Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 8022–8027, July 1999
Evolution

Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus

MARK J. GIBBS* AND GEORG F. WEILLER

Bioinformatics, Research School of Biological Sciences, The Australian National University, G.P.O. Box 475, Canberra 2601, Australia

Communicated by Bryan D. Harrison, Scottish Crop Research Institute, Dundee, United Kingdom, April 28, 1999 (received for review December 22, 1998)

ABSTRACT **There are several similarities between the small, circular, single-stranded-DNA genomes of circoviruses that infect vertebrates and the nanoviruses that infect plants. We analyzed circovirus and nanovirus replication initiator protein (Rep) sequences and confirmed that an N-terminal region in circovirus Reps is similar to an equivalent region in nanovirus Reps. However, we found that the remaining C-terminal region is related to an RNA-binding protein (protein 2C), encoded by picorna-like viruses, and we concluded that the sequence encoding this region of Rep was acquired from one of these single-stranded RNA viruses, probably a calicivirus, by recombination. This is clear evidence that a DNA virus has incorporated a gene from an RNA virus, and the fact that none of these viruses code for a reverse transcriptase suggests that another agent with this capacity was involved. Circoviruses were thought to be a sister-group of nanoviruses, but our phylogenetic analyses, which take account of the recombination, indicate that circoviruses evolved from a nanovirus. A nanovirus DNA was transferred from a plant to a vertebrate. This transferred DNA included the viral origin of replication; the sequence conservation clearly indicates that it maintained the ability to replicate. In view of these properties, we conclude that the transferred DNA was a kind of virus and the transfer was a host-switch. We speculate that this host-switch occurred when a vertebrate was exposed to sap from an infected plant. All characterized caliciviruses infect vertebrates, suggesting that the host-switch happened first and that the recombination took place in a vertebrate.**

Sometimes viruses are transmitted to a host species that they have not previously infected or that they rarely infect. Several of these atypical interspecies transmission events (host-switching events) have led to disease outbreaks in this century (1–3). Analyses of viral genomic sequences provide a historical perspective on these events; the phylogenies of families of viruses inferred from sequences often do not match those of their hosts (2–4), suggesting that there have been many host-switching events in the past. Almost all of these events involved the transfer of viruses between hosts in the same phylum or division. However, similarities between the genomes of some plant-infecting, vertebrate-infecting, invertebrate-infecting, and microbe-infecting viruses indicate that they have common ancestors (5), and suggest that at some point, ancestral viruses switched between these very different kinds of hosts. These more radical changes in host preference have led to the evolution of many new virus species. They are significant in terms of both disease and evolution; beyond that, little about them is known: neither the identity of the original hosts, nor the possibility of linkage between the change in host preference and a specific genetic change.

The history of viruses is further complicated by interspecies recombination. Distinct viruses have recombined with each other, producing viruses with new combinations of genes (6, 7); viruses have also captured genes from their hosts (8, 9). These interspecies recombinational events join sequences with different evolutionary histories; hence, it is important to test viral sequence datasets for evidence of recombination before phylogenetic trees are inferred. If a set of aligned sequences contains regions with significantly different phylogenetic signals and the regions are not delineated, errors may result.

Interspecies recombination between viruses has been linked to severe disease outbreaks (10), and some analyses suggest that it may be linked to host-switching (7, 11). Fortunately, newly generated interspecies recombinants are rarely found, suggesting that few have arisen recently. To date no evidence has been reported of recombination between viruses that infect hosts from different kingdoms, e.g., no evidence of vertebrate-infecting viruses recombining with plant-infecting viruses. Furthermore, although evidence has been reported of recombination between RNA viruses (6, 7), and between DNA viruses (10, 13), only one report has suggested recombination between an RNA virus and a DNA virus. A glycoprotein gene appears to have been transferred between a baculovirus and a Thogoto-like virus, but the direction of transfer is unclear and it is possible that both viruses acquired their gene from their arthropod hosts independently (12).

The two known circoviruses, *Porcine circovirus* (PCV) and *Psittacine beak and feather disease circovirus* (BFDV), are similar in several ways to a set of plant-infecting viruses previously known as plant circoviruses (14–16), but recently reclassified in the genus *Nanovirus* (16, 17). Circoviruses and nanoviruses have small, icosahedral particles, 17–22 nm in diameter, and small, circular, single-stranded DNA genomes; those of circoviruses are about 2 kb long, whereas nanovirus genomes are about 1 kb long. The two kinds of virus encode a replication initiator protein (Rep), and there are clear similarities between the sequences of these proteins (14). Reps initiate rolling circle replication at a nonanucleotide sequence within a longer origin-of-replication sequence (15, 18, 19), and the nonanucleotide sequences of circoviruses match those of some nanoviruses. Two recent phylogenetic analyses placed Rep sequences of circoviruses and nanoviruses in separate groups (15, 17). Here we show that these results may have been distorted by an interspecies recombinational event and that the true evolutionary history of circoviruses and nanoviruses reveals significant information about a major host-switching event.

Evidence of Interspecies Recombination

The nonredundant amino acid sequence database was searched by using the program BLASTP, version 2 (20). Initial

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

Abbreviations: BFDV, *Psittacine beak and feather disease circovirus*; PCV, *Porcine circovirus*; P-loop, phosphate-binding loop; Rep, replication initiator protein.

