The association of recombination events in the founding and emergence
of subgenogroup evolutionary lineages of human enterovirus 71

E.C. McWilliam Leitch¹, M. Cabrerizo², J. Cardosa³, H. Harvala⁴,
O.E. Ivanova⁵, S. Koike⁶, A.C.M. Kroes⁷, A. Lukashev⁵, D. Perera²
M. Roivainen⁸, P. Susi⁹, G. Trallero², D. J. Evans¹⁰, and P. Simmonds¹*  

¹Centre for Infectious Diseases, University of Edinburgh, Summerhall, Edinburgh, EH9 1QH, UK; ²Enterovirus Laboratory, National Centre for Microbiology, Carlos III Institute of Health, Majadahonda, Madrid, Spain; ³Institute of Health and Community Medicine, University Sarawak Malaysia, Malaysia; ⁴Specialist Virology Centre, Royal Infirmary of Edinburgh, Edinburgh, UK, ⁵M.P. Chumakov Institute of Poliomyelitis and Viral Encephalitides, Moscow, Russia; ⁶Tokyo Metropolitan Institute of Medical Science, Tokyo, Japan; ⁷Department of Medical Microbiology, Leiden University Medical Centre, Leiden, The Netherlands; ⁸Intestinal Viruses Unit, National Institute for Health and Welfare, Finland; ⁹Department of Virology, University of Turku, Finland; ¹⁰Department of Biological Sciences, University of Warwick, UK.

Running title: EV71 recombination and evolution

*Correspondence: Tel: 0131 650 7927  
Fax: 0131 650 6511  
Email: Peter.Simmonds@ed.ac.uk

Document totals: Word count for summary: 246  
Word count for main text: 5481
Enterovirus 71 (EV71) is responsible for frequent large scale outbreaks of hand, foot and mouth disease worldwide, and represent a major aetiological agent of severe, sometime fatal neurological disease. EV71 variants have been classified into three genogroups (GgA, GgB, GgC), the latter two further subdivided into subgenogroups, B1-B5 and C1-C5. To investigate the dual roles of recombination and evolution in the epidemiology and transmission of EV71 worldwide, we performed a large scale genetic analysis of isolates (n=308) collected from 19 countries worldwide over a forty year period. A series of recombination events occurred over this period, identified through incongruities in sequence grouping between the VP1 and 3Dpol regions. Eleven 3Dpol clades were identified, each specific to EV71 and associated with specific subgenogroups but interspersed phylogenetically with clades of Coxsackie A virus 16 and other EV species A serotypes. The likelihood of recombination increased with VP1 sequence divergence; mean half-lives for EV71 recombinant forms (RFs) of 6 and 9 years for GgB and GgC overlapped with those observed for the EV- B serotypes, echovirus (E) 9, E30 and E11 respectively (1.3 - 9.8 years). Within genogroups, further, sporadic recombination events occurred, such as the linkage of two B4 variants to RF-W instead of RF-A, and two C4 variants to RF-H. Intriguingly, recombination events occurred as a founding event of most subgenogroups immediately preceding their lineage expansion and global emergence. The possibility that recombination contributed to their subsequent spread through improved fitness requires further biological and immunological characterisation.
INTRODUCTION

Enterovirus (EV) 71 is one of the most frequently pathogenic of the human enteroviruses, responsible for large-scale epidemic occurrences of neurological disease throughout South East Asia (51, 58). EV71 contains a single-stranded, positive-sense RNA genome and is classified as a member of species A (EV-A) in the Enterovirus genus of the Picornaviridae family (60). As with other EV-A enteroviruses, EV71 is transmitted by the faecal-oral route and normally causes sub-clinical or relatively mild, self-limiting infections such as hand, foot and mouth disease (HFMD) (65). However, unlike other members of EV-A, EV71 infections are associated in a small proportion of subjects with a wide array of severe disease presentations including aseptic meningitis, encephalitis and acute flaccid paralysis (AFP) (reviewed in (29, 58)). Severe and fatal EV71 infections are predominantly found in young children, with male patients outnumbering female patients (12, 52). There has been a substantial increase in the frequency and severity of EV71 epidemics in recent years, particularly in the Asia-Pacific region (5, 13, 58, 65), prompting urgent, ongoing investigations of the virological and host factors contributing to the apparently increasing pathogenicity of the virus (5, 11, 58).

Considerable insights into the evolution and molecular epidemiology of circulating strains and genotypes of EV71 from many of the most affected countries have been obtained through analysis of the structural genome regions, principally VP1. EV71 has been classified into a total of three genogroups (Ggs), designated GgA – GgC (6), showing approximately 13-20% amino acid sequence divergence from each other in the VP1 region (6, 28) and which are estimated to have originated from a common ancestor as recently as 1941 (61). Several studies have investigated...
whether different genogroups or subgenogroups vary in their pathogenicity which
might then explain the variability in outcomes of EV71 infections in different decades
and between continents. While whole genome sequence comparisons of EV71 strains
isolated from severe or fatal cases of EV71 infection show no reproducible
differences from those causing more mild infections (54, 57), specific associations of
GgC2 variants in severe neurological disease and of B3 with HFMD or mild /
inapparent infections were observed during an outbreak in Perth, Australia in 1988
when both genogroups were co-circulating (35, 36). More recently, a greater
likelihood of GgC5 to cause neurological complications than GgC4 has been reported
(44).

In the current study we have genetically characterised a large number of EV71
strains collected from South and East Asia, Australia and Europe in the VP1 region
and a distal region of the genome (part of the 3D polymerase-encoding region; 3Dpol)
for identification of recombination events. Recombination is a frequently documented
phenomenon in picornaviruses and contributes to their evolutionary diversity as well
as a means to acquire new phenotypic traits from acquisition of novel combination of
structural and non-structural genes and 5' untranslated region (5'UTR) sequences.

