100*	0	0
Primate-Cetacea	13.41	86.24	0.35
Rabbit: Cotton tail-European	64.52	35.48	0
Human-Rabbit	0.70	82.73	16.57
Cat-Dog	0.17	98.05	1.78
Cat-Raccoon	0.02*	96.80	3.18
Dog-Raccoon	3.76	94.82	1.42
Human-Rabbit-Carnivora	5.84*	93.68	0.48
Human (beta)-Bovine	1.45	96.62	1.92
Muridae: *M.coucha-M.minutus*	71.27**	28.73	0
Human-Muridae	8.05	91.27	0.69
Human-Bovine	0	99.60	0.40*
Human-Dog	0	99.60	0.40*
Bovine-Dog	2.00	93.46	4.55
Bovine-Muridae	0	22.48	77.52
Dog-Muridae	0	22.56	77.44
Human-Porcupine	2.02	62.58	35.40
Elk-Roe deer	60.33**	39.66	0.01
Cervidae: Deer-Elk-Roe deer	100**	0	0
Cervidae-Bovidae	100**	0	0

Table 4.3: Observed distribution of diversification mechanisms at PV divergences of the E1 gene from the biased sampling analyses run with likelihood penalties of ln(0.005) for sampled times that violate the corresponding host-speciation times.

* indicates P-values < 0.05 obtained from the biased sampling analysis of simulated data generated under a model of codivergence at each node. ** indicates P-values < 0.01 from this analysis.

L1 gene PV divergence (ln(0.05))	Prior divergence	Codivergence	Later divergence
Chimp-Bonobo	0.27	99.06	0.66
Human-Chimp-Bonobo	10.83	88.83	0.35
Human-Monkey	99.50**	0.50	0
Cetacea: Dolphin-Porpoise	100*	0	0
Rabbit: Cotton tail-European	95.75**	4.25	0
Human-Rabbit	2.61	75.84	21.56
Cat-Dog	0	0.44	99.56
Dog-Raccoon	33.12	64.09	2.79
Human,Rabbit-Carnivora	21.80*	77.42	0.78
Cetacea-Bovine	100**	0	0
Muridae: *M.coucha-M.minutus*	20.30	79.29	0.42
Human-Muridae	1.03	88.90	10.07
Human-Bovine	0	99.99	0.01
Human-Dog	0	99.96	0.04
Bovine-Dog	4.65	80.82	14.53
Bovine-Muridae	0	99.94	0.06
Dog-Muridae	0	99.99	0.01
Human-Porcupine	0	0	100*
Elk-Roe deer	97.42**	2.58	0
Cervidae: Deer-Elk-Roe deer	100**	0	0
Cervidae-Bovidae	100**	0	0

Table 4.4: Observed distribution of diversification mechanisms at PV divergences of the L1 gene from the biased sampling analyses run with likelihood penalties of ln(0.05) for sampled times that violate the corresponding host-speciation times.

* indicates P-values < 0.05 obtained from the biased sampling analysis of simulated data generated under a model of codivergence at each node. ** indicates P-values < 0.01 from this analysis.

L1 gene PV divergence ln((0.005))	Prior divergence	Codivergence	Later divergence
Chimp-Bonobo	0.02	99.89**	0.09
Human-Chimp-Bonobo	0.10	98.97	0.04
Human-Monkey	85.43*	14.57	0
Cetacea: Dolphin-Porpoise	100**	0	0
Rabbit: Cotton tail-European	61.71**	38.29	0
Human-Rabbit	0.16	94.68	5.16
Cat-Dog	0	4.43	95.57
Dog-Raccoon	31.11	68.46	0.44
Human-Rabbit-Carnivora	2.68*	97.27	0.05
Cetacea-Bovine	100**	0	0
Muridae: *M.coucha-M.minutus*	2.14	97.79	0.07
Human-Muridae	0.09	98.50	1.42
Human-Bovine	0	100	0
Human-Dog	0	100	0
Bovine-Dog	0.50	97.81**	1.69
Bovine-Muridae	0	100	0
Dog-Muridae	0	100	0
Human-Porcupine	0	0	100*
Elk-Roe deer	70.38**	29.61	0.01
Cervidae: Deer-Elk-Roe deer	98.50**	1.50	0
Cervidae-Bovidae	98.31**	1.69	0

Table 4.5: Observed distribution of diversification mechanisms at PV divergences of the L1 gene from the biased sampling analyses run with likelihood penalties of ln(0.005) for sampled times that violate the corresponding host-speciation times.

* indicates P-values < 0.05 obtained from the biased sampling analysis of simulated data generated under a model of codivergence at each node. ** indicates P-values < 0.01 from this analysis.

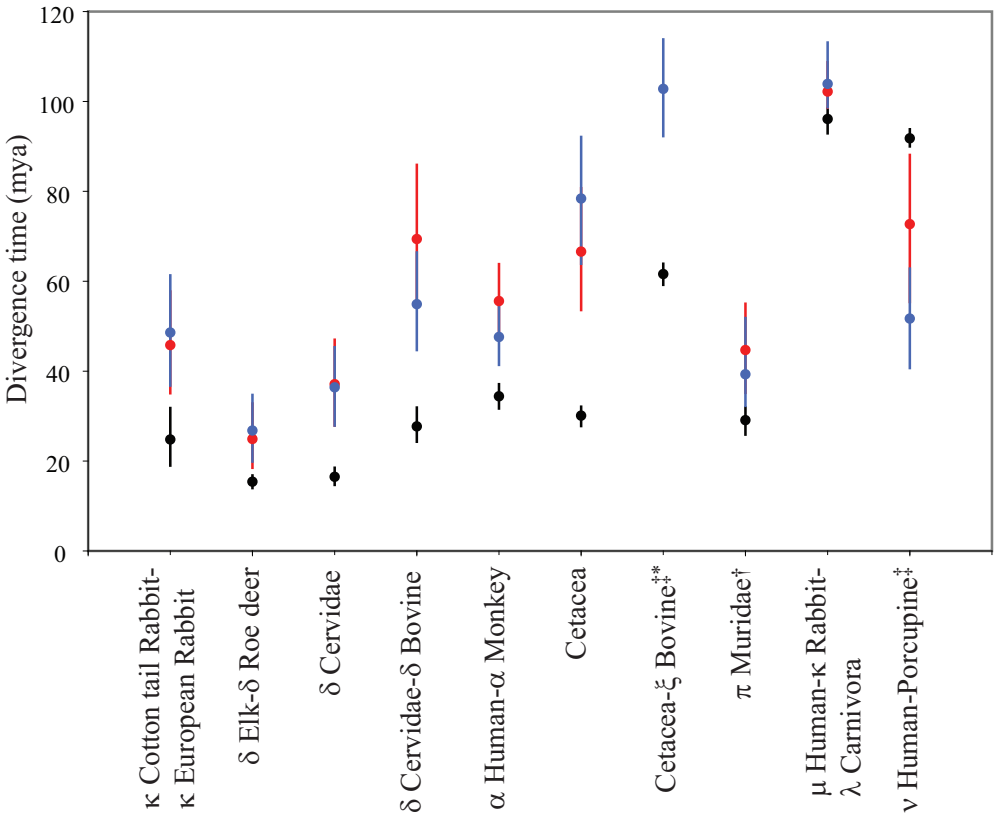


Figure 4.12: Divergence times for the host (black), E1 (red), and L1 (blue) genes. CIs for the host and viral divergence times are indicated with error bars; unseen error bars represent CIs smaller than the size of the symbols. Viral divergence times further back than host divergence times (e.g., human–monkey) represent prior divergence, whereas viral divergence times more recent than host divergence times (e.g., human–porcupine) represent likely host transfer events. †Statistically significant violation of host divergence time observed for E1 only. ‡Statistically significant violation of host divergence time observed for L1 only. *Node present in L1 gene tree only. The only host transfer event found to be statistically significant with the weaker penalty was the post-host speciation divergence of the ν HPV-porcupine (EdPV1) L1 genes. The proposed host transfer of the E1 genes of these PV lineages was not found to be significant (P ~ 0.30); however, the position of this node differs in both gene trees. For the E1 gene the posterior distribution of divergence times ranged from 55.38-88.14 mya compared with 40.70-62.82 mya for the L1 gene.

a.

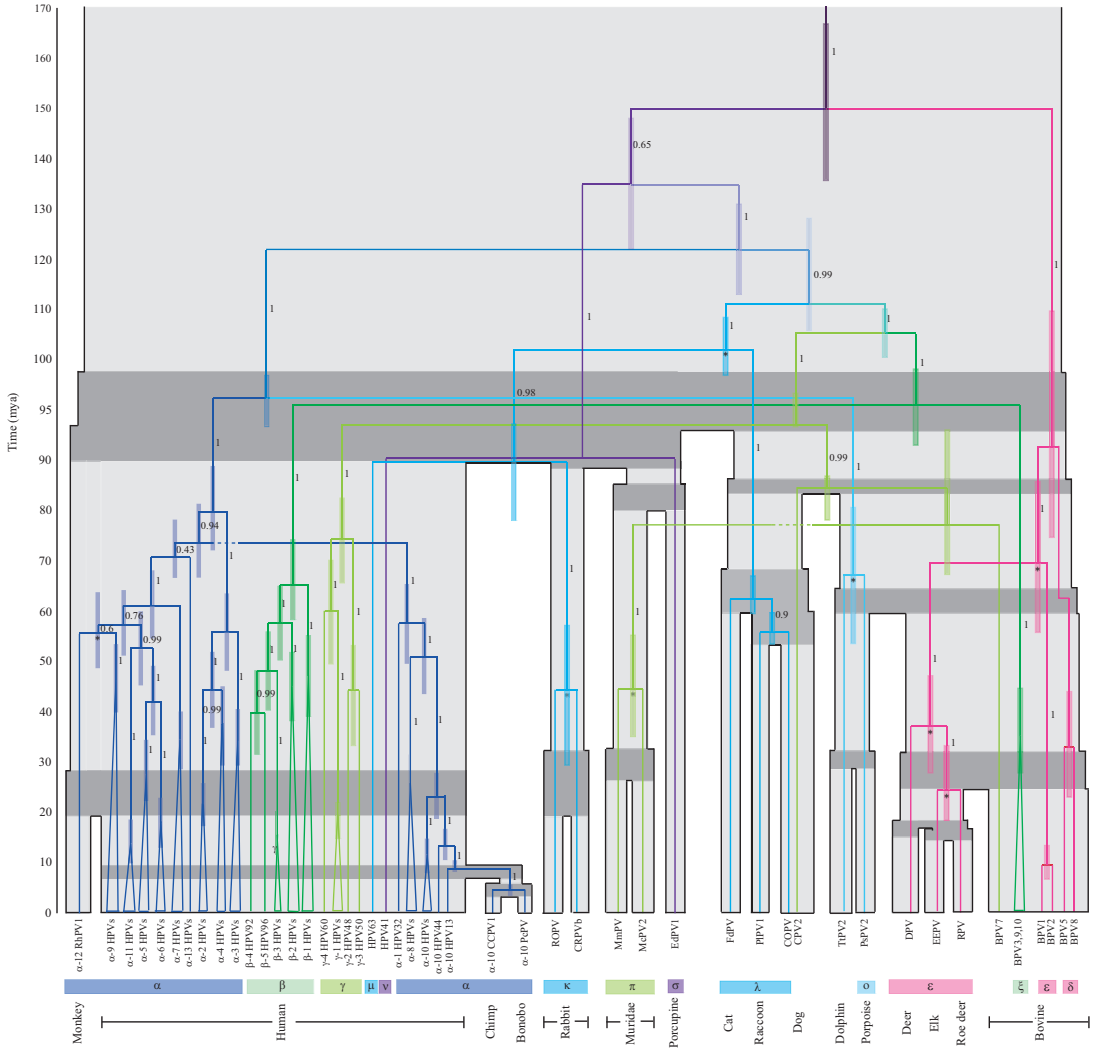


Figure 4.13 (see next page for description)

b.

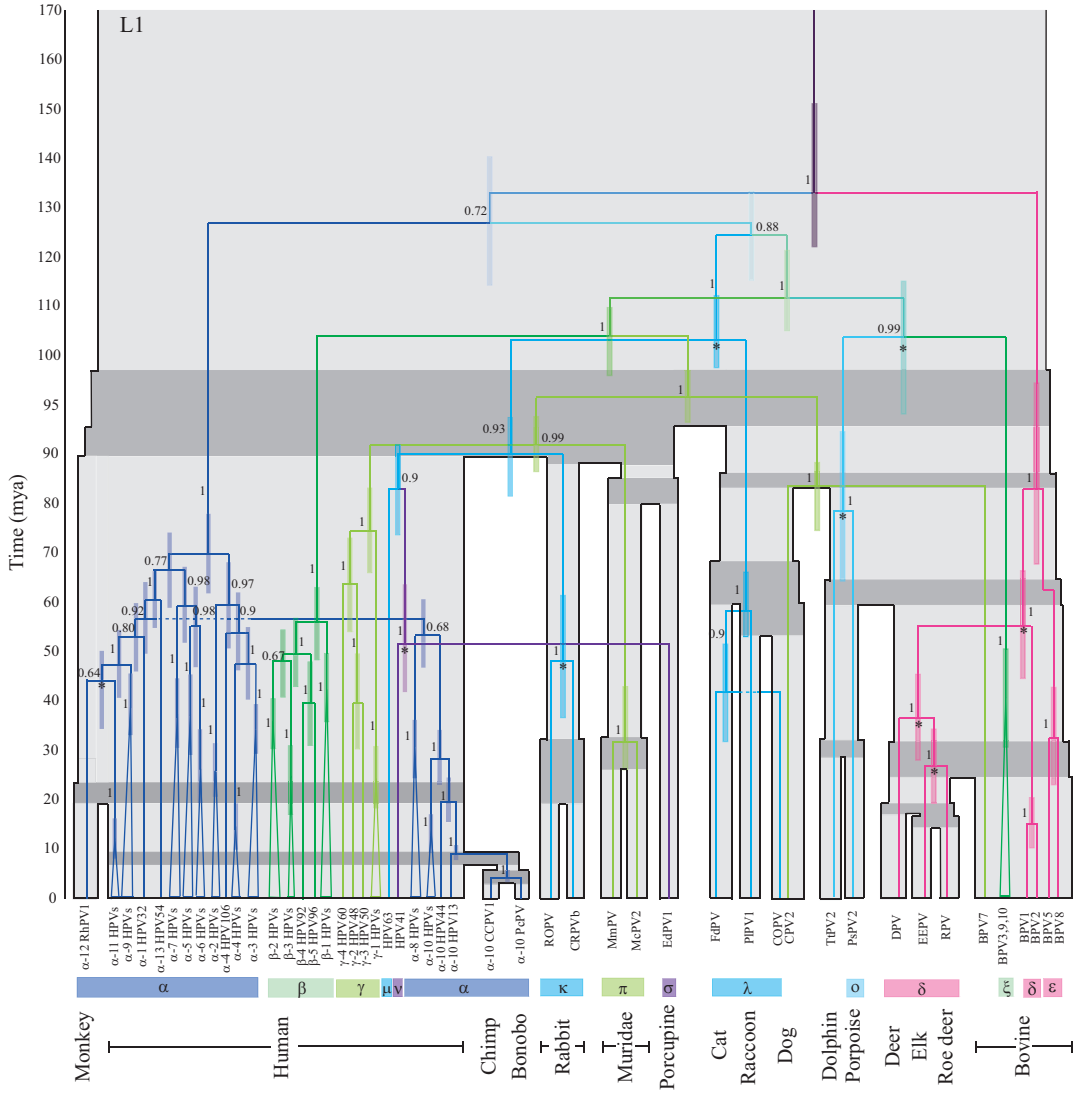


Figure 4.13 (constructed by hand): The a) E1 and b) L1 gene trees, each shown on top of the associated host tree topology (grey) derived from the mammalian phylogeny estimated by Bininda-Emonds and colleagues, and scaled according to the times of the host divergences (Ma). The timings of the PV splits correspond with the mean times sampled from the biased sampling analysis of each gene; the 95% CIs of viral divergence times at each node are represented by the colored bars. The host speciation times of related host taxa are highlighted in dark grey. Posterior probabilities of internal PV branches are indicated beside the branches. PV nodes labeled with a * indicate divergences for which cospeciation violations were found to be statistically significant. Labels below the tree indicate 1) the names of the PV taxa—‘‘α-2 HPVs’’ groups together all HPVs included in our analysis from species 2 of the α genus, 2) the genus classifications of the PV taxa, 3) the host species from which the virus was isolated. PV clades are colored according to genus classifications; for simplicity, some genera that consistently group together in both gene trees have been assigned the same color.

4.3.3 Estimated evolutionary rates

When sampling PV divergence times I allowed for rate heterogeneity across branches following rejection of a molecular clock. Branch rates were sampled from a log normal distribution; the mean evolutionary rate estimated for the E1 genes was 7.1×10^{-9} (sd = 3.1×10^{-10}) nucleotide substitutions per site per year and 9.7×10^{-9} (sd = 5.2×10^{-10}) nucleotide substitutions per site per year for the L1 genes. In order to provide a more accurate estimate of the rate I performed the BEAST analysis again for each gene, specifying constraints only for those nodes that did not show significant violations of cospeciation, for which I applied the standard uniform prior distribution of divergence times. The resulting mean rates obtained were 7.1×10^{-9} (sd = 1.5×10^{-10}) nucleotide substitutions/site/year for the E1 genes and 9.6×10^{-9} (sd = 2.1×10^{-10}) nucleotide substitutions/site/year for the L1 genes, which are in good agreement with our previous estimates. Branch-specific evolutionary rates are similar at the top and bottom of our trees suggesting against saturation having an affect on our analysis.

4.4 Discussion

By employing a novel biased sampling approach for the estimation of PV divergence times using Bayesian MCMC algorithms, I have attempted to characterise ancestral diversification mechanisms of the PVs. The need for such a characterisation arises from observations that inferred PV phylogenies do not demonstrate evidence of strictly codivergence with their vertebrate hosts. The identification of close phylogenetic relationships among PV lineages from hosts that are distantly related suggests the possibility of host transfer events, i.e., viral transmission between distinct co-existing host species, despite the lack of physical or experimental support for such events.

A common approach to studying phylogenetic incongruities between associates (parasites/viruses) and their hosts is to first determine whether there is substantial

evidence for codivergence of the associated entities. This is achieved by using reconciliation methods (e.g. TreeMap/Jungles) to determine if the optimal number of codivergence events postulated between reconciled host and associate trees are more than would be expected if the observed associations arose by chance. Alternatively, a method such as ParaFit can be used to test a global hypothesis of cospeciation based on the evolutionary distances observed in host and associate trees. The aim of this study was not specifically to determine the extent of codivergence of PVs and their hosts, as the highly host specific nature of PVs and slow rates of evolution suggest that PVs have been evolving with vertebrates. However, during this coevolutionary period, the phylogenetic inconsistencies indicate that PV divergence has occurred independently of its host on several occasions. The aim of this thesis was to characterise the diversification events of the virus lineages so as to explain the differences observed between the phylogenies of the PVs and their mammalian hosts and to understand how the observed associations were formed.

Using this biased sampling approach I was able to identify viral divergences where the evidence indicates a process other than codivergence – either prior viral divergence preceding the host divergence or host transfers following the host divergence. Based on the events inferred and the corresponding sampled divergence times, the PV gene phylogenies may be redrawn, to scale, on top of the associated host phylogeny to illustrate the evolutionary trajectory of PV lineages (Figure 4.13). There are topological uncertainties in the analysis, but we can characterize the overall picture.

There was a wide diversification of PVs among mammals starting from around 150 mya. Starting with an early divergence of the δ-ε artiodactyl PV lineage from that of the other mammalian PVs, by the time of the Euarchontoglires-Laurasiatheria divergence approximately 96 mya both genes had well-defined α primate, β primate, δ-ε artiodactyl, ξ bovine, cetacean, and λ (Carnivora excluding CPV2) lineages. In addition, the L1 gene had diverged into two lineages ancestral to the γ human-π murid PVs and BPV7-CPV2 types, while the E1 appears to demonstrate a divergence into lineages ancestral to the γ HPV and murid-BPV7-CPV2 lineages. The μ human-ν human-σ porcupine-κ rabbit clade present at this time in L1 was divided into μ human-κ rabbit PVs and ν human-σ porcupine clades in E1.

The L1 gene exhibits a codivergence of the μ HPV-ν HPV-EdPV1 and κ rabbit PV lineages in the ancestral Euarchontoglires species, followed by a divergence of μ human and ν human-σ porcupine PV lineages and a host transfer event between humans (ν HPV41) and porcupine (σ EdPV1). The E1 gene follows a different trajectory, with the ν human-σ porcupine lineage diverging from other of the PV lineages quite early; the μ human-κ rabbit PV lineage diverges from the λ clade sometime later, but still prior to the split between Euarchontoglires and Laurasiatheria.

Varsani et al (2006) analyzed various PV sequences using a suite of recombination detection methods and identified ν HPV41 as a putative recombinant sequence with the canine PV (COPV) from the λ genus being an extant relative of one of the donor sequences. Their analysis highlighted the E1 gene of ν HPV as the location of recombination. The E1 gene trees estimated in this analysis show ν HPV41 to be quite distantly related to the λ Carnivora PVs and therefore do not concur with their findings. The consistent grouping of ν HPV41 and σ EdPV1 in the estimated gene trees suggests that ν HPV41 is unlikely to be a recombinant genome but the variable position of this clade in the different gene trees may indicate a recombination event in the ancestral viral lineage.

One possible explanation involves a recombination event occurring between an unknown ancestral PV lineage and the ancestral lineage of ν HPV41 resulting in the ancestral lineage of human and porcupine PV types. Subsequent to the recombination event the PV lineage diverged with the Primate and Rodent hosts but co-infection (both virus lineages infecting the same host) in either host approximately 50 mya resulted in transfer of the late region of the genome from one species to the other, though it is not possible to deduce the direction of this transfer. A simpler scenario would be that the two human PV subtypes diverged from an ancestral Primate PV lineage and subsequently there was a host transfer of the ancestral ν HPV41 lineage to an ancestral Hystricognathi species resulting in the σ EdPV1 lineage infecting porcupines; although the E1 gene analysis does not support a common ancestor for ν HPV41 and μ HPV63 and did not demonstrate significant support a host transfer

event between the ν human and σ porcupine lineages, a proportion of the divergence times sampled outside of the host speciation range do overlap with those of L1 gene. Codivergence of the ν HPV41 and σ EdPV1 L1 gene sequences would require a substitution rate of approximately 6.7×10^{-9} nucleotide substitutions per site per year, which falls outside of the distribution of rates applied across the different branches of the L1 tree and supports a non-codivergence mechanism at this node.

For the L1 gene, codivergence is observed for the γ human and π murid PV lineages, and at the canine (CPV2) –bovine (BPV7) divergence. The scenario depicted for the E1 genes of these PV types is an early divergence of γ HPVs from the π Muridae-CPV2–BPV7 lineage, followed by codivergences in the latter. Despite the lack of statistical support for rejecting codivergence at these various nodes, the observed E1 topology is inconsistent with the host topology, as we would expect PV lineages from the Euarchontoglires (humans and Muridae) and Laurasiatheria (canine and bovine) to cluster separately.

A similar situation was observed in the λ clade of the carnivoran PVs: the E1 gene tree topology of this clade is congruent with the host topology and there is no evidence of host transfer or prior divergence of the cat, dog (COPV) and raccoon PVs. In the L1 gene tree, however, the cat PV is more closely related to the dog PV than the raccoon PV with insufficient statistical support in favor of host transfer at the cat-dog PV node. It is important to note that host transfers and prior divergences can only be detected when these events occur sufficiently far from the divergence between the hosts. It may be that the prior divergence or host transfer events occurred within the estimated time for the host divergence event, or that there is insufficient data to make a reliable identification of the process of the virus divergence. It is also possible that the topology of the viral trees is erroneous, despite the high posterior probabilities observed. More sampling from within this polyphyletic clade should help resolve these uncertainties.

The relative rarity of PV host transfer events detected in this analysis is in agreement with the practical difficulties associated with such events. There is a distinct lack of physical evidence supporting the host transfer of double-stranded DNA viruses in general. This is likely due to their high species specificity and slow

evolutionary rates, which may make it difficult to adapt to new environments quickly. When evaluating the likelihood of PVs switching host species, we must also consider that PVs may only gain entry to the basal cells of epithelial tissue via epithelial wounds and therefore zoonotic transmissions would require direct contact between the different host species at the very least. However, indications of potential host transfer events exist. For instance, the recent identification of PV types shared by two monkey species, the *Macaca mulata* and the *Macaca fascicularis* (Chen et al. 2009), provides the first indications that host transfer of PVs may be possible between different host species. The PV type isolated from the Atlantic white-sided dolphin (sp. *Lagenorhynchus acutus*) has been classified as TtPV3var - a variant of TtPV3, which was isolated from the closely related bottlenose dolphin (sp. *Tursiops truncatus*) (Gottschling et al. 2011a).

Given the ancestral association of PV types with their hosts, predicted in this analysis, the absence of PV lineages from various extant hosts can be explained by incomplete lineage sorting of the virus among the descendant host species (the virus was not vertically transmitted to all descendant hosts), extinction of virus lineages along particular hosts or a failure to detect these viruses in non-human species. The findings of the present analysis indicate the HPV radiations began tens of millions of years prior to the existence of humans – the divergence of the common ancestor of the α PVs is estimated to have occurred 70-80 mya, that of the β PVs is estimated at around 55-65 mya and that of the γ PVs is estimated at around 75 mya in our analysis. According to these timings, all three genera existed prior to the divergence of the ancestral Primate species, the α and γ PVs may even have existed prior to the divergence of the Euarchonta, which include the Dermoptera (e.g. flying lemurs) and Scandentia (e.g. tree shrews) orders as well as the Primates. However, no PVs have been isolated from the Dermoptera or the Scandentia.

The number of known hosts is gradually increasing; since this analysis was performed, new PV types have been identified in diverse species such as the house mouse (Joh et al. 2011), California sea lion (Rivera et al. 2012), Hamadryas baboon (Bergin et al. 2012), Arabian camel (Ure et al. 2011), the marsupial brush-tailed Bettong (Bennett et al. 2010), and reptiles like the Carpet python (Lange et al. 2012). However, no host species has been uncovered that boasts as extensive diversification

of PV types as is observed in humans. If similar radiations are present in other mammalian (and non-mammalian) orders then the Papillomaviridae family has the potential to be many orders larger than estimated under a strictly codiverging mechanism of PV diversification.

Topological differences between the E1 and L1 genes did not result in conflicting divergence times for the majority of viral nodes; this may serve to strengthen the argument for recombination among ancestral PV lineages, since recombination occurs between co-existing lineages.

The cetacean PVs provide an interesting example of this. Both cetacean PVs were extracted from genital warts; in the E1 gene tree they form a clade sister to the α PVs, which are the only other clade comprising of genital PVs. In the L1 gene tree, the cetacean PVs form a clade sister to the ξ bovine PVs, thus for these PVs the L1 gene tree appears to reflect the host phylogeny whereas the E1 gene tree reflects the biological properties of the virus. In Chapter 3, I considered a hypothesis of convergent evolution to explain these differences. However, temporal data may favour a hypothesis of recombination. The results of the biased sampling analysis indicate codivergence of the E1 α primate PV-cetacean PV lineages and prior divergence of the ξ bovine PV-cetacean PV L1 genes. In addition, the sampled divergence times for the E1 α primate-cetacean node are similar with those of the L1 ξ bovine-cetacean node. The results appear to suggest that the ancestral PV lineage that was passed on to the two cetacean animals may be a recombinant PV formed from the early region of the ancestral α primate PV genome and the late region of the ancestral ξ bovine PV genome. New data presents a more plausible scenario that fits in well with these results. Gottschling et al. (Gottschling et al. 2011a) proposed that the ancestor of a newly discovered cetacean PphPV3, which clusters with the α PV clade in phylogenies derived independently from the early genes and the late genes, recombined with the ancestor of the other cetacean PVs and passed on its early genes to this ancestor. Thus, the codivergence inferred at E1 α primate-cetacean node may reflect the codivergence of the E1 α -PphV3 ancestor.

Previous estimates of the rate of evolution of PVs have been obtained from PV sequences between closely related hosts under the assumption of cospeciation of host and virus. For feline PVs an initial estimate of $7.3\text{-}9.6\times10^{-9}$ nucleotide substitutions/site/year (Tachezy et al. 2002) was later revised to an overall rate of 1.95×10^{-8} (95% CI: 1.32×10^{-8}, 2.47×10^{-8}) nucleotide substitutions/site/year for the viral coding genome and with evolutionary rates for individual genes ranging from 1.44×10^{-8} (for E7) to 2.39×10^{-8} (for E6) (Rector et al. 2007). A rate of $3.3\text{-}3.6\times10^{-8}$ nucleotide substitutions/site/year was estimated from primate PV sequences (Van Ranst et al. 1995).

The Bayesian approach used to investigate cospeciation involves estimation of the evolutionary rates along each branch. The mean rate from the resulting distribution of branch rates therefore allows us to supply estimates of the overall average rate of PV evolution, as well as an estimation of how much this rate varies along various branches of the phylogenetic tree. I found different rates for the E1 genes and the L1 genes. The former are found to evolve slower than the latter with mean evolutionary rates of 7.10×10^{-9} (s.d. = 1.49×10^{-9}) nucleotide substitutions/site/year and 9.57×10^{-9} (s.d. = 2.08×10^{-9}) nucleotide substitutions/site/year, respectively.

Previous estimates for these two genes found evolutionary rates of 1.76×10^{-8} (95% CI: 1.2×10^{-8}, 2.31×10^{-8}) and 1.84×10^{-8} (95% CI: 1.27×10^{-8}, 2.35×10^{-8}), respectively, however, this analysis was restricted to feline PVs (Rector et al. 2007). Our lower evolutionary rates correlate with our observations of prior divergence of PV lineages whereas previous estimates have assumed strict correspondence with host divergence times among a small set of closely related PVs. The E1 gene codes for a protein that initiates replication whilst the L1 gene codes for the viral capsid protein. It may be expected that the L1 gene has a higher evolutionary rate than the E1 gene, as the capsid proteins must maintain diversity in order to evade recognition by the host immune system.

