Pathogen evolution and disease emergence in carnivores

Alex J. McCarthy*, Marie-Anne Shaw and Simon J. Goodman*

Institute of Integrative and Comparative Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK

Emerging infectious diseases constitute some of the most pressing problems facing human and domestic animal health, and biodiversity conservation. Currently it is not clear whether the removal of past constraints on geographical distribution and transmission possibilities for pathogens alone are sufficient to give rise to novel host–pathogen combinations, or whether pathogen evolution is also generally required for establishment in novel hosts. Canine distemper virus (CDV) is a morbillivirus that is prevalent in the world dog population and poses an important conservation threat to a diverse range of carnivores. We performed an extensive phylogenetic and molecular evolution analysis on complete sequences of all CDV genes to assess the role of selection and recombination in shaping viral genetic diversity and driving the emergence of CDV in non-dog hosts. We tested the specific hypothesis that molecular adaptation at known receptor-binding sites of the haemagglutinin gene is associated with independent instances of the spread of CDV to novel non-dog hosts in the wild. This hypothesis was upheld, providing compelling evidence that repeated evolution at known functional sites (in this case residues 530 and 549 of the haemagglutinin molecule) is associated with multiple independent occurrences of disease emergence in a range of novel host species.

Keywords: canine distemper virus; molecular adaptation; zoonoses; pathogen evolution; haemagglutinin; SLAM receptor

1. INTRODUCTION

Emerging infectious diseases (EIDs) constitute some of the most pressing problems facing human and domestic animal health, and biodiversity conservation (Daszak et al. 2000). Such diseases have obvious implications for human welfare, and also cause enormous economic impacts, and in the context of conservation, can drive rapid population declines, or even extinctions (Daszak et al. 2000; Woolhouse & Gowtage-Sequeria 2005). Recent examples, which have received high levels of attention, include epidemics of SARS disease (Stadler et al. 2003), the invasion of West Nile virus into the Americas (Nash et al. 2001), the potential for the emergence of new influenza pandemic strains (Monto 2005), and the role of the fungal disease chytridiomycosis in global amphibian declines (Daszak et al. 1999). The drivers of disease emergence are still poorly understood. Globalization of trade and travel, along with human impingement into previously undisturbed habitats, can remove past constraints on geographical distributions and transmission possibilities. However, whether such factors alone give rise to novel host–pathogen combinations, or whether pathogen evolution is generally required for establishment in novel hosts or for changes to virulence, remains a major unanswered question (Altizer et al. 2003). Resolving this question has important implications for surveillance, predicting impacts and planning responses to novel EIDs.

The powerful statistical methods now available for the analysis of molecular evolution make it possible to detect signatures of selection in DNA/RNA sequences from pathogen populations, but linking such signals or specific genetic polymorphisms to the process of disease emergence is more problematic (Yang et al. 2000; Haydon et al. 2001; Woeik et al. 2001; Anishchenko et al. 2006). Rather than simply identifying sites that might be under selection, this requires a hypothesis-testing approach where variation at sites known to be functionally significant in the infection or virulence process, for example at residues involved in interactions with specific host receptors (and therefore cell types or species), can be tested for association with host switches or modulation of virulence. To date this has been achieved for only a very small number of pathogens, generally with clear implications for human health, and most represent only single instances of host switches. For example, variation at residues 226 and 228 of the haemagglutinin protein (HA) in influenza A H2 and H3 serotypes is a determinant of receptor specificity, while variation at residues 182 and 192 of HA in influenza H5N1 is likely to be involved in the adaptation of H5N1 to human hosts (Connor et al. 1994; Yamada et al. 2006); and Anishchenko et al. (2006) have identified a mutation that allows an equine avirulent strain of Venezuelan equine encephalitis virus (VEEV, a zoonotic human pathogen) circulating in rodents to infect and amplify in horses. There are few robust examples of such evolution in emerging wildlife diseases or of evolution associated with multiple independent switches to a range of different hosts by the same pathogen (Altizer et al. 2003). The best characterized examples include the sudden emergence of canine parvovirus (CPV), which infects domesticated dogs, wolves and coyotes, in 1978 and its subsequent development into a global pandemic. CPV evolved from feline parvovirus (FPV; Hueffer & Parrish 2003) and the jump into a canine host...
range was facilitated by substitutions in the FPV capsid protein (e.g. VP2 residues K93N and D323N) which allowed CPV to use the canine transferrin receptor to infect canine host cells (Chang et al. 1992; Hueffer et al. 2003). Recently, a T249P amino acid substitution in the NS3 helicase of American isolates of West Nile Virus (WNV) has been found to modulate increased virulence of WNV in corvids (Brault et al. 2007). Diversification of lentiviruses in primates probably constitutes the best example of adaptation associated with multiple independent emergence events in different host species (Hahn et al. 2000). Other examples of positive selection and host adaptation do exist, e.g. for rabies (Holmes et al. 2002), but for most cases the specific functional roles of the residues involved have not yet been elucidated. In order to understand the broader role of pathogen evolution in disease emergence, more studies are required on pathogens that infect a diverse host range, and have well-characterized molecular and cell biology.

