Molecular evolution of types in non-polio enteroviruses

Alexander N. Lukashev1,2,* and Yulia A. Vakulenko1,3

Abstract

Non-polio enteroviruses are a ubiquitous and divergent group of non-enveloped RNA viruses. Novel types are reported regularly in addition to over 100 known types; however, mechanisms of emergence of novel types remain obscure. Here, the 33 most common types represented by 35–629 non-redundant partial VP1 sequences in GenBank were studied in parallel using Bayesian coalescent molecular clock analysis to investigate common evolutionary trends among enterovirus types. Inferred substitution rates were in the range of $0.41 \times 10^{-2}$ to $3.07 \times 10^{-2}$ substitutions per site per year. The most recent common ancestors of known isolates of each type presumably existed between 55 and 200 years ago. Phylogenetic analysis results suggested that global type populations underwent bottlenecks that could repeatedly reset the common ancestor dates. Nevertheless, species-level analysis suggested that the contemporary enterovirus types emerged within the last millennium. Analysis of 2657 complete VP1 sequences of the 24 most common types indicated that the type criterion based upon 75% nucleotide sequence identity remains generally valid, despite exponential growth of the number of known sequences and a high rate of mutation fixation. However, in a few types there was evidence that enteroviruses can drift slightly beyond the type threshold, up to 73% identity, and both amino acid and nucleotide sequences should be considered for type identification. Analysis of sequence distances within types implied that sequence-identity-based identification of genotypes is rational within some, but not all, types and distinct genotype cut-offs (9–20%) may be useful for different types.

INTRODUCTION

Enteroviruses (EVs) are small, non-enveloped RNA viruses with a non-segmented, positive-sense genome. EVs are ubiquitous and usually circulate in the population causing mild disease or no overt symptoms. Occasionally, EVs can cause poliomyelitis, meningitis, severe infection of infants, myocarditis and a wide range of less frequent syndromes [1]. In addition, in the twentieth century, EVs were a source of several pandemics of emerging infections, such as poliomyelitis caused by poliovirus types 1–3 [2], haemorrhagic conjunctivitis caused by EV 70 [3, 4], meningoencephalitis caused by EV 71 [5], and enterovirus uveitis caused by echoviruses 11 and 19 [6].

The genus Enterovirus includes rhinoviruses (three species) and ten species of EVs termed A–J. EVs A–D, formerly termed human EVs, infect humans, but are also commonly found in primates [7, 8]. Below species level, they are classified into types (formerly termed serotypes) that can be identified by either serological or molecular methods. Each type has its prototype strain, the first strain identified as a novel type. Most of the prototype strains were isolated in 1947–1955. Initially, subsequent isolates were assigned to a serotype based on an infectivity neutralization test. Currently, the EV type criterion is the nucleotide sequence distance in the VP1 genome region, which can reliably distinguish known types from each other [9]. Viruses that differ from known types by more than 25% nucleotide sequence in the full VP1 genome region are identified as novel types [9]. Amino acid sequence identity of above 85% has been suggested as an additional type criterion [9]. Identification of clinical samples is usually done by partial sequencing of the VP1 genome region (approx. 300 nt fragment, from here on termed 'the typing fragment'), which can be amplified using universal PCR primers [10]. The number of EV types is currently 317 (150 excluding rhinoviruses) and is constantly growing [11]. It is not clear if novel types emerge de novo, or are transferred from primates, which host many types of EV A–D that were not previously found in humans [12, 13]. Investigation of the emergence mechanisms of types is critical for understanding the phenomenon of emerging EV infections.
One of the key evolutionary mechanisms in EVs is recombination. Naturally circulating EVs recombine every few years within a species [14, 15], and the capsid-encoding genome region evolves virtually independently from the genome region that encodes non-structural proteins [16]. Capsid-encoding genes of distinct types recombine with each other very seldom [17, 18] and usually behave as distinct populations, while the non-structural genes recombine freely within a species that comprises many types, and the whole species is a population in the non-structural genome region [19]. Therefore, evolution of the capsid-encoding genes per se could be studied without regard to the non-structural genome regions.

Here, we applied Bayesian coalescent phylogenetic analysis for comprehensive analysis of the 33 most prevalent EV types representing species A–D that are predominantly found in humans. The three types of poliovirus belong to the species Enterovirus C; however, it is very hard to distinguish naturally circulating EVs from early vaccine-derived polioviruses that originated from the live polio vaccine and could compromise molecular dating. Therefore, poliovirus was excluded from this study.

