Sommaire

Introduction

1. Contexte général

Depuis les premières descriptions du SIDA au début des années 1980, de nombreuses études ont été réalisées sur les causes et les conséquences de l'émergence ou de la ré-émergence des maladies virales. Les virus émergents sont définis comme ceux qui sont récemment apparus ou ceux dont les populations ont récemment augmenté en prévalence, en pathogénicité et/ou en répartition géographique (Holmes and Rambaut, 2004). Ce phénomène a été mis en avant par les récentes ou plus anciennes épidémies virales telles que celles associées aux maladies émergentes de l'Homme (Grippe Espagnole 1918, Vana and Westover, 2008; SRAS, Holmes and Rambaut, 2004; maladie du Nil occidental, Nosal and Pellizzari, 2003; dengue, Gubler, 2007 et chikungunya, Schuffenecker et al., 2006), des animaux (fièvre aphteuse, Grubman and Baxt, 2004) et des plantes (enroulement foliaire et jaunissement de la tomate, Rybicki et al., 2000; Moriones and NavasCastillo, 2000; mosaïque du manioc, Legg and Fauquet, 2004). Pourtant, malgré les importants efforts de recherche visant à identifier les forces motrices liées à l'émergence, il reste difficile d'expliquer pourquoi et comment de nouvelles maladies virales continuent à apparaître régulièrement aussi bien dans le monde animal que végétal. De grands progrès restent à être accomplis pour ce qui est un des buts premiers de la recherche sur les virus émergents, à savoir prédire quels sont les virus les plus susceptibles d'apparaître à l'avenir, leurs conséquences épidémiologiques et les enrayer.

L'émergence virale est généralement associée à des facteurs écologiques et moléculaires aboutissant à de nouvelles interactions virus - vecteur - plante - environnement (Fargette et al., 2006). Pour ce qui est du facteur moléculaire, le passage à un nouvel hôte semble être l'un des éléments déterminants (Fargette et al., 2006; Holmes and Drummond, 2007; Holmes and Rambaut, 2004). L'interaction hôte virus est souvent mal comprise et difficile à décrire, mais requiert une adaptation du virus pour que celui-ci puisse réaliser une infection. La compréhension fine de l'évolution comme un des facteurs régissant la genèse d'une population virale, représente une étape essentielle à la compréhension des phénomènes d'émergence.

Au niveau de l'hôte, différentes espèces, ou individus d'une espèce, peuvent présenter des sensibilités différentes à une infection virale. Au niveau du virus, différents variants viraux peuvent apparaîtrent au sein d'une espèce et différer dans leur capacité à reconnaître les récepteurs cellulaires d'une nouvelle espèce hôte (Baranowski et al., 2001) ou dans leur capacité à être transmis avec succès entre les individus de la nouvelle espèce hôte. Dans le contexte évolutif où la survie d'un virus dépendrait de sa capacité à s'adapter à un nouvel hôte, il est

cohérent de penser que le virus génétiquement le plus variable, et donc le plus adaptable, aura le plus de facilité à franchir la barrière des espèces et à établir une infection (Woolhouse et al., 2001). Néanmoins, en terme de génétique des populations, le virus a besoin pour passer d'un pic de *fitness*[1] à un autre (saut d'espèce), de traverser une vallée de faible *fitness* où le virus est vraisemblablement mal adapté aux deux espèces en question (Poelwijk et al., 2007).

La capacité du virus à s'adapter à un hôte est conditionnée par sa variabilité génétique, qui est liée à la fois à son taux de mutation et de recombinaison. La mutation est le mécanisme évolutif par lequel du polymorphisme génétique est créé à des sites individuels du génome, tandis que la recombinaison est le mécanisme par lequel le polymorphisme est redistribué entre individus en formant de nouvelles combinaisons à partir de la variabilité préexistante. Malgré l'importance de ces deux mécanismes fondamentaux dans l'évolution du vivant en général et dans le contexte plus réduit de l'évolution virale, leurs conséquences au niveau de la *fitness* sont encore très mal comprises (Betancourt and Bollback, 2006; Otto and Gerstein, 2006).

La plupart des recherches *in natura* en évolution virale ont porté sur l'évolution de virus à ARN, qui représentent les plus importants agents pathogènes viraux humains. Au contraire chez les plantes, la plupart des émergences virales sont liées aux virus à ADN simple brin (ADNsb) du genre *Begomovirus* transmis par aleurode (Seal et al., 2006). Malgré l'utilisation des polymérases cellulaires de leur hôte, ayant un niveau de fidélité élevé, ces virus présentent un taux élevé de mutation et de recombinaison et affichent à l'échelle de la population ou de l'hôte des niveaux de diversité comparables à ceux de certains virus à ARN (Ge et al., 2007; Isnard et al., 1998; Duffy and Holmes, 2008). Le potentiel important d'évolution de ces virus ADNsb, l'existence de biotypes invasifs de leur insecte vecteur et leur prévalence en augmentation permanente font de ces virus un modèle idéal pour la description et la compréhension de l'émergence virale et des mécanismes sous-jacents.

2. Le couple *Begomovirus – Bemisa tabaci*

a. Origine des virus à ADN circulaire simple brin

Les virus à ADN circulaire simple brin comprennent plusieurs familles virales infectant animaux (*Parvoviridae*, *Anellovirius*[2], *Circoviridae*), végétaux (*Geminiviridae*, *Nanoviridae*) et bactéries (*Microviridae*, *Inoviridae*). Bien que la

[1] La Fitness correspond au nombre de descendants viables et fertiles que produit en moyenne chaque individu d'un génotype donné. Par commodité et usage, le terme anglo-saxons *fitness* est préféré ici au terme français de "valeur sélective".

[2] le genre *Anellovirus* n'est à ce jour rattaché à aucune famille virale.

gamme d'hôtes et la taille des génomes soient très différentes d'une famille à une autre (Figure 1), une origine commune pour certaines protéines d'une partie de ces virus est suspectée (Gibbs et al., 2006; Vega-Rocha et al., 2007, Koonin and Ilyina, 1992). Certains auteurs proposent que les géminivirus aient évolué à partir de composants extra chromosomiques de cellules procaryotes ou des premières ébauches de cellules eucaryotes. La très grande homologie de structure et de fonction de la protéine associée à la réplication (Campos-Olivas et al., 2002; Vega-Rocha et al., 2007), ainsi que certains caractères procaryotiques des géminivirus modernes supportent cette hypothèse (voir Rojas et al., 2005 pour revue). La présence d'une importante flore bactérienne symbiotique chez les insectes piqueurs suceurs, intervenant dans la gamme d'hôte, l'écologie et l'évolution (Chiel et al., 2007), et l'association intime entre virus, insectes vecteurs et ces symbiontes lors de la transmission circulante (Morin et al., 1999), semblent aussi avoir offert des opportunités d'acquisition de matériel génétique d'origine procaryotique (Czosnek et al., 2001).

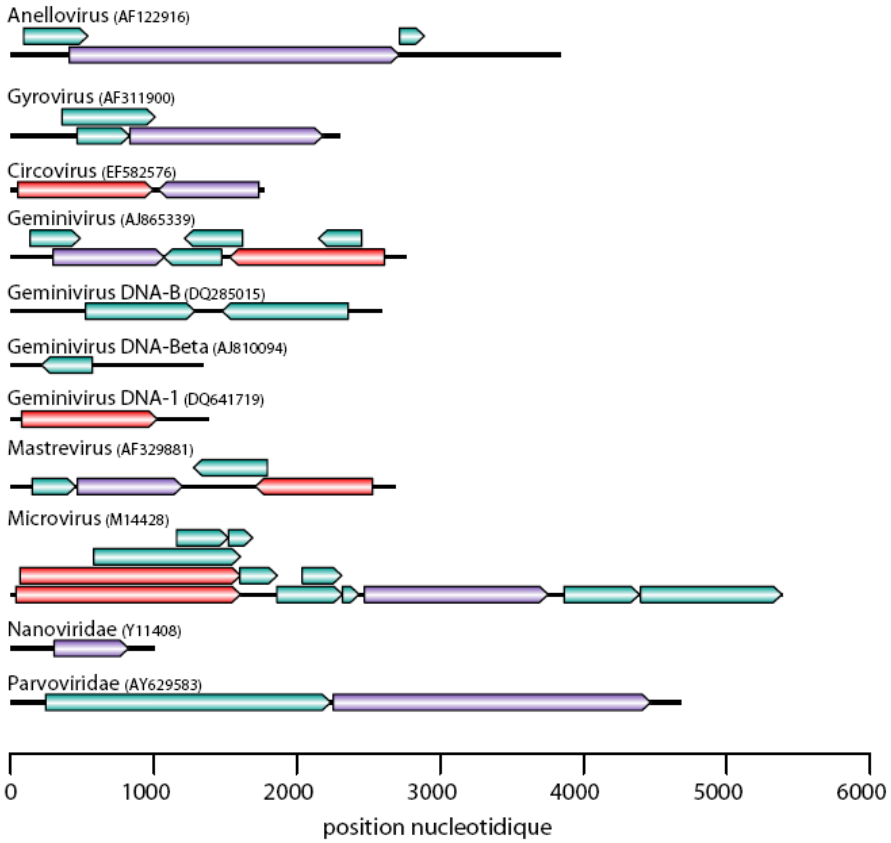


Figure 1 : Organisation génomique des principales familles et principaux genres de virus à ADN circulaire simple brin. Les flèches représentent les cadres ouverts de lecture (open reading frame ou ORF), avec en violet les ORFs codant pour une protéine de capside et en rouge les ORFs codant pour une protéine associée à la réplication. Pour chaque genre viral, les positions relatives des ORFs sont tracées en fonction de la séquence virale dont le numéro d'accession est donné entre parenthèses.

Les particules virales des virus à ADNsb sont de forme icosaédrique, avec deux exceptions : la famille des *Inoviridae* présentant une particule filamenteuse et la famille des *Geminiviridae* qui comportent des particules virales icosaédriques associées en doublets. L'architecture unique des particules de géminivirus semble être directement dépendante de la taille des génomes encapsidés. En effet, des ADN sub-génomiques d'environs la moitié de la taille du génome complet du *Maize streak virus* (MSV, 2.7kb) ont été associés à des particules simples (Casado et al., 2004), alors que des génomes de taille anormalement élevée de *l'African cassava mosaic virus* (ACMV) ont été associés à des particules virales en triplets (Frischmuth et al., 2001). Au-delà de soulever l'étonnement, cette curiosité structurale souligne une nouvelle fois le passionnant et tortueux parcours évolutif de l'ensemble de ces virus à ADN circulaire simple brin.

b. La famille des Geminiviridae

Au sein de la famille des géminivirus, quatre genres (Figure 2) ont été décrits, parmi lesquels les *Mastrevirus* (infectant principalement des plantes monocotylédones et transmis par cicadelle) et les *Begomovirus* (infectant des dicotylédones et transmis par aleurode) sont les genres les mieux caractérisés (Pour revue Rojas et al., 2005). L'espèce type du genre *Begomovirus* est le *Bean golden mosaic virus*, virus de la mosaïque dorée du haricot (van Regenmortel et al., 2000). Les bégomovirus sont responsables de viroses sur de nombreuses cultures d'importance économique dans le monde entier et particulièrement en zones tropicales, notamment sur le manioc (*Manihot esculenta*), le haricot (*Phaseolus vulgaris*), le coton (*Gossypium hirsutum*) et la tomate (*Solanum lycopersicum*).

c. Organisation génomique et fonctions des différentes protéines

Les bégomovirus sont majoritairement bipartites, c'est-à-dire que l'information génétique est portée par deux ADNs circulaires distincts (ADN A et ADN B). Quelques bégomovirus, principalement originaires de l'Ancien Monde, sont monopartites. Ils ne possèdent qu'un ADN A d'environ 2800 pb, c'est le cas notamment du *Tomato yellow leaf curl virus* (TYLCV; Antignus and Cohen, 1994) et des bégomovirus indigènes des îles du sud ouest de l'océan Indien, comme le *Tomato leaf curl Mayotte virus* (ToLCYTV; Delatte et al., 2005b). Plus récemment, des ADN satellites ont été mis en évidence en association avec certains bégomovirus (Briddon and Stanley, 2006; Dry et al., 1997; et pour revue, Stanley et al., 1997). Ces ADNs d'environs 1300 bases participent à l'infection virale, sans être pour autant nécessaires à celle-ci. Deux types majeurs sont pour le moment décrits, les ADN-Beta et les ADN-1. Les ADN-Beta semblent associés à la symptomatologie avec un possible rôle de suppression du post-transcriptional gene silencing (PTGS ; Vanderschuren et al., 2007; Briddon and Stanley, 2006; Vanitharani et al., 2005) tandis que les ADN-1, qui portent un gène codant pour une protéine associée à la réplication (protéine *Rep*) semblent avoir un rôle dans la modulation de l'accumulation du virus (Figure 3; Briddon and Stanley, 2006).