Canine distemper virus (CDV), a morbillivirus in the paramyxovirus family, has high prevalence in the world dog population, causing a range of symptoms including nasal and conjunctival discharge, respiratory congestion, fever, immunosuppression and neurological damage. CDV poses an important conservation threat to many carnivore species. Spillover resulting from interactions between domestic or feral dogs and wild species has led to mass mortalities in species ranging from wild canids (African wild dog, *Lycaon pictus*, bat-eared fox, *Otocyon megalotis*; Carpenter et al. 1998), felids (lion, *Panthera leo*; Harder et al. 1996), hyaenids (spotted hyena, *Crocuta crocuta*; Haas et al. 1996), phocids (Caspian seal, *Phoca caspica*; Baikal seal, *Phoca sibirica*; Mamanev et al. 1995; Kennedy et al. 2000), mustelids (black-footed ferret, *Mustela nigripes*; Williams et al. 1988), viverrids (palm civet, *Paracynictis larvata*; Machida et al. 1992), ailurids (red panda, *Ailurus fulgens*) to procyonids (raccoon, *Procyon lotor*; Lemberger et al. 2005). In many cases CDV is a direct threat to the continued persistence of small populations of conservation concern, having extirpated the last remnant wild population of the black-footed ferret in 1985 (Williams et al. 1988), and causing recurrent mortality among African wild dogs (Alexander & Appel 1994). In addition, CDV may have contributed to the extinction of the Tasmanian tiger (*Thylacinus cynocephalus*) at the beginning of the twentieth century (Guiler 1961). CDV therefore meets the criterion of having a diverse host range in which it is possible to test repeated independent emergence and to evaluate the role of evolutionary changes in such events.

Morbilliviruses also have the benefit of having well-characterized cell and molecular biology, mostly based on the work on measles virus (MV) and CDV. The morbilliviruses have a non-segmented single-stranded negative sense RNA genome encoding six proteins. Haemagglutinin (H) and fusion (F) glycoproteins form the virus envelope and are crucial to cellular infection via attachment to signalling lymphocyte activation molecule (SLAM) and CD46 cellular receptors (Tatsu et al. 2001). The H protein binds to one or more receptors resulting in cellular attachment and activation of the F protein by tissue-specific proteases leading to cellular infection, therefore these proteins have key roles in determining host range and tropism (Lamb & Kolakofsky 2001; Seki et al. 2003). These molecules are well characterized for MV and CDV, with predicted tertiary structures available, including the identification of key functional residues. In particular, sites 530 and 548 of the haemagglutinin protein have been shown experimentally to determine host cell tropism *in vitro* (Vongpunsawad et al. 2004; von Messling et al. 2005). The genome is packed in a ribonucleoprotein (RNP) structure composed of the nucleocapsid protein (N), phosphoprotein (P) and RNA-dependent RNA-polymerase (L). The matrix protein (M) links the RNP to the virus envelope.

Previous evolutionary studies of CDV have largely focused on phylogenetic relationships among different strains, but lack consensus, using a variety of complete and partial sequences from different genes, and a range of phylogenetic methodologies (Hashimoto et al. 2001; Martella et al. 2006, 2007). Studies of complete *H* gene sequences have identified seven distinct geographically separated clusters of CDV isolates, but some relationships among these lineages remain ambiguous. No studies have attempted to investigate patterns of selection or recombination or how these relate to the many instances of emergence in non-dog hosts.

Here we perform extensive phylogenetic and molecular evolution analyses using full-length sequences for all CDV genes in order to determine the role of selection and recombination in shaping viral genetic diversity. We test the specific hypothesis that molecular adaptations at the receptor-binding sites 530 and 548 of the haemagglutinin gene, which have a functional role in determining host cell tropism, are associated with independent incidents of the spread of CDV to non-dog hosts. This hypothesis was upheld for residue 530, and we additionally identify residue 549 in the SLAM-binding region as another strong candidate, providing compelling evidence that repeated pathogen evolution at functional sites is associated with multiple, independent incidents of disease emergence in a range of novel hosts.

2. MATERIAL AND METHODS

(a) Datasets

Sequences were extracted from the GenBank database. Datasets for the analysis of selection and recombination include CDV sequences acquired from dogs and non-dog hosts. Datasets were constructed for *F* (*n* = 13, 1782 bp), *H* (*n* = 73, 1812 bp), *L* (*n* = 9, 6552 bp), *M* (*n* = 11, 1005 bp), *N* (*n* = 12, 1569 bp) and *P* (*n* = 9, 1521 bp) genes (Accession numbers available in table S1 in the electronic supplementary material and *figure 1*) and comprise all distinct full-length sequences publicly available at the time of writing. Sequences were aligned using *Clustal X* (Thompson et al. 1997).