The derived timings of the distant viral divergences can be compromised by saturation. Examination of the sampled phylogenies found no correlation between the branch specific substitution rates and the depth of the branch on the phylogenetic tree, providing no evidence for such saturation effects. More conclusive evidence of the lack of such saturation would require a better characterization of the timing of these

deeper nodes, something that is not available given the current sequence data and available host speciation information.

In performing this analysis I am introducing a new method to investigate diversification mechanisms of viruses and other parasites. Previous methods have generally relied on a tree reconciliation approach (e.g., TreeMap), which involve counting events necessary to explain discrepancies between the calculated host and associate trees. These methods are susceptible to the problems of unknown host and parasite phylogenies, the need to assign relative weights to the different diversification events and the existence of equally parsimonious but different solutions.

The difficulties encountered with TreeMap are clearly demonstrated in a recent application of the method to resolve PV-host phylogenetic incongruities (Gottschling et al. 2011b). Since TreeMap compares only the topological structure of the host and virus trees, a number of potentially optimal reconciliations may be obtained when considering codivergence, host transfer, prior divergence and sorting events, particularly when there is no distinction in the relative weighting given to non-codiverging events. In such cases, the number of potential solutions is found to increase with the number of host transfer events that are allowed to occur and therefore Gottschling and colleagues had to limit the number of host transfer events allowed in the reconciliation. Thus, purely due to the limitations of the method, they have had to exclude a large number of potential solutions, one of which may be the correct one.

The extensively tangled nature of the topological incongruities observed between phylogenies of the PVs and their hosts makes it almost unmanageable for methods based purely on topological comparisons. Gottschling and colleagues had to deconstruct the complete PV tree, estimated for PV types comprising 30 different PV genera infecting a total of 43 different vertebrate species, into four large well-supported clades. Three of these four clades consist of PVs infecting host species from the euarchontoglires and laurasiatheria superorders of placental mammals (1 clade also contains the marsupial infecting BpPV1), whilst the other clade contains only species from the laurasiatheria superorder. Multiple optimal reconciliations were

obtained for all 4 clades, with the clade comprised of α primate PVs, ο cetacean PVs, υ cetacean PVs, a ω carnivoran PV and a dyo-δ artiodactyl PV producing 169 optimal solutions. Thus, the evolutionary history of PVs is too complex to be analysed using only the branching patterns.

I have instead implemented an approach that considers codivergence to represent the "null hypothesis" and tests for violations of codivergence by sampling viral divergence times that are biased for the host speciation times. This provides a means of inferring the different evolutionary scenarios without requiring explicit knowledge of the viral divergence times. The bias towards codivergence means that only those divergences that strongly conflict with the host speciation times will be identified. By utilizing Bayesian phylogenetic methods the analysis can accommodate topological uncertainties in the virus phylogeny, unlike other methods, and also incorporates evolutionary information present in the data set to evaluate temporal congruence.

The only assumption made in this method is that host tree and the associated divergence times are correct, which is necessary in order for the method to produce results. The robustness of the analysis to errors in the host phylogeny and speciation times requires investigation. Explicit consideration of evolutionary events along each lineage is circumvented making the biased sampling method more suitable for complex data sets with high parasite-to-host ratios than alternative methods of characterising host-parasite phylogenetic incongruities. This also presents a significant advantage over other methods since the omission of lineages can be misleading and cause an analysis to arrive at an incorrect solution. In the biased sampling approach, the more lineages that are included in the analysis, the more accurate the phylogenetic estimation and, in particular, the estimation of rates and divergence times will be.

By emphasising temporal comparisions rather than topological comparisons, this approach is better equipped to deal with instances of false congruence and hence non-cospeciating events may be inferred in virus clades that appear to track the host tree. An example of this can be found in the present analysis. PV types from the δ genus cluster together in PV phylogenies. This genus consists of ungulate-infecting PVs, with known host species being bovidae (cows), ovidae (sheep), and cervidae (deer,

roe deer, reindeer and elk). The data set analysed in this thesis did not contain the ovidae and reindeer PV types; however, previous phylogenetic estimates that have included these types have found that the topology estimated within the δ PV clade does not follow the speciation patterns of the associated hosts. The principal incongruity is the closer association of the ovine PV types with the cervidae PV types than with the bovine PV types, whereas the ovine hosts are more closely related to bovine species than to the cervidae species.

The omission of the ovine PV sequences from this analysis resulted in the topology of the δ PV clade appearing congruent with that of the associated hosts: the bovine PV types clustered together (p=1.0), the cervidae PV types clustered together (p=1.0), and the topology of PV types within the δ cerivdae clade mirrored that of the hosts (p=1.0) in the E1 and L1 gene trees. Thus, assuming cophylogeny equates to codivergence, the apparent cophylogenetic structure of the analysed δ PVs would be inferred as evidence of codivergnce of the δ PVs. However, the biased sampling approach taken, which compares viral divergence times against the corresponding host speciation times, rejected the null hypothesis of codivergence in favour of prior divergence at the split of the δ bovidae-δ cervidae PV lineages and within the cervidae PV clade.

The size of the bias applied is important. If the bias towards codivergence is not sufficiently strong, the MCMC sampling will be dominated by irrelevant timescales, and the posterior probabilities of both real and synthetic data will include negligible cospeciation posteriors, resulting in lack of statistical power. Conversely, when the bias is too strong the MCMC mixing times become inconveniently long; this is especially a problem when there is evidence rejecting cospeciation based on minimal posteriors, as occurred with the higher bias used in this paper. It is best to be suspicious of results rejecting codivergence unless the results concur across multiple MCMC threads, as in the results reported here.

The statistical power of the biased sampling analysis employed in this thesis is also reduced by the conservative nature of the assumption of the general predominance of cospeciation. An examination of the power of the method using simulated data sets is required to determine its statistical capabilities and the extent of

the effect of taxon sampling and the applied bias on the results. Comparison of viral speciation times with that of their hosts will always be conservative, however, as prior radiation and host transfer events that occur within the uncertainty of the host speciation time cannot be detected with this method.

The calculations described here are computationally intensive, as the MCMC analysis must be repeated for each of the parametric bootstrap simulations. Parametric bootstrapping to determine the statistical significance of violations of codivergence was necessary as the use of likelihood penalties and an improper prior distribution on divergence times meant that the resulting MCMC chain may not to reflect the posterior distribution. The benefits of temporal comparison and Bayesian phylogenetic analysis have already been detailed. A less computationally demanding approach that combines these aspects in host-associate cophylogenetic analysis would be to develop the method of Huelsenbeck, Rannala and Larget (2000). The statistical method employed here is to incorporate a host transfer prior into Bayesian phylogenetic analysis, to model codivergence and host transfer of a parasite/virus along a host tree. A Bayesian approach that includes all four events by including priors for prior divergence and sorting events would allow posterior probabilities for these events to be determined at each node and would circumvent the dependency on estimates of host speciation times.

Acknowledgement

I would like to thank Andrew Rambaut for advice on the use of BEAST.

Conclusion

The PVs present an interesting family of viruses for evolutionary studies: they constitute a large, continually expanding, family that has diversified to form strong associations with many different host species, to target specific anatomical sites, and though they are largely innocuous parasites, they have also evolved certain high-risk types that demonstrate the potential to cause cancer. This last discovery generated substantial medical interest towards the PVs and the efforts of much biological research have succeeded in producing two vaccines to prevent against cervical cancer-causing HPV infections.

Efforts to understand the evolutionary dynamics of the PVs have not made as much progress. A key question concerns the means by which PVs have been transmitted between species to produce the observed host range, currently comprising species from the reptilian, avian and mammalian orders. Phylogenetic estimations of the PVs present a picture that is difficult to interpret: distinct PV types isolated from different host species' do not display a branching pattern that is concordant with that of the associated hosts. The most notable incongruity observed is the failure of intra-host PV types to cluster together. Large distances are also observed between PV types from closely related hosts. This contradicts our expectations of an evolutionary scenario in which these host-specific, slow-evolving viruses simply tracked their hosts, speciating only when their hosts did.

Incongruent phylogenetic patterns between parasites and their hosts are often interpreted as symptomatic of inter-species transmissions, an event in which a parasite species has successfully crossed species boundaries and established productive infection in a new host species. The multiple incongruities observed between topologies of the PV tree and the host tree would therefore suggest the possibility of multiple host transfer events in recent history. The prevalence of PV infection in humans coupled with detected infection in various domestic species and livestock

provides ample means for the virus to jump between hosts, however the PVs have so far demonstrated an inability to establish productive infection in new hosts. The paucity of data supporting host transfer events generates further curiosity in the observed PV-host phylogenetic incongruities and there is great interest in resolving these differences by determining the true nature of the events that produced the observed PV-host associations. Phylogenetic observations have generated much speculation on this topic however analytical methods are yet to be applied on a comprehensive data set of the PVs. In this thesis, I have performed the first characterisation of ancestral diversification mechanisms of the PVs.

To characterise the evolutionary history of the PVs I devised a method in which temporal comparisons of host and virus speciation events could be made in the absence of known viral speciation times or a constant rate of evolution. Bayesian methods of phylogenetic analysis, which allow sampling of phylogenetic parameters, were utilised to sample divergence times between PV lineages. PV divergence times were biased towards those of the corresponding host, in accordance with the null hypothesis of cospeciation, by imposing a likelihood penalty on all viral divergence times sampled outside of the temporal range of the corresponding host speciation. The imposition of a penalty provides a means of identifying those PV speciation events for which the genetic data presents substantial support against cospeciation. A Bayesian MCMC chain generated under such conditions will therefore sample divergence times corresponding to cospeciation and/or non-cospeciation for each node. In performing a Bayesian analysis, one would expect to be able to make inferences based on the posterior distribution: the proportion of times sampled from within the host speciation range would be interpreted as the posterior probability of cospeciation of the virus lineage and for violations of cospeciation, the posterior probability of prior divergence (host transfer) would be derived from the proportion of times sampled before (after) host speciation. In this analysis however, the penalised distribution, which acts as a prior distribution on viral divergence times, is specified over an infinite range and therefore non-integrable. A likely consequence of using an improper prior distribution is the estimation of an improper posterior distribution thus we cannot be certain that inferences made from the sampled chain will be reflective of the posterior probabilities. To perform a statistical evaluation I employed the parametric

bootstrapping approach to determine p-values for the observed violations of cospeciation, i.e., the sampled proportions of host transfer and prior divergence.

I applied this biased sampling approach to analyse the highly conserved E1 and L1 genes of 108 PVs covering 18 different host species. For both the E1 and the L1 data set the results demonstrate substantial support in favour of an ancient association of the PVs with their hosts. There is also strong support for the theory that, despite multiple incongruities between the host and virus phylogenies, new PV-host associations have largely been acquired by descent and not by host transfer events. This is not equivalent to saying that PV lineages have cospeciated with their hosts. The sampled times indicate a number of statistically significant prior divergence events, where adaptive radiation of virus lineages resulted in multiple lineages associated with ancestral hosts. Further identification of PV types will provide the only means of determining whether these multiple lineages then cospeciated with their hosts and survived to the present day or whether they have been lost - either through extinction at some point or due to incomplete lineage sorting in the speciated hosts such that not all lineages are inherited by the new hosts. The absence of fossil data for viruses renders it impossible to discern between the last two situations however the estimated divergence times indicate the existence of at least 7 PV lineages prior to the separation of the Euarchontoglires (primates/rodents/lagomorphs) and the Laurasiatheria (artiodactyla/carnivore/cetacea), which suggests the potential for substantial PV diversity among the mammalian kingdom. It will be very interesting to see how much PV diversity is discovered in the animal kingdom given that many animal orders and thousands of species currently remain unrepresented in the PV database. An interesting avenue of future research will involve identification of the molecular changes and environments that facilitated the various within-host adaptive radiation events. Among the HPVs there have been diversifications to cutaneous tissue, mucosal tissue, specific anatomic sites and oncogenecity; however, each of these diversifications represents a large clade of PV types within which the effect of further PV radiations is not known. Identification of the molecular changes responsible for these divergences will provide the first step in elucidating the reasons behind substantial prior divergence events within the PV family.

Whilst the analysis predicts that many PV lineages existed prior to the hosts they currently associate with, there is also some sampling of viral speciation events that occurred after that of the corresponding host. Significant support for these host transfer events was only found for the human-porcupine PV divergence in the L1 gene, however, and the different positions occupied by this clade in the E1 and L1 gene trees may point to an ancestral recombination event at least 40 ma. Although continued sampling has revealed greater diversity in many PV clades, new relatives of the ν-human and σ-porcupine PV lineages are yet to be identified to offer further clarity on the likely events occurring within this clade.

The results obtained may be affected by the imbalance in the data set. Some clades, namely those corresponding to genera populated by the HPVs, have been densely sampled whilst clades formed by other genera are more sparsely sampled. This can affect the accuracy of phylogenetic estimation and the estimation of divergence times. For instance, the relatively recent divergence of the avian PV and the ancestral mammalian PV lineage estimated in this analysis is more likely a consequence of the greater evolutionary distance between the mammalian and avian PVs and an underestimation of this distance due to the lack of sequence data in this region of the tree. For the analysis of the PVs, however, I feel that the omission of sequences to provide a more balanced tree is not the best approach since the removal of data will also affect the estimation process. Increased sampling of PV types from non-human hosts will provide a more balanced tree and allow for more accurate estimation of PV phylogenies, evolutionary rates and diversification times. The PV database is continually expanding and currently covers 39 non-human host species. As the gaps in the host range of PVs begin to be filled in, re-evaluations of temporal congruence between PV and host divergences will serve to refine the evolutionary picture of PV diversification mechanisms presented here.

The tangled evolutionary history of the PVs is further complicated by the finding of different evolutionary histories for each gene. Phylogenetic incongruities with the L2 gene are supported by the identification by recombination detection methods of multiple potential recombination signals in this gene of various PVs. The findings of phylogenetic incongruities between all PV genes would therefore suggest a highly

convoluted evolutionary history of PVs involving multiple ancestral recombination events. The identification of recombinant PV types and recombination breakpoints in PV genomes would therefore form the next area of study. If recombination has been a dominant evolutionary force in the PV family, affecting multiple genes, attempts to understand the evolutionary history of the PVs will face far greater challenges than are currently realised.

Appendix A

Host species	PV Type	PV Genus-Species	GenBank Accession Numbers
Human	HPV32	α-1	X74475
	HPV10	α-2	X74465
	HPV28	α-2	U31783
	HPV29	α-2	U31784
	HPV77	α-2	Y15175
	HPV94a	α-2	AJ620211
	HPV61	α-3	U31793
	candHPV62	α-3	AY395706
	HPV72	α-3	X94164
	HPV83	α-3	AF151983
	HPV84	α-3	AF293960
	candHPV86	α-3	AF349909
	candHPV87	α-3	AJ400628
	candHPV89	α-3	AF436128
	HPV102	α-3	DQ080083
	HPV27	α-4	X74473
	HPV57	α-4	X55965
	HPV106	α-4	DQ080082
	HPV26	α-5	X74472
	HPV69	α-5	AB027020
	HPV82	α-5	AB027021
	HPV30	α-6	X74474
	HPV66	α-6	U31794
	HPV18	α-7	X05015
	HPV39	α-7	M62849
	HPV45	α-7	X74479
	HPV59	α-7	X77858
	HPV68a	α-7	DQ080079
	HPV70	α-7	U21941
	candHPV85	α-7	AF131950
	HPV97	α-7	DQ080080
	HPV7	α-8	X74463
	HPV40	α-8	X74478
	HPV43	α-8	AJ620205
	candHPV91	α-8	AF419318
	HPV16	α-9	K02718
	HPV31	α-9	J04353
	HPV33	α-9	M12732
	HPV35	α-9	M74117
	HPV52	α-9	X74481

Host species	PV Type[a]	PV Genus-Species	GenBank Accession Numbers
Homo sapien (human)	HPV58	α-9	D90400
	HPV67	α-9	D21208
	HPV6	α-10	AF092932
	HPV6b	α-10	X00203
	HPV11	α-10	M14119
	HPV13	α-10	DQ344807
	HPV44	α-10	U31788
	HPV34	α-11	X74476
	HPV73	α-11	X94165
	HPV54	α-13	U37488
	HPV5	β-1	M17463
	HPV5b	β-1	D90252
	HPV12	β-1	X74466
	HPV19	β-1	X74470
	HPV20	β-1	U31778
	HPV21	β-1	U31779
	HPV24	β-1	U31782
	HPV25	β-1	X74471
	HPV36	β-1	U31785
	HPV93	β-1	AY382778
	RTRX7	β-1	U85660
	HPV9	β-2	X74464
	HPV15	β-2	X74468
	HPV17	β-2	X74469
	HPV22	β-2	U31780
	HPV23	β-2	U31781
	HPV37	β-2	U31786
	HPV38	β-2	U31787
	HPV80	β-2	Y15176
	HPV49	β-3	X74480
	HPV75	β-3	Y15173
	HPV76	β-3	Y15174
	candHPV92	β-4	AF531420
	candHPV96	β-5	AY382779
	HPV4	γ-1	X70827
	HPV65	γ-1	X70829
	HPV95	γ-1	AJ620210
	HPV48	γ-2	U31789
	HPV50	γ-3	U31790
	HPV60	γ-4	U31792
	HPV63	μ	X70828
	HPV41	ν	X56147

Host species	PV Type[a]	PV Genus-Species	GenBank Accession Numbers
Pan paniscus (bonobo)	PcPV (PpPV1)	α-10	X62844
Pan troglodytes (common chimpanzee)	CCPV1 (PtPV1)	α-10	AF020905
Macaca mulata (Rhesus monkey)	RhPV1 (MmPV1)	α-12	M60184
Micromys minutus (Muridae)	MmPV (MmiPV1)	π	DQ269468
Mastomys coucha (Muridae)	McPV2	π	DQ664501
Sylvilagus floridanus (Cottontail rabbit)	CRPVb (SfPV1)	κ	AJ243287
Oryctolagus cuniculus (European /domestic rabbit)	ROPV (OcPV1)	κ	AF227240
Erethizon dorsatum (porcupine)	EdPV1	σ	AY684126
Bos Taurus (bovine)	BPV1	δ	X02346
	BPV2	δ	M20219
	BPV3	ξ	AF486184
	BPV5	ε	AJ620206
	BPV7	unclassified	DQ217793
	BPV8	ε	DQ098913
	BPV9	ξ	AB331650
	BPV10	ξ	AB331651
Odocoileus virginianus (deer)	DPV (OvPV1)	δ	M11910
Capreolus capreolus (Roe deer)	RPV (CcaPV1)	δ	AF443292
Alces alces (European Elk)	EePV (AaPV1)	δ	M15953
Phocoena spinipinnis (porpoise)	PsPV (PsPV1)	ο	AJ238373
Tursiops truncatus (Bottlenosed dolphin)	TtPV2	υ	AY956402

Host species	PV Type[a]	PV Genus-Species	GenBank Accession Numbers
Canis familiaris (dog)	CPV2	τ	AY722648
	COPV (CPV1)	λ	D55633
Procyon lotor (raccoon)	PlPV1	λ	AY763115
Felis domesticus (cat)	FdPV (FdPV1)	λ	AF480454

Table A.1: Data set of PV types analysed. [a] abbreviations in brackets indicate new names following reclassification by Bernard et al. (2010).

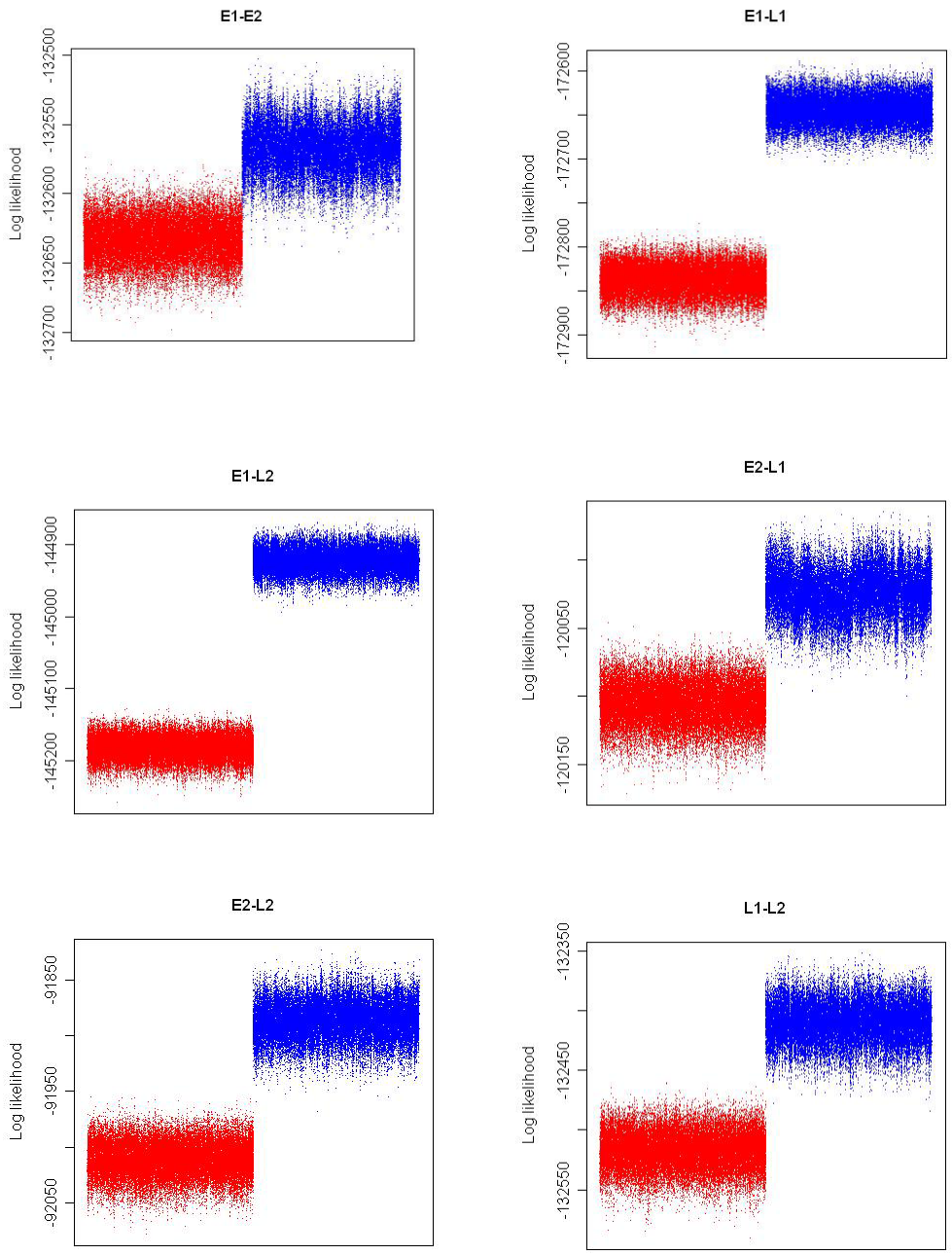

E1-E2
Log likelihood
-132500
-132550
-132600
-132650
-132700
E1-L1
Log likelihood
-172600
-172700
-172800
-172900
E1-L2
Log likelihood
-144900
-145000
-145100
-145200
E2-L1
Log likelihood
-120050
-120150
E2-L2
Log likelihood
-91850
-91950
-92050
L1-L2
Log likelihood
-132350
-132450
-132550

Figure A.2

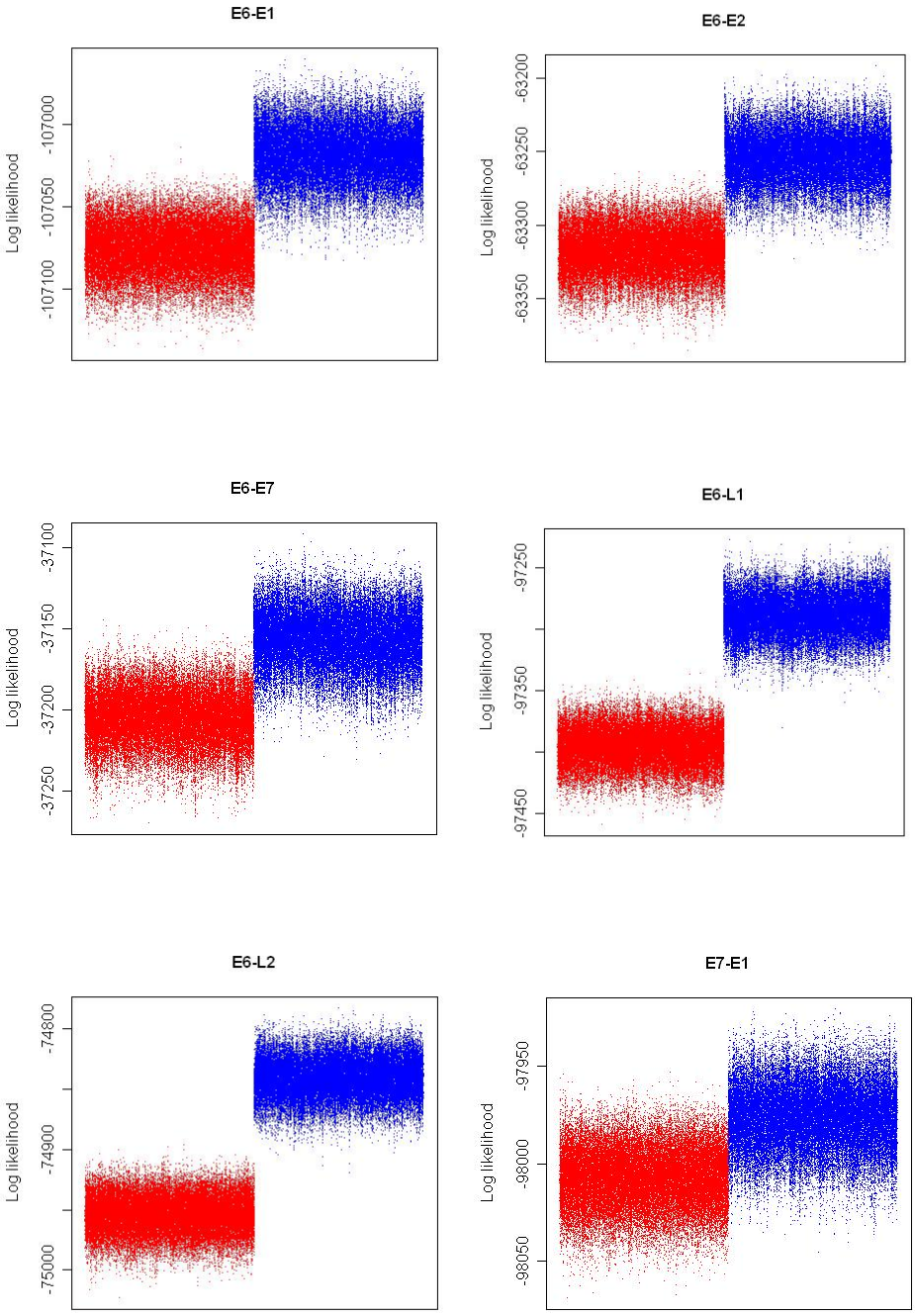

E6-E1
Log likelihood
-107100
-107050
-107000
E6-E2
Log likelihood
-63350
-63300
-63250
-63200
E6-E7
Log likelihood
-37250
-37200
-37150
-37100
E6-L1
Log likelihood
-97450
-97350
-97250
E6-L2
Log likelihood
-75000
-74900
-74800
E7-E1
Log likelihood
-98050
-98000
-97950

Figure A.2

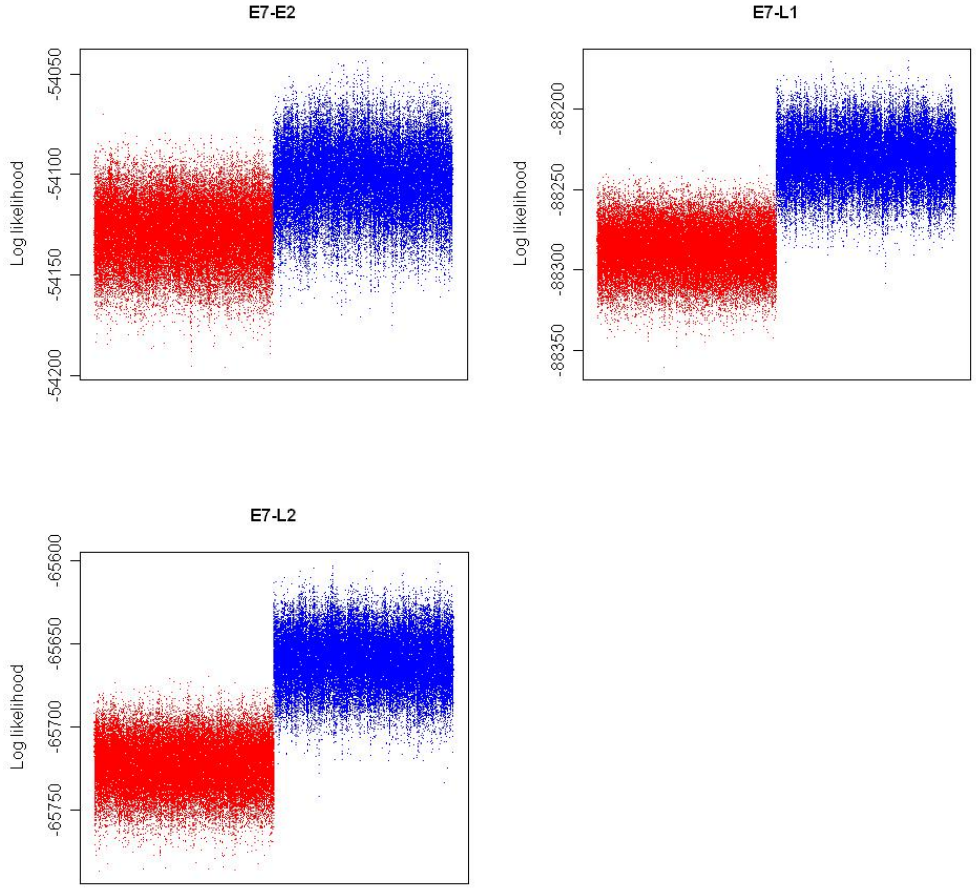


Figure A.2: Sampled likelihoods of paired-gene MCMC chains run with evolutionary parameters constrained to be the same for each gene. Red chain represents topological constraint on paired genes. Blue chain represents independent topologies for each gene.