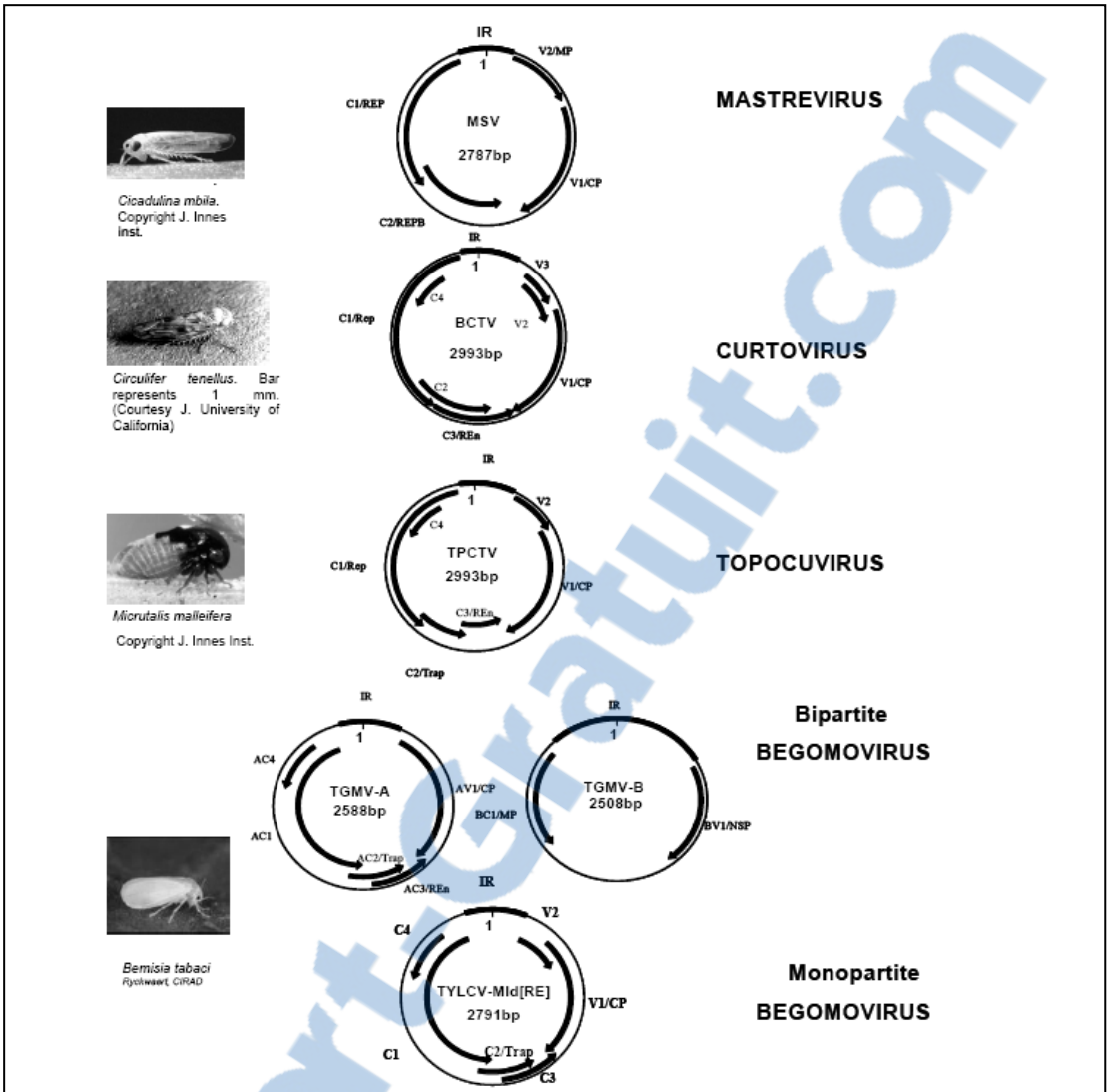


Figure 2 : Représentation des génomes d'un exemple type des différents genres de la famille des *Geminiviridae* et photos de leurs insectes vecteurs respectifs.

L'ADN A des bégomovirus monopartites présente six régions codantes ou ORFs pour « open reading frame » dont certaines sont chevauchantes et une zone intergénique (*intergenic region* ou IR). Les ORFs V1 et V2 sont codés par le brin viral alors que les ORFs C1, C2, C3 et C4 sont codés par le brin complémentaire (Figure 3). La tige boucle (5'-TAATATTAC-3') est une région conservée de l'IR chez tous les bégomovirus. Elle représente le point d'initiation de la réplication (Orozco and Hanley-Bowdoin, 1996).

L'ORF CP (V1) code pour la protéine de capside qui représente l'unité de base dans la constitution de la particule virale en doublet des géminivirus. La CP associée à la protéine de mouvement MP (V2, absente chez les bégomovirus bipartites) intervient dans les mécanismes de diffusion du virus dans la plante hôte. Lors de l'infection et de la réplication, l'ADN viral migre dans le noyau des cellules. Il doit traverser plusieurs barrières cellulaires avant d'accéder à

l'intérieur du noyau, pour s'y répliquer. C'est lors de cette migration que les protéines virales, notamment la protéine CP, sont indispensables pour protéger le matériel génétique (Rojas et al., 2001). Par ailleurs, la CP a un rôle essentiel lors du passage des particules virales à travers la paroi intestinale de l'insecte vecteur (Briddon et al., 1990). La liaison des particules virales avec des protéines chaperonnes de type GroEL, produite par des bactéries endosymbiontes, serait aussi primordiale pour empêcher la dégradation de ces particules (Morin et al., 2000).

La protéine associée à la réplication (Rep ou ORF C1) intervient lors de la multiplication virale. En association avec la protéine activatrice de la réplication REn (ORF C3), elle permet l'accumulation de l'ADN viral dans les cellules végétales (Hanley-Bowdoin et al., 2000). La protéine activatrice de la transcription TraP (ORF C2) contribue au pouvoir pathogène du virus et à l'activation de la transcription des ORFs (van Wezel et al., 2001). Par ailleurs, le rôle de cette protéine, en association ou non à la protéine C4, dans la suppression du mécanisme général de résistance des plantes aux virus par « *silencing* » (encore nommé VIGS pour *virus induced gene silencing*) a été démontré (Bisaro, 2006; Rojas et al., 2005). La protéine C4, codé par l'ORF C4, semble intervenir dans l'expression et la sévérité des symptômes, dans la gamme d'hôte (Krake et al., 1998; Latham et al., 1997), et dans le mouvement du virus (Jupin et al., 1994). L'expression de l'ORF C4 du *Tomato leaf curl virus* dans des plantes transgéniques de *Nicotiana benthamiana* a permis d'obtenir l'expression de symptômes de virose, fournissant une preuve supplémentaire de la participation de cet ORF dans l'expression des symptômes (Selth et al., 2004).

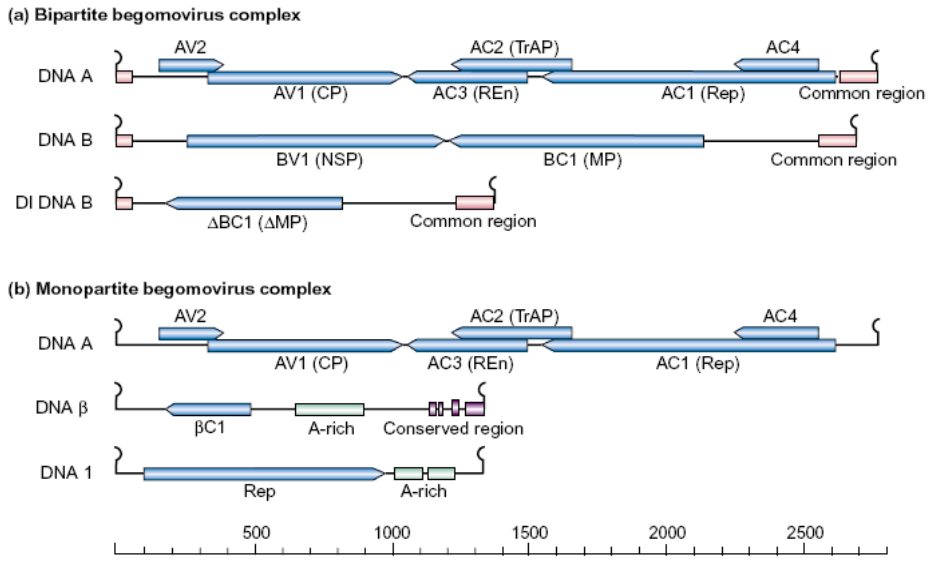


Figure 3 : Organisation génomique des bégomovirus bipartites (a) et monopartites (b) et des ADN satellites qui leur sont associés. D'après Mansoor *et al*. (2006).

L'ADN B d'environ 2600 pb des bégomovirus bipartites est divisé en deux régions codantes non chevauchantes, BV1 et BC1. Les protéines associées à ces ORFs interviennent dans le développement des symptômes et le transport de l'ADN viral dans les cellules de la plante. Elles interviennent plus particulièrement dans l'exportation de l'ADN viral hors du noyau et dans le transport inter-cellulaire via les plasmodesmes (Gafni and Epel, 2002).

d. Réplication et transcription

La réplication des bégomovirus a lieu au sein du noyau. A partir de l'ADN simple brin (forme encapsidée), la forme réplicative du virus (ADN double brin : ADNdb) est synthétisée par la polymérase cellulaire grâce à l'amorçage d'un brin d'ARN (Saunders et al., 1992). Cet ADNdb, en s'associant avec les histones de la cellule hôte, forme des minichromosomes viraux. Cette première étape est réalisée avec la seule utilisation des protéines de l'hôte (Pilartz and Jeske, 1992). Cet ADN est alors capable de se répliquer en cercle roulant (*Rolling Circle Replication* ou RCR, Figure 4) de manière analogue aux phages ADN circulaire simple brin (Novick, 1998).