*To whom reprint requests should be addressed. e-mail: mgibbs@rsbs.anu.edu.au.

searches with nanovirus Rep sequences suggested that these sequences were similar only to Rep sequences of circoviruses and other nanoviruses, but a search with a circovirus Rep sequence yielded significant similarities with sequences from caliciviruses and other picorna-like viruses. Picorna-like viruses code for a polyprotein that is cleaved proteolytically to produce an RNA-binding protein known as the 2C-protein (21), and it was the conserved region within the 2C-protein (22) that matched a C-terminal part of the circovirus Rep. Alignments between circovirus Rep sequences and calicivirus 2C-protein sequences yielded Expect values (*E* values; ref. 20) as small as 4×10^{-6}, and alignments between circovirus and nanovirus Rep sequences yielded *E* values as small as 8×10^{-17}.

E values $<1 \times 10^{-2}$ probably indicate relatedness (20), unless they are affected by unusual sequence composition. However, concern arose because both the Rep and 2C-proteins include a short glycine-rich sequence, ending with a Gly-Lys-Thr or a Gly-Lys-Ser (GKT/S) motif and forming a phosphate-binding loop (P-loop); this sequence was included in the alignment identified by BLASTP. Several kinds of nucleotide-binding protein include a P-loop; structural studies show some of them to be unrelated (23). Therefore, the presence of the P-loop in both 2C-proteins and circovirus Reps could result from convergence and this could partly explain their affinities. To confirm that the various similarities to circovirus Reps resulted from common ancestry, rather than similar compositions, we used the program ALIGN (24). This program calculates a *Z* score for a pair-wise alignment by scoring alignments of the same pair of sequences after the positions of residues in one sequence have been randomized, but not its composition. We compiled a dataset consisting of the available nanovirus and circovirus Rep amino acid sequences, the conserved regions of the 2C-protein sequences from four caliciviruses, and the conserved regions of the 2C-protein sequences from a virus in each genus in the *Picornaviridae*, *Sequiviridae*, and *Comoviridae*, i.e., the picorna-like supergroup (25, 26). We made 100 alignments from randomized sequences for each pair-wise comparison. *Z* scores >5 are generally assumed to indicate homology (27); we obtained *Z* scores >5 from several comparisons between 2C-protein and circovirus Rep sequences, but we had doubts about the relationship because the highest of these scores was only 5.9. For this reason, we repeated the tests using Rep sequences from which N-terminal sequences, including the P-loop sequence, had been deleted.

Tests with circovirus sequences that had been truncated up to and including the GKT/S motif confirmed that the similarities were not due solely to the presence of the P-loop, because they yielded *Z* scores as high as 8.6 in comparisons with calicivirus sequences. Tests with circovirus Rep sequences that had been truncated to a point 15 residues to the N-terminal side of the GKT/S motif yielded *Z* scores as high as 9.3 in comparisons with calicivirus sequences and as high as 7.3 in comparisons with sequences from other picorna-like viruses. Thus, the tests confirmed the common ancestry of the circovirus Rep and picorna-like virus 2C-proteins.

The homology of the circovirus and nanovirus Reps was similarly confirmed; we used Rep sequences from which C-terminal sequences had been deleted, and obtained *Z* scores up to 10.1. The highest *Z* score was 2.9 from a pair-wise comparison between a complete nanovirus Rep sequence and a 2C-protein sequence, 4.6 from a comparison between a 2C-protein sequence and a nanovirus C-terminal Rep sequence, and 3.2 from a comparison between circovirus and nanovirus C-terminal Rep sequences. However, most of the comparisons in these three sets yielded scores approaching zero, suggesting that these polypeptides are unrelated or are only very distantly related.

The Recombination Site

To delineate regions with different origins in circovirus Rep sequences, we made multiple alignments with the program CLUSTALW (28). We tested the accuracy of these alignments by altering the order of alignment, by altering the alignment parameters, and by aligning circovirus sequences in combined datasets with either 2C-protein sequences or nanovirus Rep sequences. We consistently found similarities between the circovirus Rep sequences and the 2C-protein sequences; the similarities extended from position 178 (relative to the alignment shown in Fig. 1) to close to the C terminus; they spanned