Recombination in picornaviruses was first observed between serotypes of
poliovirus in vaccine recipients (8, 20, 39) and more recently in a wide range of
human enteroviruses (1, 15, 16, 21, 31, 41, 43, 45, 48, 56), foot-and-mouth disease
virus (FMDV) and teschoviruses (55), and more recently parechoviruses (3, 4). In
each virus, recombination breakpoints concentrate in the 2A region although further
sites occur in P2 and P3 genes and in the 5' UTR (1, 3, 16, 30, 33, 49, 67). In contrast,
phylogenies of the more divergent capsid-encoding genes, VP1, VP2 and VP3 are
congruent with each other (and correspond to their serotypic classification) (31, 42,
49), leading to the idea that sequence diversification in structural gene regions is largely uncoupled from that of non-structural genes (31, 32, 37, 55). The contribution of recombination to the evolution and molecular epidemiology of EV71 and its relationship to diversification of capsid-encoding genes is the focus of the current study.

The study used a large dataset of newly acquired and published sequences in VP1 (capsid) and 3Dpol (non-structural) gene regions of EV71 and CAV16. The identification of recombination events was assisted through the assignment of 3Dpol sequences of EV71 isolates to individual recombinant forms, as developed for the analysis of EV species B enteroviruses (37, 38). The study provides convincing evidence for recombination events in the founding of most subgenogroup lineages over an evolutionary timescale comparable to that observed within EV-B enterovirus serotypes. These analyses further our understanding of the molecular epidemiology of EV71 and its varied clinical manifestations.

METHODS

Samples. 193 isolates from 11 countries collected between the years 1990 and 2010 (Table 1) were obtained from internationally distributed referral centres. The following convention was used to name isolates: two letter country code + isolate number / two letter city or region abbreviation / 3Dpol clade / year of collection (e.g. MY40/Sw/A/06 for isolate number 40 referred from Sarawak in Malaysia, isolated in 2006 and belonging to the 3Dpol clade A [defined below]). Sequences obtained in the current study were supplemented with 5 previously unpublished sequences from
Russia collected between 2000 and 2009 and 110 published sequences of complete genomes from 10 countries collected between 1970 and 2009, including the EV71 prototype strain BrCr-CA-70. All sequences in the current study were obtained from laboratory passaged virus stocks, as were the published complete genome sequences of EV71 and CAV16 incorporated in the analysis. This greatly reduces the likelihood of sampling mixed virus populations that might for example have originated from co-infections of the study subjects with different EV71 or CAV-16 strains.

Amplification of VP1 and 3Dpol regions and nucleotide sequencing. RNA extraction and nested reverse transcription (RT)-PCRs were performed as previously using newly-designed enterovirus primers specific for EV71 and Coxsackievirus (CV) A16. These amplified a 1055bp region of the VP1 gene and a 759bp region of the 3Dpol gene. Primers comprised the following: VP1: OS (outer sense; position 2268 in the EV71 prototype strain BRCr-CA-70, U22521): CCN TGG ATH AGY AAC CAN CAY T; OAS (outer antisense; 3411): ACR TAD ATD GCN CCN GAY TGY TG; IS (inner sense; 2332): TNA S NA TYT GGT AYC ARA CAN AYT; IAS (inner antisense; 3604): GAR AAR CTR ACY GGR TAG TGY TTT CT. 3Dpol: OS (5830) GSA CYA TGA TGT AYA AYT TYC CHA C; OAS (7045): GGN GTC ATD GTY ARN CCR TAY TCY TT; IS (6261): ATG AGY ATR GAR GAN GCN TGY TAY G; IAS (7195): TCY TTN GTC CAD CGR ATR GAY TCR T. Amplicons were directly sequenced using BigDye (ABI) and the inner sense or antisense primer and nucleotide sequences were aligned with the Simmonic sequence package version 1.9 (56); (link on http://www.virus-evolution.org/Software). Confirmatory repetitions of RT-PCR and sequencing were performed for all isolates which showed incongruences in tree positions between the VP1 and 3Dpol regions as described below.
Phylogenetic analysis. Bootstrapped maximum likelihood trees for VP1 and 3Dpol regions were generated using RAxML with the GTRGAMMA model (general time reversible + gamma distribution for rates over sites, with all model parameters estimated by RAxML) and 100 bootstraps (59). Regression analysis and investigation of geographical and temporal aspects of recombination were performed using maximum composite likelihood distances (MCL, calculated by MEGA) between sequences.

A Bayesian Markov chain Monte Carlo (MCMC) method implemented in the BEAST package version 1.53 (18) was used to estimate temporal phylogenies and rates of evolution (19). Individual datasets prepared for BEAST analysis (corresponding to genogroups, sub-genogroups or 3Dpol clades used to define RFs) were checked for recombination using the programs Genetic Algorithm Recombination Detection (GARD) or the Single Breakpoint (SBP) method for larger VP1 datasets (GgC, RF-W) in the Datamonkey package that provides an interface to the HyPhy program (27, 47). Further testing was performed on each dataset using several algorithms implemented in the RDP package (RDP, GENCONV, MaxChi, Chimaera, SiScan and 3Seq; (34)).

BEAST analysis used Constant Population size as a prior, while for selected datasets, other tree priors were used (Exponential Growth, Bayesian Skyline) to determine the effect priors on analysis outcomes. Two independent runs for each set were analysed using the SRD06 model of substitution (53), with chain lengths of 50 million or 100 million and a relaxed molecular clock model that allows evolutionary rates to vary among lineages. All other parameters were optimised during the burn-in period. Output from BEAST was analysed using the program TRACER (http://beast.bio.ed.ac.uk/Tracer) and the results of each duplicate were compared.
succession dynamics were examined using a maximum clade credibility (MCC) tree visualised in FigTree following annotation in Tree-Annotator.