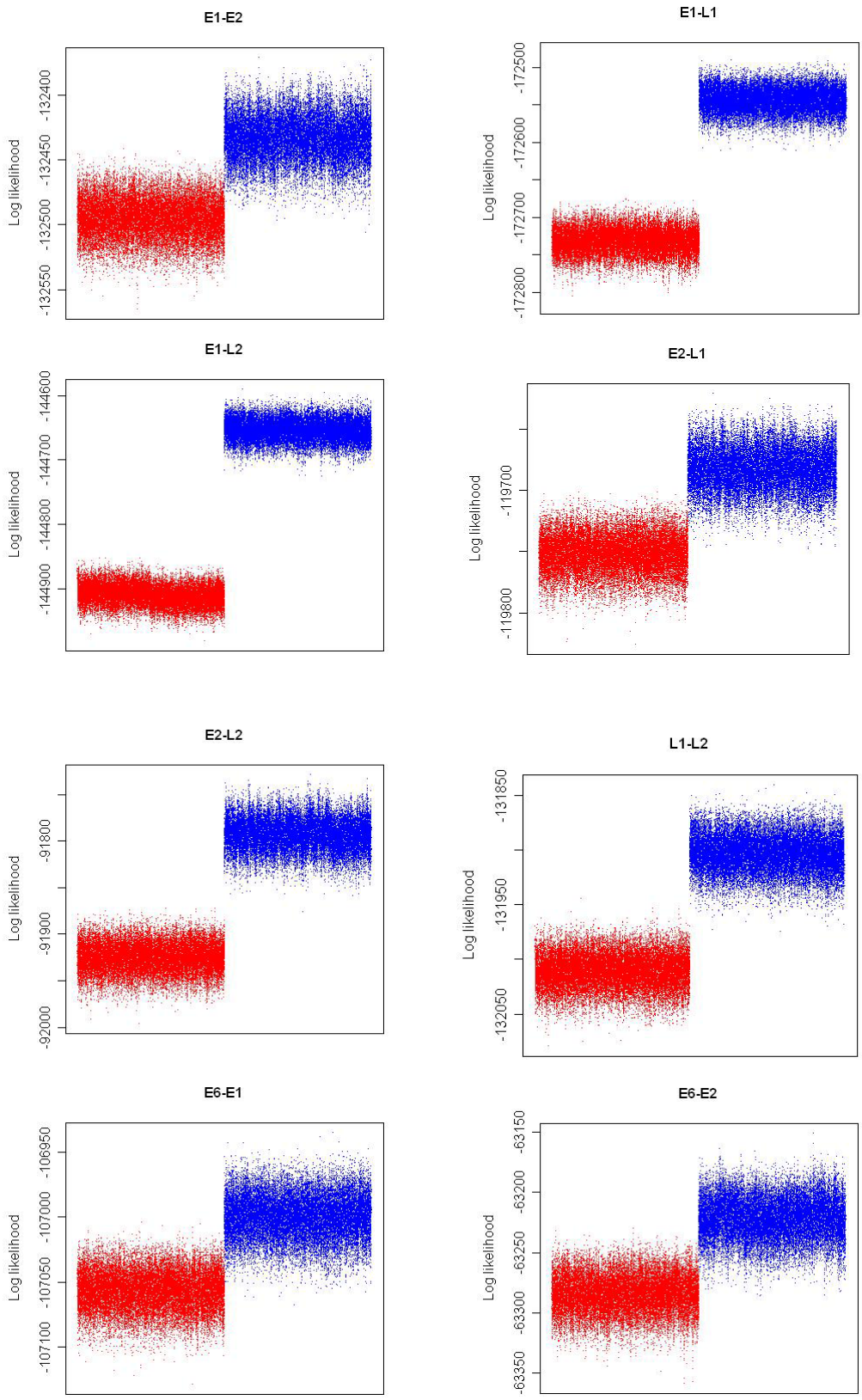

E1-E2
Log likelihood
-132400
-132450
-132500
-132550
E1-L1
Log likelihood
-172500
-172600
-172700
-172800
E1-L2
Log likelihood
-144600
-144700
-144800
-144900
E2-L1
Log likelihood
-119700
-119800
E2-L2
Log likelihood
-91800
-91900
-92000
L1-L2
Log likelihood
-131850
-131950
-132050
E6-E1
Log likelihood
-106950
-107000
-107050
-107100
E6-E2
Log likelihood
-63150
-63200
-63250
-63300
-63350

Figure A.3

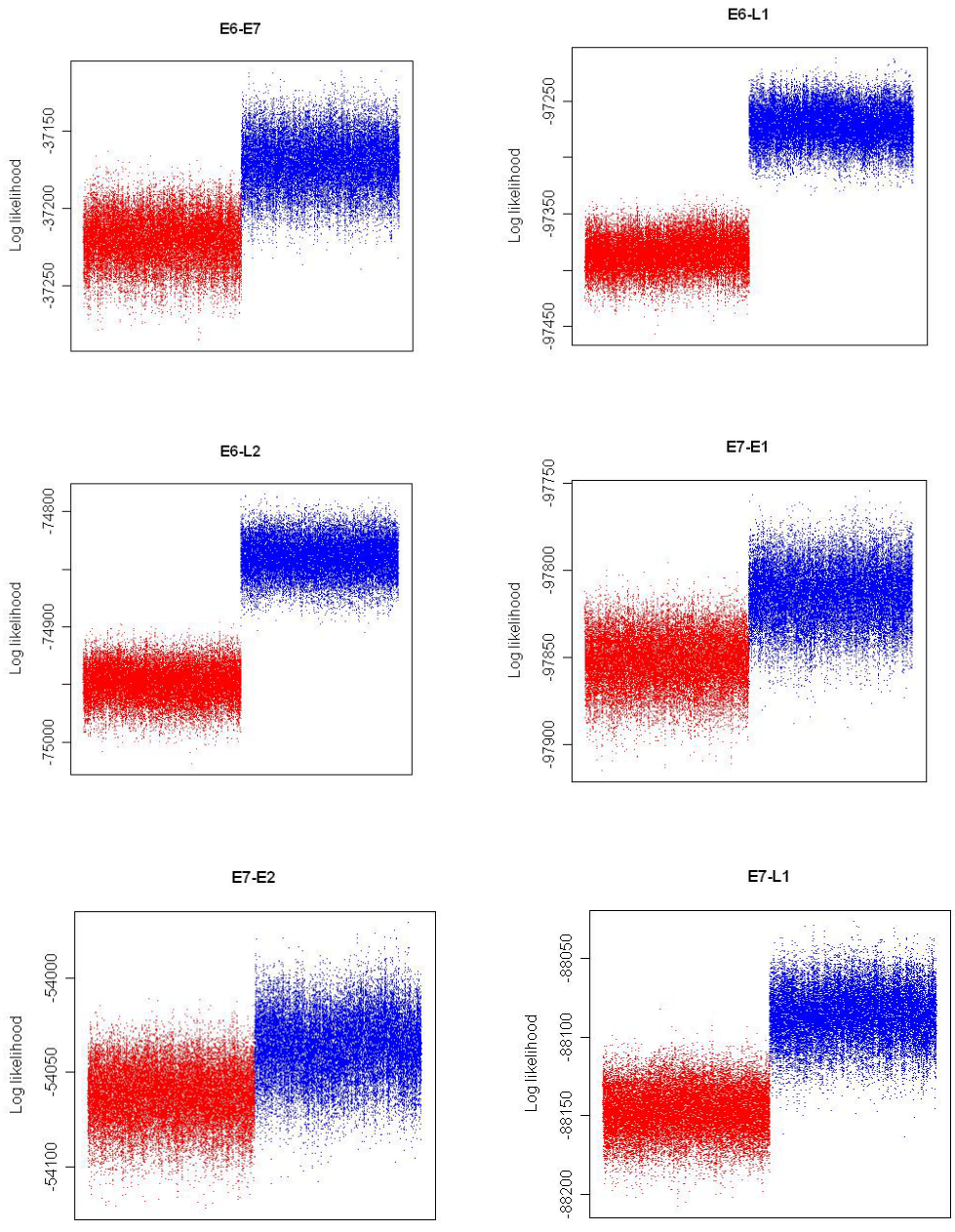

E6-E7
Log likelihood
-37150
-37200
-37250
E6-L1
Log likelihood
-97250
-97350
-97450
E6-L2
Log likelihood
-74800
-74900
-75000
E7-E1
Log likelihood
-97750
-97800
-97850
-97900
E7-E2
Log likelihood
-54000
-54050
-54100
E7-L1
Log likelihood
-88050
-88100
-88150
-88200

Figure A.3

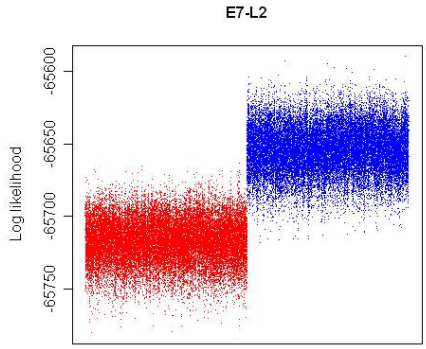


Figure A.3: Sampled likelihoods of paired-gene MCMC chains run with independent evolutionary parameters for each gene. Red chain represents topological constraint on paired genes. Blue chain represents independent topologies for each gene.

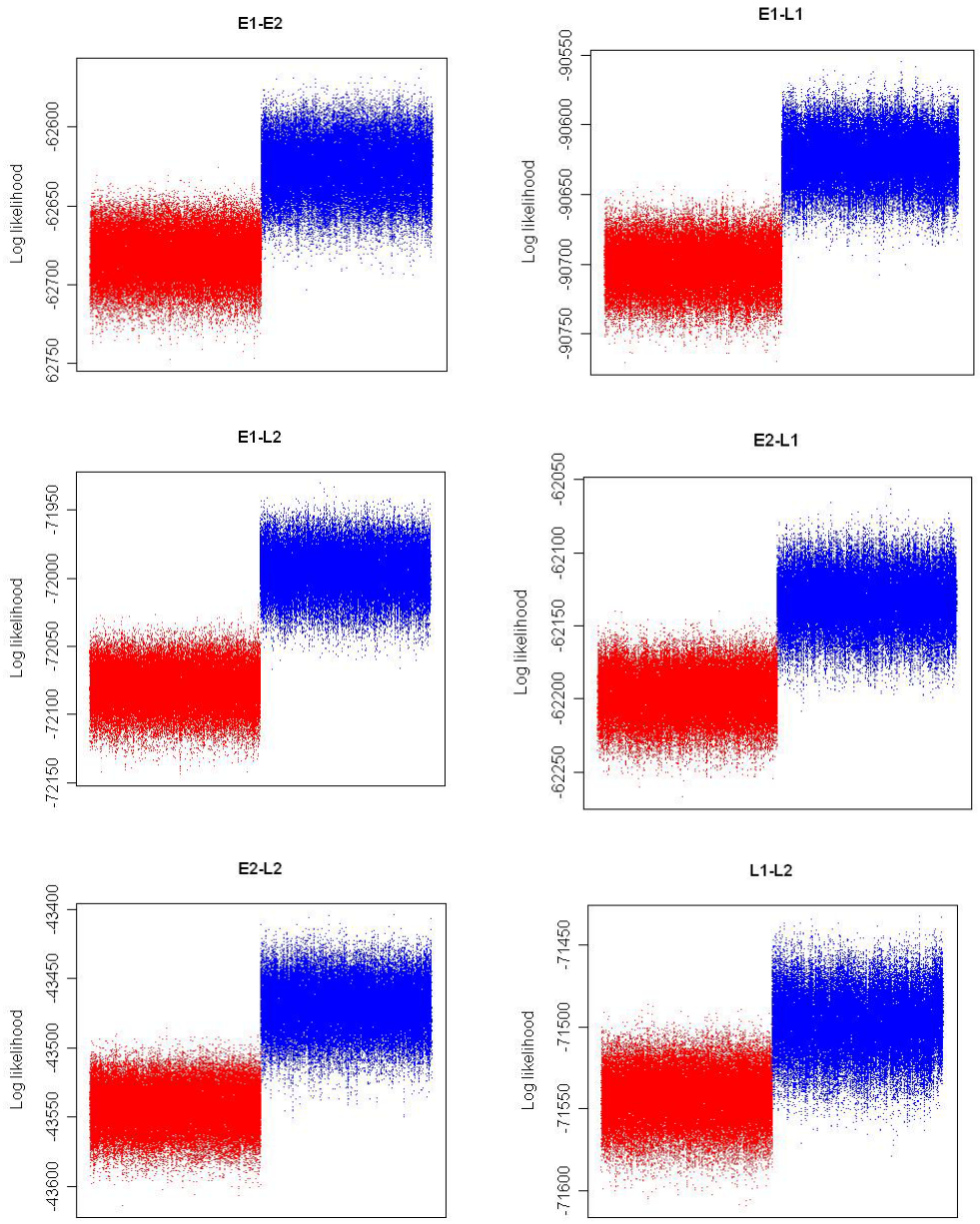


Figure A.4: Sampled likelihoods of paired-gene MCMC chains run using data from the third codon sites only (evolutionary parameters constrained across genes). Chains in which the paired genes are constrained to sample identical topologies are shown in red and chains in which independent topologies are sampled for each gene are shown in blue.

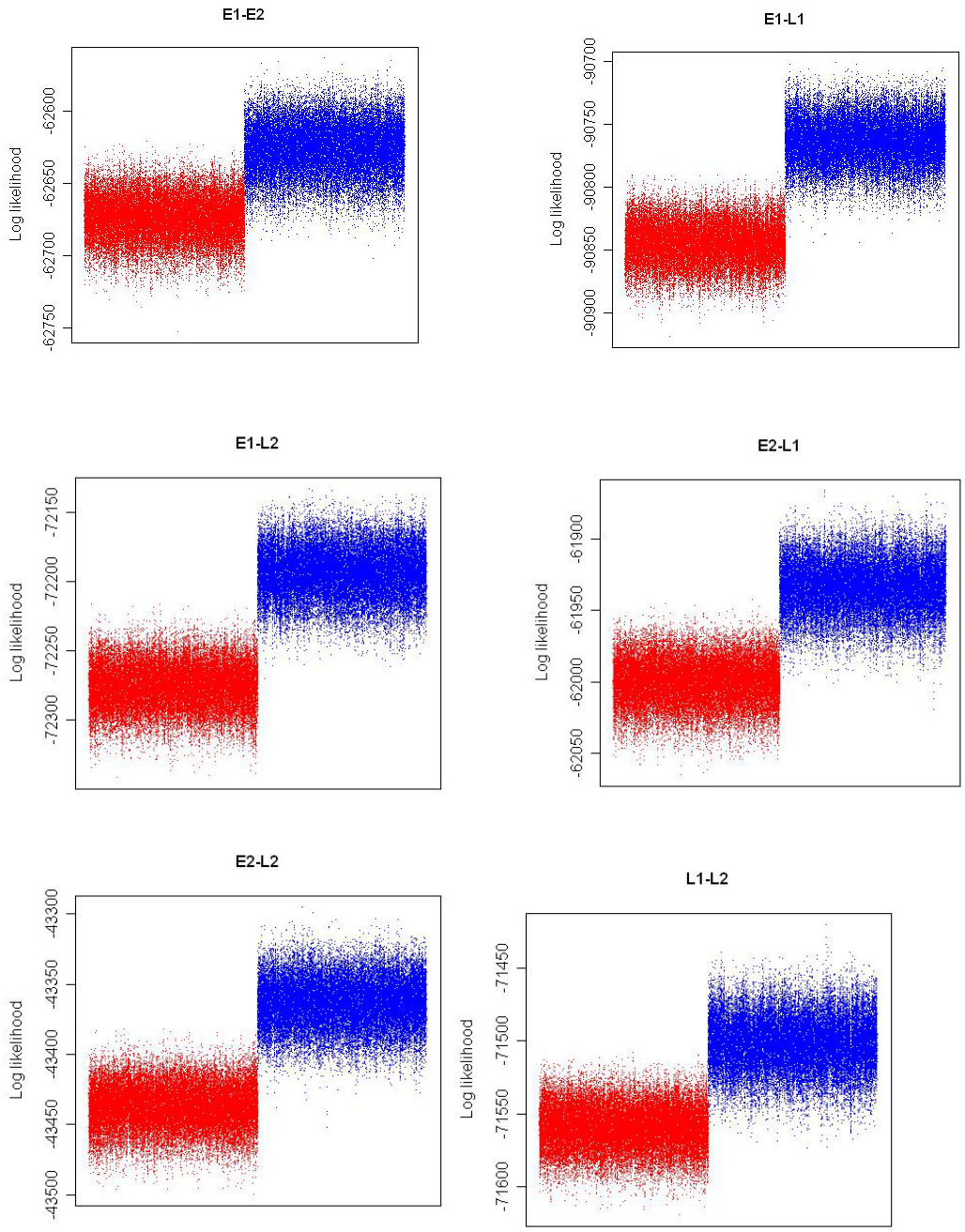


Figure A.5: Sampled likelihoods of paired-gene MCMC chains run using data from the third codon sites only (independent evolutionary parameters across genes). Chains in which the paired genes are constrained to sample identical topologies are shown in red and chains in which independent topologies are sampled for each gene are shown in blue.

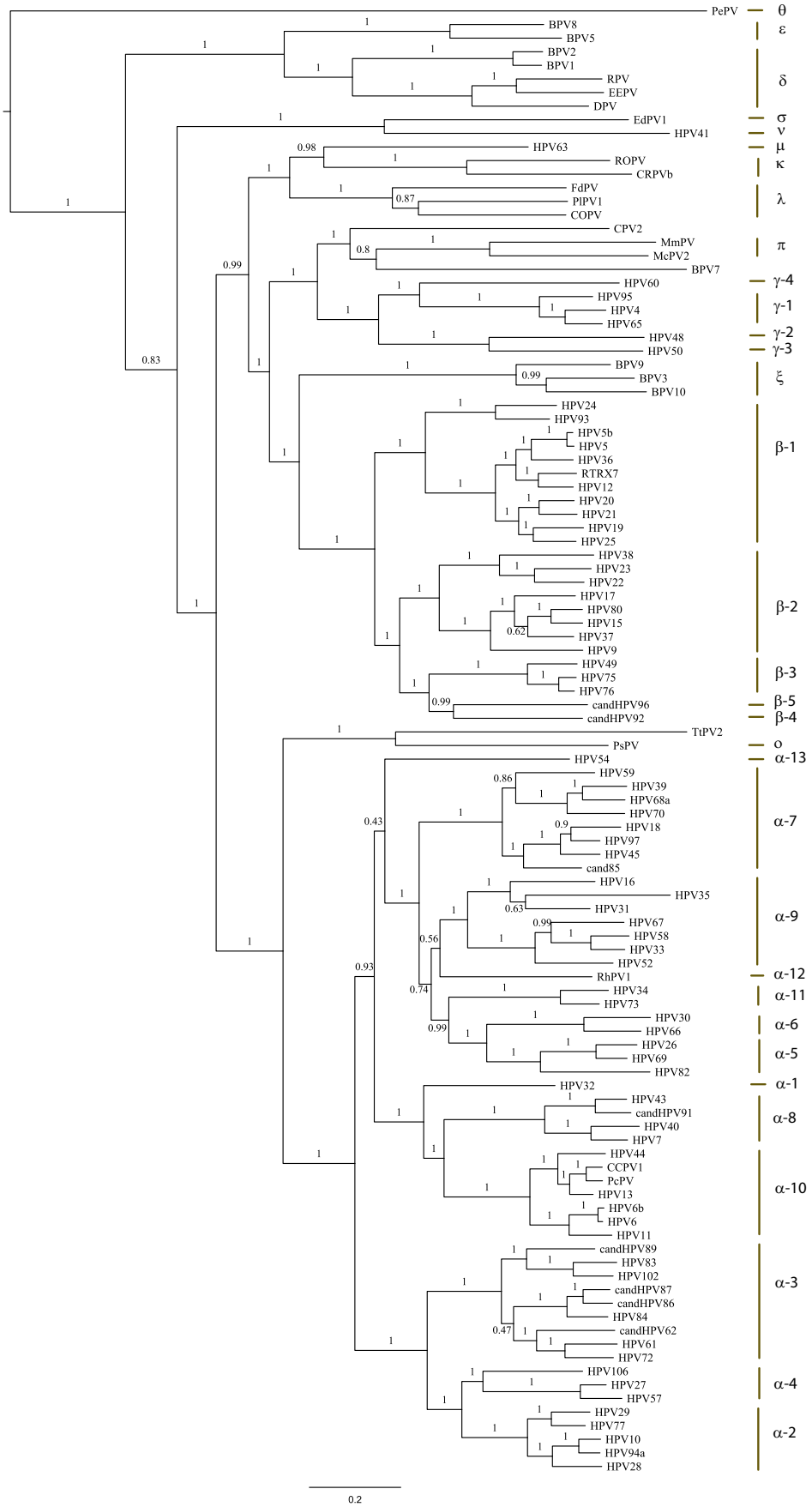


Figure A.6: MAP phylogeny for the E1 gene.

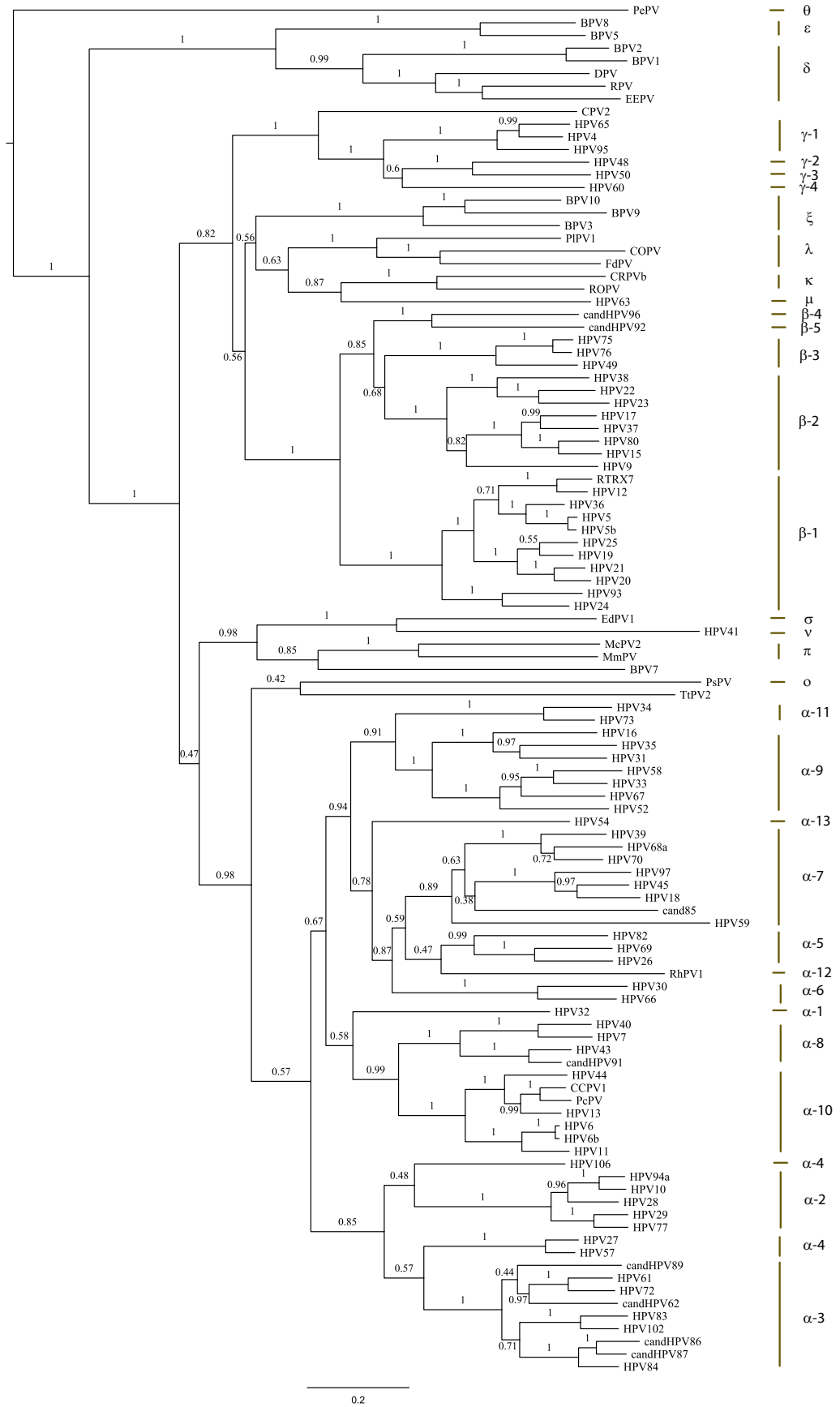


Figure A.7: MAP phylogeny for the E2 gene.

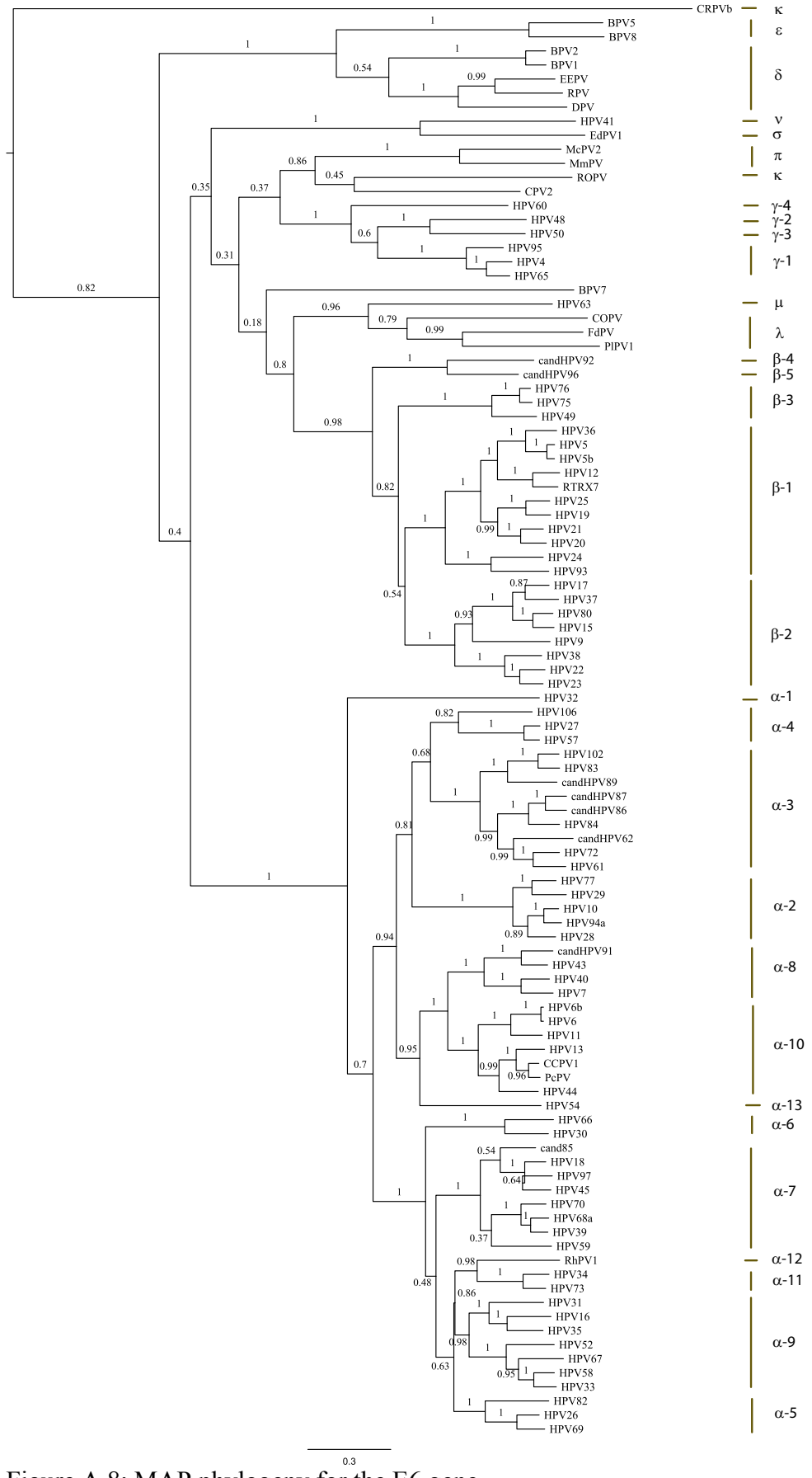


Figure A.8: MAP phylogeny for the E6 gene.