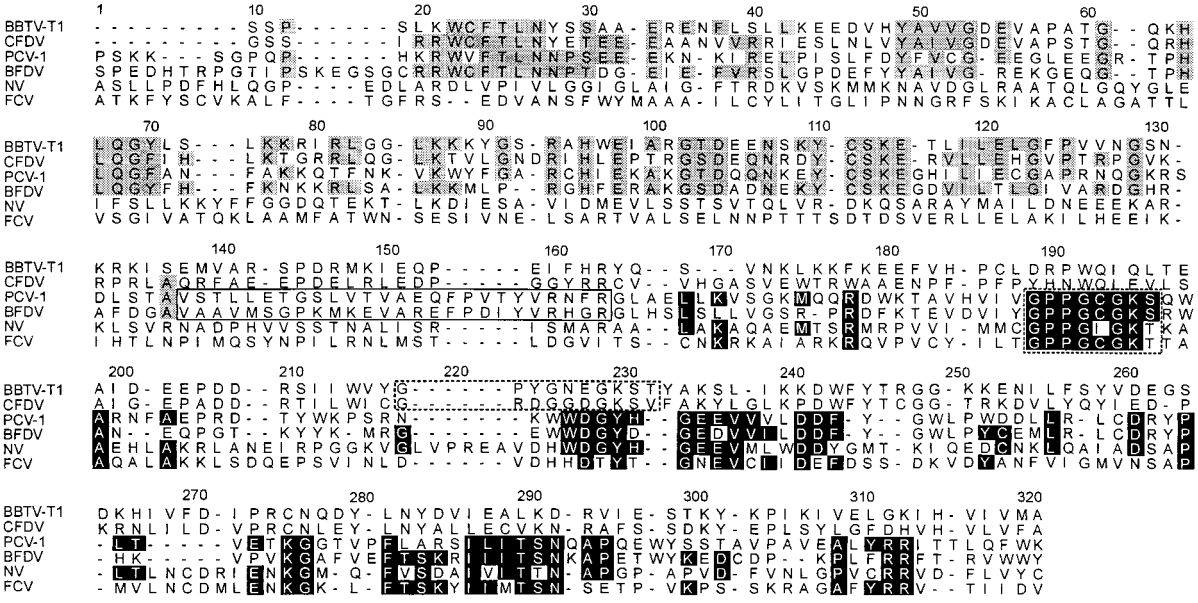

FIG. 1. An alignment of Rep sequences from *Banana bunchy top nanovirus* Taiwanese isolate DNA1 (BBTV-T1), *Coconut foliar decay nanovirus* (CFDV), PCV type 1, and BFDV, together with the 2C-protein conserved region sequences from *Norwalk calicivirus* (NV) and the *Feline calicivirus* (FCV). The region in which the recombination probably occurred is marked with a solid-line box. Identities between the N-terminal sequences of circovirus and nanovirus Reps are marked with gray blocks, and those between calicivirus 2C-proteins and the C-terminal sequences of circovirus Reps are marked with black blocks. P-loop sequences are marked with dashed-line boxes.

the entire conserved region of the 2C-protein sequences of picorna-like viruses and matched conserved positions in those sequences (Fig. 1: positions 178 to 313). Similarities between circovirus and nanovirus Reps extended from the N terminus of Rep to position 129, but few were found beyond that point, suggesting that the recombination site lay between positions 129 and 178. Phylogenetic profiles made with the program PHYLPRO (29) suggested a recombination site between positions 137 and 164.

These observations suggest that the P-loops of circovirus Reps share a common origin with the P-loops in 2C-proteins. Similarities between these P-loop sequences, especially the positions of glycine and proline residues, support this possibility (Fig. 1). The P-loops of the nanovirus Reps are not aligned with those of the circovirus Reps or those of the calicivirus 2C-proteins in the alignment (Fig. 1). These P-loops were aligned, however, when we minimized the gap costs; alignments made in this way also showed a phylogenetic inconsistency between positions 137 and 164.

Phylogeny Inference and the Choice of Outgroups

Separate phylogenies were found for the N- and C-terminal regions of the circovirus Reps and their respective homologues, and the effect of alignment order on the phylogenies was tested. Maximum likelihood trees (30) were inferred from aligned amino acid sequences by quartet puzzling by using the program PUZZLE, version 4 (31), after positions including gaps had been excluded. Likelihoods were calculated by using the BLOSUM 62 substitution matrix (32); a gamma distribution of the rates of change for variable sites with a shape parameter was estimated from the data by using a neighbor-joining tree. Maximum likelihood and most parsimonious trees (33) were also found from aligned nucleotide sequences by heuristic searching with the program PAUP, version 4d64 (written by David L. Swofford), after positions, including gaps and third codon positions, had been excluded. The parameters, including the shape parameter, of the substitution model used to calculate likelihoods from the nucleotide data were estimated and re-estimated using trees found after successive heuristic searches.

The choice of outgroups was difficult. Caliciviruses are clearly distinct from other picorna-like viruses. Hence, the root of the calicivirus 2C-protein gene cluster could be located using the sequences of other picorna-like viruses as outliers, but whether the root of a combined calicivirus–circovirus sequence cluster could also be located in this way was not clear. Therefore, we decided to leave the C-terminal tree unrooted. Geminivirus Rep (AL1) amino acid sequences were the obvious choice as outliers for the N-terminal sequence tree, because these sequences were thought to be similar to circovirus and nanovirus Rep sequences (14, 15, 34), but again, uncertainty arose because the relationship with the geminivirus sequences had not been firmly established. A BLAST alignment comparing a nanovirus Rep sequence with one from a geminivirus yielded 0.16 as the smallest *E* value, and 4.3 was the highest *Z* score obtained with ALIGN. However, we did not dismiss the relationship because there are significant similarities between the viruses. Geminiviruses, like circoviruses and nanoviruses, have small, circular, single-stranded DNA genomes, their replication is initiated at a nonanucleotide sequence (35) similar to that of circoviruses and nanoviruses (14, 34), and both geminivirus and nanovirus Reps cleave their respective single-stranded DNAs within the nonanucleotide sequence between nucleotide positions 7 and 8. To test further the relationship between nanovirus and geminivirus Reps, we made protein structure predictions. The program PHDSEC (36) was used to predict the positions of α-helices and β-strands in these proteins with separate alignments of nanovirus and geminivirus sequences. We then mapped the positions of the predicted structures from both kinds of protein onto a single alignment and found that the predicted positions of structural elements in the two kinds of protein correlated (Fig. 2, *A* and *B*; β-strand $\chi^2 = 38$, $P < 0.001$; α-helix $\chi^2 = 62$, $P < 0.001$). We thus confirmed that the proteins are homologues, and we decided to use geminivirus Rep sequences as outliers to root the N-terminal sequence tree.

