Molecular evolutionary dynamics of Ross River virus and implications for vaccine efficacy

Anita Jones,1 Kym Lowry,1 John Aaskov,1 Edward C. Holmes2,3 and Andrew Kitchen2

INTRODUCTION

Ross River virus (RRV) is a mosquito-borne member of the genus Alphavirus that causes epidemic polyarthritis in humans, costing the Australian health system at least US$10 million annually. Recent progress in RRV vaccine development requires accurate assessment of RRV genetic diversity and evolution, particularly as they may affect the utility of future vaccination. In this study, we provide new RRV genome sequences and investigate the evolutionary dynamics of RRV from time-structured E2 gene datasets. Our analysis indicates that, although RRV evolves at a similar rate to other alphaviruses (mean evolutionary rate of approx. 8×10^{-4} nucleotide substitutions per site year^{-1}), the relative genetic diversity of RRV has been continuously low through time, possibly as a result of purifying selection imposed by replication in a wide range of natural host and vector species. Together, these findings suggest that vaccination against RRV is unlikely to result in the rapid antigenic evolution that could compromise the future efficacy of current RRV vaccines.

The GenBank/EMBL/DDBJ accession numbers for the novel RRV genome sequences described in this paper are GQ433354–GQ433360.

Four supplementary tables are available with the online version of this paper.
primarily in urban cycles of transmission. This raises the possibility that positive selection could affect the antigenic profile of vector-borne RNA viruses with largely urban transmission cycles, even though these tend to be subject to greater selective constraints than directly transmitted viruses such as influenza A virus (Woelk & Holmes, 2002). Antigenic drift caused by positive selection also requires sufficient virus diversity in the host population. This can be generated either by high rates of nucleotide substitution (themselves a product of high background mutation rates) or by large effective population sizes. In addition, understanding how RRV might respond to future vaccination pressures also requires knowledge of the fitness distribution of virus mutations. This may be manifest as the relative numbers of non-synonymous and synonymous substitutions per site and the extent of virus gene flow among geographical regions, as natural selection is expected to be most pronounced in spatially well-mixed populations.

Although intensive molecular evolutionary studies have been performed on a number of alphaviruses, most recently Chikungunya virus (Cherian et al., 2009), no such studies have been performed on RRV. Indeed, there is only limited information on the genetic variability of RRV (Sammels et al., 1995), with only two RRV genomes sequenced and available (strains NB5092 and T48; Faragher et al., 1988). Whilst both strains were recovered from mosquitoes and adapted to grow in the brains of suckling mice (Faragher et al., 1988), the pedigree of the NB5092 strain is uncertain (Gard et al., 1973). Herein, we both provide novel genome sequences of RRV recovered from multiple host species (birds, marsupials and a human) and estimate the evolutionary and epidemiological dynamics of gene sequences from 61 isolates of RRV, incorporating time-structured data.

RESULTS AND DISCUSSION

The nucleotide sequences of the non-coding regions of RRV, particularly the 5′ untranslated region (UTR) and the UTR preceding the structural protein genes, were highly conserved, with only minor variations found in the human isolate (QML1) and mosquito strains T48 and NB5092. The 3′ UTRs of isolates T48 (mosquito), 2975B, 2982B, 3078B (birds) and QML1 (human) each contained the four-repeat nucleotide sequences described previously (Faragher & Dalgarno, 1986). Both isolates from wallabies (8961W and 9057W) had a 35 nt deletion after nt 54 of the 3′ UTR that removed the first 22 nt from repeat sequence III, and which has been described previously (Faragher & Dalgarno, 1986). In addition, there were single (T48, 2982) and double (QML1) nucleotide deletions in the same region (−11 and −11, −12, respectively) of the 23 nt putative promoter for negative-strand RNA synthesis (Kuhn et al., 1990), which immediately precedes the poly(A) tail of the genome. Finally, the poly(A) tail varied in length from 29 nt (2982B) to 63 nt (T48).

RRV evolutionary rate and population dynamics

The mean rate of nucleotide substitution for RRV, calculated by using the sampling dates of the partial E2 gene sequences, was $8.08 \times 10^{-3}$ nucleotide substitutions per site (subs/site) year$^{-1}$, with a 95% highest probability density (HPD) of $3.97-12.70 \times 10^{-4}$ subs/site year$^{-1}$. Importantly, these results are consistent when estimated by using a wide range of nucleotide-substitution, molecular-clock and demographic models (Table 1; complete results of these analyses, as well as marginal likelihoods, are presented in Supplementary Table S1, available in JGV Online). Our estimate is consistent with both rate estimates of closely related alphaviruses (e.g. approx. $1 \times 10^{-3}$ subs/site year$^{-1}$ for Chikungunya virus; Cherian et al., 2009) and other vector-borne RNA viruses (Hanada et al., 2004; Jenkins et al., 2002). These results require a reassessment of previous assumptions that RRV may evolve slowly (Sammels et al., 1995) and highlight the importance of independent estimates of substitution rates.