RESULTS
Molecular evolution of the 33 most common non-polio EV types was analysed. Datasets for the typing VP1 sequence (approx. 300 nt) included 2–15 times more sequences than the full VP1 datasets (Table 1). Geographical coverage of short alignments was also better. Therefore, the short alignments were used for phylogenetic analysis. EV species are genetically complex entities that contain groups with supposedly distinct evolutionary patterns and do not necessarily constitute a single population. Therefore, individual types were firstly analysed separately. The datasets for each type varied by the total number, temporal and geographical distribution of sequences. The most ancient strains of most types were isolated in 1947–1955, so the time between the most ancient and the most recent virus, or sample depth, was on average 64 years. Distribution of isolation dates was not uniform: there was a time gap between sampling dates of the prototype strains (usually 1947–1955) and other isolates (from here on termed ‘contemporary’) that were isolated later, mostly after 1970 (Fig. 1). Moreover, most samples were isolated in regions with a temperate climate. Thus, we examined to which extent sample bias could affect further analysis. Influence of several factors was evaluated.

The key parameters inferred upon Bayesian coalescent phylogenetic analysis are the substitution rate and the root height, or time of the most recent common ancestor (tMRCA) for each type. Both sample depth and inhomogeneity could affect these parameters. The tMRCA of individual EV types varied between 55 and 200 years ago (Tables 1 and S1, available in the online version of this article). Substitution rates within types were in the range of 0.41×10⁻² to 1.22×10⁻² substitutions per site per year in all types except for CV-A13, which had a substitution rate of 3.07×10⁻² substitutions per site per year. One explanation for an unusually high substitution rate in CVA13 could be the fact that it was the most prevalent type in Sub-Saharan Africa [20], and could have a circulation profile different from those of other types. The total number of sequences in a dataset affected tree height (Fig. 2a, Spearman r=0.48, P value <0.05), but not the substitution rate (Fig. 2b, Spearman r=−0.23, P value=0.09). The effect of sample size on root height was apparently absent among 11 of 33 types that were represented by more than 100 sequences (Fig. 2a). Among these types, there was no significant correlation between sample size and root height (Spearman r=0.33, P value=0.24). Therefore, sample size of over 100 non-redundant sequences could be considered sufficient for the analysis, while the root height of smaller datasets could be moderately biased.

Some types were historically more uniformly sampled over time than others. For these types, the time between isolation of the prototype strain and the next most ancient strain was short. On the contrary, for several types there was a long time gap between dates of the prototype and the next most ancient sequence. For example, the prototype isolate of CV-A2 dated from 1947, and the next known VP1 sequence dated from 2000. The time intervals between isolation dates of the prototype and the oldest non-prototype sequences were in the range of −3 (when an archive isolate not officially recognized as a prototype was sequenced recently) to 53 years. Time between the prototype and the next most ancient isolate (which vaguely represents temporal non-uniformity of the sample) weakly affected root height (Fig. 2c, Spearman r=−0.34, P value=0.048), but not the rate estimates.

Prototype strains of most types were isolated at about the same time, between 1948 and 1955. Therefore, datasets that included prototype strains were poorly suited for seeking a relationship between sample depth and root height. If there were no prototype strains available, there would have been a wide and more even distribution of sample depths among different types. Removal of prototype strains from datasets led to a decrease in sample depth by 0–53 years (mean=30 years). The inferred substitution rate did not change significantly upon removal of prototype strains from datasets. Removal of prototype isolates also reduced the tMRCA by 1–163 years (mean 24.6 years). Therefore, in the absence of a prototype strain, the inferred MRCAs of distinct EV types appeared to be about 25 years ‘younger’. There was a moderate negative correlation between the sample depth and the root age (Fig. 2d, Pearson r=−0.54, P value <0.01) in these datasets lacking the prototype strains. Types with lower sample depth (more recent oldest isolate) appeared significantly younger.

Analysis of sample bias effects on EV evolutionary estimates indicated that datasets with greater than 100 sequences are generally sufficient for analysis and that the age of the most ancient sequence (which does not vary a lot among types) was critical for root height estimates. In other words, types
generally appeared on average \( n \) years younger than their oldest isolates. The strong effect of age of the most ancient sequence on root height, but not on the substitution rate, could be explained by bottlenecks and lineage extinction that have reset the apparent tree root age (the founder effect). If such bottlenecks occurred, they should also have affected the topology of the phylogenetic trees. All isolates after a hypothetical bottleneck would group together.