Accession numbers. All newly generated sequences obtained in this study were submitted to GenBank, and were assigned the accession numbers HQ676156-HQ676288 (VP1) and HQ676289-HQ676487 (3Dpol).

RESULTS

Phylogeny of EV71 VP1 and 3Dpol genome regions. A total of 198 EV71 isolates from 11 countries and 37 CVA16 isolates (Table 1) were analysed in the current study concurrently with the prototype sequence BrCr-CA-70 (U22521, designated US01/CA/Q/70 in this study) and 109 previously published full genome sequences of EV71 and 8 of CVA16 (Supplementary Table S1). EV71 and CVA16 sequences were amplified and sequenced in the VP1 region between positions 2458 and 3345 (numbering based on BrCr-CA-70). All EV71 isolates analysed in the current study clustered with other EV71 sequences to form a monophyletic group separate from other EV-A serotypes/types (data not shown). The isolates assembled into the previously designated genogroups B2-B5, C1, C2, C4 and C5; those from Europe were exclusively GgC1, GgC2, GgC3 and GgC5 (Supplementary Data; Fig. S1).

For recombination analysis, sequences from the 3’ end of the genome (within the 3Dpol coding region) were determined for each of the 198 isolates analysed in VP1, and the dataset combined with available complete genome sequences of EV71 and CVA16 (Supplementary Data; Fig. S1). Sequences in the 3Dpol region formed a
series of bootstrap supported clades; EV71 3Dpol clades comprised groups A, D, E, G, H, L, Q, T, V, W and Y, while clades C, I, J, M, O and S were identified among CVA16 sequences. Individual genogroups were associated with specific 3Dpol clades in the majority of cases.

Supporting these phylogenetic assignments, the pairwise distributions of sequence distances between variants in both VP1 and 3Dpol fell into a series of discontinuous ranges (Supplementary data, Fig. S2A). For example, the minimum value in the VP1 distribution separating the second and third distributions of approximately 19% corresponds to the threshold value separating intra- from inter-genogroup distances between GgA, GbB and GgC (6) (Supplementary data; Fig. S2A). The lower threshold value (10-11%) corresponds to the previously designated sub-genogroup boundaries (6). An analogous division of 3Dpol sequences into the 15 clades by phylogenetic analysis was similarly supported by their distribution of pairwise distances (Supplementary data; Fig. S2B). The low point in the distribution of 0.13 corresponded closely to the threshold value dividing distances between and within the phylogenetically-defined clades.

3Dpol classification provides the means to identify and quantify recombination events within datasets through identification of phylogenetic incongruences between phylogenetically supported clades in different genome regions, as used in previous analyses of EV- B serotypes (37, 38). 3Dpol gene sequences of the 15 recombinant forms (RFs) of EV71 and CAV16 indeed imperfectly mapped onto the phylogeny of the VP1 region, and revealed several phylogeny violations indicative of recombination (Fig. 1). By branch rotation to maximise tree branching orders and by ignoring any phylogenetic grouping without 70% or more bootstrap support, trees from the two regions revealed both branching...
order differences within EV71 (red dotted lines) as well as highly interspersed
groupings of CVA16 sequences in the 3DPol tree (red sequence labels) that contrasted
with the consistent outgroup position of CVA16 in the VP1 region (Fig. 1).

As specific examples, GgB3 grouped with B4 and B5 in VP1, but adopted a
distinct tree position in 3DPol closest to C4 and CVA16 variants. The branching order
of GgA was similarly incongruent. Within genogroups, while the majority of
sequences were associated with a specific 3DPol group, such as C2 with W, and C4
with L, individual variants within each showed phylogenetically distinct 3DPol
sequences similarly indicative of recombination. Each of the CVA16 groups identified
as separate RFs originated through at least 6 further recombination events in their
evolutionary history. This analysis depicts the minimum number of recombination
events required to resolve the phylogenies of the available sequences. The
interspersed nature of other EV-A 3DPol sequences in the phylogenetic tree (data not
shown) entails a much large number of further recombination events over the longer
period of EV species A enterovirus evolution.

Sequence diversification in VP1 and 3DPol regions. EV71 VP1 variability was
predominantly restricted to synonymous sites indicative of predominantly neutral
sequence change. Using the data available on the collection dates of samples in a
Bayesian MCMC method (BEAST; (19)) with a relaxed molecular clock for
substitution rate calculations, substitution rates and times of most recent common
ancestor (TMRCAs) of different genogroups and 3DPol clades defining the largest RFs
were calculated. Each dataset was analysed for internal recombination prior to BEAST
analysis. GARD/SBP detected no recombination events in any dataset, while detection
of recombinant sequences using RDP-implemented methods was infrequent and
inconsistent between methods (Supplementary Data, Table S4). To investigate whether these possibly recombinant sequences influenced the validity of the MCMC analysis, one of the sequences in VP1 identified as recombinant by three methods (TW13) was excluded from the GgC dataset and CN01 was removed from the clade L 3Dpol sequence dataset. Substitution rates and TMRCAs were compared with the full datasets (Supplementary Data; Table S5). Effectively identical rates and dates were calculated for each comparison, indicating that intra-clade variation had no effect on the validity of the BEAST analysis.

The crude evolutionary rate of evolution for the whole dataset of VP1 sequences was \(7.2 \times 10^{-3}\) s/s/y (high-probability distribution [HPD] range, \(6.2 \times 10^{-3}\) to \(8.3 \times 10^{-3}\); Table 2), with a predicted TMRCA of all extant EV71 lineages of 80 years (HPD, 58 to 106 years). Analysis of individual genogroups within EV71, GgB and GgC revealed comparable evolutionary rates of \(6.2 \times 10^{-3}\) s/s/y (HPD range, \(5.2 \times 10^{-3}\) to \(7.3 \times 10^{-3}\)) for GgB and \(5.0 \times 10^{-3}\) s/s/y (HPD range, \(4.3 \times 10^{-3}\) to \(5.8 \times 10^{-3}\)) for GgC, and TMRCAs of 36 years (HPD, 35 to 38 years) and 34 years (HPD, 25 to 37 years) respectively. Similar substitution rates and more recent MRCA (8 – 32 years) were observed among subgenogroups (Table 2).