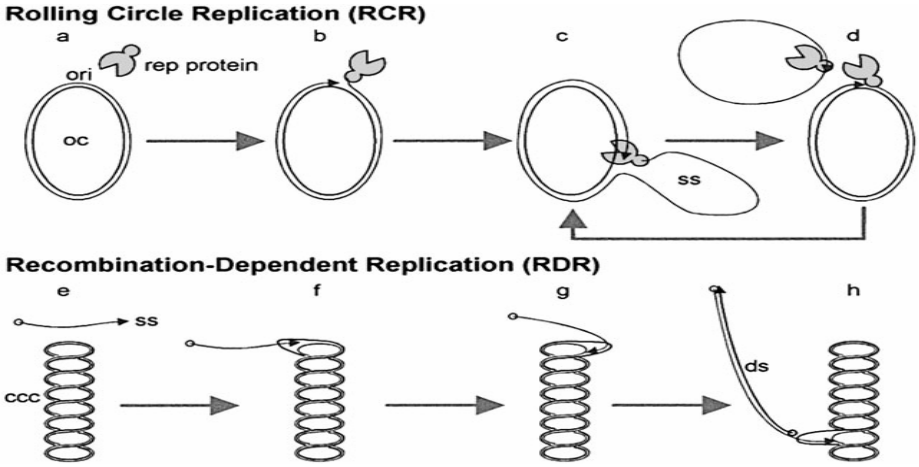


Figure 4 : Schématisation des modèles de réplication en cercle roulant (RCR) et de la réplication dépendante de la récombinaison (RDR). Etape a : accrochage de la protéine associée à la réplication (Rep) à l'origine de réplication (ori). Etape b : Ouverture de l'ADN et liaison covalente de la Rep à l'extrémité 5'. Etape c : déplacement et réplication. Etape d : Nouvelle ouverture de l'ADN, fermeture des ADNs simples brins et relargarge de la Rep. Etape e : interaction entre un ADN simple brin incomplet et la forme super-enroulée de l'ADN viral (cccDNA) à des sites homologues. Etape f : recombinaison homologue. Etape g : élongation de l'ADN simple brin. Etape h : synthèse de l'ADN complémentaire et obtention d'un ADN double

La protéine Rep seule est suffisante à l'initiation de la réplication en introduisant une ouverture (#) au sein de la séquence conservée de la tige boucle (TAATATT#AC; Hanley-Bowdoin et al., 2000). Ce modèle de réplication en cercle roulant a été confirmé par microscopie électronique (Jeske et al., 2001). Ces expériences ainsi que d'autres utilisant l'électrophorèse bidimensionnelle, ont par ailleurs permis l'identification d'intermédiaires additionnels de réplication

compatibles avec un modèle de réplication dépendante de la recombinaison (*Recombination-Dependent Replication* ou RDR, Figure 4), analogues à celui du bactériophage T4 (Mosig, 1998; Mosig et al., 2001). La transcription a aussi lieu dans le noyau de la cellule. Elle est bidirectionnelle depuis les séquences promotrices situées au sein de la zone intergénique pour tous les ORFs, excepté pour les ORFs C2 et C3 où la région promotrice est intégrée dans l'ORF C1. La transcription des bégomovirus est complexe, aboutissant fréquemment à des ARN messagers polycistroniques (Hanley-Bowdoin et al., 1989; Shivaprasad et al., 2005; Sunter and Bisaro, 1989).

e. L'insecte vecteur, Bemisia tabaci

Les bégomovirus sont transmis aux plantes par l'intermédiaire d'un insecte vecteur phloémophage, *Bemisia tabaci* (Hemiptera : Aleyrodidae) selon le mode circulant. Le virus, une fois ingéré par l'insecte, qui se nourrit essentiellement de sève élaborée, va pénétrer dans le tractus intestinal par l'intermédiaire du bol alimentaire. Il va ensuite traverser une première barrière qui est l'intestin pour se retrouver dans l'hémolymphe de l'insecte. Les particules virales vont ensuite rejoindre, s'associer puis traverser la deuxième barrière à la transmission que représentent les glandes salivaires. Le virus circulant est enfin relargué dans une nouvelle plante par salivation lors de l'alimentation de l'insecte (Czosneck, 2007).
B. tabaci a été décrit pour la première fois en 1889 en Grèce (Gennadius, 1889). Cette "mouche blanche" a été classée dans l'ordre des Hemiptera, la famille Aleyrodidae et la sous-famille Aleyrodinae. *B. tabaci* est un aleurode d'environ 1 à 1,5 millimètre de long, avec un corps blanc – jaunâtre et 2 paires d'ailes blanches disposées en forme de « toit » (Figure 5).

Figure 5 : Couple d'individus adultes du biotype B de *Bemisia tabaci* en alimentation sur une feuille de chou. Photo CIRAD.

B. tabaci est présent sur tous les continents excepté l'Antarctique (Perring, 2001). Du fait de sa plasticité écologique, la caractérisation taxonomique de l'espèce *B. tabaci* a toujours posé et pose encore de nombreuses difficultées. Pour

cette raison, les taxonomistes ont proposé d'associer cette espèce à un complexe de biotypes (Brown and Bird, 1995; Perring, 2001; Frohlich et al., 1999). L'analyse moléculaire de la diversité génétique de différentes populations mondiales de *B. tabaci* a montré l'existence de nombreux groupes biologiques, à partir desquels sept clades ont pu être définis (Boykin et al., 2007; Delatte et al., 2006). A partir des années 1990, le biotype B vraisemblablement originaire du Nord Est du continent Africain et du Moyen-Orient se serait rapidement propagé aux dépens des autres biotypes locaux (Brown and Bird, 1995). Ce biotype présente des caractéristiques biologiques particulières : (1) un taux de fertilité accru (Delatte et al., 2008; Bethke et al., 1991), (2) une mobilité qui lui permet de se déplacer à longue distance (Blackmer et al., 1995), (3) une gamme de plantes hôtes très importante (Brown and Bird, 1995) et (4) une résistance à certains insecticides (Brown and Bird, 1995; Horowitz et al., 2005). La pullulation d'autres biotypes de *B. tabaci* a également été associée à l'émergence de certaines maladies virales comme la mosaïque africaine du manioc en Afrique de l'Est (Figure 6).

Figure 6 : Pullulations d'aleurodes sur feuille de manioc (*Manihot esculenta*) directement associées à la dissémination du variant Ougandais impliquée dans l'épidémie sévère de la mosaïque du manioc en Afrique de l'Est (Legg et al. Molecular Ecology 2002). Photo James Legg.

La diversité génétique ou plasticité écologique importante de *B. tabaci* avec l'existence de nombreux biotypes lui permet d'accéder à une très large gamme de plantes hôtes. Plus de 900 espèces de plantes hôtes de *B. tabaci*, appartenant à 74 familles botaniques, ont été recensées. L'ensemble de ces espèces représente 73% des plantes cultivées (Servin et al., 1999). Ainsi, la récente propagation mondiale des biotypes B et Q de *B. tabaci* ayant des caractéristiques biologiques particulières semble avoir joué un rôle primordial dans l'émergence et les sauts d'espèces des bégomovirus. L'opportunité d'accéder à de nouvelles niches écologiques par le transport des virus d'une plante à une autre en conjonction

avec leur important potentiel d'évolution, représente probablement le trait écologique le plus important dans l'émergence des bégomovirus. De plus, l'augmentation des populations d'aleurode augmente le nombre de contacts possible entre virus et plantes, facilitant ainsi l'adaptation des virus à de nouveaux hôtes (Moriones and NavasCastillo, 2000, Rybicki and Pietersen, 1999).

3. Evolution et expansion des populations virales

Malgré leurs conditions de pathogène obligatoire, les virus ont montré une grande capacité à investir de nouvelles niches écologiques. L'adaptation à ces nouvelles niches est liée à leur capacité d'évolution dont les moteurs sont un taux de mutation très élevé et une grande capacité à échanger du matériel génétique par recombinaison.

a. La mutation

La mutation génère de la variabilité génétique, base sur laquelle l'adaptation à de nouveaux environnements est rendue possible. Les mutations sont générées lors de la réplication, par des erreurs de copie du génome par les polymérases ARN ou ADN dépendantes (réplicases). Les ARN polymérases ARN dépendantes qui répliquent les virus à ARN, ont un taux d'erreur compris entre 10^{-3} à 10^{-5} par nucléotide et par réplication (Jenkins et al., 2002). A l'inverse, les taux de mutation sont environ 1000 fois plus faibles pour les virus à ADN double brin de vertébrés qui dépendent des polymérases cellulaires pour leur réplication. Le taux de mutation de ces virus à ADN double brin est ainsi proche de celui des ADNs cellulaires estimé à 10^{-8} - 10^{-11} (Drake et al., 1998). Par contre, les virus à ADN simple brin d'animaux et de plantes, qui utilisent également les ADN polymérases à ADN cellulaires de leur hôte pour leur réplication, semblent présenter des taux de mutations et de substitutions[3] bien plus élevés que celui normalement engendré par ces enzymes (Figure 7; Duffy et al., 2008).

En effet, les travaux réalisés à la fois sur le *Maize streak virus* (*Geminiviridae, Mastrevirus*; Isnard et al., 1998) et sur le *Tomato yellow leaf curl China virus* (*Geminiviridae, Begomovirus*; Duffy and Holmes, 2008; Ge et al., 2007) ont montré des taux de mutation et de substitution anormalement élevés. Ces résultats restent totalement incompris sur un plan mécanistique. La seule protéine virale indispensable à la réplication du génome viral, est la Rep. Cette protéine multifonctionnelle interagit avec plusieurs protéines de l'hôte. Trois grandes catégories de gènes codant pour des protéines interagissant avec Rep

[3]Une des questions importante dans la compréhension de l'évolution des génomes est de savoir comment le taux de mutation généré par génome et par cycle de réplication est relié au taux de substitution au niveau de la population, c'est à dire au taux ou ce fixe les mutations dans les génomes. Quand les mutations sont neutres sélectivement, la relation entre taux de mutation et de substitution est directe. Par contre, des relations beaucoup plus complexes ont lieu lorsque la sélection naturelle opère ou bien lorsque les génomes expérimentent des variations de dynamique de réplication.

sont connues : (1) des gènes impliqués dans la division cellulaire et le développement (Kong and Hanley-Bowdoin, 2002; Selth et al., 2005; Xie et al., 1999), (2) des gènes impliqués dans le système de sumoylation à l'origine de modifications post-traductionnelles des protéines (Castillo et al., 2004), (3) des gènes impliqués dans la machinerie cellulaire de réplication de l'ADN (Castillo et al., 2003; Luque et al., 2002). Ces éléments suggèrent que le virus serait capable d'interférer dans de nombreux mécanismes de la cellule hôte et d'en détourner les fonctions à son avantage, avec une possible modification de la fidélité de la polymérase cellulaire.