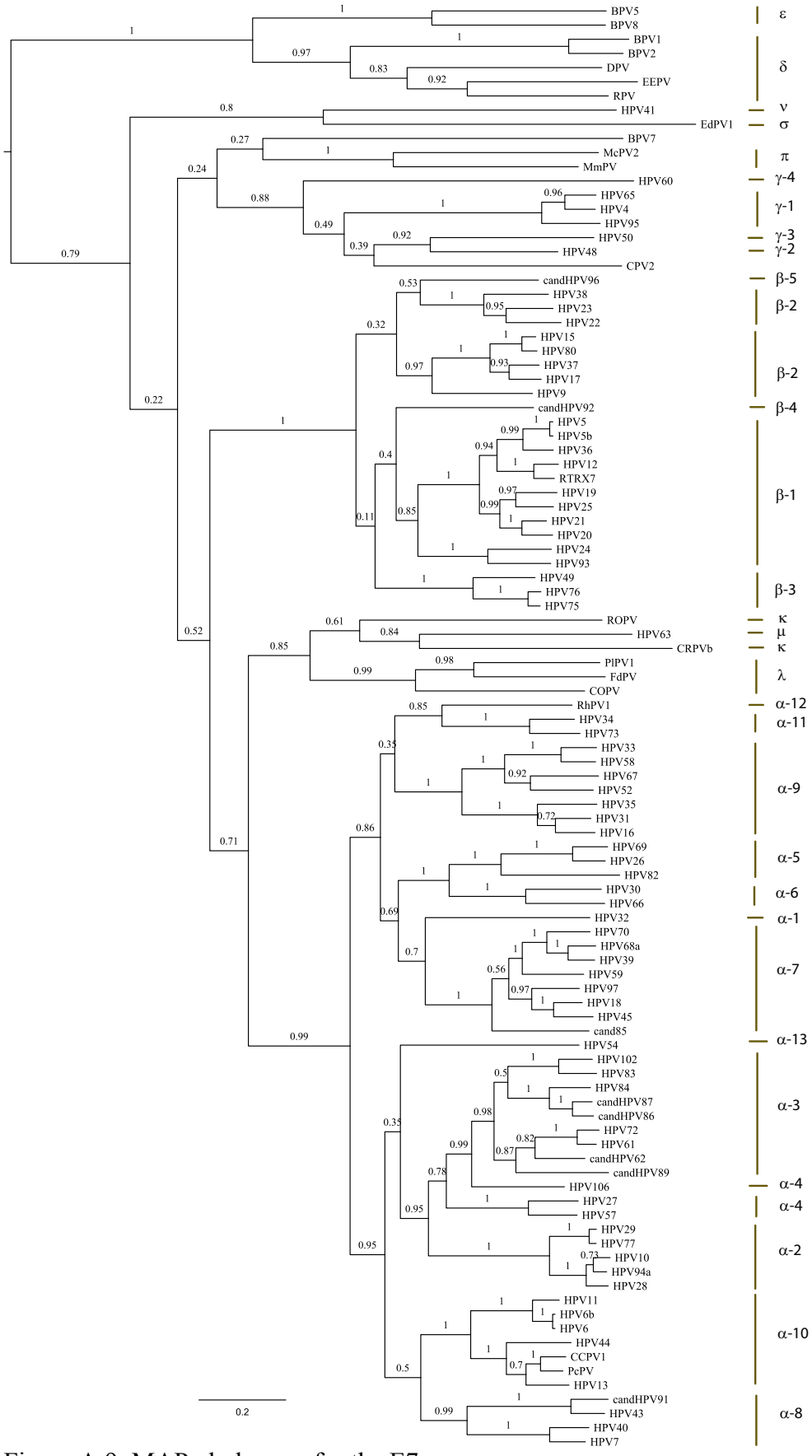


Figure A.9: MAP phylogeny for the E7 gene.

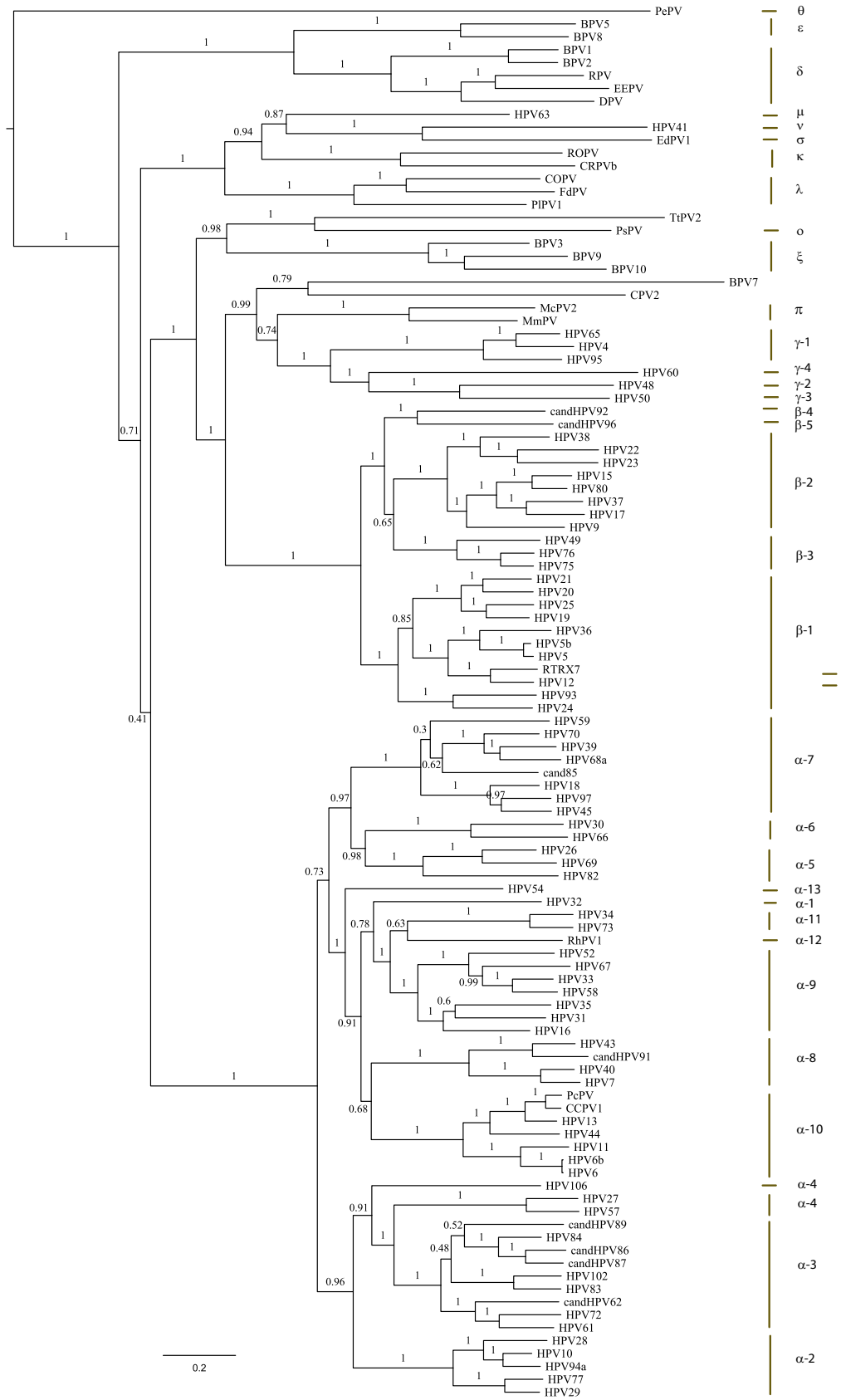


Figure A.10: MAP phylogeny for the L1 gene.

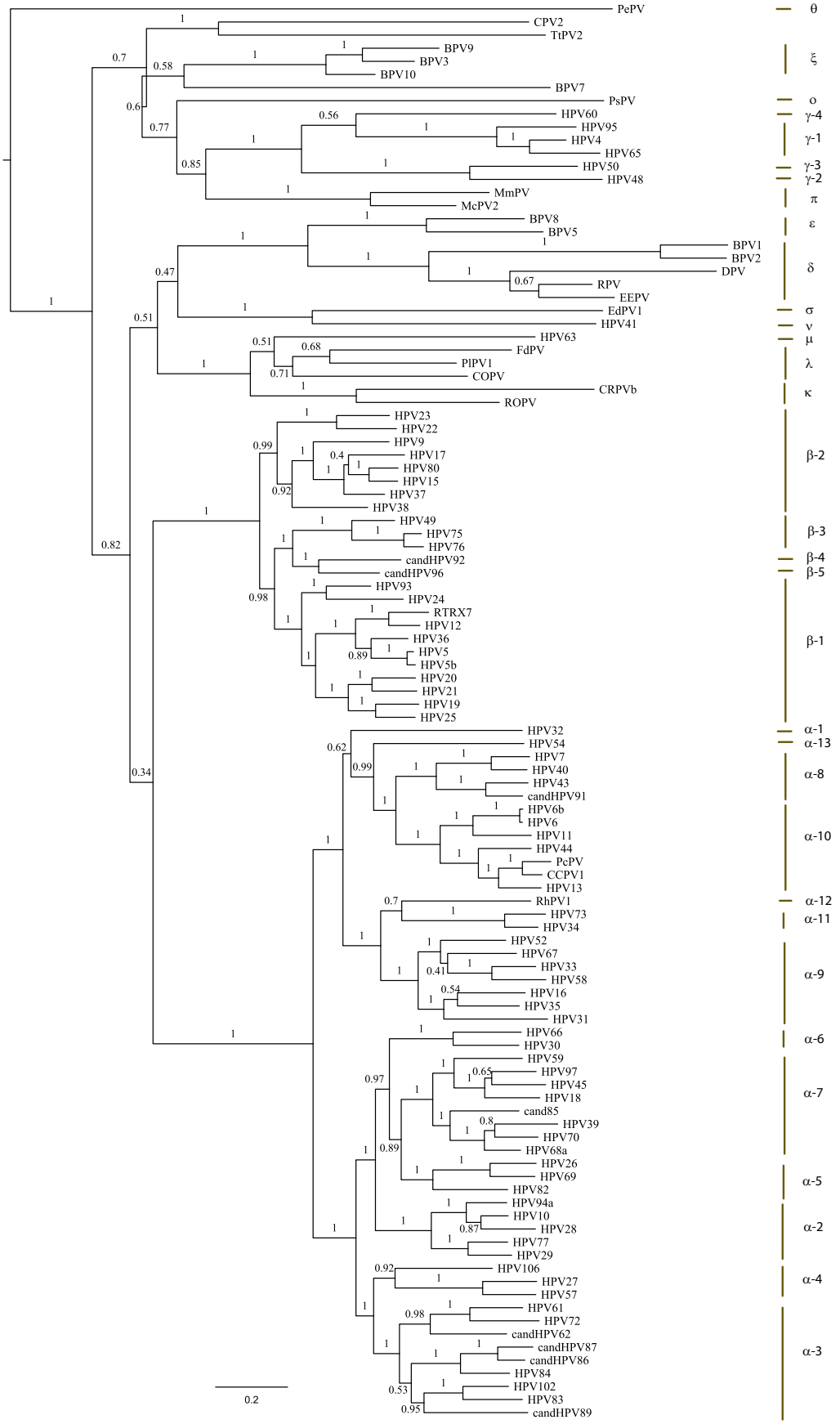


Figure A.11: MAP phylogeny for the L2 gene.

Appendix B

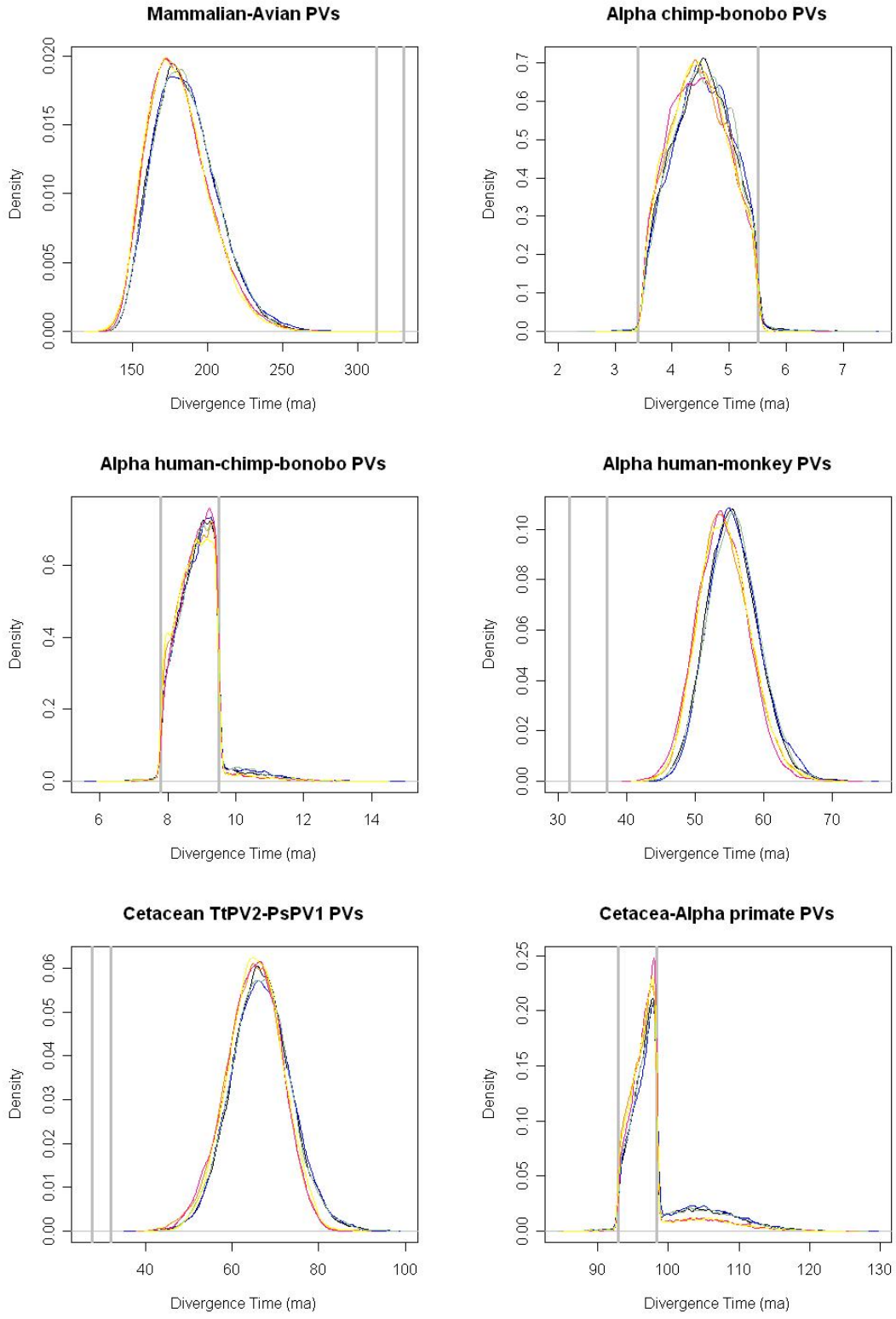


Figure B.1

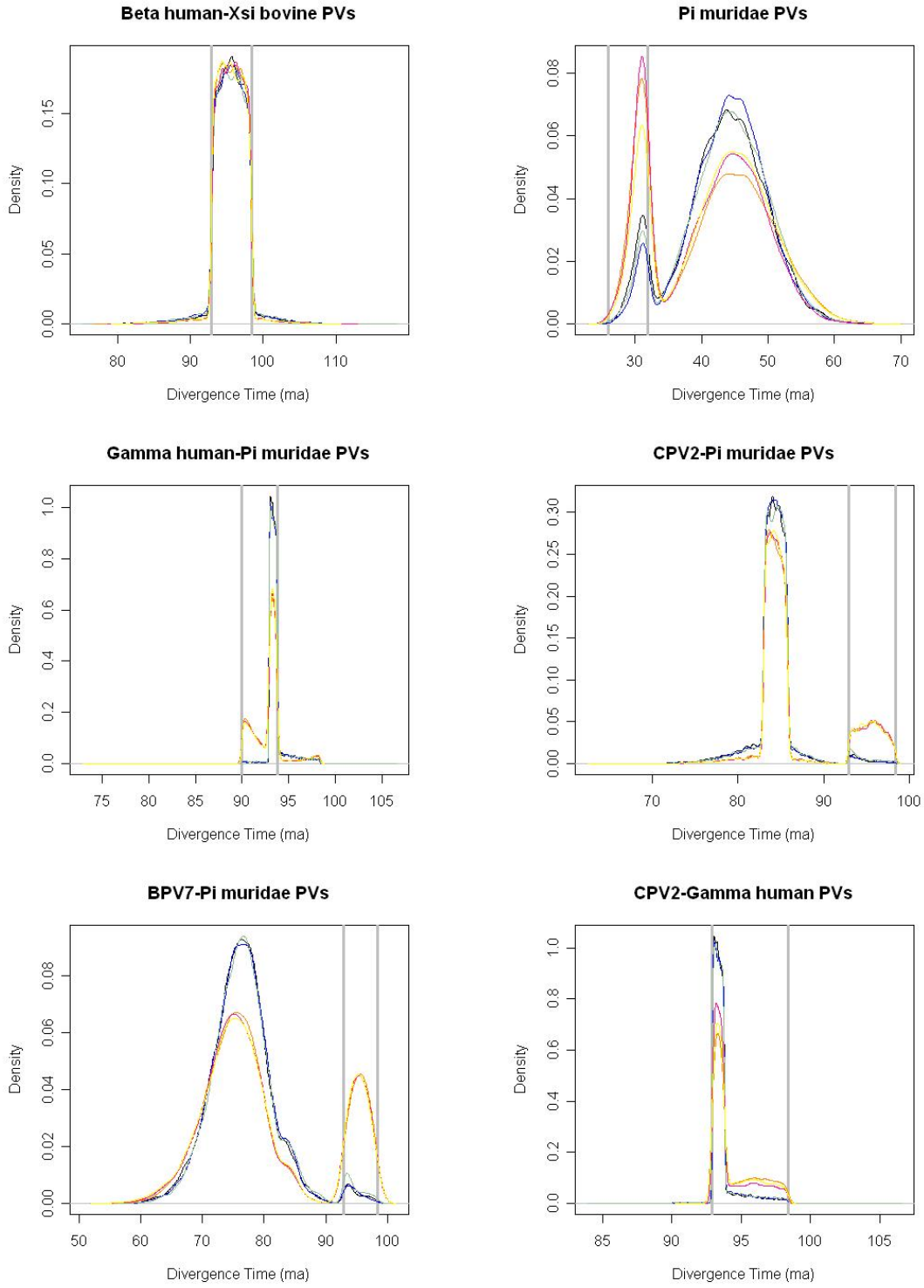

Beta human-Xsi bovine PVs
Density
0.00
0.05
0.10
0.15
80
90
100
110
Divergence Time (ma)
Pi muridae PVs
Density
0.00
0.02
0.04
0.06
0.08
30
40
50
60
70
Divergence Time (ma)
Gamma human-Pi muridae PVs
Density
0.0
0.2
0.4
0.6
0.8
1.0
75
80
85
90
95
100
105
Divergence Time (ma)
CPV2-Pi muridae PVs
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
70
80
90
100
Divergence Time (ma)
BPV7-Pi muridae PVs
Density
0.00
0.02
0.04
0.06
0.08
50
60
70
80
90
100
Divergence Time (ma)
CPV2-Gamma human PVs
Density
0.0
0.2
0.4
0.6
0.8
1.0
85
90
95
100
105
Divergence Time (ma)

Figure B.1

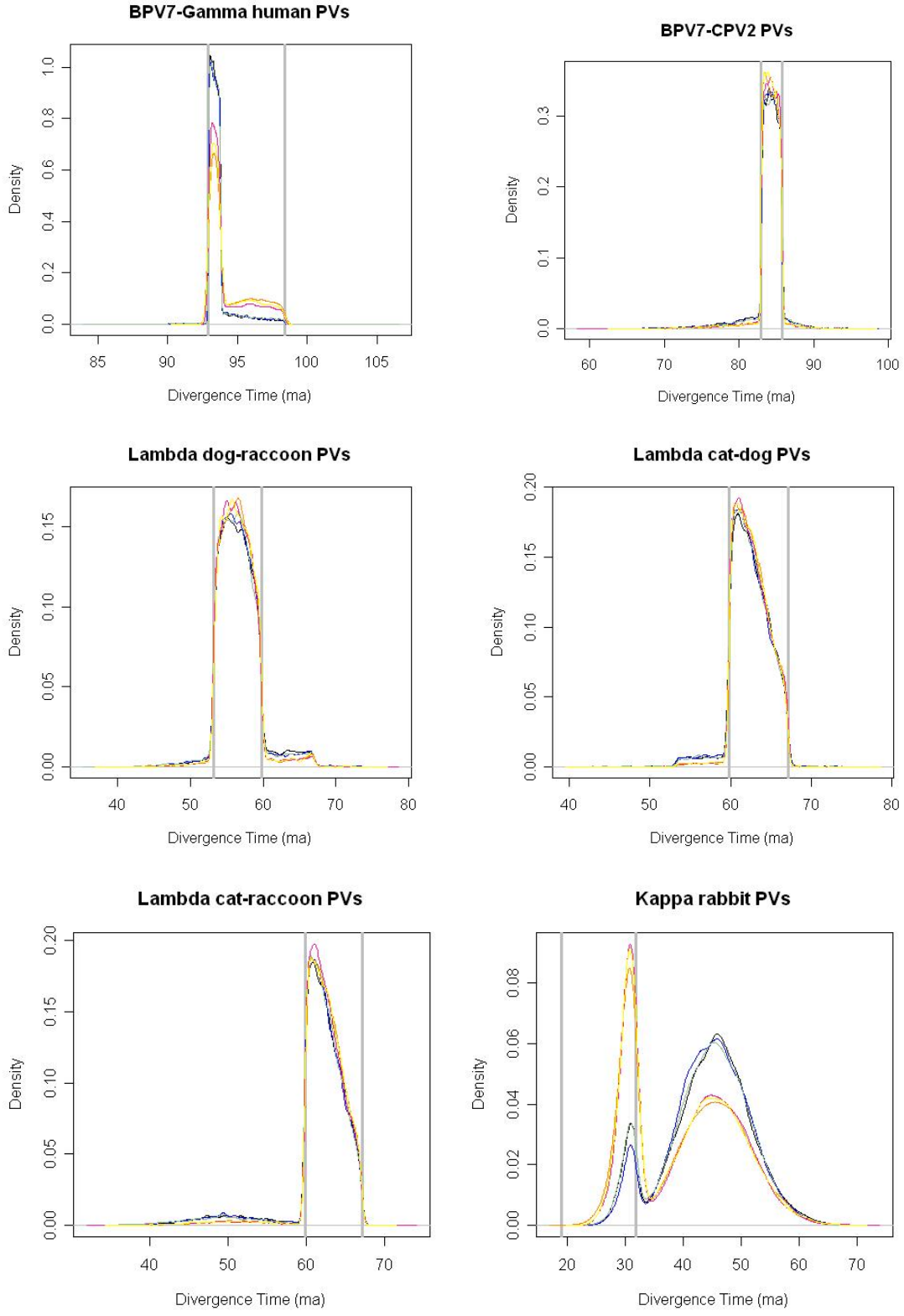

BPV7-Gamma human PVs
Density
Divergence Time (ma)
BPV7-CPV2 PVs
Density
Divergence Time (ma)
Lambda dog-raccoon PVs
Density
Divergence Time (ma)
Lambda cat-dog PVs
Density
Divergence Time (ma)
Lambda cat-raccoon PVs
Density
Divergence Time (ma)
Kappa rabbit PVs
Density
Divergence Time (ma)

Figure B.1

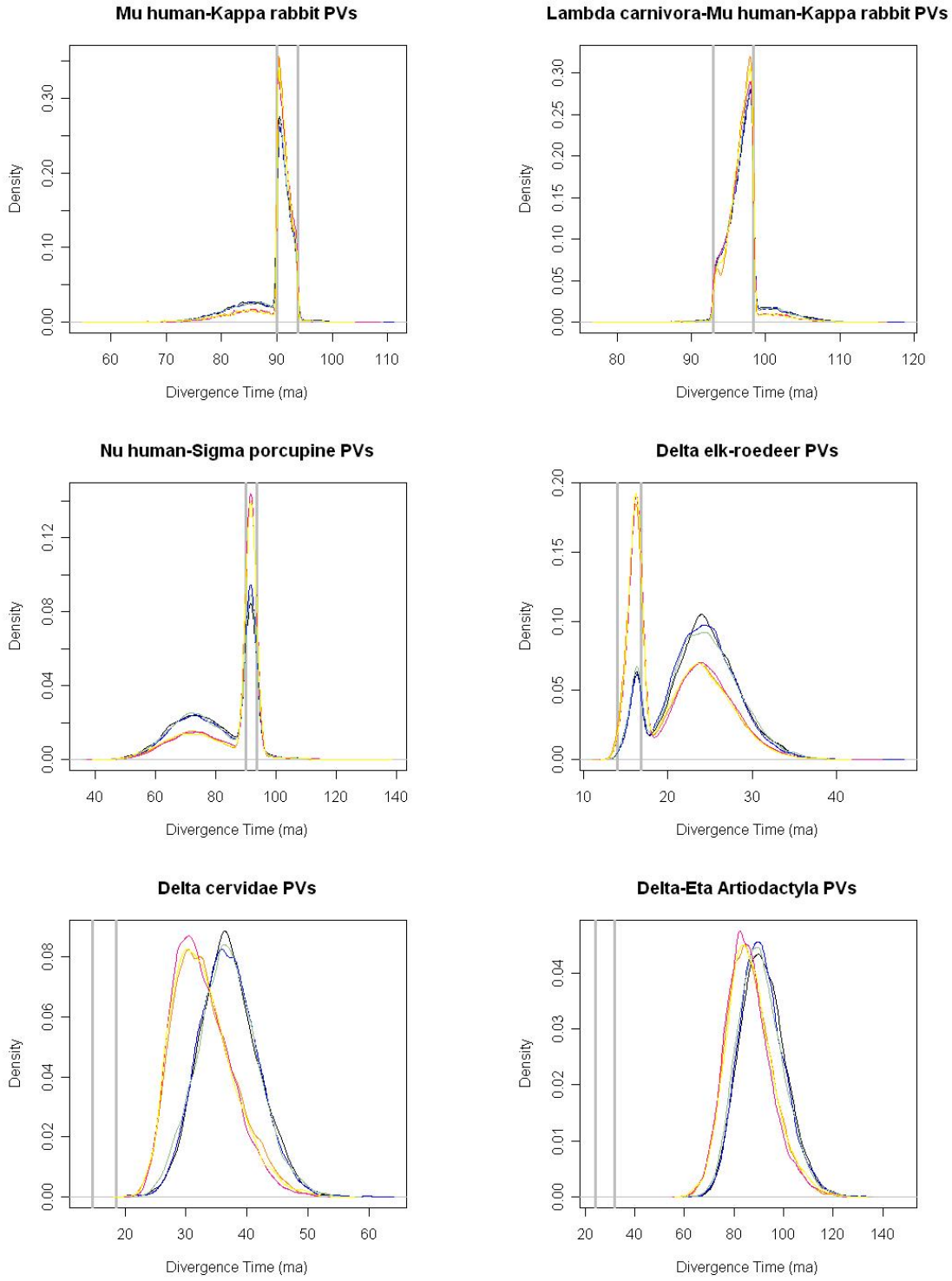


Figure B.1: The sampled times for PV divergences of the E1 gene. Pink, orange and yellow densities indicate sampled times for chains simulated under the ln(0.005) penalty; black, blue and green densities indicate sampled times for chains simulated under the ln(0.05) penalty. The vertical grey bars indicate the speciation range of the corresponding host (as estimated by Bininda-Emonds et al. 2007).