Circovirus Sequences with Distinct Origins

In all of the trees found for Rep N-terminal sequences (Fig. 3), nanovirus sequences were split into two main clusters, and circovirus sequences were placed within one of these clusters, the same cluster in each tree. Surprisingly, this indicates that the 5′-region of the circovirus Rep gene diverged sometime after the origin and early diversification of nanovirus Rep genes and thus that the circovirus Rep gene evolved from a nanovirus gene. Because all nanoviruses infect plants and all circoviruses infect vertebrates, the trees indicate that the 5′ part of the circovirus Rep gene was transferred from a plant to a vertebrate. Clearly, confidence in the location of the root of the tree is important to this conclusion. The root was located by using seven sequences from the three geminivirus genera as outliers; it was placed on the same branch in each of the trees; and the midpoint was on this same branch in each of the trees.

Database searches supported the branching pattern; they showed that some nanovirus Rep N-terminal sequences are more closely related to circovirus Rep sequences than to other nanovirus Rep sequences. They also showed that the N-terminal part of the circovirus Rep is much more closely related to nanovirus sequences than it is to geminivirus sequences, indicating that the circovirus lineage diverged sometime after the nanoviruses and geminiviruses had diverged from a common ancestor. Since both nanoviruses and geminiviruses infect plants, this also implies that part of the circovirus Rep gene originated in a plant virus.

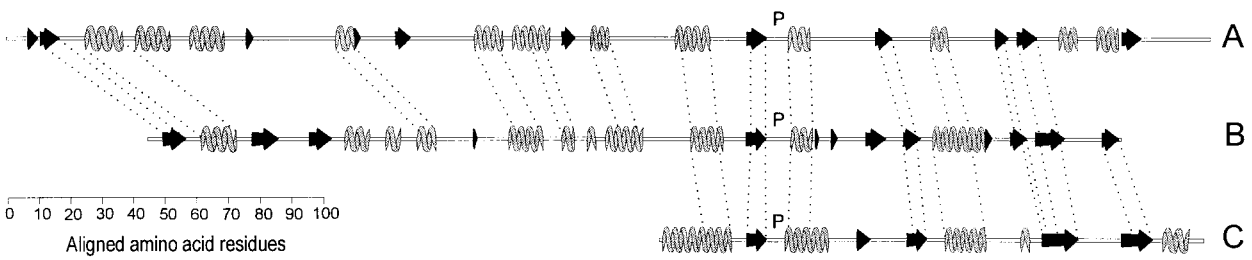

FIG. 2. Predicted secondary structures for geminivirus Reps (*A*), nanovirus Reps (*B*), and calicivirus 2C-proteins (*C*). Black arrows represent regions predicted to form β-strands, and gray helices represent regions predicted to form α-helices. Predictions were made from separate alignments of four geminivirus, ten nanovirus, and six calicivirus sequences. The position of the P-loop in each set of sequences is marked "P." Dotted lines join structural elements that had matching positions when sequences were aligned.

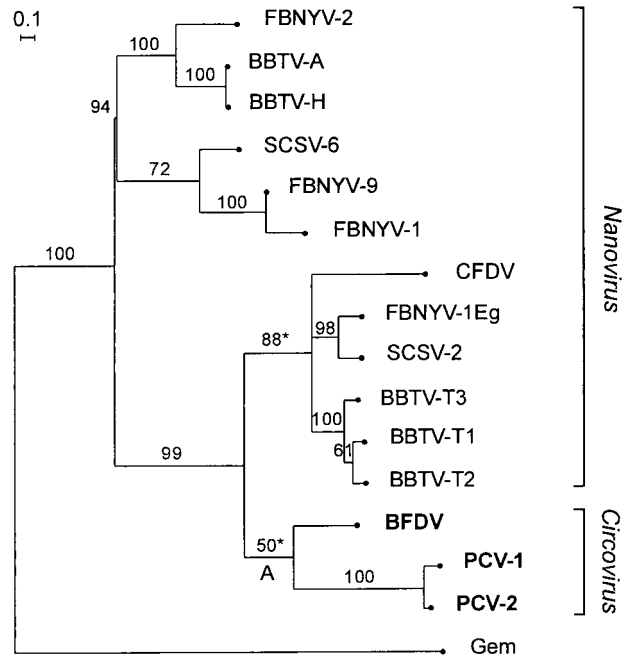

FIG. 3. A maximum likelihood tree for the N-terminal region of the available nanovirus and circovirus Rep amino acid sequences (up to position 129, see Fig. 1). The equivalent sequences from seven geminiviruses were used as an out-group (marked "Gem") to root the tree. Sequences: BBTV-A and -H, *Banana bunchy top nanovirus* isolates from Australia and Hawaii; BBTV-T1, -T2, and -T3, Taiwanese isolates DNAs 1, 2, and 3; BFDV, *Psittacine beak and feather disease circovirus*; CFDV, *Coconut foliar decay nanovirus*; FBNYV-1, -2, -9, and -1Eg, *Faba bean necrotic yellows nanovirus* components from isolates from Syria and Egypt; PCV-1 and -2, *Porcine circovirus* types 1 and 2; SCSV-2 and -6, *Subterranean clover stunt nanovirus* components 2 and 6. Bootstrap values are percentages from 10,000 neighbor-joining trees inferred from the amino acid sequences. Asterisks mark branches not found in the maximum likelihood or most parsimonious trees inferred from nucleotide sequences. Note that several nanovirus isolates have two or more distinct Rep genes carried by different genomic molecules.

