The genomes of several plant species contain endogenous geminiviral sequences

Denis Filloux[1], Sasha Murrell[2,3], Maneerat Koohapitagtam[1,4], Michael Golden[2], Charlotte Julian[1], Serge Galzi[1], Marilyne Uzest[1], Marguerite Rodier-Goud[5], Angélique D'Hont[5], Marie-Stéphanie Vernerey[1], Paul Wilkin[6], Michel Peterschmitt[1], Stephan Winter[7], Ben Murrell[2,8], Darren P. Martin[2], and Philippe Roumagnac[1]

[1] CIRAD-INRA-SupAgro, UMR BGPI, Campus International de Montferrier-Baillarguet, 34398 Montpellier Cedex-5, France
[2] Computational Biology Group, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town 4579, South Africa
[3] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA
[4] Department of Pest Management, Faculty of Natural Resources, Prince of Songkla University, Hat Yai campus, Thailand 90120
[5] CIRAD, UMR AGAP, TA A-108/03, Avenue Agropolis, F-34398 Montpellier Cedex 5, France
[6] Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AB, UK
[7] DSMZ Plant Virus Department, Messeweg 11/12, 38102, Braunschweig, Germany
[8] Department of Medicine, University of California, San Diego, La Jolla, CA

Endogenous viral sequences are essentially 'fossil records' that can sometimes reveal the genomic features of long extinct virus species. Although numerous known instances exist of single-stranded DNA (ssDNA) genomes becoming stably integrated within the genomes of bacteria and animals, there remain very few examples of such integration events in plants. The best studied of these events are those which yielded the geminivirus-related DNA elements (GRD) and the geminivirus-like elements (EGV) found respectively within the nuclear genomes of several *Nicotiana* species and various *Dioscorea* spp. of the *Enantiophyllum* clade.

Those two new classes of endogenous plant virus sequence are apparently derived from ancient geminiviruses in the genus *Begomovirus*. GRD and EGV sequences likely became integrated millions years ago. Interestingly, we found evidence of natural selection actively favouring the maintenance of EGV-expressed replication-associated protein (Rep) amino acid sequences, which clearly indicates that functional EGV Rep proteins were probably expressed for prolonged periods following endogenization.

We recently found using *in silico* searches that other ssDNA virus-like sequences are included within complete or draft genomes of various plant species, including apple tree (*Malus domestica*), black cottonwood (*Populus trichocarpa*), various *Coffea* spp, eggplant (*Solanum melongena*), lettuce (*lactuca sativa*), and Tepary bean (*Phaseolus acutifolius*), which suggests that endogenous geminiviruses may be more common in plant genomes than has previously been appreciated.