(b) Selection analysis

To investigate selection pressures on gene sequences, the maximum likelihood method of Yang et al. (2000) was implemented. The method calculates *d*~*u*~/*d*~*s*~ ratios (*ω* value) taking into account the phylogenetic relationship between taxa, and codon usage and transition/transversion biases for a number of models (Yang 1997). The M0 model averages the *ω* value between 0 and 1. The M1a model of neutral evolution calculates the proportion of sites that are conserved (*ω* = 0, *p*0) and neutrally evolving (*ω* = 1, *p*1). The M2a model extends the M1a model by incorporating a third class of sites
that can estimate $\omega > 1$, therefore taking into account positive selection. The M3 model of positive selection estimates the $\omega$ ratio for three classes of sites ($p_0$, $p_1$ and $p_2$). A second model of neutral evolution, the M7 model, calculates values of $\omega$ between 0 and 1 for 10 categories using a discrete $\beta$ distribution (controlled by parameters $p$ and $q$). The M8 model extends the M7 model of neutral evolution by the addition of two parameters that estimate $\omega > 1$ for an additional class of sites, and therefore takes into account positive selection. Comparison of likelihood values for nested models by likelihood ratio tests (LRTs) determines if models of positive selection (M2a, M3 and M8) are significantly more likely than models of neutral evolution (M1a and M7). Bayesian methods are used to locate specific sites that have $\omega > 1$ with high posterior probabilities. The analysis was implemented using the CODEML program of the PAML package (Yang 1997).

(c) Recombination analysis
The LDDHAT program was used to estimate the population recombination rate ($\rho$) and population mutation rate ($\theta$) using a composite-likelihood method from a set of aligned sequences (McVean et al. 2002). Estimations are based on biallelic sites where the frequency of alleles was higher than 0.1. The significance of recombination rates are tested against a null hypothesis of no recombination ($\rho=0$) via a likelihood permutation-based approach. Relative estimates of the population recombination rate are robust to violations in key assumptions of the model, such as the presence of selection (Richman et al. 2003; Smith & Fearnhead 2005). Alternative measures of recombination $r^2$ and $|D'|$ (Awadalla et al. 1999) were also assessed. Tests for these measures are based on the correlation of linkage disequilibrium (LD) with distance between pairs of polymorphic sites, under the expectation that a significant decline in LD with distance signifies that recombination has occurred within a population (McVean et al. 2002). The program GENECONV, v. 1.81 (Sawyer 1989, 1999), was used to identify intragenic recombination/gene conversion events.

(d) Phylogenetic analysis
The most appropriate models of molecular evolution for use in the phylogenetic analyses were identified using the MODELGENERATOR program (Keane et al. 2006). Phylogenetic trees for the $H$ gene were constructed using Bayesian and Maximum likelihood approaches. Bayesian trees were obtained using MrBayes v. 3.1.1. (Huelsenbeck & Ronquist 2001). Three replicate analyses were run for a minimum of 1 000 000 generations to ensure convergence of the MCMC chains, with an approximate 20% burn-in period assessed by the standard deviation between chains falling below 0.05. Maximum likelihood phylogenies with support derived from 1000 bootstrap replicates were constructed in TREEFINDER (Jobb et al. 2004).

3. RESULTS
(a) Detection of selection
For the CDV $H$ gene, 73 CDV sequences isolated from dogs and non-dog hosts (see figure 1) were analysed, excluding vaccine sequences AF259552, AF378705, AB212966 and Z35493. Likelihood scores were highest in models of positive selection (M2a, M3 and M8; Table S2 in the electronic supplementary material). 0.09% of sites had a $\omega$ value of 2.51, and 2.3% of sites had a $\omega$ value of 1.886 for the M3 and M8 models, respectively. This suggests that a small proportion of residues are under the influence of positive selection. LRTs were performed between nested models to identify the most probable model. The M7 model was rejected in favour of the M8 model with high probability ($\chi^2 = 12.01$, $p = 0.0025$), demonstrating that the observed $\omega$ ratio is significantly greater than 1 (table 1). Under the M8 model of positive selection, 10 individual residues were identified as being under the influence of positive selection ($\omega > 1$) by Bayesian analysis (Sites 29 ($\rho > 1 = 0.803$), 178 (0.541), 180 (0.929), 225 (0.608), 386 (0.552), 412 (0.669), 475 (0.784), 530 (0.973), 549 (0.865) and 603 (0.868)). The Bayesian posterior support for some sites is relatively low when taken alone, but support for the key 530 site and 549 (which is also in the SLAMBinding region) is among the highest identified. Site 548 was not identified as being under selection. We report all sites with $\omega > 1$ as the phylogenetic distribution of these sites among CDV strains is informative. Bayesian posterior probability values for sites with $\rho \geq 0.75$ are shown in italics.

Analysis of full-length gene sequences for CDV $F$, $L$, $M$, $N$ and $P$ genes revealed no $\omega$ ratios that were significantly greater than 1 based on LRTs, and therefore positive selection cannot be detected in these genes. Within the $F$ gene, a minority of sites (approx. 24%) were confirmed as evolving neutrally (where $\omega = 1$; table S2 in the electronic supplementary material), an observation supported by the rejection of the M0 model of purifying selection by LRT (where $\omega < 1$) in favour of models incorporating neutral evolution (table 1).