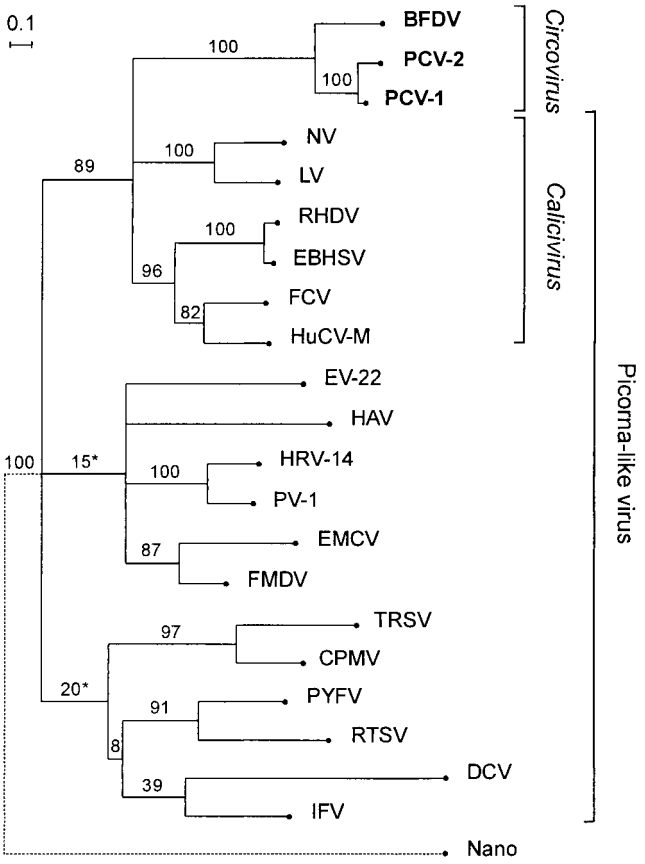

FIG. 4. A maximum likelihood tree for the 2C-protein amino acid conserved sequences of picorna-like viruses and the equivalent region from the circovirus Reps (from position 178 to 313; see Fig. 1). The equivalent C-terminal regions of six nanovirus Reps were included in the analysis and these sequences are represented by the node marked "Nano." The actual estimate for the length of the branch leading to the "Nano" node is double that shown. Sequences: BFDV, *Psittacine beak and feather disease circovirus*; CPMV, *Cowpea mosaic comovirus*, DCV, *Drosophila virus C*; EBHSV, *European brown hare syndrome calicivirus*; EMV, *Encephalomyocarditis cardiovirus*; EV-22, *Echovirus* 22; FCV, *Feline calicivirus*; FMDV, *Foot and mouth disease aphthovirus*; HAV, *Hepatitis A hepatovirus*; HRV-14, *Human rhinovirus 14*; HuCV-M, *Human calicivirus* Manchester isolate; IFV, *Infectious flacherie virus*; LV, *Lordsdale calicivirus*; NV, *Norwalk calicivirus*; PCV-1 and -2, *Porcine circovirus* types 1 and 2; PV-1, *Poliovirus* 1; PYFV, *Parsnip yellow fleck sequivirus*; RHDV, *Rabbit hemorrhagic disease virus*; RTSV, *Rice tungro spherical waikavirus*; TRSV, *Tobacco ringspot nepovirus*. Bootstrap values are percentages from 10,000 neighbor-joining trees inferred from the amino acid sequences. Asterisks mark branches not found in the maximum likelihood or most parsimonious trees inferred from nucleotide sequences.

Earlier analyses had placed the circovirus sequences as a sister-group to the nanovirus sequences. However, only one of the two nanovirus Rep clusters seems to have been represented in the first of these analyses (15), which could explain the discrepancy. Moreover, complete Rep sequences were used in the second analysis (17) and probably also in the first analysis. When we used complete Rep sequences, we too found that the circovirus and nanovirus sequences were placed as sister-groups; hence, errors were probably introduced when the different phylogenetic signals from the two parts of the protein were not identified and delineated in the earlier analyses.

Circovirus sequences were grouped with sequences from caliciviruses in all the trees found for Rep C-terminal and 2C-protein sequences (Fig. 4). This clustering suggests that the ancestral circovirus acquired its Rep gene 3′-sequence from an as-yet-uncharacterized lineage of calicivirus. The trees thus imply that this part of the circovirus DNA genome was acquired from a virus with an RNA genome. However, given the uncertainty about the root of the trees, it was important to consider whether the shared sequence was carried originally by an ancestral circovirus and was transferred to a picorna-like virus by recombination. We ruled out this possibility because it would require an additional major interspecies recombinational event to explain the relationship between circoviruses and nanoviruses. Furthermore, the trees show that 2C-protein genes are far more diverse than the equivalent circovirus sequences, confirming that the 2C-protein genes had an earlier origin.