Indeed, on the phylogenetic trees the prototype strains occupied the basal position in 16 of 33 types (CV-A2, CV-A5, CV-A10, CV-A13, CV-A16, E-1, E-2, E-7, E-14, E-18, E-19, E-20, E-25, E-29, E-33 and EV-D68), and in all these trees the sub-group of all contemporary strains was supported with a posterior probability of 1 (Fig. 3). Therefore, a single common ancestor (or few ancestors) of all contemporary isolates within most types likely existed already after isolation of the prototype strain in the 1950s. Such tree topology is compatible with the hypothesis of global-scale bottlenecks and extinction of most of the virus variants that hypothetically existed in the 1950s. This was best exemplified by CV-A2, CV-A5, CV-A10, CV-A13, CV-A16, E-19 and E-33, in which the MRCA of the contemporary viruses existed just 24–54 years ago. This topology was not uniform, and several grouping patterns of contemporary isolates could be identified among EV types.

---

**Table 1.** Enterovirus sequences used for the study, and key evolutionary estimates

<table>
<thead>
<tr>
<th>Type</th>
<th>Total number of sequences in Genbank*</th>
<th>Number of sequences in VP1 ‘typing’ fragment alignment†</th>
<th>Number of sequences in full VP1 alignment†</th>
<th>Root height (years)</th>
<th>Substitution rate ( \times 10^{-2} ) (substitutions per site per year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV-A2</td>
<td>302</td>
<td>54</td>
<td>18</td>
<td>77</td>
<td>1.12</td>
</tr>
<tr>
<td>CV-A4</td>
<td>475</td>
<td>101</td>
<td>51</td>
<td>71</td>
<td>0.71</td>
</tr>
<tr>
<td>CV-A5</td>
<td>165</td>
<td>38</td>
<td>11</td>
<td>77</td>
<td>0.83</td>
</tr>
<tr>
<td>CV-A6</td>
<td>4156</td>
<td>180</td>
<td>89</td>
<td>116</td>
<td>0.9</td>
</tr>
<tr>
<td>CV-A9</td>
<td>607</td>
<td>100</td>
<td>52</td>
<td>102</td>
<td>0.61</td>
</tr>
<tr>
<td>CV-A10</td>
<td>1809</td>
<td>126</td>
<td>71</td>
<td>116</td>
<td>0.83</td>
</tr>
<tr>
<td>CV-A13</td>
<td>199</td>
<td>70</td>
<td>30</td>
<td>65</td>
<td>3.07</td>
</tr>
<tr>
<td>CV-A16</td>
<td>3818</td>
<td>237</td>
<td>264</td>
<td>200</td>
<td>0.82</td>
</tr>
<tr>
<td>CV-A20</td>
<td>118</td>
<td>54</td>
<td>34</td>
<td>179</td>
<td>0.41</td>
</tr>
<tr>
<td>CV-A24</td>
<td>1527</td>
<td>86</td>
<td>46</td>
<td>127</td>
<td>0.42</td>
</tr>
<tr>
<td>CV-B1</td>
<td>558</td>
<td>92</td>
<td>43</td>
<td>84</td>
<td>0.91</td>
</tr>
<tr>
<td>CV-B2</td>
<td>428</td>
<td>86</td>
<td>34</td>
<td>76</td>
<td>0.62</td>
</tr>
<tr>
<td>CV-B3</td>
<td>945</td>
<td>155</td>
<td>106</td>
<td>129</td>
<td>0.68</td>
</tr>
<tr>
<td>CV-B4</td>
<td>838</td>
<td>145</td>
<td>102</td>
<td>102</td>
<td>0.67</td>
</tr>
<tr>
<td>CV-B5</td>
<td>1226</td>
<td>206</td>
<td>136</td>
<td>130</td>
<td>0.77</td>
</tr>
<tr>
<td>E-11</td>
<td>1360</td>
<td>269</td>
<td>111</td>
<td>122</td>
<td>0.65</td>
</tr>
<tr>
<td>E-13</td>
<td>754</td>
<td>110</td>
<td>50</td>
<td>100</td>
<td>1.09</td>
</tr>
<tr>
<td>E-14</td>
<td>205</td>
<td>78</td>
<td>43</td>
<td>74</td>
<td>1.21</td>
</tr>
<tr>
<td>E-18</td>
<td>335</td>
<td>52</td>
<td>29</td>
<td>82</td>
<td>0.62</td>
</tr>
<tr>
<td>E-19</td>
<td>285</td>
<td>58</td>
<td>32</td>
<td>73</td>
<td>0.95</td>
</tr>
<tr>
<td>E-1</td>
<td>129</td>
<td>51</td>
<td>31</td>
<td>78</td>
<td>1.22</td>
</tr>
<tr>
<td>E-2</td>
<td>71</td>
<td>31</td>
<td>10</td>
<td>69</td>
<td>1.12</td>
</tr>
<tr>
<td>E-3</td>
<td>286</td>
<td>66</td>
<td>36</td>
<td>81</td>
<td>0.67</td>
</tr>
<tr>
<td>E-20</td>
<td>136</td>
<td>36</td>
<td>22</td>
<td>76</td>
<td>0.69</td>
</tr>
<tr>
<td>E-21</td>
<td>101</td>
<td>31</td>
<td>17</td>
<td>113</td>
<td>0.59</td>
</tr>
<tr>
<td>E-25</td>
<td>346</td>
<td>97</td>
<td>45</td>
<td>84</td>
<td>0.77</td>
</tr>
<tr>
<td>E-29</td>
<td>113</td>
<td>35</td>
<td>25</td>
<td>65</td>
<td>1.02</td>
</tr>
<tr>
<td>E-30</td>
<td>4200</td>
<td>325</td>
<td>270</td>
<td>120</td>
<td>0.75</td>
</tr>
<tr>
<td>E-33</td>
<td>182</td>
<td>41</td>
<td>29</td>
<td>92</td>
<td>1.02</td>
</tr>
<tr>
<td>E-6</td>
<td>1727</td>
<td>329</td>
<td>185</td>
<td>105</td>
<td>0.96</td>
</tr>
<tr>
<td>E-7</td>
<td>346</td>
<td>123</td>
<td>87</td>
<td>86</td>
<td>0.83</td>
</tr>
<tr>
<td>EV-D68</td>
<td>1667</td>
<td>80</td>
<td>33</td>
<td>55</td>
<td>0.63</td>
</tr>
<tr>
<td>EV-A71</td>
<td>9517</td>
<td>399</td>
<td>512</td>
<td>67</td>
<td>0.6</td>
</tr>
</tbody>
</table>