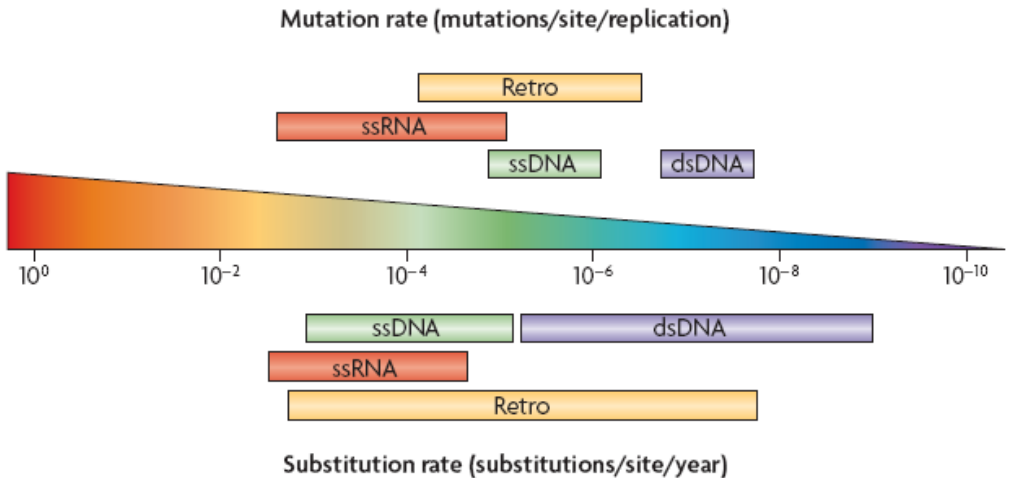


Figure 7 : Comparaison entre taux de mutation et de substitution des virus suivant la nature de leur génome. Les taux de mutations (partie haute du schéma) sont donnés en mutations par site par réplication tandis que les taux de substitutions (partie basse du schéma) sont donnés en substitution par site et par an. D'après Duffy et al., (2008).

b. La recombinaison

La recombinaison virale est un processus permettant la genèse d'un virus en créant de nouvelles combinaisons de séquences à partir de l'ADN de deux virus parents. La recombinaison présente l'énorme avantage pour un virus de permettre l'acquisition en une seule étape d'une grande variabilité génétique et ainsi de permettre de créer de nouveaux types d'arrangements au sein du génome. Conjointement, mutation et recombinaison peuvent potentiellement fournir un accès à un polymorphisme beaucoup plus importante que par mutation seule (Figure 8; Crameri et al., 1998; Stemmer, 1994). Dans le même temps, la recombinaison permet de diminuer la fréquence avec laquelle des mutations délétères s'accumulent dans la population.

En virologie, deux processus de réarrangement génomique différents peuvent être distingués : la pseudo-recombinaison et la recombinaison. La pseudo-recombinaison a lieu chez les virus multipartites et implique l'échange de composants génomiques intacts entre virus, aboutissant à la formation de

génomes présentant de nouvelles combinaisons de composants génomiques. La recombinaison implique quant à elle, l'échange de matériel génétique au sein d'un composant génomique. Néanmoins, que ce soit par pseudo-recombinaison ou recombinaison, le réarrangement de l'information génétique doit obligatoirement générer des génomes fonctionnels et raisonnablement adaptés.

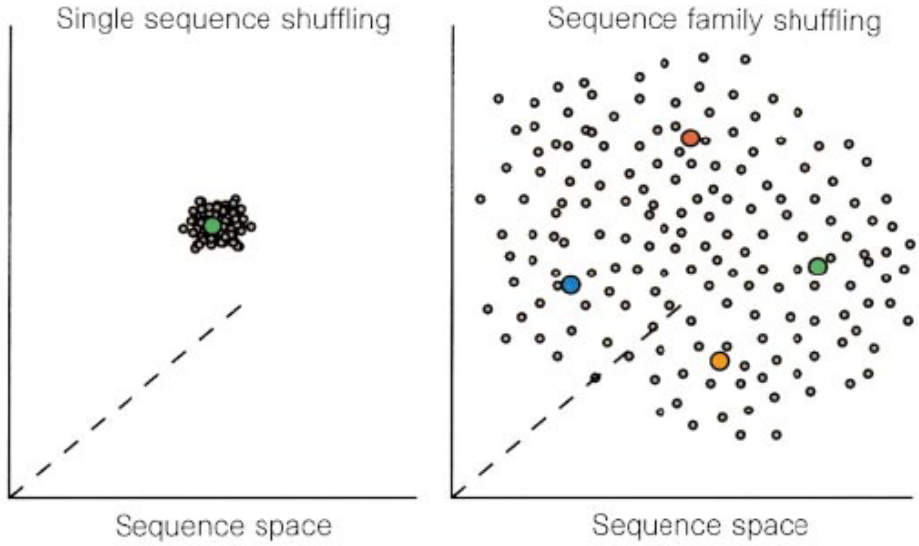


Figure 8 : Exploration de l'espace séquentiel par mutation (gauche) et recombinaison (droite). Pour des librairies de taille égale, la recombinaison permet l'exploration d'un espace de diversité plus important. D'après Crameri et al. (1998).

Depuis longtemps, le phénomène de recombinaison a été identifié chez les géminivirus avec des recombinaisons entre virus de même espèce (Zhou et al., 1997), d'espèces différentes (Briddon et al., 1996; Klute et al., 1996) et de genre différents (Saunders and Stanley, 1999). Suite à l'obtention d'un nombre croissant de séquences génomiques complètes et à l'apparition de méthodes efficaces de détection de la recombinaison, il a été montré que ce processus joue un rôle primordial dans la génération de la diversité chez les géminivirus (Padidam et al., 1999). De nombreux recombinants ont été décrits comme présentant une valeur sélective supérieure à celles de leurs parents. Pour le cas des bégomovirus, nous pouvons notamment citer en exemples (1) la souche sévère du TYLCV (TYLCV-IL), issu d'une recombinaison ancestrale entre un ancêtre des TYLCVs et un *Tomato leaf curl virus* asiatique, réputé plus agressif que la souche Mild du TYLCV (TYLCV-Mld), (2) le *Tomato leaf curl Malaga virus* (TYLCMalV; Monci et *al.*, 2002), issu d'une recombinaison entre le TYLCV-Mld et le TYLCSV, présente une gamme de plantes hôtes plus large, ou (3) le variant sévère Ougandais de l'*East African cassava mosaic virus* (EACMV-UG2), issu d'une recombinaison entre l'EACMV et l'ACMV, responsable d'une épidémie très sévère sur manioc en Afrique de l'Est avec des conséquences humanitaires et économiques dramatiques (Zhou et al., 1997).

Grâce à une meilleure description des évènements de recombinaison des séquences de géminivirus, les mécanismes aboutissant à leur émergence sont mieux compris. Le premier pas pour la genèse de recombinants est la coinfection de l'hôte par différents types de virus. De nombreuses études ont montré que ces coinfections étaient un phénomène fréquent *in natura* (Monci et al., 2002; SanchezCampos et al., 1999). Lors de coinfections expérimentales entre TYLCV et TYLCSV, il a aussi été montré que près de la moitié des noyaux de cellules infectées présentaient les deux virus coinoculés, pré-requis à la recombinaison (Morilla et al., 2004).

Dès lors, il est important d'essayer de comprendre pourquoi la recombinaison a pris tant d'importance dans l'évolution de ce genre viral. Jeske et al. (2001) ont montré que lors de la réplication des bégomovirus, de nombreux intermédiaires de réplications étaient créés. De manière intéressante, ces intermédiaires correspondent à des fragments de virus correspondant à une partie des ORFs complémentaires, jusqu'à la zone intergénique. Il a alors été proposé que ces intermédiaires de réplication étaient créés suite à un conflit entre les complexes de réplication du virus et ceux, évoluant en sens inverse sur la matrice virale, de transcription des ORFs complémentaires. Les fragments de génomes synthétisés seraient alors utilisés dans les phénomènes de réplication dépendante de la recombinaison (RDR) aboutissant à la génération de virus chimères lorsque des matrices de différentes espèces rentrent en contact. Ce travail propose une explication au grand nombre de recombinants créés, mais n'explique pas comment ceux-ci s'adaptent à de nouvelles niches ou se répandent dans la population.

Paradoxalement, malgré la multiplicité des descriptions d'évènements de recombinaison pour de nombreux genres viraux, très peu d'études se sont penchées sur une description précise des profils de recombinaison observés et encore moins sur les facteurs liés à la création de ces mêmes profils. Si dans l'absolue, la recombinaison permet l'obtention d'une diversité très importante, il est fort probable que les contraintes liés aux génomes viraux de faible taille (ORFs chevauchants, plusieurs fonctions par gène) entraînent un certain nombre de réarrangements délétères pour une fraction des recombinants créés. Les règles qui régissent la sélection d'organismes recombinants au sein de populations virales commencent seulement à être définies. Parmi les génomes viraux créés par recombinaison, certains seraient viables (avec différents niveaux de *fitness*), d'autres non. Le transfert de gènes serait d'autant plus facilité que les parents seraient proches. De plus, les gènes présentant de nombreuses interactions avec d'autres gènes seraient moins transférables que ceux présentant peu d'interactions (hypothèse de "la complexité"; Jain et al., 1999). Un virus recombinant serait d'autant plus fonctionnel que les virus parents seraient proches, aboutissant ainsi à une plus faible probabilité de rupture des interactions au sein du génome (Martin et al., 2005). De la même manière, le

réarrangement des différents ADN de virus multipartites suivraient des règles similaires (Escriu et al., 2007). La valeur évolutive de la recombinaison doit ainsi dépendre de chaque gène et des fragments de gènes transférés. La nécessité du maintien des réseaux d'interactions au sein d'un organisme vivant ou tout simplement d'une molécule a ainsi été proposée comme thème majeur des contraintes liées à la recombinaison. C'est seulement en comprenant ces contraintes qu'il sera possible de définir les possibilité et les limites de la recombinaison dans la création de virus. Un grand travail reste à être accomplis pour (1) démontrer ou non l'hypothèse de la nécessité de maintient des réseaux d'interactions pour la création de virus recombinant viables et (2) mettre en évidence la nature des facteurs et leur degré d'intervention dans le façonnage des profils de recombinaison.

4. Emergence des bégomovirus dans les îles du Sud-Ouest de l'océan Indien

Des études récentes menées dans les îles du Sud-Ouest de l'océan Indien (SWIO) ont illustré la capacité d'évolution et d'émergence des bégomovirus. En effet, l'introduction accidentelle de bégomovirus invasifs de la tomate d'origine méditerranéenne a été reportée et décrite à La Réunion. D'autre part, une grande diversité de bégomovirus indigènes a été découverte sur cultures maraîchères à Madagascar et Mayotte.

a. Expansion mondiale : l'exemple du TYLCV

La maladie du tomato yellow leaf curl (TYLCD ou maladie du jaunissement et de l'enroulement foliaire de la tomate) est associée à des complexes d'espèces dont la plus connue est celle du TYLCV, et ceci sur une large gamme d'hôte, cultivés ou non (Cohen and Nitzany, 1966; Dalmon et al., 2000; Monci et al., 2002). Deux souches du TYLCV ont successivement été introduites à La Réunion ces dernières années, ce qui souligne l'importante extension de ce groupe de virus.

D'un point de vue taxonomique, tous les virus responsables du TYLCD sont apparentés (TYLCVs-like) et appartiennent à six espèces nord africaines et méditerranéennes (TYLCV, TYLCSV), *Tomato yellow leaf curl Axiarqua virus* (TYLCAxV), TYLCMalV, *Tomato yellow leaf curl Mali virus* (TYLCMLV) et *Tomato leaf curl Soudan virus* (ToLCSDV) et 15 souches différentes (Figure 9; Abhary et al., 2007).

Les différentes espèces et souches de TYLCV-like forment un cluster de séquences qui diffèrent les unes des autres par des évènements de recombinaison le long du génome. C'est ainsi qu'il a été montré que le TYLCV-IL était le fruit d'une recombinaison entre un ancêtre du TYLCV-Mld et un ancêtre des ToLCVs asiatiques (Navas-Castillo et al., 2000). Des recombinaisons interspécifiques vraisemblablement plus récentes sont à l'origine du TYLCMalV et

du TYLCAxV entre le TYLCV-IL et le TYLCSV, et le TYLCV-Mld et le TYLCSV, respectivement[4] (Garcia-Andres et al., 2006; Monci et al., 2002).