Transfer of a Replicating Plant Virus DNA

In both circovirus and nanovirus DNA, the origin of replication (*ori*) is adjacent to the N-terminal part of the Rep gene (15, 37). This proximity and the similarities between the *ori* sequences of circoviruses and nanoviruses indicate that these sequences evolved from a common ancestral sequence, and, more importantly, that the nanovirus DNA that was transferred from a plant to a vertebrate included the *ori* sequence. In the genomes of both kinds of virus, *ori* consists of a conserved stem–loop and other less well-conserved sequences, including direct or inverted repeats (15, 19, 37). The dimensions of the stem–loop are relatively consistent, and its com-

position has been partly conserved. The stem is eight to 11 bp long and the 5′ side of the stem is guanine-rich. The loop consists of 10 to 13 bases and the nonanucleotide sequence forms part of the loop. Almost all of the circovirus DNAs have the nonanucleotide sequence 5′-TAGTATTAC, and the nanovirus DNAs that form the sister-group to the circoviruses in the N-terminal sequence tree (Fig. 3) have an identical nonanucleotide sequence. Only one of the PCV strains from this cluster has a slightly different nonanucleotide sequence (i.e., 5′-AAGTATTAC).

The similarity between circovirus and nanovirus *ori* sequences shows that the *ori* was conserved after it was transferred from a plant to a vertebrate, but the *ori* sequence does not code for a protein. This conservation can be explained only if the *ori* maintained its function as a nucleotide sequence. Rep specifically binds, cleaves, and ligates DNA at conserved sequences within the *ori* (18, 35, 38, 39); hence, the conservation of *ori* indicates that some or all of these activities were maintained. All of these replication-related activities of geminivirus Reps have been mapped to the N-terminal 130 amino acids in these proteins (39), and, given the homology, it is reasonable to assume that the same region in nanovirus and circovirus Reps performs these functions. Amino acid residue 130 in the geminivirus Rep aligns with the residues at position 132 in the nanovirus and circovirus sequence alignment (Fig. 1), which is on the N-terminal side of the recombination site. Thus, the protein encoded by the transferred nanovirus DNA probably retained all of the activities associated with the *ori*, which included almost all the processes necessary for replication; the *ori* was maintained because these activities were maintained. Rep and *ori* are the only components that are essential for the replication of these DNAs (38, 39); the host supplies the other components, including a DNA polymerase.

The only component of the nanovirus replication system that is not present in circoviruses is the 3′ part of the Rep gene that encodes the P-loop. The amino acid sequence that includes the P-loop is required for geminivirus replication, and, because nanovirus Reps have that sequence, they probably also require it; assuming that they do, the transferred nanovirus DNA probably included the sequence encoding the P-loop region. It is most likely that the recombinational event that replaced this sequence in the DNA with part of a 2C-protein gene, also encoding a P-loop, occurred sometime before or after the transfer from a plant to a vertebrate. The possibility that the transfer and the recombinational event were simultaneous is a product of the probabilities of each of the events, which is far lower than the probability that the events occurred at different times. Therefore, we conclude that when the nanovirus DNA was transferred to a vertebrate, its sequence was maintained and it survived because it was complete and could replicate. If the recombinational event occurred after the transfer, it may have improved the fitness of the DNA by improving one of the enzymatic activities that contributed to its replication.

The ability of the transferred nanovirus DNA to induce its own replication is significant because it can be directly related to the definition of a virus. Viruses are acellular parasites with nucleotide genomes that encode at least one protein involved in their own replication and that, once in a host cell, can induce their own replication. It is not essential that a parasite produce particles or has the ability to be transmitted horizontally for it to be recognized as a virus (5). Thus, the transferred nanovirus DNA was a minimal virus, and its transfer from a plant to a vertebrate represents a host-switch. The host-switch was significant because it established a completely new virus lineage.

Apart from the Rep gene, ORF C1 is the only gene conserved in the genomes of both BFDV and PCV. It is possible that ORF C1 also came from the progenitor nanovirus, but we found no significant similarities between the amino acid sequence encoded by ORF C1 and any protein sequence in the database including nanovirus sequences. The ORF C1 sequences from BFDV, PCV-1, and PCV-2 differ much more than the Rep sequences of these viruses, indicating that ORF C1 has changed more rapidly than Rep; it is therefore possible that the phylogenetic signal that could link this ORF with a nanovirus gene has been erased. Each nanovirus gene is encoded on a distinct DNA circle, whereas the Rep gene and ORF C1 are carried on the same circovirus DNA. Thus, if ORF C1 originated in the progenitor nanovirus, it was probably incorporated into the Rep-encoding DNA by recombination. It is also possible that this gene was incorporated by recombination from the genome of another virus or a host.

Timing and Mechanism

All characterized caliciviruses infect vertebrates (21). Thus, the recombinational event probably took place in a vertebrate and the host-switch event probably took place before the recombination. Caliciviruses do not have a DNA stage in their replication cycle, nor do nanoviruses have an RNA stage; and neither kind of virus encodes a reverse transcriptase. Therefore, another agent, possibly a retrovirus or retrotransposon, must have contributed a reverse transcriptase for the recombination to take place. Most vertebrates carry retroviruses, but plant retroviruses are relatively rare. Hence, the requirement for reverse transcription supports the case for recombination in a vertebrate. It is possible the 2C-protein gene was copied into cDNA and incorporated into the circovirus genome by the same reverse transcriptase. Alternatively, the gene may have been incorporated in a second reaction. Rep may have cut and joined the sequences or there may have been a retrovirus intermediate.