Our coalescent analysis of the 61 partial E2 gene sequences produced a Bayesian skyline plot, which traces relative

<table>
<thead>
<tr>
<th>Nucleotide-substitution model*</th>
<th>Clock model</th>
<th>Demographic model</th>
<th>Marginal likelihood (ln)†</th>
<th>TMRCA (years)</th>
<th>Mean substitution rate (subs/site year$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTR + $\Gamma_4 + 1$</td>
<td>Relaxed (log-normal)</td>
<td>Bayesian skyline plot</td>
<td>−647.97</td>
<td>54.5 (45.0–79.0)</td>
<td>$8.13 \times 10^{-4}$ (3.87–12.6 $\times 10^{-4}$)</td>
</tr>
<tr>
<td>GTR + $\Gamma_4 + 1$</td>
<td>Strict</td>
<td>Bayesian skyline plot</td>
<td>−648.28</td>
<td>53.4 (45.0–70.7)</td>
<td>$8.55 \times 10^{-4}$ (4.50–12.9 $\times 10^{-4}$)</td>
</tr>
<tr>
<td>GTR + $\Gamma_4 + 1$</td>
<td>Relaxed (log-normal)</td>
<td>Constant</td>
<td>−649.44</td>
<td>62.4 (45.0–89.2)</td>
<td>$6.91 \times 10^{-4}$ (3.44–10.9 $\times 10^{-4}$)</td>
</tr>
<tr>
<td>GTR + $\Gamma_4 + 1$</td>
<td>Strict</td>
<td>Constant</td>
<td>−650.27</td>
<td>58.8 (45.0–79.9)</td>
<td>$7.45 \times 10^{-4}$ (3.85–11.4 $\times 10^{-4}$)</td>
</tr>
<tr>
<td>HKY</td>
<td>Strict</td>
<td>Constant</td>
<td>−653.09</td>
<td>59.1 (45.0–79.7)</td>
<td>$6.78 \times 10^{-4}$ (3.58–10.4 $\times 10^{-4}$)</td>
</tr>
</tbody>
</table>

*GTR, General time-reversible nucleotide-substitution model with six reversible nucleotide substitution-rate categories; $\Gamma$, gamma distribution of between-nucleotide sites model of substitution-rate variation; 1, proportion of invariant sites model of substitution-rate variation; HKY, Hasegawa–Kishino–Yano substitution model with two reversible nucleotide substitution-rate categories (transitions and transversions).

†Natural log (ln) of the marginal likelihood.
genetic diversity through time, revealing that the diversity of RRV has been remarkably low over time (Fig. 1). Conservatively estimating significant changes as the occurrence of non-overlapping 95% credible intervals, we saw no significant changes in the genetic diversity of RRV over time, despite at least one well-documented RRV epidemic (Aaskov et al., 1981; Rosen et al., 1981; Tesh et al., 1981). This is supported by the lack of statistical preference for either Bayesian skyline or constant demographic models compared with each other (Table 1). Furthermore, the estimated mean time to the most recent common ancestor (TMRCA) of the sampled RRV sequences was only 55 years (95% HPD, 45–79 years), which is consistent with the low levels of genetic diversity observed in RRV over the past several decades (reviewed by Russell, 2002). Together, these results suggest that the genetic diversity of RRV is, and has been, lower than that of some other arboviruses, such as the flaviviruses dengue virus (Dunham & Holmes, 2007) and yellow fever virus (Bryant et al., 2007), despite their broadly equivalent rates of nucleotide substitution. Importantly, the 17D yellow fever vaccine has retained its efficacy and has not led to antibody-escape mutations, despite the greater genetic diversity of yellow fever virus populations and >60 years of use (Barrett, 1997; Monath, 1999). Interestingly, the genetic diversity of RRV is also slightly lower (although with overlapping 95% HPDs) than that observed in a closely related alphavirus, Chikungunya virus (Cherian et al., 2009), suggesting that similar mechanisms might underlie both. Consistent with this, the TMRCA of the sampled RRV sequences was rather more recent than that of Chikungunya virus (55 years for RRV cf. 129 years for Chikungunya; Cherian et al., 2009). This suggests either that the effective population size of the natural reservoir of RRV is smaller than that of Chikungunya, resulting in a more rapid turnover of genetic diversity in RRV by genetic drift, or that positive selection is driving the more rapid turnover of diversity in RRV.