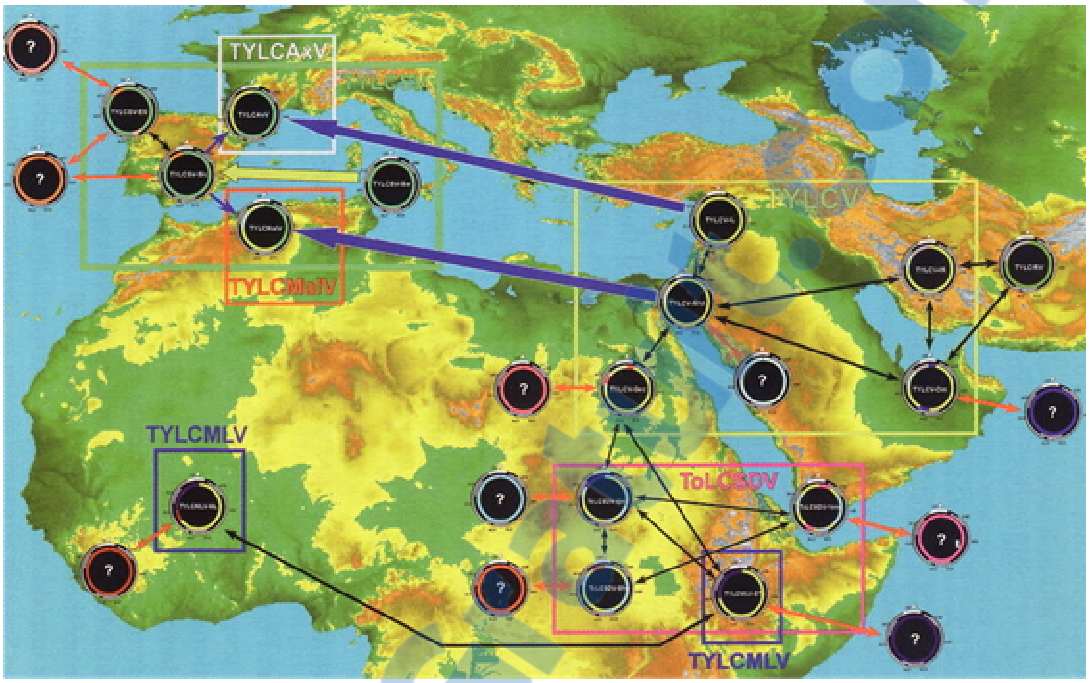


Figure 9 : Représentation des liens de parenté théoriques de représentants des six espèces de bégomovirus monopartites impliquées dans la maladie du tomato yellow leaf curl en Afrique du Nord, Méditerranée et Moyen-Orient. Les différentes couleurs sur les représentations schématiques des génomes, représentent l'origine recombinante des fragments. Les flèches bleues, noires et rouges représentent respectivement des recombinaisons récentes, anciennes et hypothétiques. TYLCV : *Tomato yellow leaf curl virus*, TYLCAxV : *Tomato yellow leaf curl Axarquia virus*, TYLCMaV : *Tomato yellow leaf curl Malaga virus*, TYLCMLV : *Tomato yellow leaf curl Mali virus*, TYLCSV : *Tomato yellow leaf curl Sardinia virus* and TYLCSDV : *Tomato yellow leaf curl Soudan virus*. D'après Abhary et al., (2007).

Les symptômes de la maladie du TYLCD (Figure 10) ont été observés pour la première fois en Jordanie dans les années 1939-1940, et décrits en 1964 sur tomate en Israël (Cohen, 1964). Le bégomovirus associé à cette maladie a été identifié dans les années 90 comme étant le TYLCV (Cohen and Antignus, 1994; Pico et al., 1996). En Afrique, il a été décrit la première fois au Soudan (Yassin and Nour, 1965), puis c'est répandu dans tout l'Est africain. Ce complexe viral s'est depuis largement répandu à travers le monde au cours des vingt dernières

[4] Il est important de considérer avec précautions la nature des parents des virus recombinants. Si un virus est proposé comme parent, il faut considérer que c'est en réalité un ancêtre de ce virus qui est le parent le plus probable. Par ailleurs, déterminer quel est le virus recombinant et quels sont les parents est parfois difficile et reste une notion toute relative. Enfin, comme souligné plus haut dans le paragraphe dédié à la recombinaison, les bégomovirus présentent une histoire évolutive très marquée par la recombinaison et l'ensemble des virus circulant dans la population ont probablement une origine recombinante, celle-ci n'étant alors plus détecté du fait de mutations masquant la recombinaison ou de l'absence de séquences parentales connues.

années. Le TYLCV a été décrit dans la plupart des régions tropicales et sub-tropicales, dans le bassin méditerranéen, en Asie, Moyen-Orient, en Afrique, dans les Caraïbes, aux Etats-Unis et en Amérique Latine (Accotto et al., 2003; Czosnek and Laterrot, 1997; Chouchane et al., 2007; Delatte et al., 2005a; Tahiri et al., 2006; Urbino et al., 2003; Bird et al., 2001; Polston et al., 1994; Ling et al., 2006; Wu et al., 2006).

Figure 10 : Symptômes de TYLCV sur tomate en plein champ à La Réunion. Photo CIRAD.

Dans de nombreuses régions du monde, une arrivée successive de différents variants de TYLCV a été constatée (Delatte et al., 2005a; Ueda et al., 2004; Idris et al., 2007; Duffy and Holmes, 2007) et disséquée sur le plan moléculaire (Duffy and Holmes, 2007) et épidémiologique (Davino et al., 2006; SanchezCampos et al., 1999). L'exemple des travaux menés en Espagne et en Italie à ce sujet est très significatif. Une dynamique d'apparition et d'expansion de variants recombinants a été constatée tout au long des 10 dernières années. Ces études ont souligné l'importance de l'ensemble des facteurs pouvant intervenir dans la dynamique virale, à savoir la gamme d'hôte, la *fitness* des virus dans l'hôte ainsi que la capacité du virus à être transmis par les différents biotypes de *B. tabaci* (Davino et al., 2006; Monci et al., 2002; Sanchez-Campos et al., 1999).

La rapidité de l'expansion géographique du TYLCV et la description de nombreux variants ces 20 dernières années, semblent liées à la dissémination mondiale du biotype B, et plus récemment du biotype Q, de *B. tabaci* (Boykin et al., 2007). Cette expansion géographique de *B. tabaci* et des bégomovirus a directement été associée à la mondialisation et à l'intensification des échanges commerciaux de plantes ornementales et de légumes (Delatte et al., 2003; Anderson et al., 2004).

b. Les bégomovirus des Iles SWIO

Suite à la première épidémie de TYLCV en 1997 à La Réunion (Peterschmitt et al., 1999), une veille sanitaire a été mise en place dans tout le Sud-Ouest de l'océan Indien. Au cours d'une campagne de prélèvement de feuilles de tomate présentant des symptômes de TYLCD, deux nouvelles espèces de begomovirus ont été découvertes à Madagascar et à Mayotte (Delatte et al., 2002; Lett et al., 2004), respectivement dénommées *Tomato leaf curl Madagascar virus* (ToLCMGV) et *Tomato leaf curl Mayotte virus* (ToLCYTV). Les premières analyses phylogénétiques ont montrées l'existence d'un groupe monophylétique, distinct de celui du TYLCV, qui semble avoir évolué en isolement à partir d'un ancêtre commun (Delatte et al., 2005b). Au delà de la découverte de nouveaux virus, cette étude a mis en évidence les larges contours de la diversité de bégomovirus existant sur le continent Africain et les îles voisines, celle-ci n'étant auparavant décrite que pour les bégomovirus isolés à partir du manioc.

5. Problématique et objectifs

L'émergence virale représente, tout particulièrement en milieu tropical, une question d'importance fondamentale pour l'agriculture de subsistance. Notamment ces 20 dernières années, le continent Africain et sa population ont du faire face à l'émergence de nombreuses épidémies virales sur des cultures aussi essentielles que le riz avec le *Rice yellow mottle virus* (RYMV; Abo et al., 1998; Fargette et al., 2004), le manioc avec le complexe d'espèces de bégomovirus associé à la maladie de la mosaïque du manioc (Legg and Fauquet, 2004) ou le maïs avec le *Maize streak virus* (MSV; Bosque-Perez, 2000). A partir de ces modèles d'études, l'analyse des données moléculaires et écologiques a permis d'identifier plusieurs facteurs impliqués dans leur émergence, dont : (1) l'évolution virale par recombinaison, (2) la synergie entre virus, (3) l'arrivée de nouveaux biotypes d'insectes vecteurs, (4) l'accès et l'adaptation à de nouvelles plantes hôtes (saut d'espèce), et (5) la dispersion à longue distance (mondialisation des échanges).

Le groupe des bégomovirus, associé à leur insecte vecteur, semble présenter l'ensemble de ces facteurs d'émergence. Il regroupe actuellement le plus grand nombre d'espèces virales émergentes d'importance économique. La capacité d'évolution rapide et d'adaptation des bégomovirus à de nouvelles niches écologiques (forts taux de mutation et de recombinaison), l'augmentation régulière de leur aire de répartition et la polyphagie de certains biotypes invasifs de leur insecte vecteur *B. tabaci* font de ce complexe virus - hôte - vecteur un modèle d'étude de l'émergence virale. La diversité virale précédemment décrite pour cette famille de virus souligne la nécessité d'avoir une vision plus globale de la diversité et des mécanismes d'évolution des complexes de bégomovirus.

Les îles du Sud-Ouest de l'océan Indien (SWIO) représentent un lieu idéal pour l'analyse de l'évolution des bégomovirus de par leurs situations géographiques et géologiques. Ainsi, Madagascar et les Seychelles (îles issues du continent Africain) et les îles des Mascareignes et des Comores (Iles volcaniques néoformées) font partie des réservoirs de diversité du vivant parmi les plus riches et les plus menacés du monde. Ces îles ont des microenvironnements hypervariables facilitant la co-occurrence d'organismes ayant des histoires de vie très différentes (Dewar and Richard, 2007; Haevermans et al., 2004). L'isolement géographique de La Réunion, associé aux récentes et sporadiques introductions d'espèces de bégomovirus exotiques sur cultures maraîchères, a créé un environnement favorable à l'étude de l'émergence virale et des facteurs épidémiologiques associés (Chapitre I). D'autre part, l'existence d'une population virale a priori indigène des îles voisines nous a permis d'étudier plus en profondeur leur diversité et les facteurs d'évolution associés ainsi que d'émettre des hypothèses sur l'origine même de ces groupes de virus (Chapitre II). Si l'importance de la recombinaison comme moteur majeur de l'évolution des bégomovirus a été montré, la manière par laquelle les recombinants sont générés et les mécanismes qui influencent la distribution des évènements de recombinaison au sein du génome sont encore mal compris. La mise à jour d'une diversité nouvelle en bégomovirus dans la région (Chapitre II) et de part le monde (base de donnée) en conjonction avec le développement de méthodes précises de détection de la recombinaison nous a permis de disséquer plus finement ce phénomène et les facteurs sélectifs associés (Chapitre III). L'existence de nombreuses autres familles virales à ADN simple brin, apparentées aux bégomovirus, nous a offert la possibilité d'un élargissement des concepts et d'une comparaison des facteurs associés à l'évolution de ces familles virales (Chapitre IV).