For the $M$ gene, the highest likelihood scores were observed in the models of positive selection (M2a, M3 and M8; table S2 in the electronic supplementary material). However, none of these models could be accepted in favour of nested models of neutral evolution (table 1). A high proportion of sites in this gene are highly conserved (approx. 98%) and likely to be constrained by purifying selection (table S2 in the electronic supplementary material). LRTs suggest that the $L$ gene is highly conserved (table 1), with a small proportion (4.5–8.2%) of sites evolving neutrally (table S2 in the electronic supplementary material). Positive selection was not significant in the analysis of the $N$ gene dataset, again the M0 model was rejected in favour of models of neutral evolution (table 1), and this suggests that the majority of sites are conserved (approx. 92%; table S2 in the electronic supplementary material). Analysis of the $P$ gene shows no significant signal of positive selection (table 1). The M0 model was rejected in favour of neutral models. A high proportion of sites are conserved (approx. 72%) and a proportion of sites could be under weak positive selection ($\omega = 1.18341$; table S2 in the electronic supplementary material), but this cannot be confirmed as the M8 model could not be accepted over the M7 model.

(b) Detection of recombination
Estimates of the population recombination rate ($\rho$) for CDV genes ranged from 1.202 to 23.835 (table 2). The strongest signal of recombination was identified in the $F$ gene (table 2), returning values of $\rho = 23.835$ and $|D'| = -0.0516$, both of which were significantly different from zero ($p < 0.01$). Permutation tests of $r^2$ yielded borderline significant results ($p = 0.049$) for the CDV $L$
3168  A. J. McCarthy et al.  Disease emergence in carnivores

Table 1. Likelihood ratio tests (LRTs) between models for selection analysis on CDV. (LRTs are calculated by taking twice the difference between the two models of codon evolution and comparing with a $\chi^2$ distribution. Significant $p$ values (in italics) indicate that the null hypothesis (first model) can be rejected in favour of the alternative hypothesis (second model), e.g. for $M_7$ versus $M_8$, $M_7$ is rejected in favour of $M_8$.)

<table>
<thead>
<tr>
<th>gene</th>
<th>LRT</th>
<th>$\chi^2$ score</th>
<th>$p$ value</th>
<th>gene</th>
<th>LRT</th>
<th>$\chi^2$ score</th>
<th>$p$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F$</td>
<td>M0vM1</td>
<td>70.790504</td>
<td>0.0001</td>
<td>$H$</td>
<td>M0vM1</td>
<td>135.373318</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>M1vM2</td>
<td>0.000676</td>
<td>0.9997</td>
<td></td>
<td>M1vM2</td>
<td>0</td>
<td>0.9999</td>
</tr>
<tr>
<td></td>
<td>M0vM3</td>
<td>70.790516</td>
<td>0.0001</td>
<td></td>
<td>M0vM3</td>
<td>140.154214</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>M1vM3</td>
<td>0.000688</td>
<td>1.0000</td>
<td></td>
<td>M1vM3</td>
<td>4.780896</td>
<td>0.3105</td>
</tr>
<tr>
<td></td>
<td>M2vM3</td>
<td>0.000012</td>
<td>1.0000</td>
<td></td>
<td>M2vM3</td>
<td>4.780896</td>
<td>0.0916</td>
</tr>
<tr>
<td></td>
<td>M7vM8</td>
<td>0.827042</td>
<td>0.6613</td>
<td></td>
<td>M7vM8</td>
<td>12.012246</td>
<td>0.0025</td>
</tr>
<tr>
<td>$L$</td>
<td>M0vM1</td>
<td>51.520978</td>
<td>0.0001</td>
<td>$M$</td>
<td>M0vM1</td>
<td>13.32445</td>
<td>0.0013</td>
</tr>
<tr>
<td></td>
<td>M1vM2</td>
<td>0</td>
<td>1.0000</td>
<td></td>
<td>M1vM2</td>
<td>1.746534</td>
<td>0.4176</td>
</tr>
<tr>
<td></td>
<td>M0vM3</td>
<td>51.958598</td>
<td>0.0001</td>
<td></td>
<td>M0vM3</td>
<td>14.716234</td>
<td>0.0053</td>
</tr>
<tr>
<td></td>
<td>M2vM3</td>
<td>0.43762</td>
<td>0.9793</td>
<td></td>
<td>M2vM3</td>
<td>3.138318</td>
<td>0.5350</td>
</tr>
<tr>
<td></td>
<td>M1vM3</td>
<td>0.43762</td>
<td>0.8035</td>
<td></td>
<td>M1vM3</td>
<td>1.391784</td>
<td>0.4986</td>
</tr>
<tr>
<td></td>
<td>M7vM8</td>
<td>0.207754</td>
<td>0.9013</td>
<td></td>
<td>M7vM8</td>
<td>3.932918</td>
<td>0.1400</td>
</tr>
<tr>
<td>$N$</td>
<td>M0vM1</td>
<td>14.062176</td>
<td>0.0009</td>
<td>$P$</td>
<td>M0vM1</td>
<td>15.371228</td>
<td>0.0005</td>
</tr>
<tr>
<td></td>
<td>M1vM2</td>
<td>0.439956</td>
<td>0.8025</td>
<td></td>
<td>M1vM2</td>
<td>0.279998</td>
<td>0.8694</td>
</tr>
<tr>
<td></td>
<td>M0vM3</td>
<td>14.062176</td>
<td>0.0009</td>
<td></td>
<td>M0vM3</td>
<td>15.371228</td>
<td>0.0005</td>
</tr>
<tr>
<td></td>
<td>M1vM3</td>
<td>0.439956</td>
<td>0.9791</td>
<td></td>
<td>M1vM3</td>
<td>0.279998</td>
<td>0.9911</td>
</tr>
<tr>
<td></td>
<td>M2vM3</td>
<td>0</td>
<td>1.0000</td>
<td></td>
<td>M2vM3</td>
<td>0</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>M7vM8</td>
<td>0.476686</td>
<td>0.7879</td>
<td></td>
<td>M7vM8</td>
<td>0.29092</td>
<td>0.8646</td>
</tr>
</tbody>
</table>