CV, Coxsackievirus; E, echovirus; EV, enterovirus.

*All genome regions, as of August 2016.

†Upon exclusion of sequences sharing over 98% identity and erroneous sequences.
In eight types (CV-A4, CV-A6, CV-A9, CV-B1, CV-B2, CV-B3, CV-B4 and E-6), extinction of mid-twentieth century genetic variants was incomplete, and a minority of modern isolates represented virus variants that branched from the population over 50 years ago (Fig. 3). In eight types (CV-A24, CV-B5, E-3, E-11, E-13, E-21, E-30 and EV-A71), modern isolates comprised several comparably prevalent diverse subgroups that split before the 1950s. These types included the most common and the best sampled types (Table 1); however, it is not clear if better sampling yielded virus variants that branched before 1950 in these types or if higher relative fitness of these types both provided high prevalence and precluded lineage extinction. It is noteworthy that in some of these types, e.g. CV-B5, E-11 and E-30, the grouping pattern observed here was similar to that described in previous publications [21–23]. However, EV-71 prototype strain BrCr was not basal in this study, but could be basal on the VP1 phylogenetic tree in other datasets [24]. Additional testing with randomly sampled EV-A71 sub-datasets showed that the position of the prototype strain can vary significantly depending on the exact sequences in the dataset (data not shown).

Grouping patterns within EV types represented a range of possible evolutionary scenarios that ranged from a single lineage balancing on the edge of extinction to diversification into several globally prevalent lineages. Moreover, evidence of several patterns could be seen within many types. This is best exemplified by CV-A4, CV-A10, E-19 and EV-D68 (but could also be suggested in certain other types), which apparently underwent severe bottlenecks (as evidenced by long branches leading from all contemporary isolates to the tree root in about 1950–1980), followed by the generation of several major lineages within each type. Therefore, bottlenecks and lineage extinction events could occur in global EV populations at any time and at any frequency, leading to prevalence of one or several contemporary virus lineages.