Chapitre I : Le TYLCV-Mld et le TYLCV-IL sur l'île de La Réunion : une association de malfaiteurs

Le TYLCV est sûrement l'espèce de bégomovirus la mieux connue avec les virus infectant le manioc. De nombreuses études ont par ailleurs décrit l'apparition de variants recombinants. Le TYLCV et son cortège de variants forment des complexes viraux dont l'extension mondiale ne cesse de s'accroître. Excepté l'Antarctique, le TYLCV est aujourd'hui présent sur tous les continents. Autour du bassin méditerranéen, zone d'origine supposée du virus, et en Espagne notamment, la description des dynamiques d'évolution et de la compétition au sein de ces complexes viraux a permis une meilleure compréhension des mécanismes participant à leur évolution. Les travaux menés ont notamment montré l'apparition de nombreux variants recombinants et ont mis en avant des phénomènes de compétition entre ceux-ci.

A la Réunion, les deux souches de TYLCV actuellement décrites (TYLCV-IL et TYLCV-Mld) ont été introduites accidentellement à quelques années d'intervalle : Le TYLCV-Mld a été détecté pour la première fois en 1997 (Peterschmitt et al., 1999) et a envahi l'ensemble du bassin maraîcher. Le TYLCV-IL (souche recombinante; NavasCastillo et al., 2000) a été détecté en 2004 dans l'ouest de l'île et l'extrémité ouest du bassin maraîcher (Delatte et al., 2005a).

Nous nous sommes attachés ici à décrire la dynamique spatio-temporelle de ce complexe viral dans un écosystème tropical et insulaire. Plusieurs scénarii étaient envisageables avec soit (1) une compétition et un déplacement d'une souche par l'autre, (2) un synergisme et une cohabitation des deux souches ou encore (3) l'apparition d'un nouveau variant recombinant plus *fit*, capable de déplacer les deux parents. L'échantillonnage réalisé sur 4 ans, en parallèle à la mise au point d'un test de détection spécifique par PCR multiplexe et de son utilisation, a permis de mettre en avant un déplacement du TYLCV-Mld par le TYLCV-IL. En seulement 4 ans, d'une situation prédominante, le TYLCV-Mld n'est plus retrouvé qu'en coinfection.

Afin de définir si l'origine de cet avantage sélectif pouvait être associée à un pouvoir pathogène différentiel, des tests comparatifs de *fitness* et de virulence ont été réalisés. Aucune différence statistique de *fitness* entre TYLCV-Mld et TYLCV-IL n'a été démontrée. En revanche, les tests de virulence ont montré que

le TYLCV-IL présentait une virulence plus élevée que le TYLCV-Mld, sur la variété de tomate la plus répandue en plein champ à La Réunion. Au-delà de la justification de l'appellation de souche sévère, jusqu'à lors jamais prouvée, cette différence de pouvoir pathogène pose des questions fondamentales sur l'origine moléculaire et les conséquences de cette virulence accrue.

Les principales différences nucléotidiques entre les souches IL et Mld du TYLCV reposent sur la nature recombinante du premier virus. Contrairement au TYLCV-Mld, l'ORF C4 du TYLCV-IL présente des liens de parenté avec les ToLCVs asiatiques (Navas-Castillo et al., 2000). Or, la protéine C4 des bégomovirus monopartites semble avoir un rôle important dans la symptomatologie (Gafni and Epel, 2002; Jupin et al., 1994; Rigden et al., 1994). D'autre part, l'expression de la protéine C4 des ToLCV et TYLCCHV asiatiques dans des plantes transgéniques de tomate a permis de générer des symptômes typiques d'enroulement foliaire de la maladie associés à ces virus (Selth et al., 2004). L'ensemble de ces éléments suggère que la nature de cette protéine jouerait un rôle primordial dans la symptomatologie et pourrait lui conférer une sévérité plus importante.

La dissémination des bégomovirus est intimement liée au comportement alimentaire de leur insecte vecteur *Bemisia tabaci*. Des travaux antérieurs ont montré que *B. tabaci* et de nombreux autres insectes phytophages, les pucerons par exemple, sont particulièrement attirés par la composante jaune du spectre lumineux (Mound, 1962). Une virulence plus élevée avec des symptômes de jaunissement plus prononcés pourrait entraîner une attirance plus forte de *B. tabaci* pour les plantes infectées par le TYLCV-IL et ainsi favoriser sa dissémination. En d'autres termes, la virulence plus importante de la souche IL du TYLCV pourrait jouer un rôle clef dans sa propagation et expliquer la dynamique virale observée à La Réunion. Néanmoins d'autres paramètres épidémiologiques pourraient également avoir contribué à l'expansion du TYLCV-IL au détriment du TYLCV-Mld. Une transmission préférentielle par *B. tabaci* ou une gamme plus élargie de plantes hôtes réservoirs seraient autant de facteurs pouvant contribuer à l'avantage sélectif du TYLCV-IL. Les travaux préliminaires effectués par notre équipe à ce sujet semblent montrer l'importance des plantes non cultivées dans la dynamique du virus pendant les cycles saisonniers de culture. Le rôle considérable des plantes réservoirs dans la dynamique des épidémies a par ailleurs déjà été montrée pour le TYLCV-IL et -Mld au détriment du TYLCSV dans le sud de l'Espagne (Garcia-Andres et al., 2006; SanchezCampos et al., 1999).

Le résultat final de cette compétition inter-souches de TYLCV dans un environnement tropical insulaire pourrait se solder à terme par l'élimination complète de la souche TYLCV-Mld, comme cela fut le cas dans le sud de l'Espagne avec le remplacement de TYLCSV par le TYLCV-IL (SanchezCampos et al., 1999). Néanmoins, ce scénario ne semble pas être inéluctable comme le suggère la cohabitation entre le TYLCV-IL et le TYLCSV dans le sud de l'Italie (Davino et al., 2006). Par conséquent, la dynamique et l'évolution future de ce complexe viral mérite toute notre attention.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Virological Methods 144 (2007) 165–168
Journal of Virological Methods
www.elsevier.com/locate/jviromet

Short communication

A multiplex PCR method discriminating between the TYLCV and TYLCV-Mld clades of *Tomato yellow leaf curl virus*

P. Lefeuvre *, M. Hoareau, H. Delatte, B. Reynaud, J.-M. Lett *

CIRAD, UMR 53 PVBMT CIRAD-Université de la Réunion, Pôle de Protection des Plantes, Ligne Paradis, 97410 Saint Pierre, La Réunion, France

Received 3 November 2006; received in revised form 29 March 2007; accepted 29 March 2007
Available online 7 May 2007

Abstract

Tomato yellow leaf curl virus (TYLCV) is one of the causal agents of tomato yellow leaf curl disease (TYLCD) and can cause up to 100% yield losses in tomato fields. As TYLCV continues to spread, many isolates have been described in different parts of the world. Recently two closely related but distinct TYLCV clades, called TYLCV and TYLCV-Mld, have been identified. Isolates from those two clades differ mainly in the nucleotide sequences of their replication associated protein genes but do not display significantly different symptomatology. In order to improve monitoring of the rapidly expanding worldwide TYLCD epidemic, a multiplex polymerase chain reaction assay (mPCR) was developed. A set of three primers were designed to detect and characterize the TYLCV and TYLCV-Mld clade isolates. The specificity and sensitivity of the mPCR were validated on TYLCV infected tomato plants and *Bemisia tabaci* whiteflies. Being cheap, fast and highly sensitive this new diagnostic tool should greatly simplify efforts to trace the global spread of TYLCV.

Keywords: Begomovirus; Emerging disease; Multiplex PCR; Detection tool

The genus *Begomovirus* (family *Geminiviridae*) is a group of emerging plant viruses with circular single-stranded DNA genomes encapsidated in twinned icosahedral virions and transmitted by the whitefly *Bemisia tabaci*. Begomoviruses cause severe diseases in a wide variety of plant species including many of great agricultural importance (Rybicki et al., 2000). Tomato yellow leaf curl disease (TYLCD) can cause total tomato (*Solanum lycopersicum*) yield losses and severely hampers production of this crop in tropical and subtropical regions (Czosnek and Laterrot, 1997). A begomovirus species complex is associated with TYLCD (Moriones and Navas-Castillo, 2000). One of the primary species in the complex is *Tomato yellow leaf curl virus* (TYLCV). Two differentially virulent variants of this species named TYLCV (accession no. X15656) and TYLCV-Mld (accession no. X76319) were isolated in Israel in the late 1980s and early 1990s (Navot et al., 1991; Antignus and Cohen, 1994). Following complete sequencing of their genomes it was realised that the TYLCV variant was in fact the product of a recombination event between the TYLCV-Mld variant and another tomato infecting begomovirus species, *Tomato leaf curl Karnataka virus* (ToLCKV; Navas-Castillo et al., 2000). TYLCV contains mostly TYLCV-Mld like sequences but the 5′-portion of its Rep gene is very ToLCKV-like.

In the past few years, many papers have reported the presence of TYLCV isolates closely related to these two Israeli variants around the world, including Florida (Polston et al., 1999), Spain (Navas-Castillo et al., 1999), Reunion (Delatte et al., 2005a,b), Morocco (Peterschmitt et al., 1999; Tahiri et al., 2006), Italy (Accotto et al., 2003), Guadeloupe (Urbino and Tassius, 2003), Cuba (Martinez-Zubiaur et al., 2004), China (Wu, 2006), and South Carolina (Ling et al., 2006). Phylogenetic analysis has shown that TYLCV isolates are separated into two major clades, one that contains the original isolate from Israel (for convenience called the IL clade) and the second that contains the "Mild" strain (for convenience called the Mild clade; Fauquet and Stanley, 2005; Navas-Castillo et al., 2000). Although the initial description of TYLCV-Mld by Antignus and Cohen (1994) suggested it induced milder symptoms than the TYLCV, other TYLCV-

* Corresponding authors. Tel.: +262 262 499243; fax: +262 262 499293.
E-mail addresses: pierre.lefeuvre@cirad.fr (P. Lefeuvre), lett@cirad.fr (J.-M. Lett).

0166-0934/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jviromet.2007.03.020

Table 1
Origin and accession numbers of the TYLCV isolates used to design the primers

TYLCV Clade	Isolate	Origin	Accession number
Mld	TYLCV-[SD]	Soudan	AY044138
	TYLCV-Mld	Israel	X76319
	TYLCV-Mld[Aic]	Japan	AB014347
	TYLCV-Mld[Atu]	Japan	AB116633
	TYLCV-Mld[ES]	Spain	AJ519441
	TYLCV-Mld[ES7297]	Spain	AF071228
	TYLCV-Mld[Kis]	Japan	AB116634
	TYLCV-Mld[PT]	Portugal	AF105975
	TYLCV-Mld[RE]	Reunion	AJ865337
	TYLCV-Mld[Shi]	Japan	AB014346
	TYLCV-Mld[Sz:Dai]	Japan	AB116635
	TYLCV-Mld[Sz:Osu]	Japan	AB116636
	TYLCV-Mld[Sz:Shi]	Japan	AB110218
	TYLCV-Mld[Sz:Yai]	Japan	AB116632
IL	TYLCV-[Alm]	Spain	AJ489258
	TYLCV-[CU]	Cuba	AJ223505
	TYLCV-[DO]	Dominican Republic	AF024715
	TYLCV-[EG]	Egypt	L12219
	TYLCV-[Flo]	Forida	AY530931
	TYLCV-[IL]	Israel	X15656
	TYLCV-[IR]	Iran	AJ132711
	TYLCV-[Mis]	Japan	AB116631
	TYLCV-[Miy]	Japan	AB116629
	TYLCV-[Nag]	Japan	AB110217
	TYLCV-[Omu]	Japan	AB116630
	TYLCV-[PR]	Puerto Rico	AY134494
	TYLCV-[RE4]	Reunion	AM409201

Mld isolates sampled worldwide are apparently as virulent as TYLCV.