Phylogenetic and geographical structure of RRV

The maximum clade credibility (MCC) phylogeny for RRV (Fig. 2), which summarizes the posterior distribution of trees produced from the BEAST analysis, displays substantial geographical and temporal structure. Most notably, this tree reveals the presence of distinct eastern and western Australian RRV lineages (geographical origins identified by colours), as well as the loss of a north-east Australian lineage. The fact that the lost lineage has not been sampled since 1977 and has been replaced in north-east Australia by another lineage suggests that, now, it either is extinct or has been reduced to a very low frequency in the RRV population. The loss of a lineage from a specific geographical distribution without a significant decrease in the overall genetic diversity of RRV (see Fig. 1) suggests that the population dynamics of RRV may differ between geographical regions. Furthermore, the basal location of sequences from north-east Australia (i.e. Queensland), combined with the continued sampling of RRV in north-east Australia from a single lineage and the relative geographical specificity of the western Australian lineage, suggests that (i) the present diversity of RRV probably originated in Queensland, and (ii) most migration of RRV within Australia is from east to west. Both of these results support epidemiological data (e.g. higher rates of RRV infection in Queensland than in other Australian provinces; Australian Department of Health and Ageing, 2009) that together support Queensland as the historical source for RRV in Australia.

Our tests of the association between the geographical origin and phylogenetic relationships of the sampled RRV sequences revealed significant associations within the tree as a whole [association index (AI), P < 0.0001; parsimony score (PS), P < 0.0001] and with respect to all geographical regions [monophyletic clade size (MC), 0.03 < P < 0.001] except the Pacific Islands (MC, P = 1.0). Hence, clear spatial structure exists within RRV populations, particularly between eastern and western Australia, confirming the results of an earlier study of RRV using distance methods (Sammels et al., 1995). Notably, these results indicate that the steady accumulation of mutations and high turnover of RRV diversity (evidenced by the high substitution rate, loss of a lineage and the recent TMRCA) characterized in these analyses are occurring simultaneously in distinct RRV populations not connected routinely by extensive gene flow.

Fig. 1. Bayesian skyline plot depicting relative genetic diversity through time for 61 dated E2 gene nucleotide sequences. Time (x-axis) is measured in calendar years, whilst genetic diversity (y-axis) is a compound variable relating the effective population size (Nₑ) and virus generation time (g). The black line traces the mean genetic diversity, whilst the grey lines indicate 95% credible intervals. The dotted vertical line indicates the lower 95% credible interval for the coalescent time of the sampled sequences (which occurs at the left-hand edge of the plot). Notably, this plot is indicative of no significant changes in RRV genetic diversity from the past to the present.
Selection pressures on RRV

Both natural selection and genetic drift can, individually or in combination, reduce the genetic diversity within a population. It is therefore possible that antigenic drift and recurrent selective sweeps have reduced genetic diversity in RRV populations. Strong temporally structured phylogenies, such as that observed here (Fig. 2), are often the hallmark of continuous selection pressure (Grenfell et al., 2004). However, the mean \( d_N/d_S \) ratio in the set of 61 RRV E2 protein sequences analysed here was 0.25, indicating that this region is largely evolving under weak purifying selection. Analyses to determine which, if any, codons in this region were evolving under positive selection failed to identify any under models that consider all \([\text{fixed effect likelihood (FEL); } P > 0.05]\) or only internal \([\text{IFEL; } P > 0.05]\) branches. Furthermore, analysis of the RRV genome dataset produced estimates of mean \( d_S/d_k \) for each gene in the RRV genome (non-structural proteins 1–4, envelope glycoproteins 1–3, C protein and 6K protein) that supported relatively stronger purifying selection acting across the RRV genome \((0.04 < d_S/d_k < 0.15)\), with no sites under significant positive selection \([\text{single likelihood ancestor counting (SLAC) model, } P > 0.05]\).