For the three of the larger RF groups (RF-A, RF-L and RF-W), remarkably similar substitution rates for individual clad es were determined in the 3Dpol region (Table 2). For example, the substitution rate for RF-A was \(6.4 \times 10^{-3}\) s/s/y (HPD range, \(4.8 \times 10^{-3}\) to \(8.1 \times 10^{-3}\)) in the VP1 region and \(6.7 \times 10^{-3}\) s/s/y (HPD range, \(5.3 \times 10^{-3}\) to \(8.3 \times 10^{-3}\)) in the 3Dpol region. Furthermore, both genome regions analysed provided consistent estimates for the TMRCAs of each individual RF group analysed; using again the example of RF-A, TMRCAs of 13.3 years (HPD, 11.7 to 15.1) and 13.9 years (HPD range, 12.3 to 15.6) for VP1 and 3Dpol respectively were observed.
These provide robust estimates for the dates of the recombination events that created each RF.

To determine whether the prior on the tree in the BEAST analysis influenced estimates of substitution rate or TMRCA, analyses were repeated using exponential growth and Bayesian skyline for the larger datasets (GgB and GgC in VP1, RF-L and RF-W in 3Dpol) (Supplementary Table S3). Each prior produced extremely similar values for substitution rate and TMRCA in each dataset. Sequence relationships between dated samples collected over a wide temporal range primarily informed these calculations rather than tree prior.

The similar dynamics of sequence drift between VP1 and 3Dpol regions and the disruptive effect of recombination is apparent through comparison of pairwise distances in the two genome regions of the better represented sub-genogroups (Fig. 2; examples of B4 and C1 shown). Discontinuities in the distributions of pairwise distances support the hypothesis of widespread recombination during EV71 evolution. Within-GgB4 and GgC1 comparisons show an approximate straight line correlation between VP1 and 3Dpol divergence (dark blue points), with a positive gradient of approximately one and a y-intercept approximating zero. This is consistent with the similar substitution rates observed between the two regions by Bayesian MCMC analysis (Table 2). Both show evidence for continuous and discontinuous sequence distributions when compared with other subgenogroups. Distributions of pairwise distances between B4 and B5 and between C1 and C2 overlapped, and their similar trajectory to within subgenogroup (B4 and C1) comparisons demonstrated a process of ongoing divergence without recombination. B4 and B5 indeed retained the same RF group, A, as did C1 and C2 (with RF-W). Comparison of B4 and C1 sequences to other subgenogroups revealed quite distinct relationships. For example, C2 and C4
were similarly divergent from C1 in VP1 (comparing the distributions of yellow and light blue points), but 3Dpol sequences from C4 were substantially divergent (pairwise distance of around 0.4), an observation that can only be accounted for by a recombination event, in this case to RF group L.

Equivalent recombination events are implied by the other inter-subgenogroup comparisons comprising B4 (A) to B1 (E), B2 (D) and B3 (G) and C1 (W) to C4 (L), C5 (T) and likely C3 (V) (RF designations in parentheses). Comparison of C1 to C2 and C4 further revealed the separate placement of RF-Y and RF-H, away from the main groups confirming the occurrence of further sporadic recombination events.

Indeed, pairwise comparison of each subgenogroup to each within gB and gC (30 comparisons; data not shown) confirmed that only B4/B5 and C1/C2 were consistently equidistant in the two genome regions; these were additionally the only two pairs of subgenogroups that shared the same 3Dpol group (A and W respectively).

**Temporal correlates of recombination in GgB and GgC.** To estimate the relationship between virus diversification and the occurrence of recombination in EV71, sequences of each isolate within EV species B and C were compared in the VP1 to estimate evolutionary divergence, and in the 3Dpol region to identify shared or different 3Dpol groupings (Fig. 3). Pairs of isolates with different 3Dpol groups were considered to have undergone recombination. The proportion of isolate comparisons with different 3Dpol groups increased with increasing VP1 divergence in both genogroups B and C (Fig. 3). As previously performed for EV-B serotypes (38), interpolation of the 50% value, combined with the substitution rate for VP1 (Table 2; 7.2 x 10^{-3} s/s/y) enabled approximate half-lives of genogroups B and C EV71 strains...
to be estimated. Values of 0.085 and 0.135 predicted half lives of 5.9 and 9.4 years for
Ggs B and C respectively. This difference is consistent with the observation of lower
divergence between recombinant sequences in GgB in the analysis of VP1 and 3Dpol
sequences (Fig. 3), and the prolonged existence of subgenogroups C1 and C2 which
retained a long term association with the 3Dpol clade, W, despite their substantial
sequence divergence in VP1 (pairwise distances ranging from approximately 0.10 to
0.17; Fig. 2).