Differentiation between the IL and "Mild" strains on the basis of symptomotology is therefore completely unreliable. Epidemiology studies focusing on these viruses depend heavily on the development of reliable assays for differentiating between them. The recombination derived differences between these isolates in the 5′-portion of the Rep gene provided a logical starting point from which to design a PCR based strategy to discriminate between the viruses. We therefore designed a universal, reliable and sensitive multiplex PCR for both the differential and simultaneous identification of isolates in both the Mild and IL TYLCV clades.

An alignment of 27 complete TYLCV nucleotide sequences available in the GenBank-EMBL-DDJP database (Table 1) was performed using the full optimal alignment method of DNAMAN2 (Lynnon Biosoft, Canada). The TYLCV-[Iran] isolate sequence was disregarded during primer design because of the high degree of divergence of its Rep ORF relative to the other TYLCV sequences (Bananej et al., 2004). Primers were designed using Oligo 5.0 (National Biosciences, USA) focusing on a conserved region of sequence for the common IL/Mild primer (TYLCV-1840F) and variable regions between IL and Mild isolates for the specific IL (IL-2642R) and Mild (Mld-2354R) primers (Table 2, Fig. 1). These primers will also identify IL- and Mild-like sequences within recombinant begomoviruses containing TYLCV derived sequences such as those described in Monci et al. (2002) and Garcia-Andres et al. (2006). Reference viruses used to standardize the PCR consisted of IL isolate, TYLCV-[RE4] (accession no. AM409201), and the Mild isolate TYLCV-Mld[RE] (accession no. AJ865337). Full-length genomes of these isolates have previously been cloned into the plasmid pGEM-T Easy Vector and transformed into the JM109 strain of *Escherichia coli* (Promega, USA; Delatte et al., 2005a). Recombinant plasmid DNA was isolated from bacteria with the Plasmid MiniPrep Spin Kit (Qiagen, Germany) and quantified with a fluorimeter (Hoeffer, Germany). The extracted plasmids were then serially diluted from 10^8 down to 10^3 copies per 2 μL in 10-fold steps and were used for the PCR specificity and sensitivity assays. These assays were performed in a 25 μl reaction mixture containing 1.25 mM $MgCl_2$, 1× PCR buffer, 0.2 mM of each dNTP, 1 μM of each primer, and 2.5 U of Red Gold Star DNA polymerase (Eurogentec, Belgium) in a GeneAmp PCR System 9600 thermocycler (Perkin-Elmer, USA) using an initial denaturation step (5 min at 95 °C), various primer concentrations (0.5–2 μM), annealing temperatures (50–60 °C), number of cycles (25, 30 or 35) and a final extension step (5 min at 72 °C). Standard gel electrophoresis with 10 μl of the final reaction mixture on a 1.5% agarose gel was used to separate PCR products. Ethidium bromide stained DNA fragments were visualised on agarose gels using ultraviolet fluorescence.

Highest PCR product yields were achieved with 0.5 μM of the Mld-2354R primer, 2 μM of the IL-2642R primer and 1 μM of the TYLCV-1840F primer, with an annealing temperature of 55 °C and 35 cycles of amplification. The expected amplification products of 514 bp for the Mild isolate and 802 bp for the IL isolate were obtained and non-specific amplification was never observed (Fig. 2).

The detection limits of the PCR were determined using a 10-fold serial dilution of purified cloned viruses containing 10^3 to 10^8 DNA copies of each genome in pure or mix samples. Detection limits with pure samples were 10^4 genome copies for TYLCV-[RE4] and 10^3 genome copies for TYLCV-Mld[RE] (Fig. 3A). Numerous papers have reported the presence of plant samples coinfected by two or more begomoviruses (Sanz et al., 2000; Garcia-Andres et al., 2006). Furthermore, in some regions, such as Reunion (Delatte et al., 2005a) and Japan (Ueda et al., 2004), viruses in both clades are co-circulating. For application of the multiplex PCR in such contexts, it is important to establish its performance when it is used to examine coinfected samples. Using a range of 10^3 to 10^8 cloned genome copies, IL:Mild ratios between $1:10^5$ and $10^5:1$ were examined with the multiplex PCR. At ratio 1:1, the detection limit of the multiplex PCR

Table 2
Primer sequences, amplicon length and location within the viral genome

Primer	Sequence	Strand	Amplicon length[a]	Location in genome[b]
TYLCV-1840F	5′-GGTCTACGTCATCAATGAC-3′	Viral	–	1840
Mld-2354R	5′-AGGGAGCTAAATCCAGTT-3′	Complementary	514	2354
IL-2642R	5′-ACACCGATTCATTTCAAC-3′	Complementary	802	2642

[a] When used in association with TYLCV-1840F.
[b] The numbering of the genome begins at the nicking site of the conserved nonanucleotide in the intergenic region of all begomoviruses, accordingly to the literature.

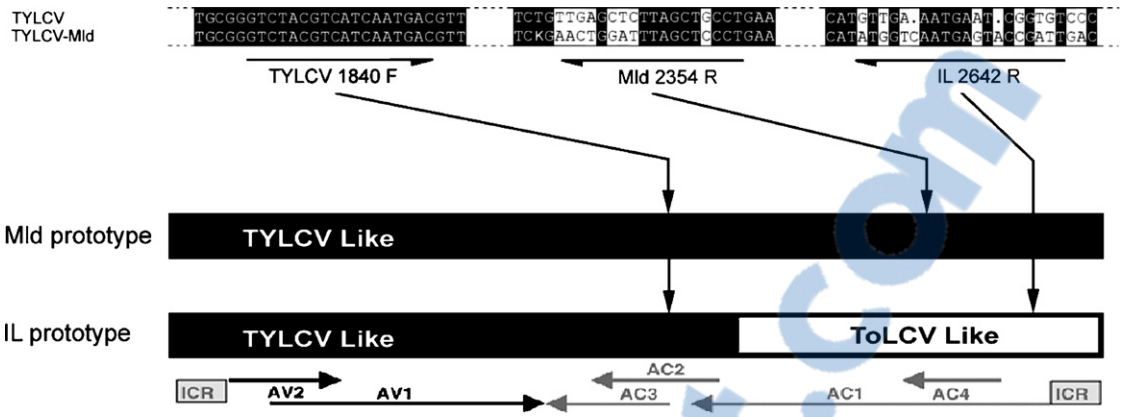


Fig. 1. Schematic representation of annealing sites of the primers designed. Black arrows represent viral strand ORF, grey arrows represent complementary strand ORF and ICR boxes represent Intergenic Coding Region. Partial alignment of consensus nucleotide sequences (viral strand) of the 13 TYLCV and 14 TYLCV-Mld isolates used to design the primers TYLCV-1840F, Mld-2354R and IL-2642R. The letter K in the Mld consensus sequence code for a G/T polymorphism. The primers Mld-2354R and IL-2642R are designed to be specific of the Mld and IL clade of TYLCV isolates respectively whereas the primer TYLCV-1840F can hybridise both clade nucleotide sequences.

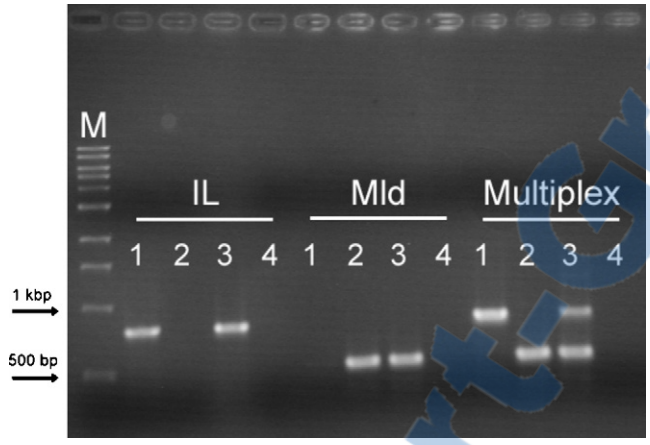


Fig. 2. Specificity of uniplex (IL and Mld) and multiplex PCR detection of TYLCV isolates. Lane (M): 1 kb DNA ladder, lane (1): 10^8 copies of TYLCV-[RE4], lane (2): 10^8 copies of TYLCV-Mld[RE], lane (3): 10^8 copies of both TYLCV-[RE4] and TYLCV-Mld[RE], lane (4): negative control.

was 10^5 total genome copies and both isolates were detected simultaneously when the IL:Mild ratio was between 10^2:1 and 1:1 (Fig. 3B).

The diagnostic test was also validated on plant and insect DNA extracts. Towards this end, 10-day-old susceptible tomato seedlings were needle agro-inoculated with the agroinfectious clones of either TYLCV-[RE4] or TYLCV-Mld[RE] (Delatte et al., 2005a). The leaves of symptomatic 21-day-old plants were harvested and preserved by dehydration with calcium chloride ($CaCl_2$) before extraction (Bos, 1977). Leaves of healthy plants were used as negative controls. Total DNA was extracted from dried samples using the DNeasy Plant Mini kit (Qiagen, Germany) according to the manufacturer's instructions. Furthermore, an epidemiological field survey was conducted for TYLCD in Reunion in 2006. Total DNA was extracted from the collected samples in the same way. To determine whether the multiplex PCR would provide sufficient sensitivity for detection of the virus within *B. tabaci*, whiteflies were fed for 3 days on either symptomatic TYLCV infected plants or healthy plants (negative control). DNA from groups of ten insects was then isolated as described previously (Delatte et al., 2005c). All

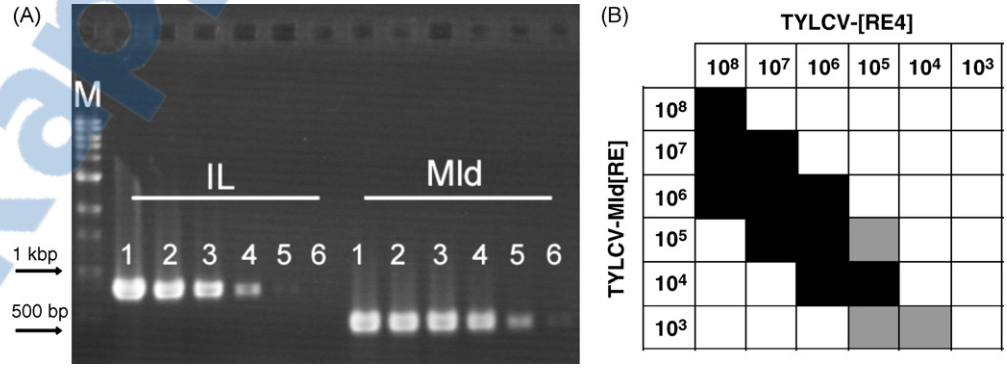


Fig. 3. (A) Comparison of sensitivity of two uniplex PCR for the detection of TYLCV-[RE4] (IL clade isolate) and TYLCV-Mld[RE] (Mld clade isolate) using 10-fold serial dilutions of viral clone. Lane (M): 1 kb DNA ladder, lane (1): 10^8 copies, lane (2): 10^7 copies, lane (3): 10^6 copies, lane (4): 10^5 copies, lane (5): 10^4 copies, lane (6): 10^3 copies. (B) Dual detection matrix, with black squares indicating an efficient dual detection, grey squares indicating a weak signal for the dual detection and white squares indicating the absence of dual detection.

these plant and insect DNA extracts were tested for coinfections with TYLCV from both clades. Both TYLCV isolates were specifically amplified and could be clearly codetected within infected plants and viruliferous insects without there being any non-specific amplification. PCR amplification from the negative controls was never observed. Effective dual detection of both TYLCV strains was also achieved when examining naturally coinfected plant samples from the field (data not shown). The aim of this work was to establish a PCR based protocol capable of specifically identifying the IL and Mild TYLCV in a single step. This new multiplex PCR test will be an extremely valuable epidemiological survey tool both for tracing the ongoing global spread of these TYLCV isolates and managing the diseases they cause.