These signatures of purifying selection are consistent with previous studies demonstrating constraints on arbovirus evolution produced by simultaneous selection for replication in two very different hosts (i.e. a range of vertebrate and mosquito hosts; Coffey et al., 2008; Scott et al., 1994; Woelk & Holmes, 2002). Importantly, such selective constraints are expected to limit the coding diversity of RRV and possibly constrain RRV’s response to selective pressures imposed by vaccination of human hosts. However, it is important to recall the limited sampling undertaken in this study and its possible effects on the conclusions drawn. For example, only five sequences were isolated from the 1979–1980 epidemic of RRV in several of the Pacific Islands, so that it is difficult to determine whether adaptive evolution occurred during this outbreak. Sampling of RRV more densely in time is clearly required to investigate the correlation of RRV diversity with historical epidemics in humans (e.g. Aaskov et al., 1981; Rosen et al., 1981; Tesh et al., 1981), with the general trend of increasing numbers of human infections (e.g. Boughton et al., 1984; Hawkes et al., 1993; Australian Department of Health and Ageing, 2009), as well as the strength of selection acting on this virus.

Collectively, our findings indicate that, whilst RRV is evolving rapidly, its observed genetic diversity is consistently low through time and has a recent origin. This is possibly due to constraints stemming from a life cycle involving replication in distinct hosts and small natural host populations, rather than antigenic drift or recurrent selective sweeps. Importantly, it is likely that these constraints limit the probability that RRV could generate enough antigenic diversity to escape vaccine-induced immunity. This is supported by historical observations of lifelong immunity provided by natural RRV infection of humans (Fraser, 1986). In the absence of substantial and unforeseen changes in the molecular evolution of RRV, our analyses suggest that the evolutionary dynamics of RRV do not preclude the success of vaccination in controlling EPA.

METHODS

Isolation and nucleotide sequencing of novel RRV genomes. Complete RRV genomes were sequenced from viruses recovered from mosquitoes (T48), wallabies (8961W, 9057W), birds (2975B, 2982B, 3078B) and a human EPA patient (QML1). RNA was extracted from RRV in the supernatant of infected cultures of Vero cells with
Determination of the nucleotide sequences of the 5′ ends of the genome was performed by using a strategy for the rapid amplification of cDNA ends (RACE) described previously (Tillet et al., 2000). Briefly, cDNA from the RRV isolates was prepared from RNA by using 30 pmol primer cP370 and SuperScript III reverse transcriptase as described above. The cDNA was treated with 1 U RNase H (Roche) for 20 min at 37 °C prior to purification using a High Pure column (Roche). cDNA was ligated to an oligonucleotide adaptor DT88 (5′-GAAGAGAGGGTGGGAAATGGCGTGG-3′), which contained a 3′ dideoxy-A and a phosphorylated 5′ terminus, using T4 RNA ligase. The resultant adaptor-ligated cDNA was amplified by using primer PDT88 (5′-CCAAACGGCATTCTCCTCTCTCC-3′, which was complementary to the adaptor and virus-specific primer cP114) and the PCR conditions described above. The PCR product was analysed and purified from a 2.5 % agarose–TAE gel and ligated into pGEM-T Easy plasmids using T4 DNA ligase (Promega). Escherichia coli DH5 cells were transformed with vector according to the manufacturer’s instructions (Invitrogen) and plasmids were purified from individual bacterial colonies that grew on Luria–Bertani agar supplemented with 100 μg ampicillin ml⁻¹, 0.5 mM IPTG and 80 μg 5-bromo-4-chloro-3-indolyl β-d-galactosidase (X-Gal) ml⁻¹, using QIAprep Miniprep kits (Qiagen). Virus cDNA inserts in plasmids were sequenced at the AGRF using plasmid-specific primers. The nucleotide sequences of the 3′ ends of the genomes of RRV T48, QML1, 2982B, 2975B, 3078B, 8961W and 9057W were determined by using a similar RACE strategy. However, adaptor cDNA DT88 was ligated to the 3′ end of the RNA using T4 RNA ligase prior to the production of cDNA using the primer PDT88 and SuperScript III reverse transcriptase as described above. PCR amplification of the resultant cDNA was performed by using the primers P11374 and PDT88, and the resultant double-stranded (ds) cDNA was purified and sequenced as above. The ds cDNA from RRV 9057W and T48 was also ligated into pGEM-T Easy plasmids and sequenced as described above. These novel genome sequences have been deposited in GenBank under accession numbers GQ433354–GQ433360, and the details of the novel RRV isolates are available in JGV Online as Supplementary Table S3.