The host-switch event occurred during the evolution of the lineage represented by the branch marked "A" in the Rep N-terminal tree (Fig. 3). This branch links the circovirus group to the rest of the tree. Viruses on one side of branch A infect vertebrates and, on the other side, they infect plants. The recombinational event can also be mapped to branch A. There is no evidence that could be used to estimate the length of time represented by branch A, and, because the trees show that the viruses are not evolving at a constant rate, it could be argued that branch A represents a long period. However, two things suggest that the opposite is true. First, the branch is relatively short. In the maximum likelihood tree inferred from amino acid sequences, in which branch A is the longest, it is only 20% of the distance from the root of the nanovirus–circovirus cluster to the nearest terminal circovirus node. Second, the virus lineage was probably evolving relatively rapidly during the time represented by branch A, because major changes occurred in the biology of this lineage, but no equivalently significant changes occurred in other lineages. Viruses from the ancestral lineage represented by branch A would have experienced dramatic changes in selection when the host-switch and recombinational events occurred, and when they invaded naive vertebrate populations. These changes in selection would have been translated into a higher mutation-fixation rate. If so, the relatively small number of substitutions represented by branch A must have occurred in a short time, and the recombination event must have occurred soon after the host-switch.

It is unclear how the ancestral circovirus switched to infect a vertebrate. There may have been several intermediate stages, in which a virus infected a plant-feeding arthropod was transmitted between arthropod species to one that fed on vertebrates, and was then transmitted to a vertebrate. However, because branch A probably represents a short period, there was probably little time available for these intermediate stages. More importantly, only aphids (*Aphididae*) or planthoppers (*Cixiidae*) naturally transmit nanoviruses; these insects feed exclusively on plants, and experiments suggest that nanovi-

ruses do not infect their insect vectors (40, 41). Therefore, it is unlikely that a nanovirus was transmitted between arthropod species. Instead, we suggest that the virus switched hosts when a vertebrate was exposed to sap from a nanovirus-infected plant, either through a wound or on ingestion.

Recombinant Protein Function and Structure

It is remarkable that the Rep of the recombinant ancestral virus was functional after a major C-terminal part of the protein had been replaced with part of a 2C-protein. 2C-proteins are involved in viral RNA replication and have little in common with Reps in terms of function (42, 43), except that both kinds of protein hydrolyze ATP (43, 44). In both kinds of protein, this activity is associated with the P-loop, and in geminivirus Reps, it is linked to replication. Hence, by replacing the P-loop and surrounding sequence in the ancestral circovirus Rep with an equivalent sequence from a 2C-protein, the essential ATPase function was probably preserved. However, simply to combine the same set of enzymatic activities was probably not sufficient, because some Rep functions may depend on the structure of the C-terminal region of the protein and its interactions with the N-terminal part. To test the possibility of structural similarity between the nanovirus Rep C-terminal region and calicivirus 2C-proteins, we mapped onto an alignment the predicted positions of α-helices and β-strands. Predicted positions of these structural elements in the two kinds of protein correlated from approximately position 165 (Fig. 1), suggesting that the polypeptides form equivalent structures (Fig. 2, *B* and *C*; β-strand $\chi^2 = 35$, $P < 0.001$; α-helix $\chi^2 = 71$, $P < 0.001$). Given some structural similarity, the recombinational event probably did not radically affect interactions between the N- and C-terminal parts of Rep. This set of circumstances could explain the maintenance of Rep functions, and, as the recombinant ancestral virus would not have been viable if the N- and C-terminal parts of the recombinant Rep were incompatible, it partly explains the survival of the virus after the recombinational event.

We thank Adrian Gibbs, Bryan Harrison, Edward Holmes, John Trueman and an anonymous reviewer for their comments on the paper. The work was funded by an Australian Commonwealth Government block grant to the Australian National University.