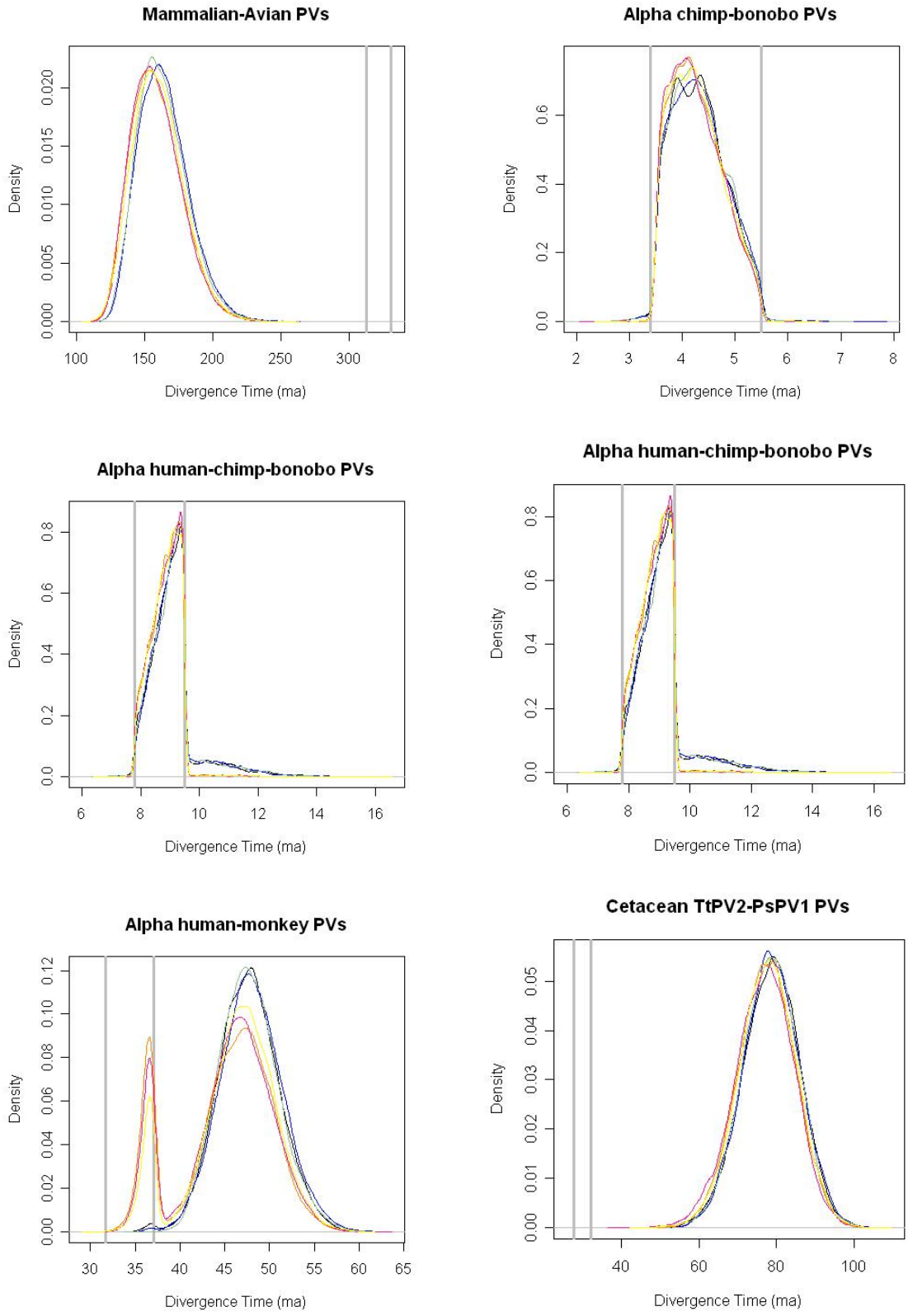

Mammalian-Avian PVs
Density
0.000
0.005
0.010
0.015
0.020
100
150
200
250
300
Divergence Time (ma)
Alpha chimp-bonobo PVs
Density
0.0
0.2
0.4
0.6
2
3
4
5
6
7
8
Divergence Time (ma)
Alpha human-chimp-bonobo PVs
Density
0.0
0.2
0.4
0.6
0.8
6
8
10
12
14
16
Divergence Time (ma)
Alpha human-chimp-bonobo PVs
Density
0.0
0.2
0.4
0.6
0.8
6
8
10
12
14
16
Divergence Time (ma)
Alpha human-monkey PVs
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
30
35
40
45
50
55
60
65
Divergence Time (ma)
Cetacean TtPV2-PsPV1 PVs
Density
0.00
0.01
0.02
0.03
0.04
0.05
40
60
80
100
Divergence Time (ma)

Figure B.2

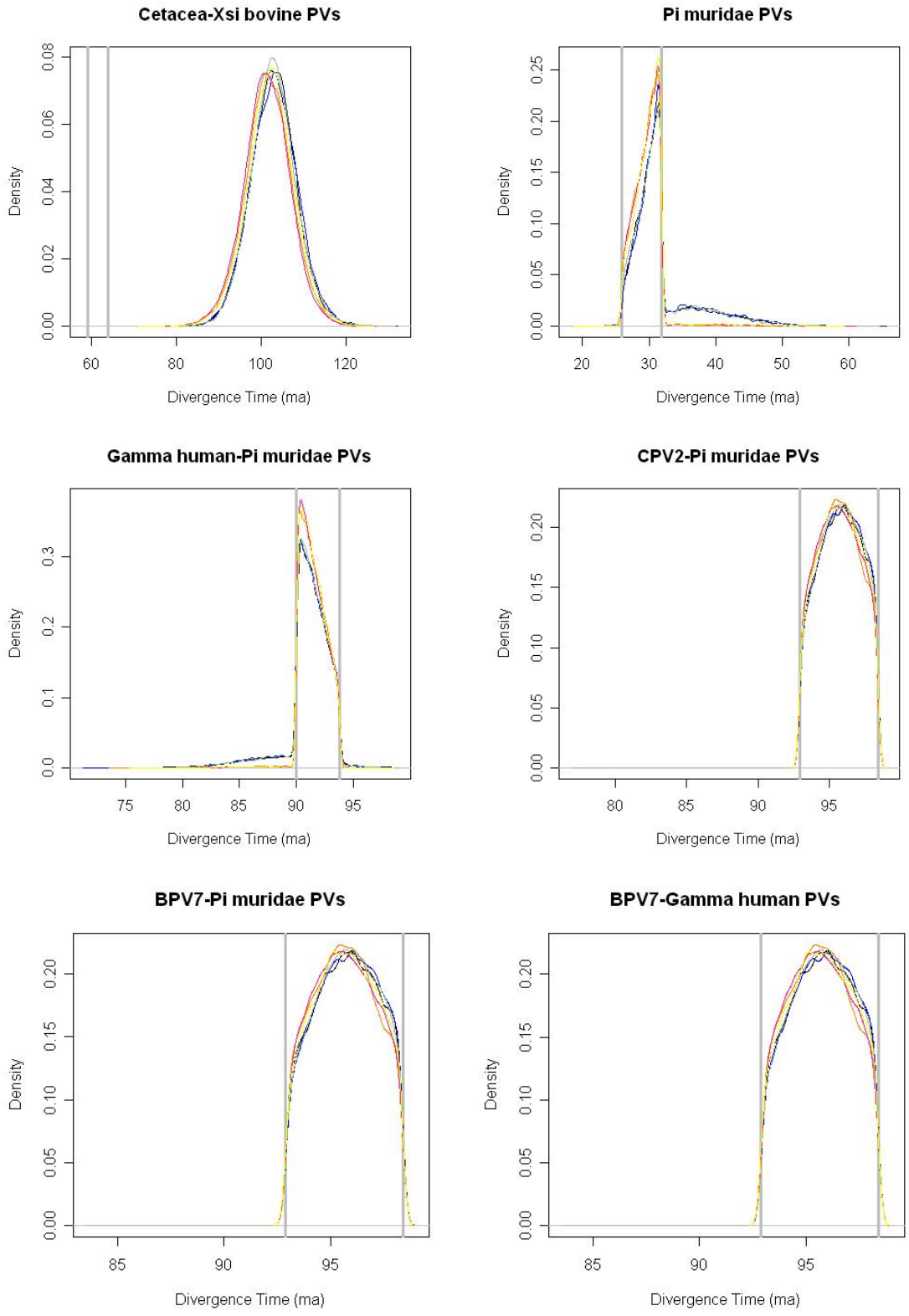

Cetacea-Xsi bovine PVs
Density
Divergence Time (ma)
Pi muridae PVs
Density
Divergence Time (ma)
Gamma human-Pi muridae PVs
Density
Divergence Time (ma)
CPV2-Pi muridae PVs
Density
Divergence Time (ma)
BPV7-Pi muridae PVs
Density
Divergence Time (ma)
BPV7-Gamma human PVs
Density
Divergence Time (ma)

Figure B.2

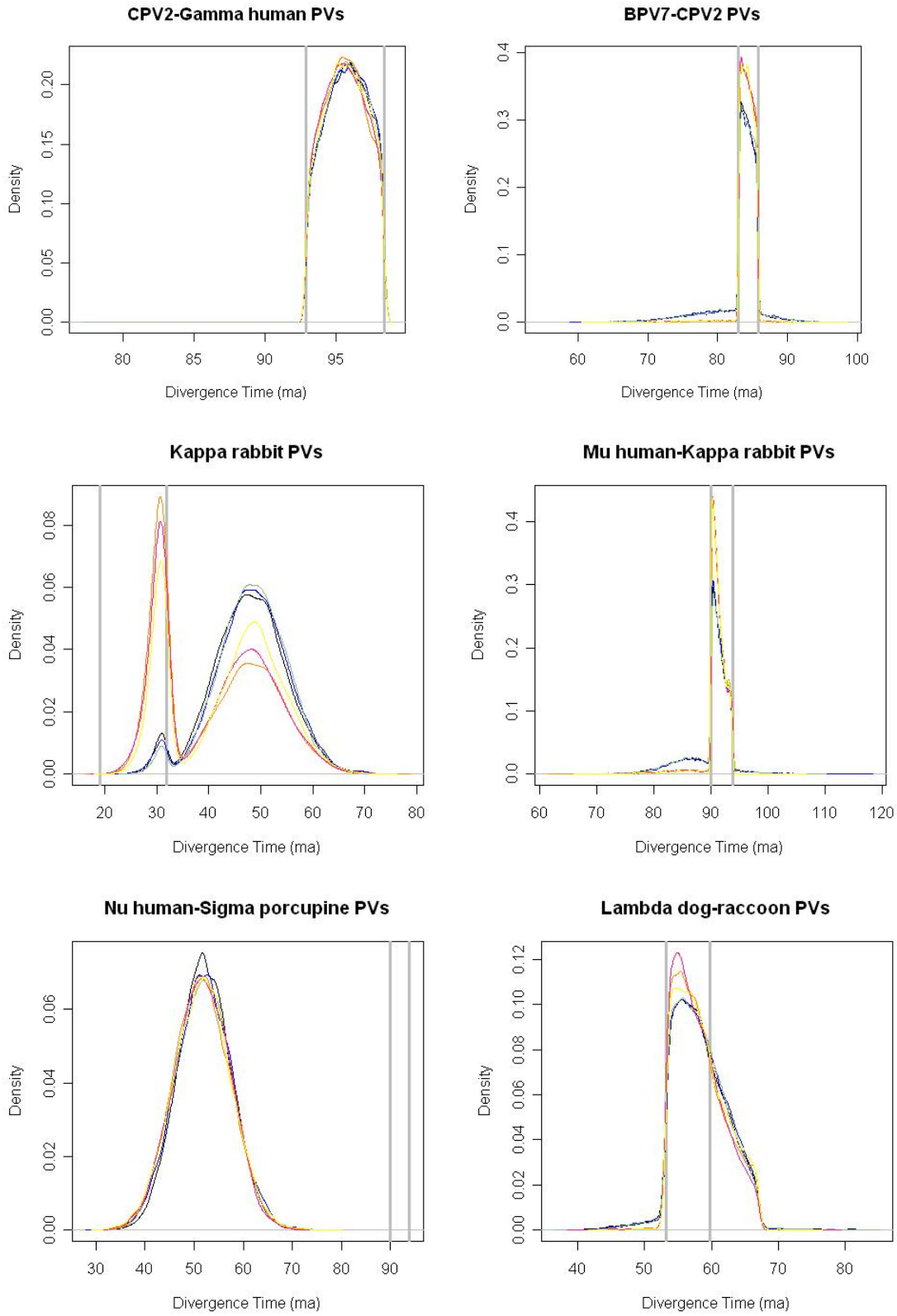

CPV2-Gamma human PVs
Density
Divergence Time (ma)
BPV7-CPV2 PVs
Density
Divergence Time (ma)
Kappa rabbit PVs
Density
Divergence Time (ma)
Mu human-Kappa rabbit PVs
Density
Divergence Time (ma)
Nu human-Sigma porcupine PVs
Density
Divergence Time (ma)
Lambda dog-raccoon PVs
Density
Divergence Time (ma)

Figure B.2

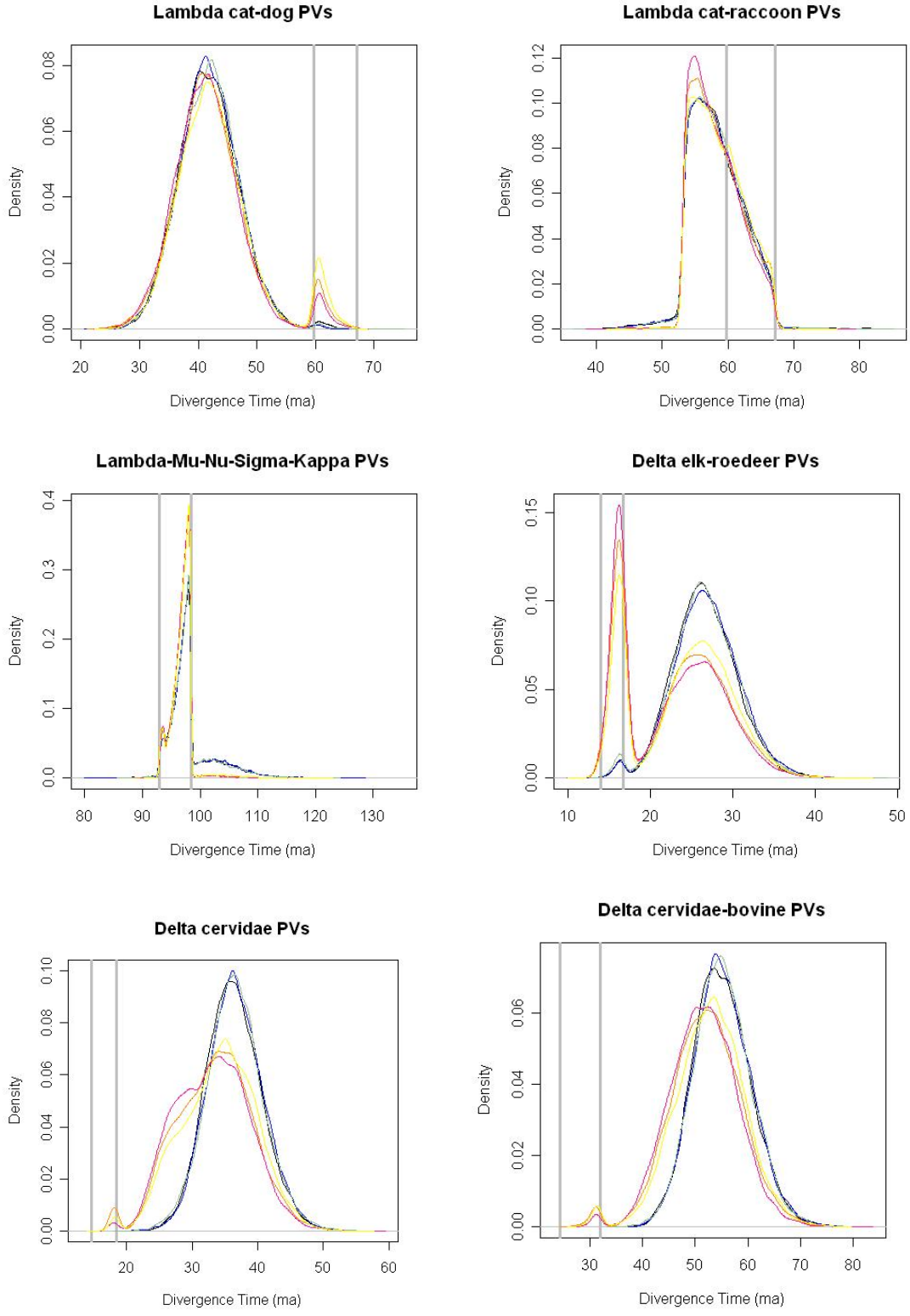

Lambda cat-dog PVs
Density
Divergence Time (ma)
Lambda cat-raccoon PVs
Density
Divergence Time (ma)
Lambda-Mu-Nu-Sigma-Kappa PVs
Density
Divergence Time (ma)
Delta elk-roedeer PVs
Density
Divergence Time (ma)
Delta cervidae PVs
Density
Divergence Time (ma)
Delta cervidae-bovine PVs
Density
Divergence Time (ma)

Figure B.2

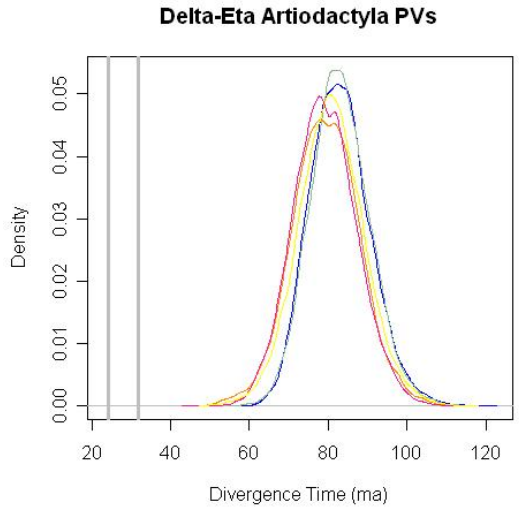


Figure B.2: The sampled times for PV divergences of the L1 gene. Pink, orange and yellow densities indicate sampled times for chains simulated under the ln(0.005) penalty; black, blue and green densities indicate sampled times for chains simulated under the ln(0.05) penalty. The vertical grey bars indicate the speciation range of the corresponding host (as estimated by Bininda-Emonds et al. 2007).

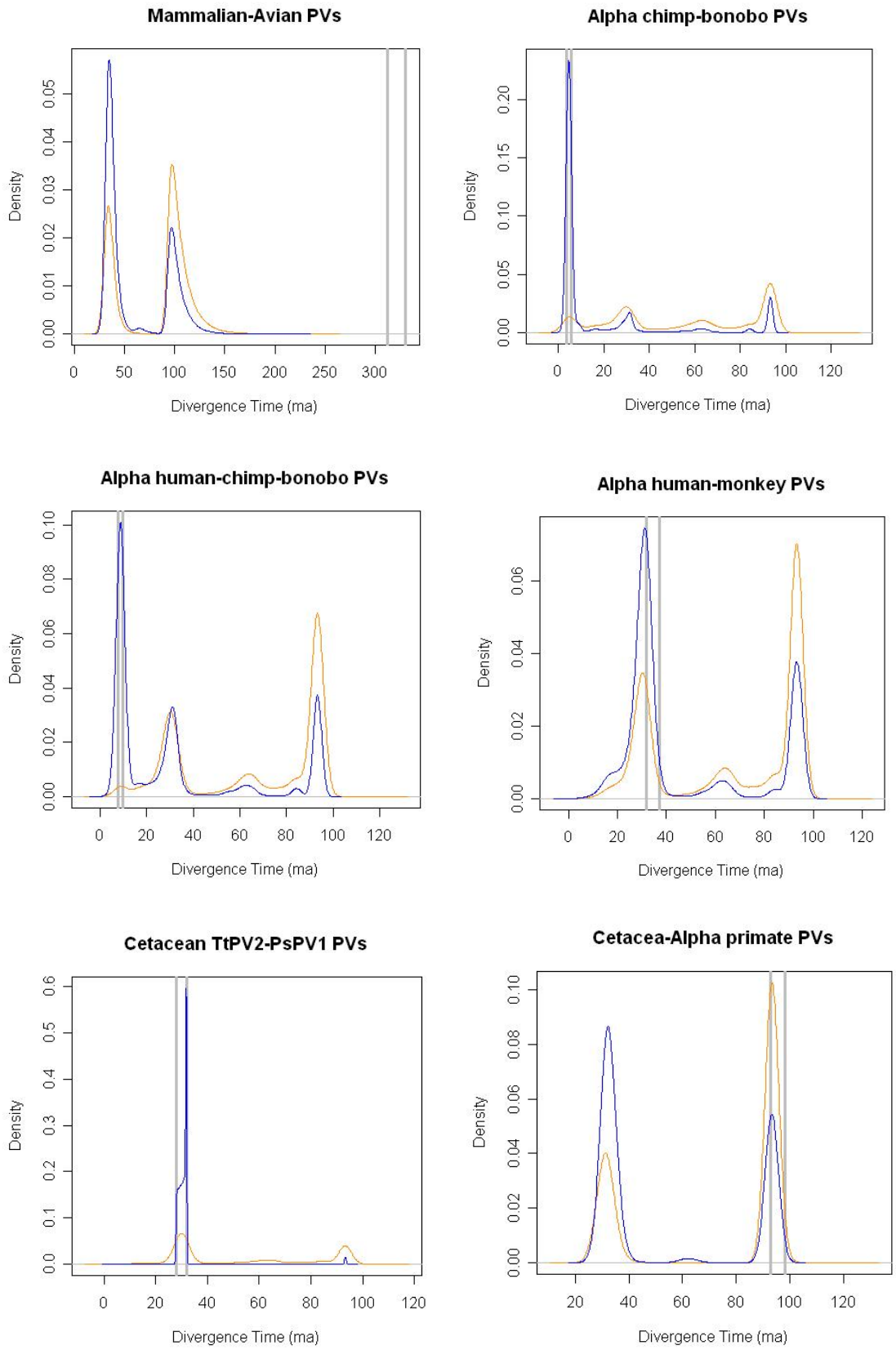

Mammalian-Avian PVs
Density
Divergence Time (ma)
Alpha chimp-bonobo PVs
Density
Divergence Time (ma)
Alpha human-chimp-bonobo PVs
Density
Divergence Time (ma)
Alpha human-monkey PVs
Density
Divergence Time (ma)
Cetacean TtPV2-PsPV1 PVs
Density
Divergence Time (ma)
Cetacea-Alpha primate PVs
Density
Divergence Time (ma)

Figure B.3

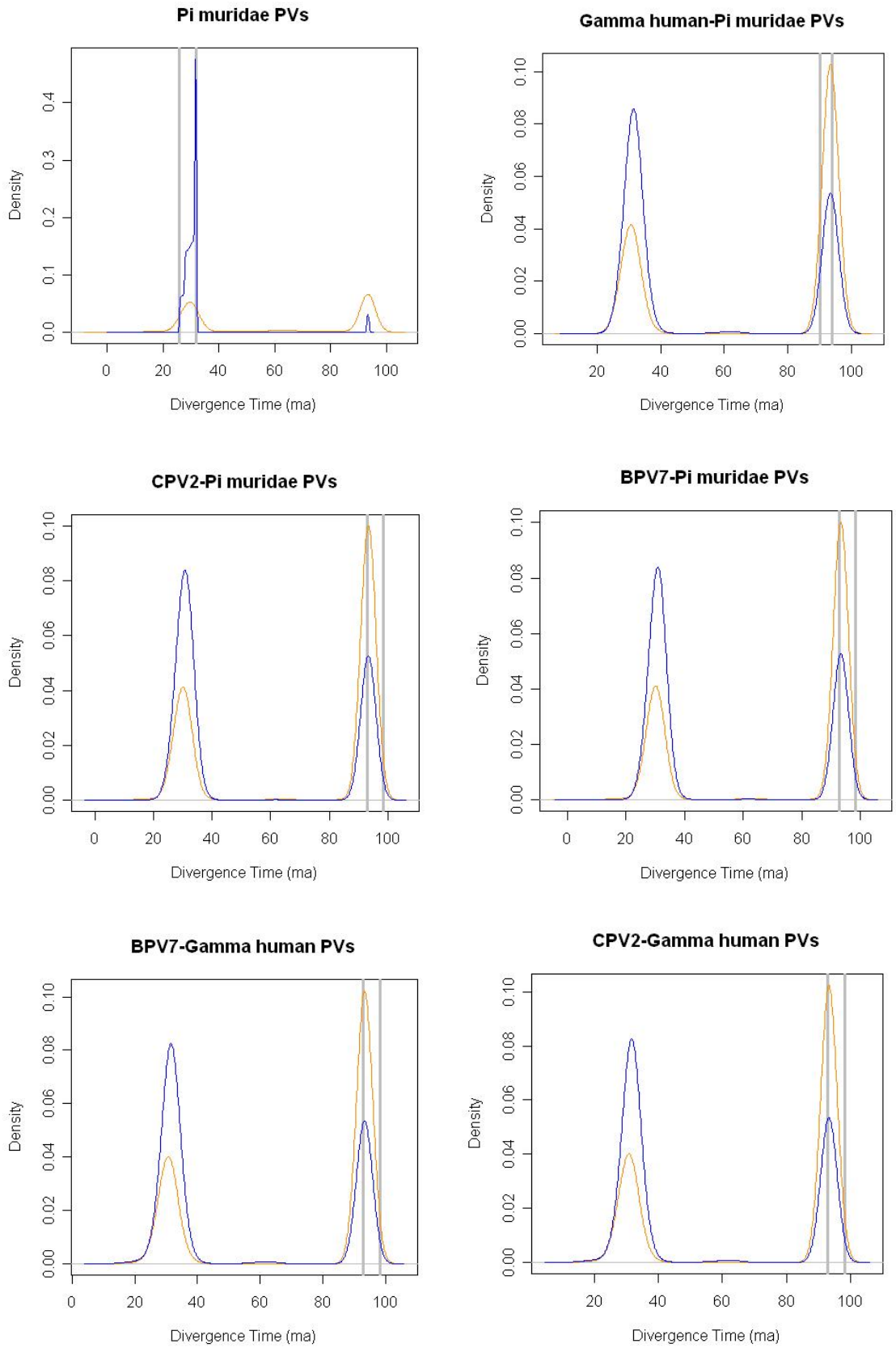

Pi muridae PVs
Density
Divergence Time (ma)
Gamma human-Pi muridae PVs
Density
Divergence Time (ma)
CPV2-Pi muridae PVs
Density
Divergence Time (ma)
BPV7-Pi muridae PVs
Density
Divergence Time (ma)
BPV7-Gamma human PVs
Density
Divergence Time (ma)
CPV2-Gamma human PVs
Density
Divergence Time (ma)

Figure B.3

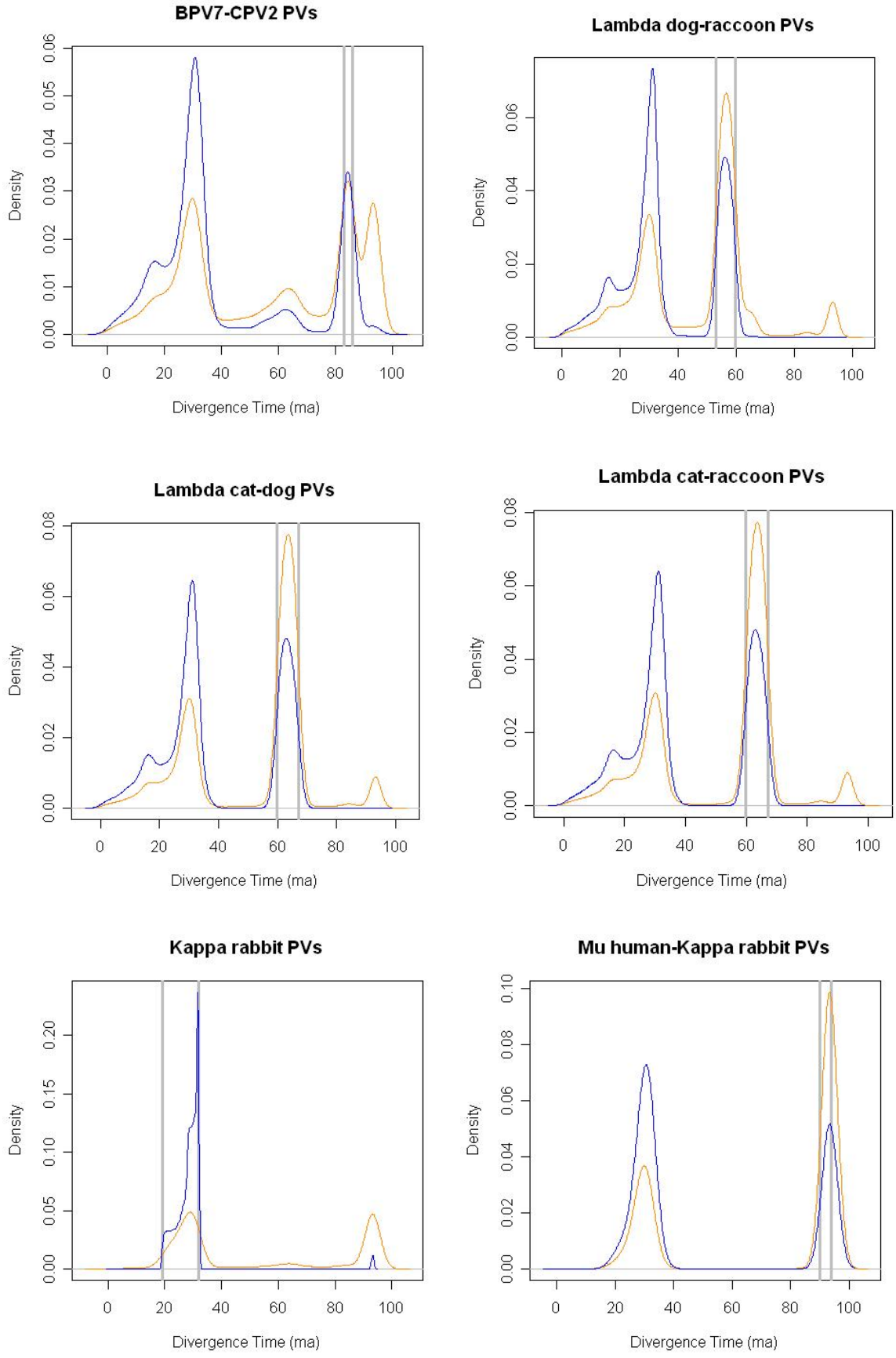

BPV7-CPV2 PVs
Density
Divergence Time (ma)
Lambda dog-raccoon PVs
Density
Divergence Time (ma)
Lambda cat-dog PVs
Density
Divergence Time (ma)
Lambda cat-raccoon PVs
Density
Divergence Time (ma)
Kappa rabbit PVs
Density
Divergence Time (ma)
Mu human-Kappa rabbit PVs
Density
Divergence Time (ma)

Figure B.3

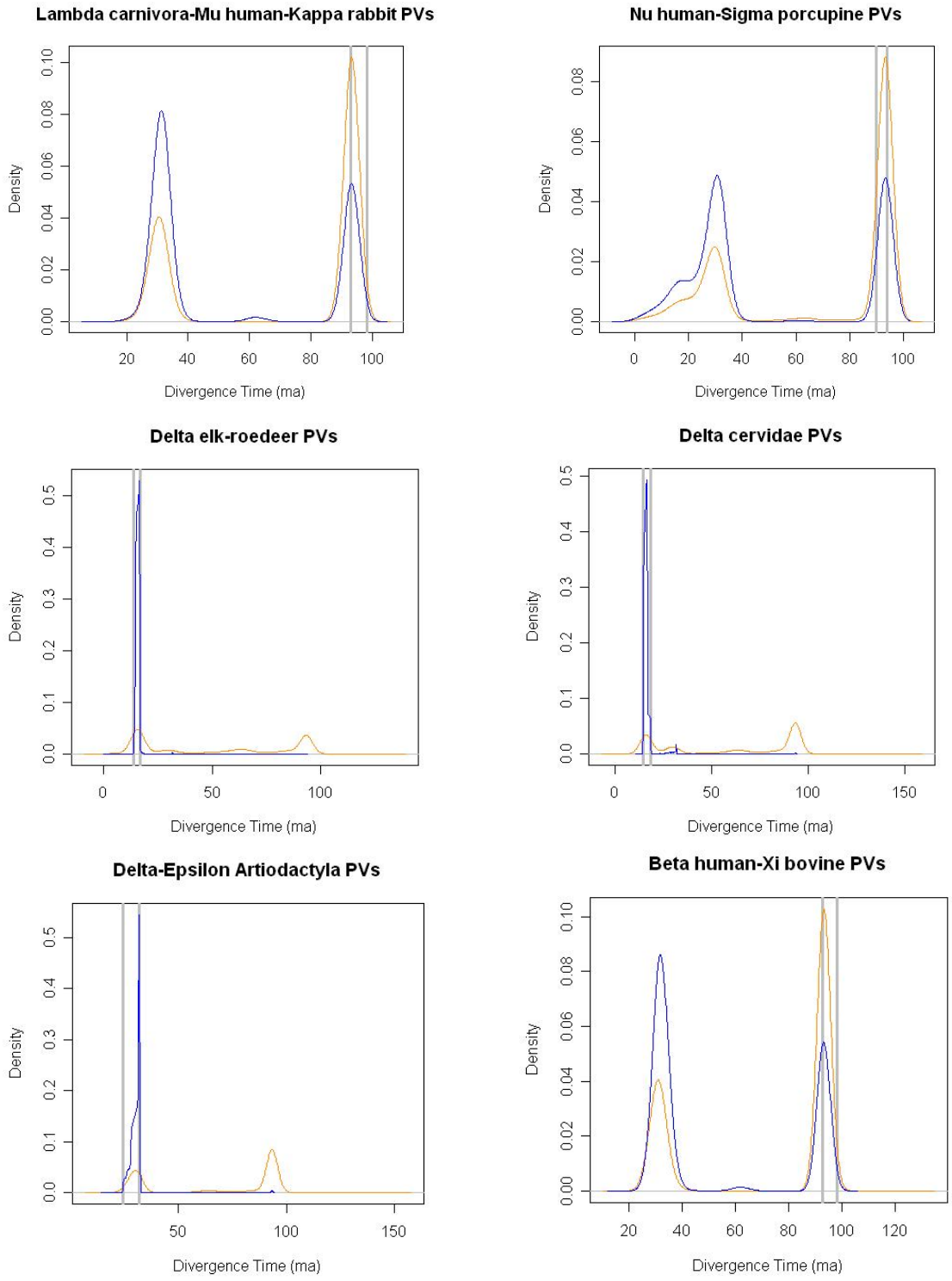


Figure B.3: The prior distributions of times for PV divergences of the E1 gene obtained by performing an MCMC simulation sampling from the prior. The orange distribution indicates the prior distribution of times for chains simulated under the ln(0.005) penalty; the blue distribution indicates the prior distribution of times for chains simulated under the ln(0.05) penalty. The vertical grey bars indicate the speciation range of the corresponding host (as estimated by Bininda-Emonds et al. 2007).