**Sequence alignments.** The seven novel RRV genome sequences and one additional RRV genome sequence taken from GenBank (strain NB5092; accession no. M20162) were aligned by using CLUSTAL_X (Thompson et al., 1994), and the resulting alignment of 11 900 nucleotide positions was edited by hand to maintain reading-frame integrity. A second dataset was assembled from the envelope glycoprotein (E2) genes of the eight viruses for which the entire nucleotide sequences have been determined and an additional 53 partial E2 sequences collected from GenBank. The sampling dates, ranging from 1959 to 2004, and geographical origins for all partial E2 sequences were also collected and recorded (available in JGV Online as Supplementary Table S4). The 61 E2 sequences were aligned as above, resulting in an alignment of 252 nt (both alignments are available from the authors on request). The passage histories of the viruses from which the E2 gene sequences were derived are described by Sammels et al. (1995).

**Estimation of evolutionary rates and population dynamics in RRV.** A Bayesian skyline coalescent analysis, which depicts changing levels of relative genetic diversity [i.e. effective population size (Nₑ) × virus generation time (g)] through time using an unconstrained demographic model (Drummond et al., 2005), was performed on the dataset of 61 dated RRV E2 sequences by using BEAST v1.4.8 (Drummond & Rambaut, 2007). In this analysis, the sampling dates of the E2 sequences alone provided the necessary calibrations for estimating the rate of nucleotide substitution in RRV. The Bayesian skyline analysis was performed by using the most general GTR+Γ+I model of nucleotide substitution, assuming a relaxed molecular clock with an uncorrelated log-normal rate distribution (Drummond et al., 2006). Additional analyses were performed with combinations of nucleotide-substitution models [i.e. Hasegawa–Kishino–Yano (HKY) with or without site-specific rate variation, strict and relaxed clocks, and constrained (i.e. constant) demographic models to assess the robustness of our results. All Markov chains were run for 100 000 000 generations and sampled every 1000 generations, with the first 10 000 samples discarded as burn-in. Unless otherwise specified, default priors were used for all parameters. The program Tracer v1.4 (http://evolve.zoo.ox.ac.uk) was used to inspect sampled posterior probabilities visually for chain stationarity, to calculate summary statistics for genetic parameters and to generate Bayesian skyline plots. Sampling error is represented by values for the 95 % HPD.

**Phylogenetic analysis of RRV geographical structure.** The posterior set of trees was used to estimate the overall MCC phylogeny for the E2 sequence data. To determine the extent of geographical structure in RRV populations, we estimated the strength of association between phylogenetic relationships and sampling locations. The degree of association was estimated by using the PS (Slatkin & Maddison, 1989) and AI (Wang et al., 2001) statistics to determine the association with sampling locations across the entire tree, and the maximum MC (Parker et al., 2008) statistic to assess the association with particular sampling locations. Each E2 sequence was coded as coming from one of five geographical regions: north-east Australia (Queensland), south-east Australia (New South Wales, Tasmania or Victoria), Western Australia, north-central Australia (Northern Territory) or Pacific Islands (Cook Islands, Fiji or Samoa) (see Supplementary Table S4). Using the program BaTS v0.90 beta (Parker et al., 2008), distributions of the three statistics (PS, AI and MC) were calculated from the posterior samples of trees and compared with null distributions generated by randomizing the RRV geographical origins and recalculating across the tree sample to determine the significance of the empirical distributions (Parker et al., 2008). The first 10 000 sampled trees (i.e. 10 %) were discarded as burn-in and 1000 randomizations were performed to estimate null distributions for the statistics of interest.

**Analysis of selection pressures.** An FEL model (Kosakovsky Pond & Frost, 2005c) was used to determine the strength of selection pressures on the dataset of dated E2 sequences, manifest as the ratio of non-synonymous (dₑ) to synonymous (dₛ) substitutions per site. The analysis was performed first on all branches of the tree (i.e. FEL) and subsequently on only internal branches (i.e. IFEL; Kosakovsky Pond et al., 2006). An additional analysis of selective pressures acting on all genes individually was performed on the coding genome data by using the SLAC model (Kosakovsky Pond & Frost, 2005c). All analyses were performed in HyPhy (Kosakovsky Pond & Frost, 2005a) on the web server Datamonkey (Kosakovsky Pond & Frost, 2005b), with all input branch lengths and tree topologies estimated by using a
neighbour-joining algorithm and HKY85-based codon-substitution model.

ACKNOWLEDGEMENTS

This study was supported by grants from the National Health and Medical Research Council of Australia and from the Cook Estate.

REFERENCES