The occurrence of recombination was mapped onto MCMC-generated time-
correlated trees (Fig. 4) to estimate when individual recombination events occurred.
GgB and GgC displayed different temporal dynamics and patterns of RF succession.
For GgB (Fig. 4A) a step-wise, time-related correlation of recombination with VP1
divergence was observed, encompassing isolates collected over a 35 year period. As
identified previously, each VP1 subgenogroup corresponded to a single RF group
with the exception of GgB4 and GgB5 (both RF-A). The consecutive replacement of
B1 by B2, B2 by B3 and B3 by B4 each involved viruses bearing distinct 3Dpol
region sequences (RF-E to RF-D to RF-G to RF-A). The date of recombination events
could be estimated as lying between the MRCA shared by the original RF and the new
RF, and the isolation date of the first clinical sample with the new RF. Thus it can be
estimated that recombination of the most recent RF group, RF-A, occurred between
1992 and 1997. The subsequent emergence of the associated sub-genogroups, GgB4
and GgB5 was characterised by large-scale, short-term outbreaks occurring with a
periodicity of approximately 3 years. However unlike earlier outbreaks (1973 to 1998)
involving RF groups E, D and G, the 3-yearly outbreaks were not immediately
preceded by a recombination event, even when VP1 divergence was extensive enough
to designate a separate subgenogroup (GgB4 to GgB5).
GgC comprised three major lineages which diverged from a common ancestor in 1983 (HPD, 1978 to 1986; Fig. 4B). Two of the VP1 lineages in GgC contained one or two sporadic monophyletic RF groups interspersed within one major RF group. It is likely that these sporadic RF groups originated from a single (datable) recombination event and each were minor and short-lived components of the circulating virus population. The lineage containing GgC1 (RF-W) persisted for at least 17 years (1990 to 2007) and GgC5 (RF-T) emerged as a sporadic RF group in 2006, having undergone a recombination event sometime between 1994 and 2006. The first clinical sample of the second lineage (mainly GgC2, RF-W) was collected in 1997 and this lineage has persisted for 13 years, up to the end of the study period in 2010. This lineage also contained two GgC3 samples belonging to RF-V which were isolated in 2000 and likely recombined between 1989 and 2000. A further sporadic RF group (Y) which appeared in 2010 probably recombined with its unknown second parental strain between 1996 and 2010, exhibiting a long quiescent period before emergence in 2010. The third lineage was comprised entirely of GgC4 sequences which persisted for 11 years (between 1998 and 2009), displayed time-correlated divergence and consisted predominantly of RF-L with the exception of two isolates of the sporadic group RF-H (isolated in 2008); this group probably recombined between 2005 and 2008, emerging in 2008 after a short quiescent period. RF-H did not appear to replace the parental RF-L group up to the end of the study in 2010.
DISCUSSION

Detection of recombination in EV71. This study applied methods developed in previous genetic analyses of EV-B enteroviruses to re-examine the occurrence and dynamics of recombination in EV71. EV71 is considered to be the most pathogenic of the currently circulating enteroviruses worldwide, infections being associated with outbreaks of HFMD and serious neurological disease in South East Asia. Understanding the underlying interactions between virus evolution and population susceptibility and the biological basis for its severe disease associations are major research priorities, as are the development of preventative or treatment strategies for its control.

In the current study, detection of recombination events was achieved through identification of bootstrap-supported clades by phylogenetic analysis of the 3Dpol region. These groupings were used to categorise EV71 variants into a series of RFs whose assignments were supported by parallel analyses of pairwise distances (Fig. S2); variants within the same RF group showed pairwise distances in the 3Dpol region of <0.19, and >0.19 between those in different groups. By these criteria, the EV71 3Dpol sequences in this study assembled into 11 clades and five of these clades comprised over 96% of the total non-structural sequences analysed. Therefore, the repertoire of 3D-pol RFs was much more restricted in EV71 than found in species B EVs (37, 38), where 119 E11 variants collected over a 14 year period could be assigned to 43 RF groups, 89 E9 samples into 23 and 240 European E30 samples into 26 RF groups. The underlying reasons for this difference in diversity between viruses in EV species A and B are unclear.
Using this classification, comparison of phylogenetic trees from VP1 and 3Dpol regions showed a large number of phylogenetic incongruities indicative of recombination (Fig. 1). Most striking was the interspersed position of 3Dpol sequences from CVA16 variants included in the analysis. As these by definition take an outlier position in the VP1 region, each occurrence therefore represents a recombination event in the evolutionary history of either EV71 or CVA16. Also striking was the difference of tree position of GgB3 and GgC4 in VP1 and in 3Dpol; both grouped away from other members of the same genogroup (10, 17, 24, 66). However, the assertion that these variants show evidence for intertypic recombination, such as GgC4 with CVA16 is not supported by the analysis in the current study. GgC4 variants are, with one exception, assigned to RF-H, a 3Dpol grouping that is distinct from its closest neighbour, RF-L, assigned to CVA16. Although occupying neighbouring positions in the tree, their assignment to different RF groups therefore provides evidence against the frequently proposed specific recombination event between GgC4 and CVA16. Subgenogroup B3 show an analogous change in tree position as previously described (10), although again without evidence for intertypic recombination since its 3Dol group (G) is also not shared with any other EV71 or CVA16 variants (Fig. 1) or other EV-A serotype (data not shown).

This tree comparison also revealed the existence of two sporadic recombinants, variants of EV71 with 3Dpol assignments different from the rest of the subgenogroup within which they are classified. These comprise the C2 variant characterised in the current study, JP17/Ac/Y/10, assigned as RF-Y instead of the majority RF-W (Fig. 1), and the C4 variants CN23/Sz/H/08 and CN19/Bj/H/08 assigned as RF-H instead of RF-L (Fig. 1). The other evident phylogeny violation is
the outlier position of the single GgA sequence in VP1 and its inlier position in 3Dpol.

Indeed, rather than representing defined recombination events between serotypes, or indeed recombination between lineages within EV71, the discordant sequence relationships between genome regions in EV71 closely resemble a quite different process of largely independent evolution of genome regions that has been previously observed in species B enteroviruses (32, 37, 38). For echoviruses 9, 11 and 30, evolutionary lineages identified within the capsid region show evidence for recombination with a larger pool of non-structural region variants that show no systematically closer genetic relationship with one serotype than with any other. In the case of E9, E11 and E30, the 92 RF-groups were thus fully interspersed with each other and with 3Dpol groups of other EV-B serotypes, with only very rare occurrences of shared RF groups between different serotypes or indeed lineages within a serotype (37, 38). The observed scattered positions of CVA16 and EV71 3Dpol-associated clades and the lack of shared RF assignments of different subgenogroups observed in the current study is indeed a precise mirror of the pattern observed in EV-B.