Table 2. Recombination analysis in CDV. (Calculation of $4N_cD_r$ ($\rho$), correlation scores for $r^2$ and $|D'|$ against distance, and permutation tests to evaluate if each statistic was significantly different from zero were performed in the LDHAT package. Significant $p$ values from permutations tests, indicating observed test statistics are significantly greater than zero, are shown in italics.)

| gene | $\rho(4N_cD_r)$ | $r^2$ | $|D'|$ | permutation tests |
|------|-----------------|------|------|------------------|
|      |                 |      |      | $P_{18}$     | $P_{12}$ | $P_{|D'|}$ |
| $F$  | 23.835          | -0.01458 | -0.05126 | 0.0001 | 0.1210 | 0.0070 |
| $H$  | 6.000           | -0.03171 | -0.03335 | 0.9090 | 0.9470 | 0.9230 |
| $L$  | 3.653           | -0.01171 | -0.01131 | 0.1290 | 0.0490 | 0.1160 |
| $M$  | 0.000           | 0.00527  | 0.00717 | 0.0720 | 0.5100 | 0.2560 |
| $N$  | 1.202           | -0.02523 | -0.3226 | 0.8370 | 0.1870 | 0.0440 |
| $P$  | 15.158          | 0.01471  | 0.03176 | 0.4550 | 0.7880 | 0.9460 |

*gene*, and a similar borderline result was returned for the $N$ gene with the $|D'|$ statistic ($p=0.044$). No test statistics were significantly different from zero for the $H$, $M$, and $P$ genes. **GeneConv** identified two pairs of recombinant sequences in the $F$ gene from dog isolates (AY964108 :AY964114 and AY964112 : AY964114), and none for the $H$ gene. In the $F$ gene, the recombinant sections lay between nucleotides 907 and 1361 (amino acids 302–453) of the coding domain in both pairs of recombinant sequences.

### (c) Phylogenetic analysis of CDV H gene sequences and distribution of sites under positive selection

In order to map the phylogenetic distribution of sites identified as being under positive selection in the $H$ gene, we reconstructed the phylogenetic relationships between CDV strains based on the nucleotide alignment of complete $H$ gene sequences (figure 1). Bayesian posterior probabilities and maximum likelihood bootstrap values within continental clades are not shown for clarity of the figure, but were consistently high.

Our analysis identifies the same geographical clades for CDV isolates that have been previously reported (Hashimoto *et al.* 2001; Martella *et al.* 2006, 2007). However, the topology derived in our analysis, rooted with phocine distemper virus (PDV-1) strains, has consistently higher support values than previous studies and places the Arctic-like clade basal to the Asia-2 clade.

Mapping the distribution of substitutions at sites identified to be under the influence of positive selection on to the phylogeny showed that the majority of amino acid substitutions were concentrated in non-dog host isolates (figure 1). In particular, substitutions at sites 530 and 549 predominantly occur in CDV isolates obtained from novel host species, indicating that the spread of CDV into these non-dog hosts is associated with evolution at these sites. Six different residues (D, E, G, N, R and S) are observed at residue 530 (530E) is distributed in all Asia-2 isolates and 530G is observed in Asia-1 and European isolates, as well as the entire America-2 clade excluding two raccoon isolates (figure 1). Out of 21 non-dog hosts, 9 (host species: raccoon; ferret; mink, *Mustela vison*; red fox, *Vulpes vulpes*; red panda and Baikal seal) have substitutions of E/G530 to R, D or N residues. This provides further support along with experimental evidence that site 530 could be important in determining host tropism (Seki *et al.* 2003).
Figure 1. Phylogenetic relationship between CDV strains based on H gene sequences and the distribution of H protein residues at sites under positive selection across lineages. PDV-1 strains were used as out-group and only the tree topology is shown for clarity. Bayesian posterior probabilities and maximum likelihood bootstrap support values (in parentheses) for the major lineages are given at the nodes. Host species, accession number and isolate location are also shown. CDV strains form seven distinct clusters: America-1 (vaccines), America-2, Arctic-Like, Asia-1, Asia-2, Europe and European wildlife. Residues 530 and 549 are located in the regions of SLAM receptor-binding importance. Circled residues are substitutions at residues 530 and 549 of isolates from non-dog hosts.