The global-scale population bottlenecks could repeatedly reset the molecular clock within types. Therefore, tMRCAs of each single type cannot be used to infer the type emergence date. To distinguish between the hypothesis of recent emergence of types (as suggested by tMRCAs of types) versus the hypothesis of infinitely repeating global bottlenecks that reset the molecular clock many times over centuries, we conducted Bayesian phylogenetic analysis at the species level. Such analysis should be interpreted with care because we cannot be sure that in the capsid-encoding genome region used for analysis, EV species represent holistic populations, and because the type emergence mechanisms are unknown and could include explosive generation of diversity over a short time (reviewed by Smura et al. [25]). Full VP1 sequence datasets were created for the types analysed of species EV-A and EV-B. To obtain manageable datasets, sequences that shared over 95 % (EV-A) or over 93 % (EV-B) identity were omitted, resulting in 163 and 441 sequences, respectively. Bayesian coalescent analysis inferred substitution rates of $0.4 \times 10^{-2}$ $(0.34 \times 10^{-2}$ to $0.45 \times 10^{-2}$) and $0.41 \times 10^{-2}$ $(0.37 \times 10^{-2}$ to $0.45 \times 10^{-2}$) substitutions per site per year, and root heights of 687 (547–830) and 574 (473–685) years for EV-A and EV-B, respectively (Fig. 4). The tree was well resolved; most nodes above individual types were supported with posterior probabilities of 1. Dates of the MRCAs of individual types inferred from full VP1 sequence analysis at the species level usually fit within the 95 % highest posterior density, but were up to 30 % higher than those inferred from separate analysis of types using partial VP1 sequences (see above). This could be because the substitution rates inferred in species datasets were on the lower edge of the substitution rate distribution observed upon analysis of individual types (see above). Branches leading to individual types were very long (80–427 years) compared with branches within individual types. This tree topology was compatible with the hypothetical bottlenecks that led to extinction of most virus lineages that predated contemporary viruses. On the other hand, the most recent separation of two distinct EV types (the MRCAs of E-21 and E-30) dated just 214 (167–264) years back, and many more tMRCAs of two types were below 300 years. Therefore, while regular bottlenecks could significantly affect evolution of types and apparent tMRCAs of types, there were indications that all known diversity of EV types could have emerged within the few last centuries.

Diversity of the genome region that encodes capsid proteins is currently the basis for identification of virus types by sequence analysis (molecular typing) of EVs. It was shown in 1999 that the maximum diversity of VP1 nucleotide sequence distance over the complete VP1 genome region within a type is 25 % [26], and this cut-off was accepted by the International Committee on Taxonomy of Viruses (ICTV) as the type criterion [11]. The number of known EV sequences has increased more than 100 times since 1999. Moreover, EVs accumulate on average $0.9 \times 10^{-2}$ substitutions per site per year in VP1 (Table 1), and could diverge significantly after the typing criteria are
implemented. We used the datasets of the most prevalent EV types to check if the type criterion is still reliable.

Firstly, the frequency distribution of full VP1 pairwise identity scores was calculated within and between all types studied using the alignment that contained 2657 sequences (Fig. 5). There were two peaks corresponding to intertypic distances (<74 %) and intratypic distances (>75 %), and a distinct valley between two peaks (74–75 %). Therefore, despite increased sampling and continuous diversification, types could be distinguished by the VP1 distance criterion. However, there were 856 sequence pairs with distances of between 74 and 75 %. Amino acid sequences could better discriminate the types. The intratypic and intertypic distance peaks were more distinct, and there were just 205 pairwise values in the range of 86–87 %. Therefore, there were only a few exclusions that challenged the typing criterion.

To investigate reasons behind the identity scores in the range of 74–75 %, we analysed the full VP1 pairwise distance distribution within 24 types that were represented by more than 30 full VP1 sequences (Fig. 6). In a number of EV types, there were intratypic pairwise distances of above 25 %, which contradicted the ICTV type criterion. In most instances, these were single outlier sequences that had identity values of between 74 and 75 % to other representatives of a type. Within five types, CV-A13, CV-A20, E-1, E-13 and E-30, there were multiple sequence pairs that exceeded the type threshold. To gain further insight into possible reasons, correspondence of pairwise nucleotide and amino acid sequence distances was plotted for typical cases (Fig. 7). In E-1, for some sequences there was a marked disparity between nucleotide and amino acid distances (Fig. 7a). Detailed analysis indicated substitutions in highly conserved VP1 regions and discordant nucleotide/amino acid phylogenies. Such isolated findings may not be regarded as novel types unless confirmed by an independent sequencing experiment, and ideally an independent isolation. Some single sequences exceeded either only the nucleotide sequence distance threshold of their type (e.g. CV-A4 AB457644 differed from other CV-A4 sequences by 24.7–26.8 % nt and 15.4 % aa, Fig. 7b), or both nucleotide and amino acid sequence thresholds (CV-A6 KF412903 differed by 25.2–27.2 % nt and 15.0–17.0 % aa from other CV-A6 sequences, C...
Fig. 3. Phylogenetic trees of VP1 gene typing fragment (294–339 nt, positions 2617–2976 according to the genome of PV1 strain Mahoney, Genbank acc. no. J02281). Prototype strains are indicated by grey branches and circles. Tip positions and branch lengths correspond to time. Bars correspond to ten years.
Fig. 3. (cont.)