Distributions of pairwise distances in the two genome regions (Fig. 2) provide evidence for further recombination events in the evolution of EV71. This analysis method demonstrates abrupt discontinuities in distributions of pairwise distances between the two genome regions indicative of recombination events accompanying the founding of most subgenogroups within GgB and GgC. This conclusion is consistent with independent evolutionary pathways of structural and non-structural genome regions identified in EV-B enterovirus (32). Indeed, only pairwise comparisons between members of the same subgenogroup, or between B1/B2 and
C4/C5 show equality in divergence (i.e. the gradient of the set of pairwise distances approximates to unity) that is entailed by their similar substitution rates (Table 2). All other comparisons between subgenogroups revealed disproportionately high and non-linear greater divergence in the 3Dpol region indicative of recombination. These plots of pairwise distances help visualise the differing degree of overall divergence in different genome regions described previously (64) and which are so evident on divergence scans and bootscanning (10, 17, 24, 66).

**Timescale of recombination events.** By measuring the relationship between VP1 divergence and occurrence of recombination, as previously carried out for EV-B serotypes (38), it was possible to calculate approximate half-lives for individual RFs of EV71 (Fig. 3). Estimates of 5.9 and 9.4 years for Gg B and GgC overlapped the range observed previously in the EV-B serotypes. E9, E30 and E11 (half-lives of 1.3, 3.1 and 9.8 years respectively (38)). However, the major EV71 RFs (A, E, G, L and W) showed decades-long circulation without recombination, a pattern also observed for the E11 RF-DU that continued to circulate throughout a study period from 1996-2008 (38). Variability in recombination frequency is likely shaped by differing viral epidemiologies that govern opportunities for coinfection and generation of hybrid viruses to occur. They may also be influenced by differing compatibility restrictions that would dictate the likelihood of replication competent viruses being generated by recombination. The marked differences in recombination frequency observed between human parechovirus (HPeV) type 1 (4 years) and HPeV type 3 (20 years) (7) may indeed be a manifestation of constraints limiting compatibility between viruses with differing cell entry mechanisms and potentially cellular tropisms (22).
More precise identification of the individual recombination events in EV71 473 was achieved through the use of time-correlated trees and superimposition of 474 branching points in the VP1 tree that most parsimoniously accounts for the RF 475 designations in descendant sequences (Fig. 4). These phylogenetic reconstructions for 476 genogroups B and C additionally place recombination events into the differing 477 evolutionary trajectories of GgB and GgC that have been characterised previously 478 (61, 63). GgB is characterised by a series of successive emergence and extinctions of 479 the B1 through B5 subgenogroups over the period from 1970 to the present day, with 480 outbreaks occurring throughout South and East Asia in an approximate 3 yearly cycle 481 (1993, 1997, 2000, 2003, 2006 and 2008) (9, 23, 46, 58, 62). The association between 482 the founding of each lineage and recombination is clearly evident from the 483 phylogenetic analysis. This provides the means to estimate dates within fairly narrow 484 windows for their occurrence. These tree-based estimates are consistent with the 485 TMRCA estimates from diversity / substitution rate calculations in both VP1 and 486 3Dpol regions (Table 2) (61).

In contrast, it has been previously established that different subgenogroups of 489 GgC EV71 variants have co-existed for at least 25 years, with three separate lineages 490 (C1, C2 and C5) emerging in the 1980’s and continuing to circulate to the present (5, 491 58, 61). This contrasting pattern is exemplified in a previous characterisation of 492 isolates of EV71 in Malaysia; GgB variants were isolated only during periodic major 493 outbreaks whereas GgC isolates were detected sporadically, both during and between 494 epidemics (46). Whether or by what mechanism these contrasting epidemiologies 495 contribute to the differing recombination frequencies of GgB and GgC (Fig. 3) 496 remains to be determined.
Evolution of EV71. It is well established that the longer term genetic and antigenic diversification of human enteroviruses and the existence of serologically distinct EV types are key factors in their evolutionary success and ongoing ubiquitous presence in human populations. What is rather less clear is the underlying mechanisms and selection pressures involved in the generation of new serotypes, as this process has to date not been directly observed. In the case of EV71, it has been hypothesised that the evolution of capsid genes, particularly VP1, represents an equivalent, immunologically driven process of diversification (61, 63), and that the successive appearance of novel genogroups and subgenogroups is favoured by an absence of pre-existing immunity to them. The genetically diverse variants of EV71 may thus be precursors in their eventual further, future diversification into new serotypes. This hypothesis is, however, not clearly supported by existing genetic and antigenic comparisons of EV71 genogroups (14, 24-26, 40, 61, 63). Firstly, sequence divergence between genogroups and subgenogroups occurs overwhelmingly at silent sites, with extremely low dN/dS ratios indicative of purifying or neutral evolutionary drift (Table 2; (14, 24, 61)). Screening of large datasets of VP1 sequences provides little if any evidence for any sites in the capsid coding region being subjected to positive selection that would typically be observed in sites under strong immunological pressure. Consistent with these analyses, there is little evidence from serological cross-neutralisation experiments for genogroup- or subgenogroup-specific antibodies, nor indeed the existence of measurable antigenic diversity between EV71 isolates (24, 26, 28, 40, 63). However, infections with GgB variants induce higher levels of neutralising antibodies than GgC, supporting an idea that the emergence of GgC may have been assisted by its intrinsically lower immunogenicity than GgB (40).
The alternative hypothesis is that the emergence and turnover of EV71 genogroups simply represents random processes of emergence and extinction of lineages without underlying natural selection, as proposed for other enteroviruses (2, 37, 50, 50). Random fixation of individual variants with no fitness advantage within a population is indeed more likely when population sizes are small. Such conditions may occur during the periodic bottlenecks in population size that inevitably occur in viruses that cause acute infections and show epidemic cycles of transmission. These instances of very rapid turnover and complete population replacements are extensively documented with a similar lack of evidence for antigenic replacement. However, as observed in EV71, such turnover is frequently associated with recombination events occurring during the founding of new evolutionary lineages. Understanding whether the partial or complete replacement of non-structural gene regions associated with such recombination events provides a replicative or immunological selective advantage to the virus and thus drives the diversification of EV71 and other enteroviruses is a key unanswered question. Future biological and immunological investigation of both capsid region sequence change and recombination are clearly required if we are really to understand the evolution of enteroviruses and indeed other non-enveloped RNA viruses.
ACKNOWLEDGEMENTS