We assessed the role of evolution in driving host switches and the emergence of CDV, one of the most important carnivore pathogens, in non-dog host species. CDV is an ideal candidate for studying the role of pathogen evolution in host switches, not only because it infects a diverse host range and is an important threat to wild carnivores of conservation concern (Mamaev et al. 1995; Harder et al. 1996; Kennedy et al. 2000; Cleaveland et al. 2006), but also because its molecular and cell biology are well characterized with tertiary structures predicted for key virus molecules. This allows a hypothesis-testing approach to be used where evidence for evolution at specific virus residues with a known function can be assessed. Using phylogenetic and molecular evolution approaches, we tested the specific hypothesis that molecular adaptation at residues 530 and 548 of the CDV haemagglutinin protein, which are involved in receptor tropism, is associated with host switches and the spread of CDV to non-dog hosts. We found that site 530 is indeed under strong positive selection, along with site 549, another residue in the SLAM receptor-binding region, but not site 548. The great majority of amino acid substitutions at these sites map to CDV isolates derived from non-dog hosts, indicating that adaptation at these receptor-binding sites is associated with the spread of CDV from dogs to non-dog hosts. The study provides compelling evidence for pathogen evolution driving multiple independent emergences of disease in novel hosts, and suggests that pathogen evolution is likely to be an important general driver of the spread of diseases to novel host ranges alongside other human mediated drivers.

Our analysis reveals signals of both selection and recombination in different genes for the global CDV population. Analysis of the CDV H gene reveals both that the overall \( \omega \) ratio is significantly greater than 1 under the M8 model (\( p = 0.0025 \)), and that a small number of specific residues are important in the adaptive evolution of CDV through positive selection, owing to their interactions with cellular receptors or with the host immune system.

Experimental evidence from in vitro receptor-binding studies shows that residues 527, 528, 529 and 552, conserved between all morbilliviruses, are crucial for CDV SLAM-dependent fusion (von Messling et al. 2005). Residues 526, 547 and 548 are also important in efficient SLAM-dependent fusion. These clusters are located in \( \beta \)-sheet 5 of the predicted H protein tertiary structure, and are predicted to be accessible for receptor interactions (figure 2; von Messling et al. 2005). It has previously been determined that mutation of site 530, in addition to substitution at site 548, of the H protein was critical in the adaptation of CDV from vero cells to marmoset B cells in culture, signifying adaptation from CD46-dependent fusion to SLAM-dependent fusion (Seki et al. 2003).

Residue 530 (Pr \( \omega > 1 = 0.969 \)) was identified by our analysis as being under the influence of positive selection. No signal of positive selection was detected at site 548, suggesting that this residue is not an important determinant of tropism in the wild; however, positive selection was

Figure 2. CDV H protein predicted tertiary structure (von Messling et al. 2005). Residues 527, 528, 529 and 552 (required for SLAM-dependent fusion and conserved between all morbilliviruses), as well residues 526, 547 and 548 (also important in SLAM-dependent fusion), are shown in black. Residues under the influence of positive selection are shown in grey. (a) Top view. (b) Side view. The illustrations of the models were prepared with Pymol (DeLano 2002).
detected at site 549. Residues 530 and 549 both fall into receptor-binding domains located on propeller β-sheet 5 of the haemagglutinin molecule (figure 2). This evidence indicates that molecular adaptation of residues 530 and 549 may change the affinity of interaction between CDV H protein and SLAM in different host species, and supports our hypothesis that molecular adaptation in receptor-binding domains is associated with independent incidents of spread of CDV to non-dog hosts. MV H residues 534 and 553 (corresponding to CDV sites 530 and 549) are not under the influence of positive selection (Woelk et al. 2002). In addition to receptor interactions, B-cell epitopes (BCEs) and T-cell epitopes (TCEs) are also reported to localize to the CDV H protein, and the antigenic region of the H protein has been reported to have changed in recent CDV field isolates (Orvell et al. 1985; Iwatsuki et al. 2000). However, conclusions regarding CDV H residues under the influence of immune selection cannot be drawn at this time as predictive methodologies currently only describe human and mouse epitopes.