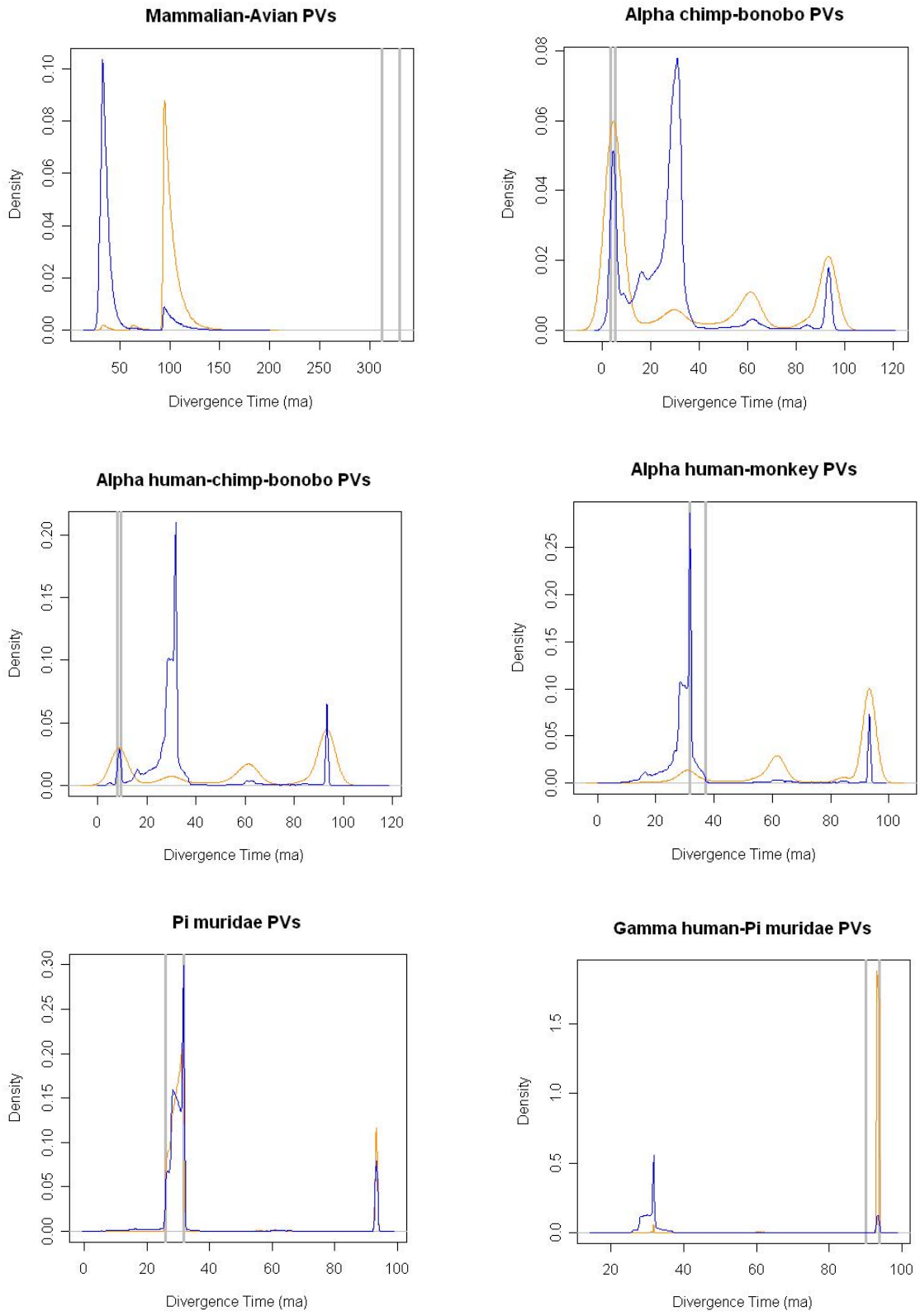

Mammalian-Avian PVs
Density
Divergence Time (ma)
Alpha chimp-bonobo PVs
Density
Divergence Time (ma)
Alpha human-chimp-bonobo PVs
Density
Divergence Time (ma)
Alpha human-monkey PVs
Density
Divergence Time (ma)
Pi muridae PVs
Density
Divergence Time (ma)
Gamma human-Pi muridae PVs
Density
Divergence Time (ma)

Figure B.4

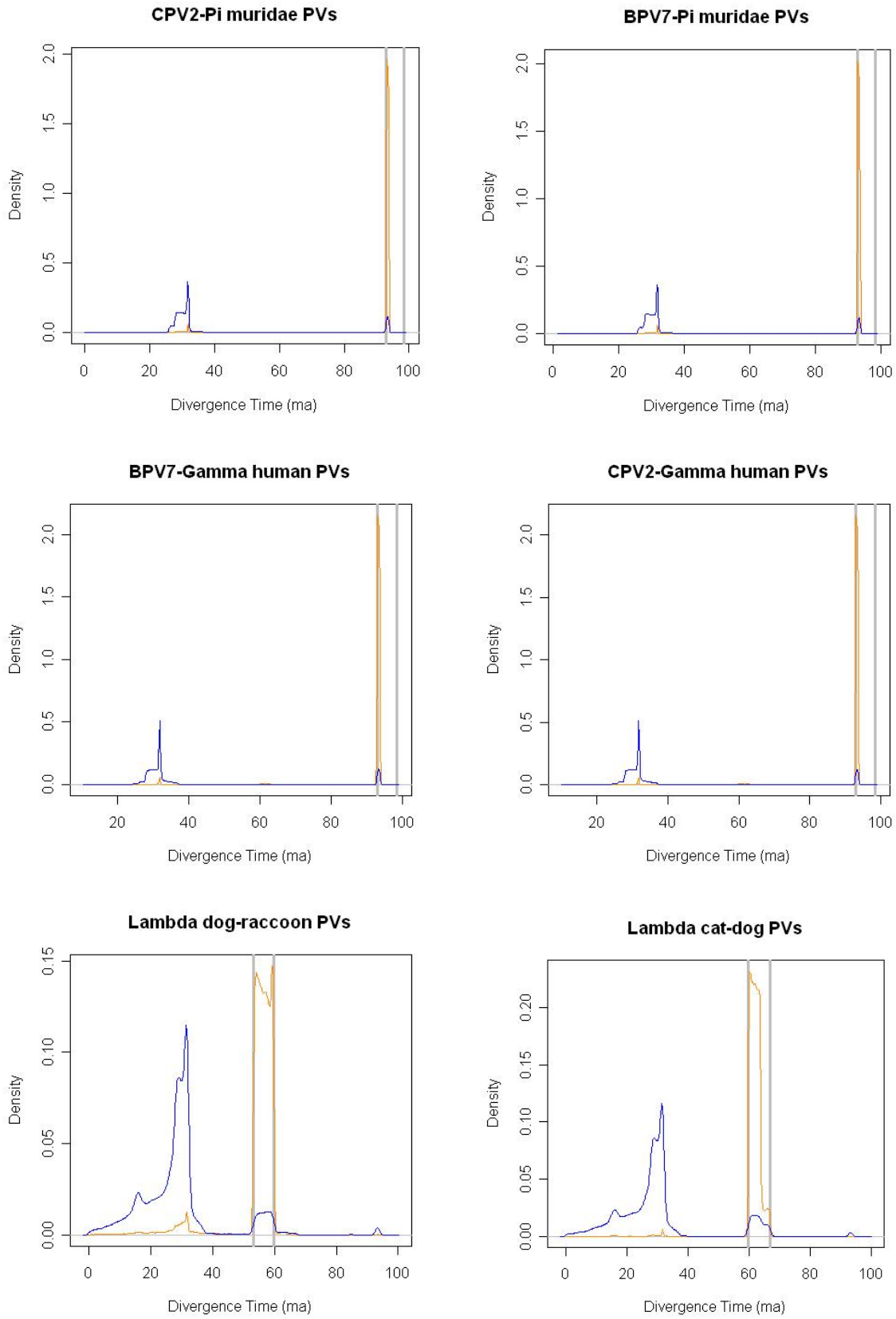

CPV2-Pi muridae PVs
BPV7-Pi muridae PVs
BPV7-Gamma human PVs
CPV2-Gamma human PVs
Lambda dog-raccoon PVs
Lambda cat-dog PVs
Density
Divergence Time (ma)

Figure B.4

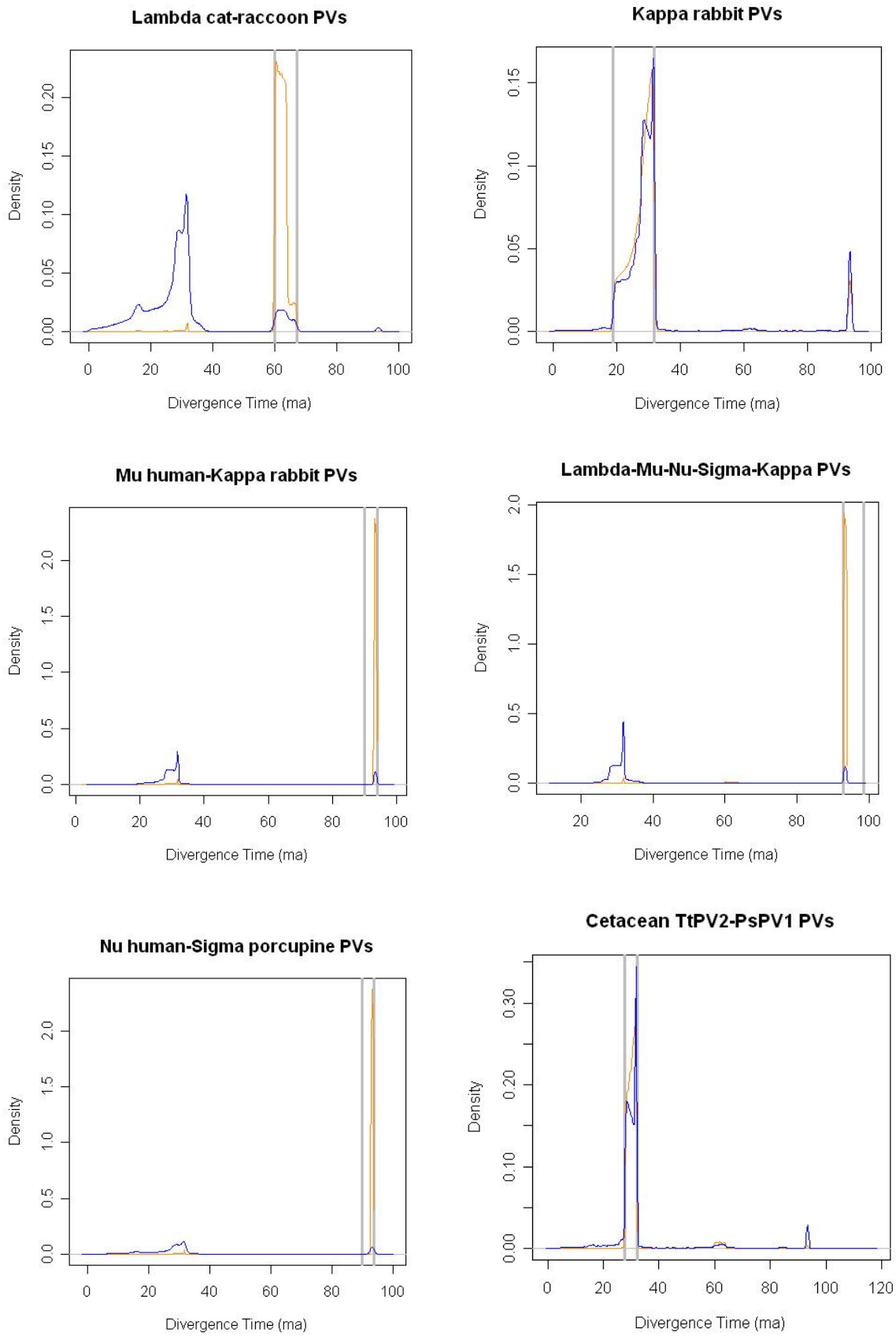

Lambda cat-raccoon PVs
Density
Divergence Time (ma)
Kappa rabbit PVs
Density
Divergence Time (ma)
Mu human-Kappa rabbit PVs
Density
Divergence Time (ma)
Lambda-Mu-Nu-Sigma-Kappa PVs
Density
Divergence Time (ma)
Nu human-Sigma porcupine PVs
Density
Divergence Time (ma)
Cetacean TtPV2-PsPV1 PVs
Density
Divergence Time (ma)

Figure B.4

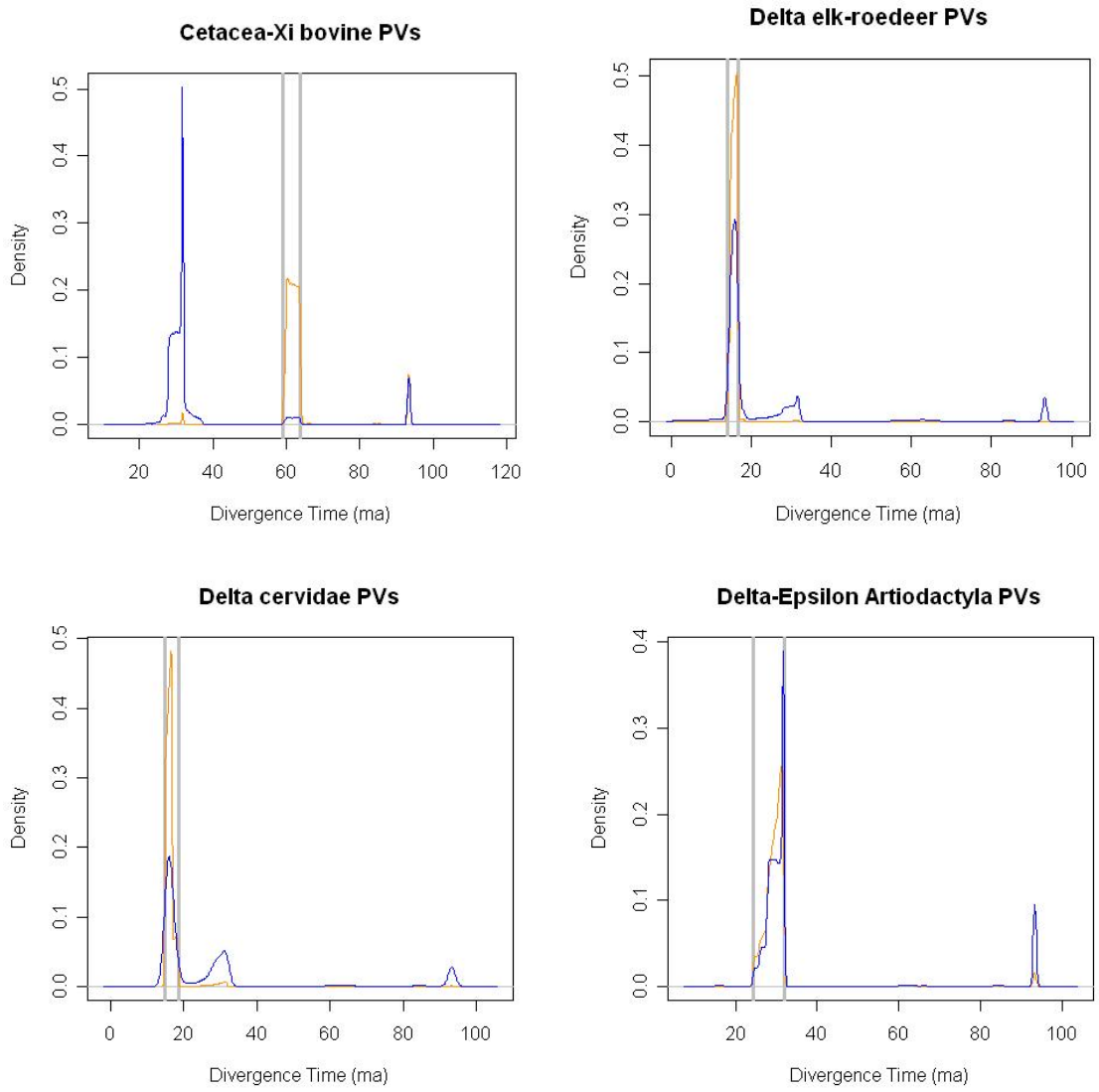


Figure B.4: The prior distributions of times for PV divergences of the L1 gene obtained by performing an MCMC simulation sampling from the prior. The orange distribution indicates the prior distribution of times for chains simulated under the ln(0.005) penalty; the blue distribution indicates the prior distribution of times for chains simulated under the ln(0.05) penalty. The vertical grey bars indicate the speciation range of the corresponding host (as estimated by Bininda-Emonds et al. 2007).

References

Ahola H, Stenlund A, Moreno-Lopez J, Pettersson U. 1983. Sequences of bovine papillomavirus type 1 DNA--functional and evolutionary implications. Nucleic Acids Res 11:2639-2650.

Alfaro ME, Zoller S, Lutzoni F. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol Biol Evol 20:255-266.

Angulo M, Carvajal-Rodriguez A. 2007. Evidence of recombination within human alpha-papillomavirus. Virol J. 4:33.

Antonsson A, Forslund O, Ekberg H, Sterner G, Hansson BG. 2000. The Ubiquity and Impressive Genomic Diversity of Human Skin Papillomaviruses Suggest a Commensalic Nature of These Viruses. Pp. 11636-11641.

Antonsson A, Hansson BG. 2002. Healthy skin of many animal species harbors papillomaviruses which are closely related to their human counterparts. J Virol. 76:12537-12542.

Antonsson A, Karanfilovska S, Lindqvist PG, Hansson BG. 2003. General acquisition of human papillomavirus infections of skin occurs in early infancy. J Clin Microbiol 41:2509-2514.

Antonsson A, McMillan NA. 2006. Papillomavirus in healthy skin of Australian animals. J Gen Virol. 87:3195-3200.

Antonsson A. 2012. Review: antibodies to cutaneous human papillomaviruses. J Med Virol 84:814-822.

Aris-Brosou S, Yang Z. 2002. Effects of Models of Rate Evolution on Estimation of Divergence Dates with Special Reference to the Metazoan 18S Ribosomal RNA Phylogeny. Syst Biol 51:703-714.

Ashrafi GH, Tsirimonaki E, Marchetti B, O'Brien PM, Sibbet GJ, Andrew L, Campo MS. 2002. Down-regulation of MHC class I by bovine papillomavirus E5 oncoproteins. Oncogene 21:248-259.

Barbosa MS, Lowy DR, Schiller JT. 1989. Papillomavirus polypeptides E6 and E7 are zinc-binding proteins. J Virol 63:1404-1407.

Barker D. 2004. LVB: parsimony and simulated annealing in the search for phylogenetic trees. Bioinformatics 20:274-275.

Barker FK, Lutzoni FM. 2002. The utility of the incongruence length difference test. Syst Biol 51:625-637.

Barnard P, McMillan NA. 1999. The human papillomavirus E7 oncoprotein abrogates signaling mediated by interferon-alpha. Virology 259:305-313.

Batista MVA, Ferreira TAE, Freitas AC, Balbino VQ. 2011. An entropy-based approach for the identification of phylogenetically informative genomic regions of Papillomavirus. Infection Genetics and Evolution 11:2026-2033.

Baxter MK, McPhillips MG, Ozato K, McBride AA. 2005. The mitotic chromosome binding activity of the papillomavirus E2 protein correlates with interaction with the cellular chromosomal protein, Brd4. J Virol 79:4806-4818.

Bayes T. 1763. An essay towards solving a problem in the doctrine of chances. Phil Trans Roy Soc 53:370-418.

Beer BE, Bailes E, Goeken R, et al. 1999. Simian immunodeficiency virus (SIV) from sun-tailed monkeys (Cercopithecus solatus): evidence for host-dependent evolution of SIV within the C. lhoesti superspecies. J Virol 73:7734-7744.

Bennett MD, Reiss A, Stevens H, et al. 2010. The First Complete Papillomavirus Genome Characterized from a Marsupial Host: a Novel Isolate from Bettongia penicillata. Journal of Virology 84:5448-5453.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2005. GenBank. Nucleic Acids Res 33:D34-38.

Berg M, Stenlund A. 1997. Functional interactions between papillomavirus E1 and E2 proteins. J Virol 71:3853-3863.

Bergin IL, Bell JD, Chen Z, et al. 2012. Novel Genital Alphapapillomaviruses in Baboons (Papio hamadryas Anubis) With Cervical Dysplasia. Vet Pathol.

Bernard H-U. 1994. Coevolution of papillomaviruses with human populations. Trends Microbiol. 2:140-143.

Bernard H-U, Burk RD, Chen Z, van Doorslaer K, Hausen Hz, de Villiers E-M. 2010. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. Virology 401:70-79.

Bernard HU, Chan SY, Manos MM, Ong CK, Villa LL, Delius H, Peyton CL, Bauer HM, Wheeler CM. 1994. Identification and assessment of known and novel human papillomaviruses by polymerase chain reaction amplification, restriction fragment length polymorphisms, nucleotide sequence, and phylogenetic algorithms. J Infect Dis 170:1077-1085.

Bininda-Emonds OR, Cardillo M, Jones KE, MacPhee RD, Beck RM, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. Nature 446:507-512.

Bloch N, Breen M, Spradbrow PB. 1994. Genomic sequences of bovine papillomaviruses in formalin-fixed sarcoids from Australian horses revealed by polymerase chain reaction. Vet Microbiol. 41:163-172.

Bloomquist EW, Dorman KS, Suchard MA. 2009. StepBrothers: inferring partially shared ancestries among recombinant viral sequences. Biostatistics 10:106-120.

Bofkin L, Goldman N. 2007. Variation in evolutionary processes at different codon positions. Mol Biol Evol 24:513-521.

Bogaert L, Martens A, Van Poucke M, Ducatelle R, De Cock H, Dewulf J, De Baere C, Peelman L, Gasthuys F. 2008. High prevalence of bovine papillomaviral DNA in the normal skin of equine sarcoid-affected and healthy horses. Vet Microbiol. 129:58-68.

Borcherds PH. 2000. Importance sampling: an illustrative approach. Eur J Phys 21:405.

Borzacchiello G, Iovane G, Marcante ML, Poggiali F, Roperto F, Roperto S, Venuti A. 2003. Presence of bovine papillomavirus type 2 DNA and expression of the viral oncoprotein E5 in naturally occurring urinary bladder tumours in cows. J Gen Virol 84:2921-2926.

Bosch FX, Lorincz A, Munoz N, Meijer CJ, Shah KV. 2002. The causal relation between human papillomavirus and cervical cancer. J Clin Pathol 55:244-265.

Boshart M, Gissmann L, Ikenberg H, Kleinheinz A, Scheurlen W, zur Hausen H. 1984. A new type of papillomavirus DNA, its presence in genital cancer biopsies and in cell lines derived from cervical cancer. Embo J. 3:1151-1157.

Bousarghin L, Touze A, Sizaret PY, Coursaget P. 2003. Human papillomavirus types 16, 31, and 58 use different endocytosis pathways to enter cells. J Virol 77:3846-3850.

Boxman IL, Russell A, Mulder LH, Bavinck JN, ter Schegget J, Green A. 2001. Association between epidermodysplasia verruciformis-associated human papillomavirus DNA in plucked eyebrow hair and solar keratoses. J Invest Dermatol 117:1108-1112.

Brauer MJ, Holder MT, Dries LA, Zwickl DJ, Lewis PO, Hillis DM. 2002. Genetic algorithms and parallel processing in maximum-likelihood phylogeny inference. Mol Biol Evol 19:1717-1726.

Bravo IG, Alonso A. 2004. Mucosal human papillomaviruses encode four different E5 proteins whose chemistry and phylogeny correlate with malignant or benign growth. J Virol. 78:13613-13626.

Bravo IG, Alonso A. 2007. Phylogeny and evolution of papillomaviruses based on the E1 and E2 proteins. Virus Genes 34:249-262.

Bravo IG, de Sanjose S, Gottschling M. 2011. The clinical importance of understanding the evolution of papillomaviruses. Trends in Microbiology 18:432-438.

Brooks DR. 1981. Hennig's Parasitological Method: A Proposed Solution. Systematic Zoology 30:229-249.

Brooks DR. 1990. Parsimony analysis in historical biogeogrphy and coevolution: Methodological and theoretical update. Systematic Zoology 39:14-30.

Brown DR, McClowry TL, Woods K, Fife KH. 1999. Nucleotide sequence and characterization of human papillomavirus type 83, a novel genital papillomavirus. Virology 260:165-172.

Brown JM, Hedtke SM, Lemmon AR, Lemmon EM. 2010. When trees grow too long: investigating the causes of highly inaccurate bayesian branch-length estimates. Syst Biol 59:145-161.

Bruno WJ, Socci ND, Halpern AL. 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol Biol Evol 17:189-197.

Buckley TR. 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Syst Biol 51:509-523.

Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Waddell PJ. 1993. Partitioning and Combining Data in Phylogenetic Analysis. Systematic Biology 42:384-397.

Burk RD, Terai M, Gravitt PE, et al. 2003. Distribution of human papillomavirus types 16 and 18 variants in squamous cell carcinomas and adenocarcinomas of the cervix. Cancer Res 63:7215-7220.

Caldeira S, de Villiers EM, Tommasino M. 2000. Human papillomavirus E7 proteins stimulate proliferation independently of their ability to associate with retinoblastoma protein. Oncogene 19:821-826.

Calleja-Macias IE, Kalantari M, Allan B, et al. 2005. Papillomavirus subtypes are natural and old taxa: phylogeny of human papillomavirus types 44 and 55 and 68a and -b. J Virol 79:6565-6569.

Camin JH, Sokal RR. 1965. A method for deducing branching sequences in phylogeny. Evolution 19:311-326.

Campbell V, Legendre P, Lapointe F-J. 2011. The performance of the Congruence Among Distance Matrices (CADM) test in phylogenetic analysis. BMC Evol Biol 11:64.

Campo MS. 1997. Bovine papillomavirus and cancer. Vet J 154:175-188.

Campo MS. 2002. Animal models of papillomavirus pathogenesis. Virus Res. 89:249-261.

Canadas MP, Videla S, Darwich L, et al. 2010. Human papillomavirus HPV-16, 18, 52 and 58 integration in cervical cells of HIV-1-infected women. Journal of Clinical Virology 48:198-201.

Carvajal-Rodriguez A. 2008. Detecting recombination and diversifying selection in human alpha-papillomavirus. Infect Genet Evol. 8:689-692.

Cavalli-Sforza LL, Edwards AWF. 1967. Phylogenetic analysis: Models and estimation procedures. Evolution 21:550-570.

Chambers G, Ellsmore VA, O'Brien PM, Reid SW, Love S, Campo MS, Nasir L. 2003. Association of bovine papillomavirus with the equine sarcoid. J Gen Virol. 84:1055-1062.

Chan S-Y, Bernard H-U, Ong C-K, Chan S-P, Hofmann B, Delius H. 1992a. Phylogenetic analysis of 48 papillomavirus types and 28 subtypes and variants: a showcase for the molecular evolution of DNA viruses. J Virol. 66:5714-5725.

Chan S-Y, Delius H, Halpern AL, Bernard H-U. 1995. Analysis of genomic sequences of 95 papillomavirus types: uniting typing, phylogeny, and taxonomy. J Virol. 69:3074-3083.

Chan S-Y, Bernard H-U, Ratterree M, Birkebak TA, Faras AJ, Ostrow RS. 1997a. Genomic diversity and evolution of papillomaviruses in rhesus monkeys. J Virol 71:4938-4943.

Chan S-Y, Ostrow RS, Faras AJ, Bernard H-U. 1997b. Genital papillomaviruses (PVs) and epidermodysplasia verruciformis PVs occur in the same monkey species: implications for PV evolution. Virology 228:213-217.

Chan SY, Ho L, Ong CK, Chow V, Drescher B, Durst M, ter Meulen J, Villa L, Luande J, Mgaya HN. 1992b. Molecular variants of human papillomavirus type 16 from four continents suggest ancient pandemic spread of the virus and its coevolution with humankind. Pp. 2057-2066.

Chang F, Syrjanen S, Shen Q, Ji HX, Syrjanen K. 1990. Human papillomavirus (HPV) DNA in esophageal precancer lesions and squamous cell carcinomas from China. Int J Cancer 45:21-25.

Chang YE, Laimins LA. 2000. Microarray analysis identifies interferon-inducible genes and Stat-1 as major transcriptional targets of human papillomavirus type 31. J Virol 74:4174-4182.

Charleston M. 1998. Jungles: A new solution to the hostparasite phylogeny reconciliation problem. Math Biosci 149:191 - 223.

Charleston MA, Robertson DL. 2002. Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. Systematic Biology 51:528-535.

Chen EY, Howley PM, Levinson AD, Seeburg PH. 1982. The primary structure and genetic organization of the bovine papillomavirus type 1 genome. Nature 299:529-534.

Chen XS, Garcea RL, Goldberg I, Casini G, Harrison SC. 2000. Structure of small virus-like particles assembled from the L1 protein of human papillomavirus 16. Mol Cell 5:557-567.