We would like to thank staff at the following universities and hospitals for technical assistance with virus isolation: Department of Virology, University of Turku, Finland; Intestinal Viruses Unit, National Institute for Health and Welfare (THL), Helsinki, Finland; Gurutze Rubio, Cruces Hospital, Bilbao, Spain; Manuel Omeñaca, Miguel Servet Hospital, Zaragoza, Spain; Nuria Rabella, Santa Cruz y San Pablo Hospital, Barcelona, Spain, Carmen Perez, Dr. Negrin Hospital, Las Palmas de Gran Canaria, Spain, and T.P. Eremeeva, M.P. Chumakov Institute of Poliomyelitis and Viral Encephalitides, Moscow, Russia (the latter supported in part by the Polio Eradication Initiative through the European Office of World Health Organization). Seiya Yamayoshi (Neurovirology Project, Tokyo Metropolitan Institute of Medical Science, 2-1-6, Kamikitazawa, Setagaya-ku, Tokyo 156-8506, Japan). We would additionally like to thank Setsuko Iizuka (Shimane prefectural Institute of Public Health and Environmental Science, 582-1, Nishihamasadacho, Matsue, Shimane 690-0122, Japan), Teruo Yamashita and Hiroko Minagawa (Laboratory of Virology, Department of Microbiology and Medical Zoology, Aichi Prefectural Institute of Public Health, 7-6, Nagare, Tsujimachi, Kita-ku, Nagoya, Aichi 462-8576, Japan), Katsumi Mizuta (Department of Microbiology, Yamagata Prefectural Institute of Public Health, 1-6-6, Tokamachi, Yamagata 990-0031, Japan) and Hidekazu Nishimura (Virus Research Center, Sendai Medical Center, 2-8-8, Miyagino, Miyagino-ku, Sendai, Miyagi 983-8520, Japan) for providing samples.

The authors are very grateful to Dr Sam Lycett and Dr Andrew Rambaut (Institute of Evolutionary Biology, University of Edinburgh) for valuable discussion.
and assistance with phylogenetic and BEAST analysis. This study was funded by a project grant from the Wellcome Trust.

REFERENCES


of intratypic recombination of enterovirus 71 in Taiwan from 2002 to 2005. 

Virus Res. 131:250-259


phylogenies of the VP1 and 3CD regions indicates frequent recombination within the species. J. Gen. Virol. 84:1223-1235


from fatal and non-fatal cases of hand, foot and mouth disease during an

55. Simmonds, P. 2006. Recombination and selection in the evolution of
80:11124-11140

56. Simmonds, P. and J. Welch. 2006. Frequency and dynamics of
recombination within different species of human enteroviruses. J. Virol.
80:483-493

57. Singh, S., C. L. Poh, and V. T. Chow. 2002. Complete sequence analyses of
enterovirus 71 strains from fatal and non-fatal cases of the hand, foot and

58. Solomon, T., P. Lewthwaite, D. Perera, M. J. Cardoso, P. McMinn, and
M. H. Ooi. 2010. Virology, epidemiology, pathogenesis, and control of
enterovirus 71. Lancet Infectious Diseases 10:778-790

59. Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based
phylogenetic analyses with thousands of taxa and mixed models.
Bioinformatics. 22:2688-2690

60. Stanway, G., F. Brown, P. Christian, T. Hovi, T. Hyypia, A. M. Q. King,
N. J. Knowles, S. M. Lemon, P. D. Minor, M. A. Pallansch, A. C.
M. Fauquet, M. A. Mayo, J. Maniloff, U. Desselberger, and L. A. Ball (eds.),


### TABLE 1

SOURCES AND COLLECTION DATES OF SURVEY SPECIMENS

<table>
<thead>
<tr>
<th>Country</th>
<th>Code</th>
<th>No. of Isolates</th>
<th>Subgenogroup</th>
<th>Year(s) of isolation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EV71</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Croatia</td>
<td>HR</td>
<td>2</td>
<td>C4</td>
<td>2005</td>
</tr>
<tr>
<td>Georgia</td>
<td>GE</td>
<td>1</td>
<td>C2</td>
<td>2007</td>
</tr>
<tr>
<td>Great Britain</td>
<td>GB</td>
<td>5</td>
<td>C2</td>
<td>2010</td>
</tr>
<tr>
<td>Iceland</td>
<td>IS</td>
<td>2</td>
<td>C1</td>
<td>2004</td>
</tr>
<tr>
<td>Latvia</td>
<td>LV</td>
<td>1</td>
<td>C1</td>
<td>2003</td>
</tr>
<tr>
<td>The Netherlands</td>
<td>NE</td>
<td>15</td>
<td>C1, C2</td>
<td>2007, 2010</td>
</tr>
</tbody>
</table>

| **CVA16**     |      |                 |              |                      |
| Great Britain | GB   | 1               | -            | 2009                 |
| Iceland       | IS   | 2               | -            | 2002, 2004           |
| Japan         | JP   | 1               | -            | 2004                 |
| Latvia        | LV   | 1               | -            | 2007                 |
| The Netherlands| NL  | 4               | -            | 2008                 |
| Russia        | RU   | 5               | -            | 2008-2010            |
| Slovak Republic | SK | 2               | -            | 2004-2005            |
| **Total**     |      | 9               | 37           | 2000-2010            |
### TABLE 2