E-11 — E-13 — E-14 —

E-18 — E-19 — E-20 —

E-21 — E-25 — E-29 —

E-30 — E-33 — CV-A13 —

CV-A20 — CV-A24 — EV-D68 —
In E-30, nucleotide distances of above 25% and amino acid sequence distances of above 15% were consistently observed between genotype A [21] and other genotypes. There was a good correspondence between amino acid and nucleotide distances throughout the dataset, and a continuous distribution of distances that extended beyond the 25% nt/15% aa ICTV type cut-off threshold (Fig. 7d, e) and up to 28% nt/21% aa distance. Apparently, E30 has gradually diverged beyond the type cut-off. In CV-A13 (Fig. 7f), and also in CV-A20 and E-13 (data not shown), nucleotide distances consistently exceeded 25%, but amino acid distances were below 15%. Therefore, nucleotide sequence alone is a poor type criterion in these types.

VP1 nucleotide sequence distance can be used to assign genotypes within an EV type, which may be useful to facilitate surveillance and analysis. However, there are no genotype designation standards for EVs. Pairwise distance distributions...
within types were generally multimodal (Fig. 6). There were distinct peaks in distance distributions within 11 out of 24 types (CV-A4, CV-A6, CV-A10, E-3, E-11, E-13, E-14, E-19, E-30, EV-A71 and CV-B3); hence, clear distance criteria for subdividing these types into genotypes could be proposed. The nucleotide distance, which could potentially distinguish virus subgroups, ranged between 80 and 91% among types; therefore, there cannot be a universal genotype cut-off value in EVs. Within other types, identification of genotypes was not clearly supported by sequence distance distribution.

**DISCUSSION**

Bayesian coalescent phylogenetic analysis yielded relatively recent tMRCAs for known sequences of each EV type. On the other hand, picornaviruses (or their ancestors) have likely existed since the eve of life on earth [27]. To solve this controversy, the impact of sample bias on the analysis first had to be evaluated. The number of non-redundant sequences of the types studied available for analysis was between 35 and 629. Usually, the most ancient isolate dated around 1950, and all other isolates were much more recent. The sample size and depth (age of the oldest isolate) had no effect on substitution rate estimates. Root age (tMRCA of a type), on the contrary, was affected by sample size and depth. However, sample size of over 100 non-redundant sequences was generally sufficient to exclude the effect of sample size on root height, and smaller sample size had only a moderate effect on root height estimates. Decreasing sample depth by removing the prototype strain from the sample led to a corresponding decrease in the root age. Therefore, the age of the oldest known sequence had the major effect on the inferred root height.

The absence of a relationship between sample depth and rate spoke against a significant effect of mutation saturation on Bayesian evolutionary estimates. Different effects of sample size/depth on the inferred rate and the root height suggested that the root age shift upon omitting the prototype strain was likely explained by biological reasons rather than a computational error. It was observed previously that all contemporary viruses of certain types, most notably E-30, are less divergent than the early isolates from the 1950s to 1980s [28, 29]. It was suggested that the virus population experienced global bottlenecks that drove most lineages to extinction and decreased virus diversity [29]. Similarly, it was noticed that at any given time the global diversity of the dominant EV-A71 genotypes was rather limited [30]. Analysis of tree topology indicated that the prototype strain was basal in half of the types studied (16/33). In nine more types, most of the contemporary viruses represented discrete clusters with relatively recent common ancestors. This topology suggested that the global populations of these types indeed underwent bottlenecks within the last decades. From an epidemiological point of view, such bottlenecks are compatible with annual virus population expansion–extinction cycles in countries with a temperate climate [31]. Such expansion–extinction cycles may also correspond to the emergence and extinction of E-30 recombinant forms observed in Europe [32]. Another mechanism that may provide such bottlenecks could be the ‘epidemic’ (i.e. expansion–extinction) epidemiological cycles of many EV types [33, 34] that take several years and might be explained by herd immunity fluctuations. Our analysis suggested that the global-scale bottlenecks occurred in most EV types and were one of the main driving forces of EV evolution.