Chen Z, Fu L, Herrero R, Schiffman M, Burk RD. 2007a. Identification of a novel human papillomavirus (HPV97) related to HPV18 and HPV45. Int J Cancer 121:193-198.

Chen Z, Schiffman M, Herrero R, Burk RD. 2007b. Identification and characterization of two novel human papillomaviruses (HPVs) by overlapping PCR: HPV102 and HPV106. J Gen Virol 88:2952-2955.

Chen Z, DeSalle R, Schiffman M, Herrero R, Burk RD. 2009. Evolutionary dynamics of variant genomes of human papillomavirus types 18, 45, and 97. J Virol 83:1443-1455.

Cheng S, Schmidt-Grimminger DC, Murant T, Broker TR, Chow LT. 1995. Differentiation-dependent up-regulation of the human papillomavirus E7 gene reactivates cellular DNA replication in suprabasal differentiated keratinocytes. Genes Dev 9:2335-2349.

Chow VTK, Leong PWF. 1999. Complete nucleotide sequence, genomic organization and phylogenetic analysis of a novel genital human papillomavirus type, HLT7474-S. J Gen Virol 80:2923-2929.

Christensen ND, Cladel NM, Reed CA, Han R. 2000. Rabbit oral papillomavirus complete genome sequence and immunity following genital infection. Virology. 269:451-461.

Ciccolini F, Di Pasquale G, Carlotti F, Crawford L, Tommasino M. 1994. Functional studies of E7 proteins from different HPV types. Oncogene 9:2633-2638.

Clertant P, Seif I. 1984. A common function for polyoma virus large-T and papillomavirus E1 proteins? Nature 311:276-279.

Cole ST, Danos O. 1987. Nucleotide sequence and comparative analysis of the human papillomavirus type 18 genome. Phylogeny of papillomaviruses and repeated structure of the E6 and E7 gene products. J Mol Biol 193:599-608.

Cooper K, Taylor L, Govind S. 1995. Human papillomavirus DNA in oesophageal carcinomas in South Africa. J Pathol 175:273-277.

Cueille N, Nougarede R, Mechali F, Philippe M, Bonne-Andrea C. 1998. Functional interaction between the bovine papillomavirus virus type 1 replicative helicase E1 and cyclin E-Cdk2. J Virol 72:7255-7262.

Danos O, Katinka M, Yaniv M. 1982. Human papillomavirus 1a complete DNA sequence: a novel type of genome organization among papovaviridae. Embo J 1:231-236.

Darlu P, Lecointre G. 2002. When does the incongruence length difference test fail? Mol Biol Evol 19:432-437.

Davies R, Hicks R, Crook T, Morris J, Vousden K. 1993. Human papillomavirus type 16 E7 associates with a histone H1 kinase and with p107 through sequences necessary for transformation. J Virol 67:2521-2528.

Davy CE, Jackson DJ, Wang Q, Raj K, Masterson PJ, Fenner NF, Southern S, Cuthill S, Millar JB, Doorbar J. 2002. Identification of a G(2) arrest domain in the E1 wedge E4 protein of human papillomavirus type 16. J Virol 76:9806-9818.

Davy CE, Jackson DJ, Raj K, et al. 2005. Human papillomavirus type 16 E1 E4-induced G2 arrest is associated with cytoplasmic retention of active Cdk1/cyclin B1 complexes. J Virol 79:3998-4011.

Day PM, Roden RB, Lowy DR, Schiller JT. 1998. The papillomavirus minor capsid protein, L2, induces localization of the major capsid protein, L1, and the viral transcription/replication protein, E2, to PML oncogenic domains. J Virol 72:142-150.

Day PM, Lowy DR, Schiller JT. 2003. Papillomaviruses infect cells via a clathrin-dependent pathway. Virology 307:1-11.
De Queiroz A. 1993. For consensus (sometimes). Syst Biol 42:368-372.
De Vienne DM, Giraud T, Shykoff JA. 2007. When Can Host Shifts Produce Congruent Host and Parasite Phylogenies? A Simulation Approach. Pp. 1428 - 1438.
de Villiers EM. 1989. Heterogeneity of the human papillomavirus group. Pp. 4898-4903.
de Villiers EM, Fauquet C, Broker TR, Bernard HU, zur Hausen H. 2004. Classification of papillomaviruses. Virology 324:17-27.
del Moral-Hernandez O, Lopez-Urrutia E, Bonilla-Moreno R, Martinez-Salazar M, Arechaga-Ocampo E, Berumen J, Villegas-Sepulveda N. 2010. The HPV-16 E7 oncoprotein is expressed mainly from the unspliced E6/E7 transcript in cervical carcinoma C33-A cells. Arch Virol 155:1959-1970.
Delius H, Saegling B, Bergmann K, Shamanin V, de Villiers EM. 1998. The genomes of three of four novel HPV types, defined by differences of their L1 genes, show high conservation of the E7 gene and the URR. Virology 240:359-365.
Deng W, Lin BY, Jin G, Wheeler CG, Ma T, Harper JW, Broker TR, Chow LT. 2004. Cyclin/CDK regulates the nucleocytoplasmic localization of the human papillomavirus E1 DNA helicase. J Virol 78:13954-13965.
Disbrow GL, Sunitha I, Baker CC, Hanover J, Schlegel R. 2003. Codon optimization of the HPV-16 E5 gene enhances protein expression. Virology 311:105-114.
Dolphin K, Belshaw R, Orme CD, Quicke DL. 2000. Noise and incongruence: interpreting results of the incongruence length difference test. Mol Phylogenet Evol 17:401-406.
Doorbar J, Ely S, Sterling J, McLean C, Crawford L. 1991. Specific interaction between HPV-16 E1-E4 and cytokeratins results in collapse of the epithelial cell intermediate filament network. Nature 352:824-827.
Doorbar J. 2005. The papillomavirus life cycle. J Clin Virol. 32 Suppl 1:S7-15.
Doorbar J. 2006. Molecular biology of human papillomavirus infection and cervical cancer. Clin Sci (Lond). 110:525-541.
Dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PC, Yang Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc Biol Sci.
Dowling APG. 2002. Testing the accuracy of TreeMap and Brooks parsimony analyses of coevolutionary patterns using artificial associations. Cladistics 18:416-435.
Drummond AJ, Ho SY, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.
Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 7:214.
Duensing S, Lee LY, Duensing A, Basile J, Piboonniyom S, Gonzalez S, Crum CP, Munger K. 2000. The human papillomavirus type 16 E6 and E7 oncoproteins cooperate to induce mitotic defects and genomic instability by uncoupling centrosome duplication from the cell division cycle. Proc Natl Acad Sci U S A 97:10002-10007.
Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9:267-276.

Durst M, Gissmann L, Ikenberg H, zur Hausen H. 1983. A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. Proc Natl Acad Sci U S A. 80:3812-3815.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792-1797.
Edwards AWF, Cavalli-Sforza LL. 1963. The reconstruction of evolution. Ann Hum Genet 27:105.
Edwards AWF. 1970. Estimation of the branch points of a branching diffusion process. J Roy Stat Soc B Met 32:155-174.
Efron B. 1979. Bootstrap methods: another look at the jackknife. Ann Statist 7:1-26.
Egelkrout EM, Galloway DA. 2007. The Humoral Immune Response to Human Papillomavirus. In: L. GR, and Daniel D, editors. The Papillomaviruses. New Haven, CT: Springer. p. 277-312.
Eichler W. 1942. Die Entfaltungsregel und andere Gesetzmaßigkeiten in den parasito genetischen Beziehungen der Mallophagen und anderer ständiger Parasiten zu ihren Wirten. Zoologischer Anzeiger 137:77-83.
Ekstrom J, Forslund O, Dillner J. Three novel papillomaviruses (HPV109, HPV112 and HPV114) and their presence in cutaneous and mucosal samples. Virology 397:331-336.
Enemark EJ, Chen G, Vaughn DE, Stenlund A, Joshua-Tor L. 2000. Crystal structure of the DNA binding domain of the replication initiation protein E1 from papillomavirus. Mol Cell 6:149-158.
Evander M, Frazer IH, Payne E, Qi YM, Hengst K, McMillan NA. 1997. Identification of the alpha6 integrin as a candidate receptor for papillomaviruses. J Virol 71:2449-2456.
Excoffier L, Yang Z. 1999. Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. Mol Biol Evol 16:1357-1368.
Fahrenholz H. 1913. Ectoparasiten und Abstammungslehre. Zoologischer Anzeiger 41:371-374.
Fan X, Chen JJ. 2004. Regulation of cell cycle progression and apoptosis by the papillomavirus E6 oncogene. Crit Rev Eukaryot Gene Expr 14:183-202.
Fan Y, Wu R, Chen MH, Kuo L, Lewis PO. 2011. Choosing among partition models in Bayesian phylogenetics. Mol Biol Evol 28:523-532.
Farris JS, Källersjö M, Kluge AG, Bult C. 1994. Testing Significance of Incongruence. Cladistics 10:315-319.
Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA. 2005. Virus Taxonomy. The Eighth Report of the International Committee on Taxonomy of Viruses. Family *Papillomaviridae*. Elsevier. p. 239-255.
Favre-Bonvin A, Reynaud C, Kretz-Remy C, Jalinot P. 2005. Human papillomavirus type 18 E6 protein binds the cellular PDZ protein TIP-2/GIPC, which is involved in transforming growth factor beta signaling and triggers its degradation by the proteasome. J Virol 79:4229-4237.
Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401-410.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 17:368-376.
Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791.

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet 22:521-565.
Felsenstein J. 2004. Inferring Phylogenies. Sunderland, MA: Sinauer Associates.
Ferenczy A, Franco E. 2002. Persistent human papillomavirus infection and cervical neoplasia. Lancet Oncol. 3:11-16.
Finnen RL, Erickson KD, Chen XS, Garcea RL. 2003. Interactions between papillomavirus L1 and L2 capsid proteins. J Virol 77:4818-4826.
Fitch WM, Margolish E. 1967. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. Biochem Genet 1:65-71.
Fitch WM. 1971. Towards defining the course of evolution: minimum change for a specific tree topology. Syst Zool 20:406-416.
Flores ER, Allen-Hoffmann BL, Lee D, Lambert PF. 2000. The human papillomavirus type 16 E7 oncogene is required for the productive stage of the viral life cycle. J Virol 74:6622-6631.
Florin L, Becker KA, Sapp C, Lambert C, Sirma H, Muller M, Streeck RE, Sapp M. 2004. Nuclear translocation of papillomavirus minor capsid protein L2 requires Hsc70. J Virol 78:5546-5553.
Forslund O, Hansson BG. 1996. Human papillomavirus type 70 genome cloned from overlapping PCR products: complete nucleotide sequence and genomic organization. J Clin Microbiol 34:802-809.
Forslund O, Antonsson A, Nordin P, Stenquist B, Hansson BG. 1999. A broad range of human papillomavirus types detected with a general PCR method suitable for analysis of cutaneous tumours and normal skin. J Gen Virol 80 (Pt 9):2437-2443.
Frazer IH. 2009. Interaction of human papillomaviruses with the host immune system: a well evolved relationship. Virology 384:410-414.
Frisch M, Biggar RJ, Goedert JJ. 2000. Human papillomavirus-associated cancers in patients with human immunodeficiency virus infection and acquired immunodeficiency syndrome. J Natl Cancer Inst 92:1500-1510.
Fu L, Terai M, Matsukura T, Herrero R, Burk RD. 2004. Codetection of a mixed population of candHPV62 containing wild-type and disrupted E1 open-reading frame in a 45-year-old woman with normal cytology. J Infect Dis 190:1303-1309.
Fu L, Van Doorslaer K, Chen Z, Ristriani T, Masson M, Trave G, Burk RD. 2010. Degradation of p53 by Human Alphapapillomavirus E6 Proteins Shows a Stronger Correlation with Phylogeny than Oncogenicity. Plos One 5.
Futuyma DJ, Moreno G. 1988. The evolution of ecological specialization. Ann Rev Ecol Syst 19:207-233.
Gambhira R, Karanam B, Jagu S, Roberts JN, Buck CB, Bossis I, Alphs H, Culp T, Christensen ND, Roden RB. 2007. A protective and broadly cross-neutralizing epitope of human papillomavirus L2. J Virol 81:13927-13931.
Gamble T, Berendzen PB, Bradley Shaffer H, Starkey DE, Simons AM. 2008. Species limits and phylogeography of North American cricket frogs (Acris: Hylidae). Mol Phylogenet Evol 48:112-125.
Garcea RL, Chen X. 2007. Papillomavirus Structure and Assembly. In: Garcea RL, and DiMaio D, editors. The Papillomaviruses. New York: Springer. p.
Garcea RL, DiMaio D. 2007. The Papillomaviruses. Springer, New York.
Garcia-Vallve S, Alonso A, Bravo IG. 2005. Papillomaviruses: different genes have different histories. Trends Microbiol. 13:514-521.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14:685-695.

Gauthier JM, Dillner J, Yaniv M. 1991. Structural analysis of the human papillomavirus type 16-E2 transactivator with antipeptide antibodies reveals a high mobility region linking the transactivation and the DNA-binding domains. Nucleic Acids Res 19:7073-7079.

Gelman A, Rubin DB. 1992. Inference from iterative simulation using multiple sequences (with discussion). Statist Sci 7:457-511.

Genther Williams SM, Disbrow GL, Schlegel R, Lee D, Threadgill DW, Lambert PF. 2005. Requirement of epidermal growth factor receptor for hyperplasia induced by E5, a high-risk human papillomavirus oncogene. Cancer Res 65:6534-6542.

Geyer CJ. 1991. Markov chain Monte Carlo maximum likelihood in Keramidas EM, ed. Computing Science and Statistics: Proc. 23rd Symp. Interface Foundation, Fairfax Station, VA.

Gibbs A, Calisher CH, Garcia-Arenal F. 1995. Molecular basis of virus evolution. Cambridge: Cambridge University Press.

Gillison ML. 2004. Human papillomavirus-associated head and neck cancer is a distinct epidemiologic, clinical, and molecular entity. Semin Oncol 31:744-754.

Giri I, Danos O, Yaniv M. 1985. Genomic structure of the cottontail rabbit (Shope) papillomavirus. Proc Natl Acad Sci U S A. 82:1580-1584.

Giribet G. 2007. Efficient tree searches with available algorithms. Evol Bioinform Online 3:341-356.

Giroglou T, Florin L, Schafer F, Streeck RE, Sapp M. 2001. Human papillomavirus infection requires cell surface heparan sulfate. J Virol 75:1565-1570.

Goldman N. 1993. Statistical tests of models of DNA substitution. J Mol Evol 36:182-198.

Goloboff PA. 1999. Analysing large data sets in reasonable times: solutions for composite optima. Cladistics 15:415-428.

Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G. 1979. Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. Systematic Zoology 28:132-163.

Gottschling M, Kohler A, Stockfleth E, Nindl I. 2007a. Phylogenetic analysis of beta-papillomaviruses as inferred from nucleotide and amino acid sequence data. Mol Phylogenet Evol. 42:213-222.

Gottschling M, Stamatakis A, Nindl I, Stockfleth E, Alonso A, Bravo IG. 2007b. Multiple evolutionary mechanisms drive papillomavirus diversification. Mol Biol Evol. 24:1242-1258.

Gottschling M, Bravo IG, Schulz E, Bracho MA, Deaville R, Jepson PD, Van Bressem MF, Stockfleth E, Nindl I. 2011a. Modular organizations of novel cetacean papillomaviruses. Mol Phylogenet Evol 59:34-42.

Gottschling M, Goker M, Stamatakis A, Bininda-Emonds OR, Nindl I, Bravo IG. 2011b. Quantifying the phylodynamic forces driving papillomavirus evolution. Mol Biol Evol 28:2101-2113.

Greenspan D, de Villiers EM, Greenspan JS, de Souza YG, zur Hausen H. 1988. Unusual HPV types in oral warts in association with HIV infection. J Oral Pathol 17:482-488.

Grievink LS, Penny D, Hendy MD, Holland BR. 2010. Phylogenetic tree reconstruction accuracy and model fit when proportions of variable sites change across the tree. Syst Biol 59:288-297.

Guess JC, McCance DJ. 2005. Decreased migration of Langerhans precursor-like cells in response to human keratinocytes expressing human papillomavirus type 16 E6/E7 is related to reduced macrophage inflammatory protein-3alpha production. J Virol 79:14852-14862.

Hafner MS, Nadler SA. 1988. Phylogenetic trees support the coevolution of parasites and their hosts. Nature 332:258-259.

Halpert R, Fruchter RG, Sedlis A, Butt K, Boyce JG, Sillman FH. 1986. Human papillomavirus and lower genital neoplasia in renal transplant patients. Obstet Gynecol 68:251-258.

Harper DM, Franco EL, Wheeler C, et al. 2004. Efficacy of a bivalent L1 virus-like particle vaccine in prevention of infection with human papillomavirus types 16 and 18 in young women: a randomised controlled trial. Lancet 364:1757-1765.

Harper DM, Franco EL, Wheeler CM, Moscicki AB, Romanowski B, Roteli-Martins CM, Jenkins D, Schuind A, Costa Clemens SA, Dubin G. 2006. Sustained efficacy up to 4.5 years of a bivalent L1 virus-like particle vaccine against human papillomavirus types 16 and 18: follow-up from a randomised control trial. Lancet 367:1247-1255.

Hasegawa M, Yano T, Kishino H. 1984. A new molecular clock of mitochondrial DNA and the evolution of Hominoids. Proc. Japan Acad. B 60:95-98.

Hasegawa M, Kishino H, Yano T. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160-174.

Hasegawa M, Kishino H, Yano T. 1989. Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidae. J Hum Evol 18:461-476.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their application. Biometrika 57:97-109.

Hatama S, Nobumoto K, Kanno T. 2008. Genomic and phylogenetic analysis of two novel bovine papillomaviruses, BPV-9 and BPV-10. J Gen Virol 89:158-163.

Hatama S, Ishihara R, Ueda Y, Kanno T, Uchida I. 2011. Detection of a novel bovine papillomavirus type 11 (BPV-11) using xipapillomavirus consensus polymerase chain reaction primers. Arch Virol 156:1281-1285.

Heath TA, Zwickl DJ, Kim J, Hillis DM. 2008. Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. Syst Biol 57:160-166.

Heilman SA, Nordberg JJ, Liu Y, Sluder G, Chen JJ. 2009. Abrogation of the postmitotic checkpoint contributes to polyploidization in human papillomavirus E7-expressing cells. J Virol 83:2756-2764.

Hendy MD, Penny D. 1989. A framework for the quantitative study of evolutionary trees. Syst Biol 38:297-309.

Herbst LH, Lenz J, Van Doorslaer K, Chen Z, Stacy BA, Wellehan Jr JFX, Manire CA, Burk RD. 2009. Genomic characterization of two novel reptilian papillomaviruses, Chelonia mydas papillomavirus 1 and Caretta caretta papillomavirus 1. Virology 383:131-135.

Hillis DM, Bull JJ. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 42:182-192.

Ho GY, Burk RD, Klein S, Kadish AS, Chang CJ, Palan P, Basu J, Tachezy R, Lewis R, Romney S. 1995. Persistent genital human papillomavirus infection as a risk factor for persistent cervical dysplasia. J Natl Cancer Inst. 87:1365-1371.

Ho L, Chan SY, Burk RD, et al. 1993. The genetic drift of human papillomavirus type 16 is a means of reconstructing prehistoric viral spread and the movement of ancient human populations. J Virol 67:6413-6423.
Holmgren SC, Patterson NA, Ozbun MA, Lambert PF. 2005. The minor capsid protein L2 contributes to two steps in the human papillomavirus type 31 life cycle. J Virol 79:3938-3948.
Hommola K, Smith JE, Qiu Y, Gilks WR. 2009. A Permutation Test of Host-Parasite Cospeciation. Pp. 1457-1468.
Hoory T, Monie A, Gravitt P, Wu TC. 2008. Molecular epidemiology of human papillomavirus. J Formos Med Assoc 107:198-217.
Hopfl R, Heim K, Christensen N, et al. 2000. Spontaneous regression of CIN and delayed-type hypersensitivity to HPV-16 oncoprotein E7. Lancet 356:1985-1986.
Hopman AH, Smedts F, Dignef W, Ummelen M, Sonke G, Mravunac M, Vooijs GP, Speel EJ, Ramaekers FC. 2004. Transition of high-grade cervical intraepithelial neoplasia to micro-invasive carcinoma is characterized by integration of HPV 16/18 and numerical chromosome abnormalities. J Pathol 202:23-33.
Horvath CA, Boulet GA, Renoux VM, Delvenne PO, Bogers JP. 2010. Mechanisms of cell entry by human papillomaviruses: an overview. Virol J 7:11.
Huelsenbeck J, Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst Biol 53:904-913.
Huelsenbeck JP, Hillis DM. 1993. Success of Phylogenetic Methods in the Four-Taxon Case. Systematic Biology 42:247-264.
Huelsenbeck JP, Bull JJ. 1996. A Likelihood Ratio Test to Detect Conflicting Phylogenetic Signal. Syst Biol 45:92-98.
Huelsenbeck JP, Bull JJ, Cunningham CW. 1996. Combining data in phylogenetic analysis. Trends in Ecology & Evolution 11:152-158.
Huelsenbeck JP, Rannala B. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science 276:227-232.
Huelsenbeck JP, Rannala B, Yang Z. 1997. Statistical Tests of Host-Parasite Cospeciation. Evolution 51:410-419.
Huelsenbeck JP, Larget B, Swofford D. 2000. A compound poisson process for relaxing the molecular clock. Genetics 154:1879-1892.
Huelsenbeck JP, Rannala B, Larget B. 2000. A Bayesian framework for the analysis of cospeciation. Evolution 54:352-364.
Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755.
Huibregtse JM, Scheffner M, Howley PM. 1991. A cellular protein mediates association of p53 with the E6 oncoprotein of human papillomavirus types 16 or 18. Embo J 10:4129-4135.
Husmeier D, McGuire G. 2003. Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. Mol Biol Evol 20:315-337.
Huson DH, Bryant D. 2006. Application of Phylogenetic Networks in Evolutionary Studies. Mol Biol Evol. 23:254-267.
Jackson AP. 2005. The Effect of Paralogous Lineages on the Application of Reconciliation Analysis by Cophylogeny Mapping. Pp. 127-145.

Jackson JA. 1999. Analysis of parasite host-switching: limitations on the use of phylogenies. Parasitology 119:S111-S123.
Jackson ME, Pennie WD, McCaffery RE, Smith KT, Grindlay GJ, Campo MS. 1991. The B subgroup bovine papillomaviruses lack an identifiable E6 open reading frame. Mol Carcinog. 4:382-387.
Janz N, Nyblom K, Nylin S. 2001. Evolutionary dynamics of host-plant specialization: a case study of the tribe Nymphalini. Evolution 55:783-796.
Jeffreys H. 1935. Some tests of significance, treated by the theory of probability. Proc Cambridge Philos Soc 31:203-222.
Jin L, Nei M. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol Biol Evol 7:82-102.
Joh J, Jenson AB, King W, Proctor M, Ingle A, Sundberg JP, Ghim SJ. 2011. Genomic analysis of the first laboratory-mouse papillomavirus. J Gen Virol 92:692-698.
Jones DL, Thompson DA, Munger K. 1997. Destabilization of the RB tumor suppressor protein and stabilization of p53 contribute to HPV type 16 E7-induced apoptosis. Virology 239:97-107.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editors. Mammalian protein metabolism. New York: Academic Press. p. 22-123.
Kass RE, Raftery AE. 1995. Bayes Factors. Journal of the American Statistical Association 90:23.
Kawana K, Yoshikawa H, Taketani Y, Yoshiike K, Kanda T. 1999. Common neutralization epitope in minor capsid protein L2 of human papillomavirus types 16 and 6. J Virol 73:6188-6190.
Kawana Y, Kawana K, Yoshikawa H, Taketani Y, Yoshiike K, Kanda T. 2001. Human papillomavirus type 16 minor capsid protein l2 N-terminal region containing a common neutralization epitope binds to the cell surface and enters the cytoplasm. J Virol 75:2331-2336.
Kawashima M, Jablonska S, Favre M, Obalek S, Croissant O, Orth G. 1986. Characterization of a new type of human papillomavirus found in a lesion of Bowen's disease of the skin. J Virol 57:688-692.
Kidd KK, Sgaramella-Zonta LA. 1971. Phylogenetic analysis: concepts and methods. Am J Hum Genet 23:235-252.
Kino N, Sata T, Sato Y, Sugase M, Matsukura T. 2000. Molecular cloning and nucleotide sequence analysis of a novel human papillomavirus (Type 82) associated with vaginal intraepithelial neoplasia. Clin Diagn Lab Immunol 7:91-95.
Kishino H, Thorne JL, Bruno WJ. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol Biol Evol 18:352-361.
Klassen GJ. 1992. Coevolution: A History of the Macroevolutionary Approach to Studying Host-Parasite Associations. The Journal of Parasitology 78:573-587.
Klingelhutz AJ, Foster SA, McDougall JK. 1996. Telomerase activation by the E6 gene product of human papillomavirus type 16. Nature 380:79-82.
Kloster BE, Manias DA, Ostrow RS, Shaver MK, McPherson SW, Rangen SR, Uno H, Faras AJ. 1988. Molecular cloning and characterization of the DNA of two papillomaviruses from monkeys. Virology 166:30-40.
Kluge AG. 1989. A Concern for Evidence and a Phylogenetic Hypothesis of Relationships Among Epicrates (Boidae, Serpentes). Systematic Zoology 38:7-25.

Knight GL, Grainger JR, Gallimore PH, Roberts S. 2004. Cooperation between different forms of the human papillomavirus type 1 E4 protein to block cell cycle progression and cellular DNA synthesis. J Virol 78:13920-13933.
Koonin EV, Wolf YI. 2012. Evolution of microbes and viruses: a paradigm shift in evolutionary biology? Front Cell Infect Microbiol 2:119.
Koutsky LA, Ault KA, Wheeler CM, Brown DR, Barr E, Alvarez FB, Chiacchierini LM, Jansen KU. 2002. A controlled trial of a human papillomavirus type 16 vaccine. N Engl J Med 347:1645-1651.
Kumar S, Hedges SB. 1998. A molecular timescale for vertebrate evolution. Nature 392:917-920.
Lai CC, Henningson C, DiMaio D. 2000. Bovine papillomavirus E5 protein induces the formation of signal transduction complexes containing dimeric activated platelet-derived growth factor beta receptor and associated signaling proteins. J Biol Chem 275:9832-9840.
Lange CE, Tobler K, Ackermann M, Favrot C. 2011. Identification of two novel equine papillomavirus sequences suggests three genera in one cluster. Vet Microbiol 149:85-90.
Lange CE, Favrot C, Ackermann M, Gull J, Vetsch E, Tobler K. 2012. Novel snake papillomavirus does not cluster with other non-mammalian papillomaviruses. Virol J 8:436.
Lapointe F-J, Legendre P. 1990. A statistical framework to test the consensus of two nested classifications. Syst Zool 39:1-13.
Lapointe F-J, Legendre P. 1992. A statistical framework to test the consensus among additive trees (cladograms). Syst Zool 41:158-171.
Larget B, Simon DL. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol 16:750-759.
Lartillot N, Philippe H. 2006. Computing Bayes Factors Using Thermodynamic Integration. Syst Biol 55:195-207.
Lazarczyk M, Cassonnet P, Pons C, Jacob Y, Favre M. 2009. The EVER Proteins as a Natural Barrier against Papillomaviruses: a New Insight into the Pathogenesis of Human Papillomavirus Infections. Pp. 348-370.
Lee SJ, Cho YS, Cho MC, et al. 2001. Both E6 and E7 oncoproteins of human papillomavirus 16 inhibit IL-18-induced IFN-gamma production in human peripheral blood mononuclear and NK cells. J Immunol 167:497-504.
Lefeuvre P, Lett JM, Varsani A, Martin DP. 2009. Widely conserved recombination patterns among single-stranded DNA viruses. J Virol 83:2697-2707.
Legendre P. 1997. Relating behaviour to habitat: solutions to the fourth-corner problem. Ecology 78:547-562.
Legendre P, Desdevises Y, Bazin E. 2002. A statistical test for host-parasite coevolution. Syst Biol 51:217-234.
Legendre P, Lapointe F-J. 2004. Assessing congruence among distance matrices: Single-malt Scotch whiskies revisited. Aust NZ J Stat 46:615-629.
Lemmon AR, Moriarty EC. 2004. The importance of proper model assumption in bayesian phylogenetics. Syst Biol 53:265-277.
Lewis PO. 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. Mol Biol Evol 15:277-283.
Lewis PO, Holder MT, Holsinger KE. 2005. Polytomies and Bayesian phylogenetic inference. Syst Biol 54:241-253.
Li M, Beard P, Estes PA, Lyon MK, Garcea RL. 1998. Intercapsomeric disulfide bonds in papillomavirus assembly and disassembly. J Virol 72:2160-2167.

Lin BY, Makhov AM, Griffith JD, Broker TR, Chow LT. 2002. Chaperone proteins abrogate inhibition of the human papillomavirus (HPV) E1 replicative helicase by the HPV E2 protein. Mol Cell Biol 22:6592-6604.

Lio P, Goldman N. 1998. Models of molecular evolution and phylogeny. Genome Res 8:1233-1244.

Liu X, Dakic A, Zhang Y, Dai Y, Chen R, Schlegel R. 2009. HPV E6 protein interacts physically and functionally with the cellular telomerase complex. Proc Natl Acad Sci U S A 106:18780-18785.

Lloyd DG, Calder VL. 1991. Multi-residue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses. J Evol Biol 4:9-21.

Longworth MS, Laimins LA. 2004a. Pathogenesis of human papillomaviruses in differentiating epithelia. Microbiol Mol Biol Rev 68:362-372.

Longworth MS, Laimins LA. 2004b. The binding of histone deacetylases and the integrity of zinc finger-like motifs of the E7 protein are essential for the life cycle of human papillomavirus type 31. J Virol 78:3533-3541.

Maddison D. 1991. The discovery and importance of multiple islands of most-parsimonious trees. Syst Zool 33:83-103.

Madison KC. 2003. Barrier function of the skin: "la raison d'etre" of the epidermis. J Invest Dermatol 121:231-241.