**CALCULATION OF RATES OF SEQUENCE CHANGE AND TMRCAs BY MCMC ANALYSIS**

<table>
<thead>
<tr>
<th>Geog. set</th>
<th>Gg/RF</th>
<th>N</th>
<th>Divergence 1</th>
<th>Regression</th>
<th>MCMC (BEAST) 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>nucleotide</td>
<td>aa</td>
<td>Substitution rate (x10^-3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>VPI 3D</td>
</tr>
<tr>
<td>Whole dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>All</td>
<td>308</td>
<td>0.18</td>
<td>0.02</td>
<td>0.14</td>
</tr>
<tr>
<td>Europe</td>
<td>All</td>
<td>58</td>
<td>0.09</td>
<td>0.01</td>
<td>0.58</td>
</tr>
<tr>
<td>Asian</td>
<td>All</td>
<td>244</td>
<td>0.17</td>
<td>0.02</td>
<td>0.32</td>
</tr>
<tr>
<td>Individual genogroups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>GgB</td>
<td>108</td>
<td>0.07</td>
<td>0.01</td>
<td>0.94</td>
</tr>
<tr>
<td>All</td>
<td>GgC</td>
<td>199</td>
<td>0.13</td>
<td>0.02</td>
<td>0.77</td>
</tr>
<tr>
<td>Asian</td>
<td>GgC</td>
<td>137</td>
<td>0.11</td>
<td>0.01</td>
<td>0.77</td>
</tr>
<tr>
<td>Individual subgenogroups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>GgB4</td>
<td>28</td>
<td>0.03</td>
<td>0.01</td>
<td>0.7</td>
</tr>
<tr>
<td>All</td>
<td>GgB5</td>
<td>60</td>
<td>0.02</td>
<td>0.00</td>
<td>0.88</td>
</tr>
<tr>
<td>All</td>
<td>GgC1</td>
<td>55</td>
<td>0.05</td>
<td>0.01</td>
<td>0.86</td>
</tr>
<tr>
<td>All</td>
<td>GgC2</td>
<td>65</td>
<td>0.05</td>
<td>0.01</td>
<td>0.75</td>
</tr>
<tr>
<td>All</td>
<td>GgC4</td>
<td>74</td>
<td>0.04</td>
<td>0.01</td>
<td>0.75</td>
</tr>
<tr>
<td>Individual RF groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>RF-A</td>
<td>88</td>
<td>0.05</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>All</td>
<td>RF-W</td>
<td>119</td>
<td>0.10</td>
<td>0.09</td>
<td>0.01</td>
</tr>
<tr>
<td>Asia</td>
<td>RF-W</td>
<td>61</td>
<td>0.10</td>
<td>0.09</td>
<td>0.02</td>
</tr>
<tr>
<td>Europe</td>
<td>RF-W</td>
<td>55</td>
<td>0.08</td>
<td>0.08</td>
<td>0.01</td>
</tr>
<tr>
<td>All</td>
<td>RF-L</td>
<td>72</td>
<td>0.04</td>
<td>0.04</td>
<td>0.01</td>
</tr>
</tbody>
</table>

1 Mean pairwise $P$ distances. Values were not calculated in the 3Dpol region for groups where 3Dpol sequences were not monophyletic.
2 The mean value is given from two independent analyses; substitution rates and TMRCAs were not calculated for groups where 3Dpol sequences were not monophyletic.
3 The number of sequences analysed in each set.
4 Frequency of substitutions per site per year.
5 Time before the present of the most recent common ancestor (years).
**FIGURE LEGENDS**

**Fig. 1.** Phylogeny of VP1 and 3Dpol regions of EV71 for study subjects and those previously determined (listed in Tables 1 and S1). Clades were identified by bootstrap analysis (values $\geq 70\%$) by maximum likelihood as implemented by RAxML. The size of the triangles is proportional to the number of sequences within each clade (bootstrap resampling values shown on branches). By branch rotation to maximise visual congruence of the two trees and the use of bootstrap values of $\geq 70\%$ to define phylogenetic groupings, the minimum number of incongruent phylogeny relationships (depicted by red dotted lines) were determined (clades showing congruent branching orders labelled with blue dotted lines). The 3Dpol region includes available sequences from CVA16; these are labelled red as their interspersed positions in 3Dpol is invariably incongruent with their outgroup position in VP1 (not included in the left hand tree).

**Fig. 2.** Relationship between sequence divergence (MCL pairwise distances) in the VP1 region (x-axis) and in 3Dpol (y-axis) among (A) genogroup B and (B) genogroup C EV71 sequences. In the examples shown, sets of pairwise distances in both regions between GgB4 and C1 to other variants within GgB and GgC are depicted. Note that subgenogroups C2 and C4 both contained within them single sporadic recombinants; these account for the additional groupings of data points encircled by the dotted line.

**Fig. 3.** Association between VP1 sequence divergence (maximum value shown for each bar on the x axis) and the proportion of recombinant comparisons (i.e. belonging to different 3Dpol clades) for GgB and GgC.
Fig. 4. MCMC tree of the VP1 sequences of GgB (A) and GgC (B) from the Asia Pacific region visualised in FigTree and plotted on a temporal y-axis scale using their sampling dates. Branches are colour coded (see Key) according to the recombination group of individual sequences and their reconstructed ancestors.
Congruent tree positions

Incongruent positions

(A) VP1 Region
(B) 3Dpol Region