1. Morse, S. S. (1994) in *The Evolutionary Biology of Viruses,* ed. Morse, S. S. (Raven, New York), pp. 325–335.
2. Sharp, P. M., Robertson, D. L. & Hahn, B. H. (1995) *Philos. Trans. R. Soc. London B* **349,** 41–47.
3. Webster, R. G., Bean, W. J. & Gorman, O. T. (1995) in *Molecular Basis of Virus Evolution*, eds. Gibbs, A. J., Calisher, C. H. & Garcia-Arenal, F. (Cambridge Univ. Press, Cambridge, U.K.), pp. 531–543.
4. Gibbs, A. J., Keese, P. L., Gibbs, M. J. & Garcia-Arenal, F. (1998) in *Origin and Evolution of Viruses*, eds. Domingo, E., Webster, R. & Holland, J. (Academic, London), pp. 263–285.
5. Murphy, F. A., Fauquet, C. M., Bishop, D. H. L., Ghabrial, S. A., Jarvis, A. W., Martelli, G. P., Mayo, M. A. & Summers, M. D. (1995) *Virus Taxonomy: Classification and Nomenclature of Viruses* (Springer, Vienna), p. 586.
6. Lai, M. M. C. (1995) in *Molecular Basis of Virus Evolution*, eds. Gibbs, A. J., Calisher, C. H. & Garcia-Arenal, F. (Cambridge Univ. Press, Cambridge, U.K.), pp. 119–132.
7. Gibbs, M. J., Armstrong, J., Weiller, G. & Gibbs, A. J. (1997) in *Potential Ecological Impact of Transgenic Plants Expressing Viral Sequences*, eds. Balázs E. & Tepfer M. (Springer, Berlin), pp. 1–19.
8. Johnson, G. P., Goebel, S. J. & Paoletti, E. (1993) *Virology* **196,** 381–401.
9. Gorbalyena, A. E. (1992) *Semin. Virol.* **3,** 359–371.
10. Zhou, X., Liu, Y., Calvert, L., Munoz, C., Otim-Nape, G. W., Robinson, D. J. & Harrison, B. D. (1997) *J. Gen. Virol.* **78,** 2101–2111.
11. Zhou, X., Liu, Y., Robinson, D. J. & Harrison, B. D. (1998) *J. Gen. Virol.* **79,** 915–923.
12. Morse, M. A., Marriot, A. C. & Nuttall, P. A. (1992) *Virology* **186,** 640–646.
13. Botstein, D. A (1980) *Ann. N. Y. Acad. Sci.* **354,** 484–490.
14. Meehan, B. M., Creelan, J. L., McNulty, M. S. & Todd, D. (1997) *J. Gen. Virol.* **78,** 221–227.
15. Niagro, F. D., Forsthoefel, A. N., Lawther, R. P., Kamalanathan, L., Ritchie, B. W., Latimer, K. S. & Lukert, P. D. (1998) *Arch. Virol.* **143,** 1723–1744.
16. Rohde, W., Randles, J. W., Langridge, P. & Hanold, D. (1990) *Virology* **176,** 648–651.
17. Katul, L., Timchenko, T., Gronenborn, B. & Vetten, H. J. (1998) *J. Gen. Virol.* **79,** 3101–3109.
18. Hafner, G. J., Stafford, M. R., Wolter, L. C., Harding, R. M. & Dale, J. L. (1997) *J. Gen. Virol.* **78,** 1795–1799.
19. Mankertz, A., Persson, F., Mankertz, J., Blaess, G. & Buhk, H.-J. (1997) *J. Virol.* **71,** 2562–2566.
20. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
21. Clarke, I. N. & Lambden, P. R. (1997) *J. Gen. Virol.* **78,** 291–301.
22. Gorbalenya, A. E., Koonin, E. V. & Wolf, Y. I. (1990) *FEBS Lett.* **262,** 145–148.
23. Saraste, M., Sibbald, P. R. & Wittinghofer, A. (1990) *Trends Biochem. Sci.* **15,** 430–434.
24. Dayhoff, M. O., Barker, W. C. & Hunt, L. T. (1983) *Methods Enzymol.* **91,** 524–545.
25. Goldbach, R. & de Haan, P. (1994) in *The Evolutionary Biology of Viruses,* ed. Morse, S. (Raven, New York), pp. 105–119.
26. Zanotto, P. M. d. A., Gibbs, M. J., Gould, E. A. & Holmes, E. C. (1996) *J. Virol.* **70,** 6083–6096.
27. Barton, G. J. (1996) in *Protein Structure Prediction: A Practical Approach*, ed. Sternberg, M. J. E. (IRL Press at Oxford Univ. Press, Oxford, U.K.), pp. 31–63.
28. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
29. Weiller, G. F. (1998) *Mol. Biol. Evol.* **15,** 326–335.
30. Felsenstein, J. (1981) *J. Mol. Evol.* **17,** 368–376.
31. Strimmer, K. & von Haeseler, A. (1996) *Mol. Biol. Evol.* **13,** 964–969.
32. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.
33. Fitch, W. M. (1971) *Syst. Zool.* **20,** 406–416.
34. Boevink, P., Chu, P. W. G. & Keese, P. (1995) *Virology* **207,** 354–361.
35. Laufs, J., Traut, W., Heyraud, F., Matzeit, V., Rogers, S. G., Schell, J. & Gronenborn, B. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 3879–3883.
36. Rost, B. & Sander, C. (1994) *Proteins* **19,** 55–72.
37. Katul, L., Timchenko, T., Gronenborn, B. & Vetten, H. J. (1998) *J. Gen. Virol.* **79,** 3101–3109.
38. Mankertz, A., Mankertz, J., Wolf, K. & Buhk, H-J. (1998) *J. Gen. Virol.* **79,** 381–384.
39. Orozco, B. M. & Hanley-Bowdoin, L. (1998) *J. Biol. Chem.* **273,** 24448–24456.
40. Katul, L., Maiss, E., Morozov, S. Y. & Vetten, H. J. (1997) *Virology* **233,** 247–259.
41. Hu, J. S., Wang, M., Sether, D., Xie, W. & Leonhardt, K. W. (1996) *Ann. Appl. Biol.* **128,** 55–64.
42. Echeverri, A. C. & Dasgupta, A. (1995) *Virology* **208,** 540–553.
43. Rodriguez, P. & Carrasco, L. (1995) *J. Biol. Chem.* **270,** 10105–10112.
44. Desbiez, C., David, C., Mettouch, A., Laufs, J. & Gronenborn, B. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 5640–5644.