Makalowski W, Boguski MS. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci U S A 95:9407-9412.

Mansky KC, Batiza A, Lambert PF. 1997. Bovine papillomavirus type 1 E1 and simian virus 40 large T antigen share regions of sequence similarity required for multiple functions. J Virol 71:7600-7608.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res 27:209-220.

Mao C, Koutsky LA, Ault KA, Wheeler CM, Brown DR, Wiley DJ, Alvarez FB, Bautista OM, Jansen KU, Barr E. 2006. Efficacy of human papillomavirus-16 vaccine to prevent cervical intraepithelial neoplasia: a randomized controlled trial. Obstet Gynecol 107:18-27.

Marchetti B, Ashrafi GH, Tsirimonaki E, O'Brien PM, Campo MS. 2002. The bovine papillomavirus oncoprotein E5 retains MHC class I molecules in the Golgi apparatus and prevents their transport to the cell surface. Oncogene 21:7808-7816.

Marks P, Rifkind RA, Richon VM, Breslow R, Miller T, Kelly WK. 2001. Histone deacetylases and cancer: causes and therapies. Nat Rev Cancer 1:194-202.

Marshall DC. 2010. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. Syst Biol 59:108-117.

Martins Lde O, Leal E, Kishino H. 2008. Phylogenetic detection of recombination with a Bayesian prior on the distance between trees. PLoS One 3:e2651.

Masterson PJ, Stanley MA, Lewis AP, Romanos MA. 1998. A C-terminal helicase domain of the human papillomavirus E1 protein binds E2 and the DNA polymerase alpha-primase p68 subunit. J Virol 72:7407-7419.

Matsukura T, Sugase M. 2001. Relationships between 80 human papillomavirus genotypes and different grades of cervical intraepithelial neoplasia: association and causality. Virology 283:139-147.

Matthews K, Leong CM, Baxter L, Inglis E, Yun K, Backstrom BT, Doorbar J, Hibma M. 2003. Depletion of Langerhans cells in human papillomavirus type

16-infected skin is associated with E6-mediated down regulation of E-cadherin. J Virol 77:8378-8385.
McIntyre MC, Ruesch MN, Laimins LA. 1996. Human papillomavirus E7 oncoproteins bind a single form of cyclin E in a complex with cdk2 and p107. Virology 215:73-82.
McMillan NA, Payne E, Frazer IH, Evander M. 1999. Expression of the alpha6 integrin confers papillomavirus binding upon receptor-negative B-cells. Virology 261:271-279.
Menzo S, Monachetti A, Trozzi C, Ciavattini A, Carloni G, Varaldo PE, Clementi M. 2001. Identification of six putative novel human papillomaviruses (HPV) and characterization of candidate HPV type 87. J Virol 75:11913-11919.
Meredith RW, Janecka JE, Gatesy J, et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science 334:521-524.
Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equations of state calculations by fast computing machines. J Chem Phys 21:1087-1092.
Minin VN, Dorman KS, Fang F, Suchard MA. 2005. Dual multiple change-point model leads to more accurate recombination detection. Bioinformatics 21:3034-3042.
Miyamoto MM, Koop BF, Slightom JL, Goodman M, Tennant MR. 1988. Molecular systematics of higher primates: genealogical relations and classification. Proc Natl Acad Sci U S A 85:7627-7631.
Miyamoto MM, Fitch WM. 1995. Testing species phylogenies and phylogenetic methods with congruence. Syst Biol 44:64-76.
Modis Y, Trus BL, Harrison SC. 2002. Atomic model of the papillomavirus capsid. Embo J 21:4754-4762.
Mole S, Milligan SG, Graham SV. 2009. Human papillomavirus type 16 E2 protein transcriptionally activates the promoter of a key cellular splicing factor, SF2/ASF. J Virol 83:357-367.
Morozov A, Shiyanov P, Barr E, Leiden JM, Raychaudhuri P. 1997. Accumulation of human papillomavirus type 16 E7 protein bypasses G1 arrest induced by serum deprivation and by the cell cycle inhibitor p21. J Virol 71:3451-3457.
Mossadegh N, Gissmann L, Muller M, Zentgraf H, Alonso A, Tomakidi P. 2004. Codon optimization of the human papillomavirus 11 (HPV 11) L1 gene leads to increased gene expression and formation of virus-like particles in mammalian epithelial cells. Virology 326:57-66.
Muller M. 2005. Codon optimization of papillomavirus genes. Methods Mol Med 119:433-444.
Munger K, Phelps WC, Bubb V, Howley PM, Schlegel R. 1989a. The E6 and E7 genes of the human papillomavirus type 16 together are necessary and sufficient for transformation of primary human keratinocytes. J Virol 63:4417-4421.
Munger K, Werness BA, Dyson N, Phelps WC, Harlow E, Howley PM. 1989b. Complex formation of human papillomavirus E7 proteins with the retinoblastoma tumor suppressor gene product. Embo J 8:4099-4105.
Munoz N, Bosch FX, de Sanjose S, Herrero R, Castellsague X, Shah KV, Snijders PJ, Meijer CJ. 2003. Epidemiologic classification of human papillomavirus types associated with cervical cancer. N Engl J Med 348:518-527.
Munoz N, Castellsague X, de Gonzalez AB, Gissmann L. 2006. Chapter 1: HPV in the etiology of human cancer. Vaccine 24 Suppl 3:S3/1-10.

Myers G, Bernard HU, Delius H, Favre M, Icenogle JP, Van Ranst M, Wheeler C. 1994. Human Papillomaviruses 1994: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences. Theoretcial Biology and Biophysics, Los Alamos National Laboratory, Los Alamos.

Myers G, Lu H, Calef C, Leitner T. 1996a. Heterogeneity of papillomaviruses. Semin Cancer Biol 7:349-358.

Myers G, Lu H, Calef C, Leitner T. 1996b. Heterogeneity of papillomaviruses. Seminars in Cancer Biology 7:349-358.

Nakagawa S, Huibregtse JM. 2000. Human scribble (Vartul) is targeted for ubiquitin-mediated degradation by the high-risk papillomavirus E6 proteins and the E6AP ubiquitin-protein ligase. Mol Cell Biol 20:8244-8253.

Narechania A, Terai M, Chen Z, DeSalle R, Burk RD. 2004. Lack of the canonical pRB-binding domain in the E7 ORF of artiodactyl papillomaviruses is associated with the development of fibropapillomas. J Gen Virol 85:1243-1250.

Narechania A, Chen Z, DeSalle R, Burk RD. 2005. Phylogenetic incongruence among oncogenic genital alpha human papillomaviruses. J Virol. 79:15503-15510.

Nasir L, Campo MS. 2008. Bovine papillomaviruses: their role in the aetiology of cutaneous tumours of bovids and equids. Vet Dermatol 19:243-254.

Nees M, Geoghegan JM, Hyman T, Frank S, Miller L, Woodworth CD. 2001. Papillomavirus type 16 oncogenes downregulate expression of interferon-responsive genes and upregulate proliferation-associated and NF-kappaB-responsive genes in cervical keratinocytes. J Virol 75:4283-4296.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418-426.

Nelson M, Detmer SE, Wentworth DE, et al. 2012. Genomic reassortment of influenza A virus in North American swine, 1998-2011.

Newton MA, Raftery AE. 1994. Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. Journal of the Royal Statistical Society. Series B (Methodological) 56:3-48.

Nonnenmacher M, Salmon J, Jacob Y, Orth G, Breitburd F. 2006. Cottontail rabbit papillomavirus E8 protein is essential for wart formation and provides new insights into viral pathogenesis. J Virol 80:4890-4900.

Nylander JA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL. 2004. Bayesian phylogenetic analysis of combined data. Syst Biol 53:47-67.

Oelze I, Kartenbeck J, Crusius K, Alonso A. 1995. Human papillomavirus type 16 E5 protein affects cell-cell communication in an epithelial cell line. J Virol 69:4489-4494.

Ogawa T, Tomita Y, Okada M, Shinozaki K, Kubonoya H, Kaiho I, Shirasawa H. 2004. Broad-spectrum detection of papillomaviruses in bovine teat papillomas and healthy teat skin. J Gen Virol 85:2191-2197.

Okun MM, Day PM, Greenstone HL, Booy FP, Lowy DR, Schiller JT, Roden RB. 2001. L1 interaction domains of papillomavirus l2 necessary for viral genome encapsidation. J Virol 75:4332-4342.

Ong C-K, Chan S-Y, Campo MS, et al. 1993. Evolution of human papillomavirus type 18: an ancient phylogenetic root in Africa and intratype diversity reflect coevolution with human ethnic groups. J Virol. 67:6424-6431.

Orth G, Jablonska S, Favre M, Croissant O, Obalek S, Jarzabek-Chorzelska M, Jibard N. 1981. Identification of papillomaviruses in butchers' warts. J Invest Dermatol 76:97-102.

Otten N, von Tscharner C, Lazary S, Antczak DF, Gerber H. 1993. DNA of bovine papillomavirus type 1 and 2 in equine sarcoids: PCR detection and direct sequencing. Arch Virol. 132:121-131.

Page RD. 1995. TreeMap. Pp. Computer program distributed by the author. University of Glasgow, Glasgow.

Page RD. 2003. Tangled Trees: Phylogeny, Cospeciation, and Coevolution. The University of Chicago Press, Chicago and London.

Page RDM. 1990a. Component analysis: A valiant failure? Cladistics 6:119-136.

Page RDM. 1990b. Temporal Congruence and Cladistic Analysis of Biogeogrphy and Cospeciation. Systematic Zoology 39:205-226.

Page RDM. 1994a. Maps between Trees and Cladistic Analysis of Historical Associations Among Genes, Organisms, and Areas. Systematic Biology 43:58-77.

Page RDM. 1994b. Parallel Phylogenies: Reconstructing the History of Host-Parasite Assemblages. Cladistics 10:155-173.

Palumbi SR. 1989. Rates of molecular evolution and the fraction of nucleotide positions free to vary. J Mol Evol 29:180-187.

Park JS, Kim EJ, Kwon HJ, Hwang ES, Namkoong SE, Um SJ. 2000. Inactivation of interferon regulatory factor-1 tumor suppressor protein by HPV E7 oncoprotein. Implication for the E7-mediated immune evasion mechanism in cervical carcinogenesis. J Biol Chem 275:6764-6769.

Penny D, Hendy MD. 1985. The Use of Tree Comparison Metrics. Systematic Zoology 34:75-82.

Petry KU, Scheffel D, Bode U, Gabrysiak T, Kochel H, Kupsch E, Glaubitz M, Niesert S, Kuhnle H, Schedel I. 1994. Cellular immunodeficiency enhances the progression of human papillomavirus-associated cervical lesions. Int J Cancer 57:836-840.

Phelps WC, Yee CL, Munger K, Howley PM. 1988. The human papillomavirus type 16 E7 gene encodes transactivation and transformation functions similar to those of adenovirus E1A. Cell 53:539-547.

Phelps WC, Munger K, Yee CL, Barnes JA, Howley PM. 1992. Structure-function analysis of the human papillomavirus type 16 E7 oncoprotein. J Virol 66:2418-2427.

Pietenpol JA, Stein RW, Moran E, Yaciuk P, Schlegel R, Lyons RM, Pittelkow MR, Munger K, Howley PM, Moses HL. 1990. TGF-beta 1 inhibition of c-myc transcription and growth in keratinocytes is abrogated by viral transforming proteins with pRB binding domains. Cell 61:777-785.

Posada D, Crandall KA. 2002. The effect of recombination on the accuracy of phylogeny estimation. J Mol Evol 54:396-402.

Raj K, Berguerand S, Southern S, Doorbar J, Beard P. 2004. E1 empty set E4 protein of human papillomavirus type 16 associates with mitochondria. J Virol 78:7199-7207.

Rambaut A, Bromham L. 1998. Estimating divergence dates from molecular sequences. Mol Biol Evol 4:442-448.

Rannala B. 2002. Identifiability of Parameters in MCMC Bayesian Inference of Phylogeny. Syst Biol 51:754-760.

Rannala B, Zhu T, Yang Z. 2012. Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. Mol Biol Evol 29:325-335.

Rector A, Lemey P, Tachezy R, et al. 2007. Ancient papillomavirus-host co-speciation in Felidae. Genome Biol. 8:R57.

Rehtanz M, Ghim SJ, Rector A, Van Ranst M, Fair PA, Bossart GD, Jenson AB. 2006. Isolation and characterization of the first American bottlenose dolphin papillomavirus: Tursiops truncatus papillomavirus type 2. J Gen Virol 87:3559-3565.

Rivera R, Robles-Sikisaka R, Hoffman EM, Stacy BA, Jensen ED, Nollens HH, Wellehan JFX, Jr. 2012. Characterization of a novel papillomavirus species (ZcPV1) from two California sea lions (Zalophus californianus). Veterinary Microbiology 155:257-266.

Roberts S, Kingsbury SR, Stoeber K, Knight GL, Gallimore PH, Williams GH. 2008. Identification of an arginine-rich motif in human papillomavirus type 1 E1^E4 protein necessary for E4-mediated inhibition of cellular DNA synthesis in vitro and in cells. J Virol 82:9056-9064.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Mathematical Biosciences 53:131-147.

Robl MG, Olson C. 1968. Oncogenic action of bovine papilloma virus in hamsters. Cancer Res 28:1596-1604.

Robles-Sikisaka R, Rivera R, Nollens HH, Leger JS, Durden WN, Stolen M, Burchell J, Wellehan JFX, Jr. 2012. Evidence of recombination and positive selection in cetacean papillomaviruses. Virology 427:189-197.

Roden RB, Lowy DR, Schiller JT. 1997. Papillomavirus is resistant to desiccation. J Infect Dis 176:1076-1079.

Roden RB, Day PM, Bronzo BK, Yutzy WHt, Yang Y, Lowy DR, Schiller JT. 2001. Positively charged termini of the L2 minor capsid protein are necessary for papillomavirus infection. J Virol 75:10493-10497.

Rodrigo AG, Kelly-Borges M, Bergquist PR, Bergquist PL. 1993. A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. N. Z. J. Bot. 31:257-268.

Ronco LV, Karpova AY, Vidal M, Howley PM. 1998. Human papillomavirus 16 E6 oncoprotein binds to interferon regulatory factor-3 and inhibits its transcriptional activity. Genes Dev 12:2061-2072.

Ronquist F, Nylin S. 1990. Process and Pattern in the Evolution of Species Associations. Systematic Zoology 39:323-344.

Ronquist F. 1995. Reconstructing the history of host-parasite associations using generalised parsimony. Cladistics 11:73-89.

Ronquist F. 2002. Parsimony Analysis of Coevolving Species Associations. In: Page RD, editors. Tangled Trees: Phylogeny, Cospeciation and Coevolution. Chicago: University of Chicago Press. p.

Rous P, Beard JW. 1935. The progression to carcinoma of virus-induced rabbit papillomas (Shope). J Exp Med. 62:523-554.

Rzhetsky A, Nei M. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol Biol Evol 10:1073-1095.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-425.

Salter LA, Pearl DK. 2001. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. Syst Biol 50:7-17.

Sanderson MJ. 1997. A nonparameteric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol 12:1218-1232.

Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalised likelihood approach. Mol Biol Evol 19:101-109.

Sankoff D. 1975. Minimal mutation trees of sequences. SIAM J Appl Math 28:35-42.

Scheffner M, Werness BA, Huibregtse JM, Levine AJ, Howley PM. 1990. The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. Cell 63:1129-1136.

Scheffner M, Huibregtse JM, Vierstra RD, Howley PM. 1993. The HPV-16 E6 and E6-AP complex functions as a ubiquitin-protein ligase in the ubiquitination of p53. Cell 75:495-505.

Schiffman M, Herrero R, Desalle R, et al. 2005. The carcinogenicity of human papillomavirus types reflects viral evolution. Virology 337:76-84.

Schmitt A, Harry JB, Rapp B, Wettstein FO, Iftner T. 1994. Comparison of the properties of the E6 and E7 genes of low- and high-risk cutaneous papillomaviruses reveals strongly transforming and high Rb-binding activity for the E7 protein of the low-risk human papillomavirus type 1. J Virol 68:7051-7059.

Schneider C, Weisshart K, Guarino LA, Dornreiter I, Fanning E. 1994. Species-specific functional interactions of DNA polymerase alpha-primase with simian virus 40 (SV40) T antigen require SV40 origin DNA. Mol Cell Biol 14:3176-3185.

Schulz E, Gottschling M, Bravo IG, Wittstatt U, Stockfleth E, Nindl I. 2009. Genomic characterization of the first insectivoran papillomavirus reveals an unusually long, second non-coding region and indicates a close relationship to Betapapillomavirus. Journal of General Virology 90:626-633.

Schwarz TF, Leo O. 2008. Immune response to human papillomavirus after prophylactic vaccination with AS04-adjuvanted HPV-16/18 vaccine: improving upon nature. Gynecol Oncol 110:S1-10.

Sedman J, Stenlund A. 1995. Co-operative interaction between the initiator E1 and the transcriptional activator E2 is required for replicator specific DNA replication of bovine papillomavirus in vivo and in vitro. Embo J 14:6218-6228.

Shadan FF, Villarreal LP. 1993. Coevolution of persistently infecting small DNA viruses and their hosts linked to host-interactive regulatory domains. Proc Natl Acad Sci U S A 90:4117-4121.

Shafti-Keramat S, Handisurya A, Kriehuber E, Meneguzzi G, Slupetzky K, Kirnbauer R. 2003. Different heparan sulfate proteoglycans serve as cellular receptors for human papillomaviruses. J Virol 77:13125-13135.

Shah SD, Doorbar J, Goldstein RA. 2010. Analysis of host-parasite incongruence in papillomavirus evolution using importance sampling. Mol Biol Evol 27:1301-1314.

Sichero L, Ferreira S, Trottier H, Duarte-Franco E, Ferenczy A, Franco EL, Villa LL. 2007. High grade cervical lesions are caused preferentially by non-European variants of HPVs 16 and 18. Int J Cancer 120:1763-1768.

Sichero L, Simao Sobrinho J, Villa LL. 2012. Oncogenic potential diverge among human papillomavirus type 16 natural variants. Virology 432:127-132.

Siddall ME. 1996. Phylogenetic Covariance Probability: Confidence and Historical Associations. Syst Biol 45:48-66.

Siddall ME, Perkins SL. 2003. Brooks Parsimony Analysis: a valiant failure. Cladistics 19:554-564.

Simon-Loriere E, Holmes EC. 2011. Why do RNA viruses recombine? Nat Rev Micro 9:617-626.

Sokal RR, Sneath PHA. 1963. Numerical Taxonomy. San Francisco, CA: W.H. Freeman and Co.

Spink KM, Laimins LA. 2005. Induction of the human papillomavirus type 31 late promoter requires differentiation but not DNA amplification. J Virol 79:4918-4926.

Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc Natl Acad Sci U S A 100:1056-1061.

Steger G, Corbach S. 1997. Dose-dependent regulation of the early promoter of human papillomavirus type 18 by the viral E2 protein. J Virol 71:50-58.

Stenlund A. 2003. E1 initiator DNA binding specificity is unmasked by selective inhibition of non-specific DNA binding. Embo J 22:954-963.

Stevenson B, Choy HA, Pinne M, et al. 2007. Leptospira interrogans endostatin-like outer membrane proteins bind host fibronectin, laminin and regulators of complement. PLoS One 2:e1188.

Stoler MH, Wolinsky SM, Whitbeck A, Broker TR, Chow LT. 1989. Differentiation-linked human papillomavirus types 6 and 11 transcription in genital condylomata revealed by in situ hybridization with message-specific RNA probes. Virology 172:331-340.

Stoler MH, Rhodes CR, Whitbeck A, Wolinsky SM, Chow LT, Broker TR. 1992. Human papillomavirus type 16 and 18 gene expression in cervical neoplasias. Hum Pathol 23:117-128.

Suchard MA, Weiss RE, Sinsheimer JS, Dorman KS, Patel M, McCabe ERB. 2003. Evolutionary similarity among genes. J Amer Statist Assoc 98:653-662.

Sundberg JP, Van Ranst M, Montali R, et al. 2000. Feline papillomas and papillomaviruses. Vet Pathol 37:1-10.

Surti T, Klein O, Aschheim K, DiMaio D, Smith SO. 1998. Structural models of the bovine papillomavirus E5 protein. Proteins 33:601-612.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609-612.

Suzuk L, Noffsinger AE, Hui YZ, Fenoglio-Preiser CM. 1996. Detection of human papillomavirus in esophageal squamous cell carcinoma. Cancer 78:704-710.

Swofford DL. 1998. PAUP*: Phylogenetic analysis using parsimony (* and other methods), version 4.0, Sinauer, Sunderland, Massachusetts.

Tachezy R, Van Ranst MA, Cruz Y, Burk RD. 1994. Analysis of short novel human papillomavirus sequences. Biochem Biophys Res Commun 204:820-827.

Tachezy R, Duson G, Rector A, Jenson AB, Sundberg JP, Van Ranst M. 2002. Cloning and genomic characterization of Felis domesticus papillomavirus type 1. Virology 301:313-321.

Tang S, Tao M, McCoy JP, Jr., Zheng ZM. 2006. The E7 oncoprotein is translated from spliced E6*I transcripts in high-risk human papillomavirus type 16- or type 18-positive cervical cancer cell lines via translation reinitiation. J Virol 80:4249-4263.

Tavare S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect Math Life Sci 17:57-86.

Terai M, Burk RD. 2001a. Characterization of a novel genital human papillomavirus by overlapping PCR: candHPV86 identified in cervicovaginal cells of a woman with cervical neoplasia. J Gen Virol 82:2035-2040.

Terai M, Burk RD. 2001b. Complete nucleotide sequence and analysis of a novel human papillomavirus (HPV 84) genome cloned by an overlapping PCR method. Virology 279:109-115.

Terai M, Burk RD. 2002. Identification and characterization of 3 novel genital human papillomaviruses by overlapping polymerase chain reaction: candHPV89, candHPV90, and candHPV91. J Infect Dis 185:1794-1797.

Terai M, DeSalle R, Burk RD. 2002. Lack of canonical E6 and E7 open reading frames in bird papillomaviruses: Fringilla coelebs papillomavirus and Psittacus erithacus timneh papillomavirus. J Virol. 76:10020-10023.

Thomas JT, Laimins LA. 1998. Human papillomavirus oncoproteins E6 and E7 independently abrogate the mitotic spindle checkpoint. J Virol 72:1131-1137.

Thompson DA, Belinsky G, Chang TH, Jones DL, Schlegel R, Munger K. 1997. The human papillomavirus-16 E6 oncoprotein decreases the vigilance of mitotic checkpoints. Oncogene 15:3025-3035.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15:1647-1657.

Thorne JL, Kishino H. 2002. Divergence Time and Evolutionary Rate Estimation with Multilocus Data. Syst Biol 51:689-702.

Thornton JW, DeSalle R. 2000. A new method to localize and test the significance of incongruence: detecting domain shuffling in the nuclear receptor superfamily. Syst Biol 49:183-201.

Ure AE, Elfadl AK, Khalafalla AI, Gameel AA, Dillner J, Forslund O. 2011. Characterization of the complete genomes of Camelus dromedarius papillomavirus types 1 and 2. J Gen Virol 92:1769-1777.

Van Bressem M-Fo, Cassonnet P, Rector A, Desaintes C, Van Waerebeek K, Alfaro-Shigueto J, Van Ranst M, Orth Gr. 2007. Genital warts in Burmeister's porpoises: characterization of Phocoena spinipinnis papillomavirus type 1 (PsPV-1) and evidence for a second, distantly related PsPV. Pp. 1928-1933.

Van Doorslaer K, Sidi AOMhO, Zanier K, Rybin V, Deryckere F, Rector A, Burk RD, Lienau EK, van Ranst M, Trave G. 2009. Identification of Unusual E6 and E7 Proteins within Avian Papillomaviruses: Cellular Localization, Biophysical Characterization, and Phylogenetic Analysis. Journal of Virology 83:8759-8770.

Van Doorslaer K, Burk RD. 2012. Association between hTERT activation by HPV E6 proteins and oncogenic risk. Virology 433:216-219.

Van Ranst M, Fuse A, Sobis H, De Meurichy W, Syrjanen SM, Billiau A, Opdenakker G. 1991. A papillomavirus related to HPV type 13 in oral focal epithelial hyperplasia in the pygmy chimpanzee. J Oral Pathol Med 20:325-331.

Van Ranst M, Kaplan JB, Sundberg JP, Burk RD. 1995. Molecular evolution of the human papillomaviruses. In: Gibbs AJ, Calisher CH, and Garcia-Arenal F, editors. Molecular Basis of Virus Evolution. Cambridge: Cambridge Univ. Press. p. 455-476.

van Regenmortel MHV, Fauquet CM, Bishop DHL, et al. 2002. Virus Taxonomy. Seventh Report of the International Committee for the Taxonomy of Viruses. Academic Press, New York, San Diego.

Varsani A, van der Walt E, Heath L, Rybicki EP, Williamson AL, Martin DP. 2006. Evidence of ancient papillomavirus recombination. J Gen Virol. 87:2527-2531.

Villa LL, Sichero L, Rahal P, Caballero O, Ferenczy A, Rohan T, Franco EL. 2000. Molecular variants of human papillomavirus types 16 and 18 preferentially associated with cervical neoplasia. J Gen Virol 81:2959-2968.

Villa LL, Costa RL, Petta CA, et al. 2005. Prophylactic quadrivalent human papillomavirus (types 6, 11, 16, and 18) L1 virus-like particle vaccine in young women: a randomised double-blind placebo-controlled multicentre phase II efficacy trial. Lancet Oncol 6:271-278.

Villa LL, Costa RL, Petta CA, et al. 2006. High sustained efficacy of a prophylactic quadrivalent human papillomavirus types 6/11/16/18 L1 virus-like particle vaccine through 5 years of follow-up. Br J Cancer 95:1459-1466.

Volter C, He Y, Delius H, Roy-Burman A, Greenspan JS, Greenspan D, de Villiers EM. 1996. Novel HPV types present in oral papillomatous lesions from patients with HIV infection. Int J Cancer 66:453-456.

Wakeley J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J Mol Evol 37:613-623.

Webb A, Hancock JM, Holmes CC. 2009. Phylogenetic inference under recombination using Bayesian stochastic topology selection. Pp. 197-203.

Webby R, Hoffmann E, Webster R. 2004. Molecular constraints to interspecies transmission of viral pathogens. Nat Med 10:S77-81.

Wildy P. 1971. Monographs in Virology. Basel ; London: Karger.

Wu X, Zhang C, Feng S, et al. 2009. Detection of HPV types and neutralizing antibodies in Gansu province, China. J Med Virol 81:693-702.

Xie W, Lewis PO, Fan Y, Kuo L, Chen MH. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst Biol 60:150-160.

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396-1401.

Yang Z. 1994a. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306-314.

Yang Z. 1994b. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst Biol 43:329-342.

Yang Z, Goldman N, Friday A. 1994. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Mol Biol Evol 11:316-324.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367-372.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555-556.

Yang Z, Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol Biol Evol 14:717-724.

Yang Z, Rannala B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst Biol 54:455-470.

Yang Z. 2006. Computational Molecular Evolution. New York: Oxford University Press.

Yang Z, Rannala B. 2006. Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. Pp. 212-226. Mol Biol Evol.

Yang Z. 2007. Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics. Mol Biol Evol 24:1639-1655.

Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. Nat Rev Genet 13:303-314.

You J, Croyle JL, Nishimura A, Ozato K, Howley PM. 2004. Interaction of the bovine papillomavirus E2 protein with Brd4 tethers the viral DNA to host mitotic chromosomes. Cell 117:349-360.

Yuan H, Ghim S, Newsome J, Apolinario T, Olcese V, Martin M, Delius H, Felsburg P, Jenson B, Schlegel R. 2007. An epidermotropic canine papillomavirus with malignant potential contains an E5 gene and establishes a unique genus. Virology 359:28-36.

Yule G. 1925. A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. Phil Trans R Soc Lon Biol 213:21-87.

Zhang B, Li P, Wang E, Brahmi Z, Dunn KW, Blum JS, Roman A. 2003. The E5 protein of human papillomavirus type 16 perturbs MHC class II antigen maturation in human foreskin keratinocytes treated with interferon-gamma. Virology 310:100-108.

Zhang P, Nouri M, Brandsma JL, Iftner T, Steinberg BM. 1999. Induction of E6/E7 expression in cottontail rabbit papillomavirus latency following UV activation. Virology 263:388-394.

Zhao KN, Hengst K, Liu WJ, Liu YH, Liu XS, McMillan NA, Frazer IH. 2000. BPV1 E2 protein enhances packaging of full-length plasmid DNA in BPV1 pseudovirions. Virology 272:382-393.

Zhao KN, Gu W, Fang NX, Saunders NA, Frazer IH. 2005. Gene codon composition determines differentiation-dependent expression of a viral capsid gene in keratinocytes in vitro and in vivo. Mol Cell Biol 25:8643-8655.

Zhao KN, Chen J. 2011. Codon usage roles in human papillomavirus. Rev Med Virol 21:397-411.

Zheng ZM, Baker CC. 2006. Papillomavirus genome structure, expression, and post-transcriptional regulation. Front Biosci 11:2286-2302.

Zhou J, Sun XY, Stenzel DJ, Frazer IH. 1991. Expression of vaccinia recombinant HPV 16 L1 and L2 ORF proteins in epithelial cells is sufficient for assembly of HPV virion-like particles. Virology 185:251-257.

Zhou J, Liu WJ, Peng SW, Sun XY, Frazer I. 1999. Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. J Virol 73:4972-4982.

Zhu W, Dong J, Shimizu E, Hatama S, Kadota K, Goto Y, Haga T. 2012. Characterization of novel bovine papillomavirus type 12 (BPV-12) causing epithelial papilloma. Arch Virol 157:85-91.

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. in Bryson V, and Vogel HJ, eds. Evolving Genes and Proteins. Academic Press, New York.

zur Hausen H. 1989. Papillomaviruses as carcinomaviruses. Adv Viral Oncol. 8:26.

zur Hausen H. 2000. Papillomaviruses causing cancer: evasion from host-cell control in early events in carcinogenesis. J Natl Cancer Inst. 92:690-698.