We determined the phylogenetic distribution of substitutions at all sites identified to be under the influence of positive selection (figure 1). Strikingly, of the 21 non-dog host isolates 9 have an E/G530D/N/R substitution, 8 have an Y549H substitution and 2 have substitutions at both residue positions (figure 1). In addition, partial H gene sequences from CDV isolates from the year 2000 epidemic in Caspian seals contain both E/G530N and Y549H substitutions (data not shown—Barrett & Banyard 2007, unpublished data). Phylogenetic analysis (not shown) indicates that these sequences form a single cluster that diverges after the America-1 clade, but basal to the Arctic-like clade. As these were not full-length sequences, we did not include them in the full analysis.

Substitutions at residues 530 and 549 are associated with CDV isolated from novel host species, and in the light of our analysis, such substitutions can also be observed directly in experimental infections. In a previous study by von Messling et al. (2003), CDV strain 5804 was passaged experimentally through ferrets yielding the 5804P ferret-adapted strain (AY386316) which differed from the parental 5804 dog strain (AY396315) at residues 106 and 549 of the haemagglutinin protein. This provides strong experimental evidence for a functional role of residue 549 in host switches. Further analysis of the role of residues 530 and 549 by manipulation of viral sequences and experimental infections would provide more insight into the mechanisms which determine host tropism for this virus. In general analyses which combine phylogenetic, molecular evolution and experimental approaches are likely to be productive routes for investigating the evolutionary basis of host switches in many viral pathogens.

Statistically significant evidence for positive selection was not detected in analyses of the F, L, M, N and P genes. These analyses suggest that a proportion of sites in these genes are conserved, and that a small proportion of sites are neutrally evolving. Notably, analysis of the L and M genes suggest that a high proportion of sites are conserved, and likely to be subject to purifying selection, as previously identified in measles virus L and M genes (Woelk et al. 2002). The high sequence conservation identified in the L and M proteins could be due to crucial roles in viral transcription and replication, and lytic formation, respectively (Lamb & Kolakofsky 2001). The lack of selection signal from the CDV F gene is surprising when previous studies identified immune selection in the MV F gene and mapped selected residues to BCEs and TCEs, but could be due to the small number of full-length gene sequences currently available for analysis (Muller et al. 1993, 1996; van Binnendijk et al. 1993; Woelk et al. 2002). Recombination could also be a confounding factor but this is more likely to generate a false signal of selection (Anisimova et al. 2003). Our analyses suggest that the CDV N and P genes are not under the influence of positive selection, despite their known interactions with the host immune response (Tipold et al. 1999; Yoshida et al. 1999; Gotoh et al. 2001; Palosaari et al. 2003; Shaffer et al. 2003; Kerdiles et al. 2006). The availability of CDV gene sequences isolated from a wider range of host species would increase the power of analysis and allow a more detailed assessment of the selection pressures acting on the CDV F, N and P genes. The present work assessed all publicly available full-length H gene sequences at the time of writing. However, there are CDV isolates from epizootics in non-dog hosts for which no sequence covering the SLAM-binding region, or the full length of the F gene is available, which would be very informative to assess in the future. Of special interest would be strains from Serengeti lions and Caspian seals (Roelke-Parker et al. 1996; Carpenter et al. 1998; Kennedy et al. 2000).

Previous studies have revealed the importance of recombination in shaping RNA virus diversity and identified that recombination acts at lower levels in single-stranded negative-sense RNA viruses compared with other RNA viruses (Worobey & Holmes 1999; McVean et al. 2002; Chare et al. 2003; Schierup et al. 2005). In our study, although there is evidence of recombination in the F gene, in which two alternative permutation tests yielded recombination estimates that were significantly different from 0 and two pairs of recombinant sequences were identified in dog isolates, results for other genes should be treated with caution. For the L and N genes, the marginally significant results may be type I errors. Recombination was not detected in an informative site test of CDV F gene sequences in a previous study (Chare et al. 2003). Theoretically, variation in the F gene provided by recombination could change interactions between host tissue-specific proteases and the F protein, thus influencing tropism (Lamb & Kolakofsky 2001). However, the recombinant segments identified by the Geneconv analysis did not map to regions with a known role in host specificity. Alternatively, recombination could be driven by immune selection, but again without knowledge of epitopes generated from this gene this must remain as speculation at this time. Our H gene ρ estimate (6.00) is slightly lower than previous estimates in the MV H gene (7.2–15.8; Chare et al. 2003; Schierup et al. 2005). The underlying reason behind the apparent low levels of recombination in paramyxoviruses is not currently understood. The paramyxovirus template for replication abides to the ‘the rule of six’ which dictates that each nucleocapsid monomer is tightly associated with six nucleotides by hydrophobic bonds. Though this prevents the presence of naked viral RNA, it also raises the question of how template recognition occurs in order to allow recombination to take place, possibly offering an explanation for the apparent low levels of recombination in negative sense RNA viruses (Egelman et al. 1989). Our
analysis detected higher levels of recombination in glycoprotein genes; this could increase variation in proteins that are determinants of cellular tropism and subjected to neutralizing antibodies, an obvious benefit to a virus population (Orvell et al. 1985; Lamb & Kolakofsky 2001; Hirama et al. 2002).