It is not clear how the global virus populations of distinct types could regularly undergo such severe bottlenecks and yet not have become extinct. One explanation may be the existence of EV reservoirs in primates that provide long-term preservation of virus variants. The other possibility is that such emergence–extinction cycles are a feature of EV epidemiology and evolution in regions with a temperate climate.
Fig. 6. Frequency distribution of uncorrected pairwise identity scores among VP1 nucleotide sequences of individual serotypes. The y-axis indicates the total number of pairwise distances within a distance matrix for the range shown on the x-axis. The type threshold is indicated by dashed lines.
climate. Known EV sequences were mostly obtained in these regions (North Atlantic, China), and sampling in other regions may yield different results. In this study, many divergent sequences of different types were observed among Indian isolates from a previous study [35]. In CV-A4, 112 of 113 contemporary isolates were distinct from the prototype strain High Point isolated in 1948, while only one isolate from Kenya, dated 1999, grouped with the prototype strain (Fig. 3). Screening in Madagascar yielded the previously unknown EV-A71 genotype F [24], which was not present among thousands of samples screened in other countries. Highly divergent novel EV-A types were found in Bangladesh [36]. Echovirus 30 genotype A, which appeared extinct in early studies [29, 37, 38], was found to be endemic (or regularly introduced from neighbouring countries) in Russia [23]. These observations highlight the importance of isolates from distant locations for EV molecular evolution studies and differences in molecular evolution patterns between countries and continents. Further studies may reveal that what we identified as ‘global bottlenecks’ based on currently known sequences applies only to countries with seasonal EV infection cycles, temperate climate or certain hygiene levels.

The hypothetical global population bottlenecks precluded precise timing of the emergence of EV types. While the MRCA dates of known isolates may be rather reliable, they only relate to these isolates and do not indicate EV type emergence times. Using coalescent Bayesian analysis on several types at once may be problematic because distinct types are not necessarily a single population and because quantum evolutionary events (multiple almost instantaneous non-synonymous mutations) cannot be ruled out at type emergence (reviewed by Smura et al. [25]). Nevertheless, such analysis was performed to identify upper boundaries of type emergence times. Long branches leading to distinct EV types were compatible with global population bottlenecks and a relatively low chance of a population splitting into two distinct groups (novel types). This suggests that EV types correspond to fitness peaks, and populations normally wander around these fitness peaks. Indeed, it has been suggested previously that type emergence involves complete remodelling of the capsid rather than a number of gradual changes [39]. By contrast, rather recent dates of tree nodes that included several types were compatible with emergence of types within the last few hundred years. These dates should be treated with caution, because saturation of synonymous sites can become significant among EVs that diverged over 100 years ago [40], and dating of events that go too far beyond the sample depth can be wrong by orders of magnitude [41]. On the other hand, the possibility of a recent emergence of types is compatible with phylogenetic relations within several known types. Most clearly, the population of E-30 apparently split into two parts (genotype A vs all other genotypes) and has already diverged beyond the ICTV type criterion (Fig. 7). Phylogenetic relations within a few more types (e.g. CV-B5) were also compatible with a split into a few lineages that are sufficiently fit to persist in
circulation. These examples show that types (at least by the ICTV definition, if not in a serological sense) can emerge by genetic drift, in addition to genetic shifts that may be assisted by isolation of populations due to host cell/organism switch.

Most EV molecular epidemiology studies have suggested designation of genotypes within a type. In some types, these genotypes were distinct upon phylogenetic analysis, as in E-11 [42], in others they were in general poorly distinguishable, and were arbitrarily assigned to facilitate discussion of virus epidemiology, as in E-30 [21]. Distribution of phylogenetic distances within EV types indicated that designation of sequence-distance-based genotypes is possible within roughly half of types. The distance criteria of these genotypes could be in the range of 9–20% nucleotide distance, but would be different for each type. Existence of distinct genotypes/clusters within most types could be explained by the founder effect and/or extinction of ancestral virus variants. Such events at a subtype level reflect the previously discussed bottlenecks of global type populations.