Few studies have detected substitutions at specific functional sites in pathogen proteins mediating the molecular adaptation of a pathogen to different host species. In order for avian influenza A to jump hosts to humans, it must adapt from using the avian form (SAx2,3Gal) to the human form (SAx2,6Gal) of the sialic acid (SA) receptor. Variation at HA residues 226 and 228 in H2 and H3 serotypes have long been known to be associated with host receptor specificity (Connor et al. 1994). More recently, substitutions N182L and Q192R of H5N1 influenza HA molecule were identified to convert HAs to recognize SAx2,6Gal instead of SAx2,3Gal (Yamada et al. 2006). Though mutation at these sites may not alone be sufficient to cause a full pandemic, they may be selected in the early phases of human infection and serve as good markers for the assessment of the ability of avian field H5N1 isolates to replicate in humans. The potential of H5N1 to adapt to humans and emerge as a pandemic strain can be gauged by surveillance of sites 182 and 192 in populations of avian field isolates (Yamada et al. 2006). In the case of CDV, if experimental evidence confirmed the influential role of sites 530 and 549 of CDV H protein in the molecular adaptation of CDV into novel hosts, then surveillance of these sites would be an informative technique to assess the potential of CDV to spread to novel wildlife populations. This could prove to be a valuable tool for combating an infectious disease that has already been associated with declines in wild carnivore populations, and the impact of which is likely to increase in the future.

Our analysis shows that pathogen evolution plays a crucial role in the establishment of CDV in novel carnivore host species. Our findings constitute one of only a small number of cases where repeated independent emergence of a pathogen in novel hosts is associated with evolution of the pathogen at key functional residues that determine host specificity. Probably the best other example comes from lentiviruses (which are also RNA viruses and include the simian (SIV) and human (HIV) immunodeficiency viruses) in African primates. Here, multiple species jumps by different SIV viruses have been documented, with evidence for infection in at least 26 species, associated with polymorphisms in the V2 and V3 loop regions of the env gene, which are involved in interactions with host receptors and determine cell tropism (Hahn et al. 2000). Adaptive variation has also been implicated in species jumps for feline lentiviruses (Poss et al. 2006). Together with other examples such as avian influenza and canine parvovirus, these present a growing body of evidence for the general importance of pathogen evolution in disease emergence, and in the case of host switches by viruses, that subtle adaptations in receptor-binding regions which determine host cell tropism may be an important step for establishment in new hosts. Of course other functional adaptation may also be important for, or arise after establishment in new hosts (Webby et al. 2004), as evidenced by the substitutions which modulate virulence in the VEEV and WNV examples. The relative rarity of these examples so far likely reflects lack of sequence data for viral isolates from multiple hosts species and detailed information about the cell biology and structure of viral proteins which preclude the hypothesis-testing approach taken here and in the other highlighted studies. The future challenges are to increase our understanding of how these evolutionary processes interact with new selection pressures on pathogens arising from human drivers such as globalization, agricultural intensification and impingement into new habitats, to promote disease emergence. Unifying identification of functional adaptation with molecular estimates of the time of disease emergence events and linking these to known changes in human drivers will likely offer a productive way forward.

CDV belongs to the genus *Morbillivirus*, other members of which have also caused mass mortalities among wildlife and livestock. The introduction of rinderpest virus to East Africa in the 1880s caused mortality levels of up to 90% among wild bovid populations such as buffalo *Syncerus caffer*, lesser kudu *Tragelaphus imberbis* and eland *Taurotragus oryx* (Plowright 1982; Kock et al. 1999), and PDV has emerged in harbour seals, *Phoca vitulina* and grey seals, *Halichoerus grypus* of Europe causing mass mortalities (Harkonen et al. 2006). Assessing the role of orthologous *H* gene residues of 530 and 549, in other morbilliviruses, in molecular adaptation to novel hosts would indicate if the same evolutionary mechanisms operate in the host switches of members of the genus *Morbillivirus*. These sites and similar sites in other RNA viruses involved in receptor binding may be potential targets for antiviral therapies. This research shows how combining molecular evolutionary and phylogenetic analysis with findings on the functional and structural properties of pathogen proteins can yield important insights into the process of disease emergence, and could prove crucial in accurately predicting the impact of and planning response to disease emergence.

We thank Roberto Cattaneo for providing the predicted tertiary structure of the CDV *H* protein, and Tom Barrett and Ashley Banyard of the Institute for Animal Health, Pirbright, UK, for giving us access to their unpublished sequence for the *H* gene of the Capssian seal CDV isolate. Two anonymous referees provided valuable comments which helped improve the original manuscript. A.J.M. was supported by a BBSRC PhD studenship.

**REFERENCES**


Disease emergence in carnivores  A. J. McCarthy et al.  3173


Muller, C. P., Ammerlaan, W., Fleckenstein, B., Krauss, S., Kalbacher, H., Schneider, F., Jung, G. & Wiesmuller, K. H.