Sequence-based criteria of EV types, i.e. the 25% nucleotide sequence distance and the 15% amino acid sequence cut-off, were suggested almost 20 years ago based on a relatively small number of sequences [9, 26]. EVs accumulate about 1% substitutions per year, and the number of known sequences has increased about 100-fold since these studies. These factors could potentially challenge the type criterion. In the dataset that mainly consisted of EV-A and EV-B sequences, the distinction between intra- and intertypic nucleotide sequence distance values was still clear (Fig. 4) and generally corresponded to the ICTV criterion. This was in agreement with recently published findings for EV-C, which also found generally good applicability of the typing criteria to contemporary viruses [43]. Our analysis suggested that the VP1 nucleotide distance identity cut-off for type designation could be conveniently extended to 74% instead of the current criterion of 75%, and even lower identity values (70–74%) do not necessarily justify designation of novel types without a compound analysis of amino acid sequence and phylogenetic grouping. It is noteworthy that discrimination of types in EV-A and EV-B observed here was apparently more robust than that in EV-C (Fig. 6), which corresponds to previous reports [20, 43].

It is clear that type designation can be hampered by both sequencing errors and biological variation. Therefore, we concur with a previous study [43] that designation of novel types should rely upon both nucleotide and amino acid sequence, and on phylogenetic grouping in the most complicated cases. To exclude sequencing errors, confirmatory sequencing in an independent laboratory and ideally an independent isolation is highly desirable.

**Conclusion**

The most recent ancestors of known isolates of individual EV types existed 55–200 years before now. Multiple lines of evidence have suggested that global-scale population bottlenecks could regularly reset the molecular clock. On the other hand, it is likely that all known diversity of EV types emerged in the last millennium. In a few types, there might be evidence of type emergence occurring by genetic drift, but relative stability of types over time suggests that they correspond to fitness peaks, which, in turn, implies the necessity of genetic shifts to found a new lineage. Understanding of EV emergence mechanisms is critical for prediction of appearance of novel EV types with altered virulence, such as poliovirus that evolved from EV-C coxsackieviruses [44]. The relative importance of genetic shift and drift, as well as the role of isolated virus reservoirs, such as secluded areas and non-human primates, in EV type emergence awaits further investigation.

**METHODS**

Among human EV sequences deposited in GenBank, 53 types were represented by more than 50 sequences and were used for the initial screening. The number of short (294–339 nt) sequences of the typing fragment was much higher than the number of complete (840–930 nt) VP1 sequences (Table 1); therefore, two alignments were created for each type. The dataset creation included the following steps:

1. All GenBank entries that contained the type designation were extracted and each type was handled separately.
2. Too-short sequences (threshold of 800 nt for full VP1 and 250 nt for the short fragment) and too-long sequences (over 8000 nt – erroneous entries) were omitted.
3. Sequences were aligned using MAFFT v.7.304 [45].
4. The full or typing VP1 sequence was excised from an alignment automatically according to a reference sequence.
5. Redundant GenBank sequences sharing more than 98% overall nucleotide sequence identity with any other sequence in the alignment and erroneously annotated sequences that shared less than 60% identity with a prototype sequence were omitted.
6. Degenerate nucleotide positions were resolved according to the consensus sequence.
7. Sequences with obvious errors (e.g. frameshifts) were excluded manually.
8. The short typing sequence alignments could be additionally truncated by up to 20 nt at each end if there were many GenBank sequences slightly shorter than the whole typing fragment. Only 100% complete VP1 sequences were used for full VP1 alignments. Sequences that only overlapped partially with the fragments analysed were excluded.

The 33 types were represented by more than 30 short VP1 sequences after alignment preparation and were used for phylogenetic analysis (Table 1).

Sequence handling was performed using Python scripts (available upon request) and BioEdit v.7.2 software [46]. Phylogenetic distances were calculated with BioPython.
Distance distribution was plotted using the Matplotlib Python module.

Phylogenetic analysis was done using a Bayesian likelihood-based algorithm implemented in Beast v.1.8.3 [47]. Short typing VP1 fragments were used for this analysis. Trial runs suggested that the SRD06 substitution model optimized for coding sequences should be used with the lognormal relaxed clock setting, which was preferred over assumption of strict and relaxed exponential clock models upon Bayes factor test (difference of log_{10} Bayes factors of tree likelihood >10). Each analysis was run over 50 million generations, and trees were sampled every 10 000 generations. Maximum clad credibility trees were annotated with TreeAnnotator v1.8.2 using a burn-in of 5 million generations. Trees were visualized with FigTree v1.4.2.

Funding information
The study was supported by the Russian Science Foundation grant 14-15-00619 and core funding of Chumakov FSCRDIBP (project 227). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Acknowledgements
We are grateful to Anatoly Gmyl and Andrei Deviatkin for valuable suggestions.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

Five reasons to publish your next article with a Microbiology Society journal

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as ‘excellent’ or ‘very good’.
5. Your article will be published on an interactive journal platform with advanced metrics.

Find out more and submit your article at microbiologyresearch.org.