Acknowledgements

We are grateful to Nathalie Becker (MNHN, Saint-Pierre, Ile de la Réunion, France) for critical readings of the manuscript. This work was funded by the Conseil Régional de la Réunion and CIRAD. English was kindly corrected by Darren Martin (IIDMM, Cape Town, South Africa).

References

Accotto, G.P., Bragaloni, M., Luison, D., Davino, S., Davino, M., 2003. First report of *Tomato yellow leaf curl virus* (TYLCV) in Italy. Plant Pathol. 52, 799–1799.

Antignus, E.Y., Cohen, S., 1994. Cloning of Tomato yellow leaf curl virus (TYLCV) and the complete nucleotide sequence of a mild infectious clone. Phytopathology 84, 707–712.

Bananej, K., Kheyr-Pour, A., Salekdeh, G.H., Ahoonmanesh, A., 2004. Complete nucleotide sequence of Iranian tomato yellow leaf curl virus isolate: further evidence for natural recombination amongst begomoviruses. Arch. Virol. 149, 1435–1443.

Bos, L., 1977. Persistance of infectivity of three viruses in plant material dried over $CaCl_2$ and stored under different conditions. Neth. J. Plant Pathol. 83, 217–220.

Czosnek, H., Laterrot, H., 1997. A worldwide survey of tomato yellow leaf curl viruses. Arch. Virol. 142, 1391–1406.

Delatte, H., Holota, H., Naze, F., Peterschmitt, M., Reynaud, B., Lett, J.M., 2005a. The presence of both recombinant and non recombinant strains of *Tomato yellow leaf curl virus* on tomato in Réunion Island. Plant Pathol. 54, 262.

Delatte, H., Martin, D.P., Naze, F., Golbach, R.W., Reynaud, B., Peterschmitt, M., Lett, J.M., 2005b. South West Indian Ocean islands tomato begomovirus populations represent a new major monopartite begomovirus group. J. Gen. Virol. 86, 1533–1542.

Delatte, H., Reynaud, B., Granier, M., Thornary, L., Lett, J.M., Goldbach, R., Peterschmitt, M., 2005c. A new silverleaf-inducing biotype Ms of *Bemisia tabaci* (Hemiptera: Aleyrodidae) indigenous of the islands of the south-west Indian Ocean. Bull. Entomol. Res. 95, 29–35.

Fauquet, C.M., Stanley, J., 2005. Revising the way we conceive and name viruses below the species level: a review of geminivirus taxonomy calls for new standardized isolate descriptors. Arch. Virol. 150, 2151–2179.

Garcia-Andres, S., Monci, F., Navas-Castillo, J., Moriones, E., 2006. Begomovirus genetic diversity in the native plant reservoir *Solanum nigrum*: evidence for the presence of a new virus species of recombinant nature. Virology 350, 433–442.

Ling, K.-S., Simmons, A.M., Hassell, R.L., Keinath, A.P., Polston, J.E., 2006. First report of *Tomato yellow leaf curl virus* in South Carolina. Plant Dis. 90, 379.

Martinez-Zubiaur, Y., Fonseca, D., Quiñones, M., Palenzuela, I., 2004. Presence of *Tomato yellow leaf curl virus* infecting squash (*Curcubita pepo*) in Cuba. Plant Dis. 88, 572.

Monci, F., SanchezCampos, S., NavasCastillo, J., Moriones, E., 2002. A natural recombinant between the geminiviruses *Tomato yellow leaf curl Sardinia virus* and *Tomato yellow leaf curl virus* exhibits a novel pathogenic phenotype and is becoming prevalent in Spanish populations. Virology 303, 317–326.

Moriones, E., Navas-Castillo, J., 2000. *Tomato yellow leaf curl virus*, an emerging virus complex causing epidemics worldwide. Virus Res. 71, 123–134.

Navas-Castillo, J., Sanchez-Campos, S., Diaz, J.A., 1999. *Tomato yellow leaf curl virus-Is* causes a novel disease of common bean and severe epidemics in tomato in Spain. Plant Dis. 83, 29–32.

Navas-Castillo, J., Sanchez-Campos, S., Noris, E., Louro, D., Accotto, G.P., Moriones, E., 2000. Natural recombination between *Tomato yellow leaf curl virus-Is* and *Tomato leaf curl virus*. J. Gen. Virol. 81, 2797–2801.

Navot, N., Pichersky, E., Zeidan, M., Zamir, D., Czosnek, H., 1991. *Tomato yellow leaf curl virus*: a whitefly-transmitted geminivirus with a single genomic component. Virology 185, 151–161.

Peterschmitt, M., Granier, M., Aboulama, S., 1999. First report of Tomato yellow leaf curl geminivirus in Morocco. Plant Dis. 83, 1074.

Polston, J.E., McGovern, R.J., Brown, L.G., 1999. Introduction of *Tomato yellow leaf curl virus* in Florida and implications for the spread of this and other geminiviruses of tomato. Plant Dis. 83, 984–988.

Rybicki, E.P., Briddon, R., Brown, J.K., Fauquet, C., Maxwell, D.P., Stanley, J., Harrison, B.D., Markham, P., Bisaro, D.M., Robinson, D.J., 2000. Family Geminiviridae. In: Van Regemnmortel, M.H.V., Fauquet, C., Bishop, D.H.L., Carstens, E., Estes, M., Lemon, S., Maniloff, J., Mayo, M.A., McGeoch, D., Pringle, C., Wickner, R. (Eds.), Virus Taxonomy, Seventh Report of the International Committee on Taxonomy of Viruses. Academic Press, New York, pp. 285–297.

Sanz, A.I., Fraile, A., Garcia-Arenal, F., Zhou, X.P., Robinson, D.J., Khalid, S., Butt, T., Harrison, B.D., 2000. Multiple infection, recombination and genome relationships among begomovirus isolates found in cotton and other plants in Pakistan. J. Gen. Virol. 81, 1839–1849.

Tahiri, A., Sekkat, A., BenNani, A., Granier, M., Delvare, G., Peterschmitt, M., 2006. Distribution of tomato-infecting begomoviruses and *Bemisia tabaci* biotypes in Morocco. Ann. Appl. Biol. 149, 175–186.

Ueda, S., Kimura, T., Onuki, M., Hanada, K., Iwanami, T., 2004. Three distinct groups of isolates of *Tomato yellow leaf curl virus* in Japan and construction of an infectious clone. J. Gen. Plant Pathol. 70, 232–238.

Urbino, C., Tassius, K., 2003. First report of *Tomato yellow leaf curl virus* in tomato in Guadeloupe. Plant Dis. 87, 1397.

Wu, J.B., 2006. First report of *Tomato yellow leaf curl virus* in China. Plant Dis. 90 (10), 1359–11359.

Rapid displacement and mixed infection as a result of direct interaction between strains of TYLCV presenting differential severity in a tropical insular environment

P. Lefeuvre[a], N. Becker[b], C. Vincent[a], M. Hoareau[a], L. Brient[a†], M. Thierry[a], S. Boutry[b], H. Delatte[a], B. Reynaud[a] and J.-M. Lett[a]*

[a]CIRAD, UMR 53 PVBMT CIRAD-Université de la Réunion, Pôle de Protection des Plantes, 7 Chemin de l'IRAT, 97410, Saint Pierre, La Réunion, France

[b]Museum National d'Histoire Naturelle, Dept. RDDM, USM 501, CNRS UMR 5166, Evolution des Régulations Endocriniennes, 57 rue Cuvier, CP 32, 75005 Paris, France

*Author for correspondence; E-mail: lett@cirad.fr

†This paper is dedicated to the memory of our respected colleague, friend and co-author Laurent Brient, who left us suddenly at the age of 25 years.

Abstract

TYLCV is one of the major constraints for tomato production in Mediterranean and tropical environment. During the last decades, TYLCV has spread worldwide and exhibit strong pattern of evolution with the emergence of several recombinant variant. Understanding the evolution of the ensuing viral complexes is one of the clues for the comprehension of viral emergence. In Réunion Island, after a first description of the mild strain of TYLCV (TYLCV-Mld) in 1998, the introduction of the Israel strain of TYLCV (TYLCV-IL) was reported in 2004. In this study, a four years survey of the viral population, following the introduction of TYLCV-IL, was performed. A displacement of TYLCV-Mld by TYLCV-IL was shown. TYLCV-Mld was only recovered in coinfection with TYLCV-IL. To understand the factors associated with this displacement, virulence and fitness of both strains were measured. A higher virulence for the TYLCV-IL was measured while fitness remains identical. Even if other factors have to be assessed as host range and insect vector transmission for both viruses, the higher virulence of TYLCV-IL is proposed as an important epidemiological advantage. Increasing yellowing symptoms due to higher virulence could influence insect vector feeding and so results in a preferential transmission of the more virulent strain.

Keywords: TYLCV, Fitness, Virulence, displacement

Introduction

During the last two decades, new begomovirus species (Geminividae) have emerged worldwide, probably as a consequence of the spread of one or more highly polyphageous biotypes of their insect vector, *Bemisia tabaci* (for review Fargette *et al.,* 2006; Seal *et al.*, 2006). Usually in a given region, multiple begomovirus species have emerged simultaneously, with the ensuing species complexes causing diseases in a wide variety of plant species including many of great agricultural importance.

Tomato yellow leaf curl disease (TYLCD) is one of the most devastating plant diseases in warm and temperate regions of the world. This disease is caused by a complex of begomovirus species, referred to as TYLCV-like viruses, including the best-known, most severe and emergent species *Tomato yellow leaf curl virus* (TYLCV). TYLCV is becoming a prevalent pest on tomato crops worldwide. Taxonomically TYLCV-like viruses all belong to the same cluster and to at least six species and 15 strains have been described (Abhary et al., 2007). Interesting studies have been performed on TYLCD situation and the spread of TYLCV-like viruses in temperate regions (Davino *et al.,* 2006; Navas-Castillo *et al.*, 2000). Although many local emergence and co-occurrence of multiple species were reported throughout the world (Delatte et al, 2005a; Ueda et al., 2004), nothing has been described in tropical insular ecosystems.

Different scenarios are expected in such epidemiological situation with competition: synergism, co-existence, displacement or emergence of new variant (Garcia-Andres *et al.,* 2007a; Monci *et al.*, 2002). Indeed, exemples are available in the literature. The displacement of one species by the other (Sanchez-Campos *et al.*, 1999), coexistence of both species (Davino *et al.*, 2006) and emergence of new severe recombinant variant (Davino *et al.,* 2008; Garcia-Andres *et al.*, 2007; Monci *et al.*, 2002) have been described in the literature. Somehow, a few is known about what are the factor shaping those viral complex.

Understanding the mechanism involved in virus complexes evolution is a key factor of the understanding of viral emergence. Islands are particularly well adapted places to study such concept because of limited migration of viruses and so isolated evolution. TYLCV was first reported in Réunion Island in 1997 as the causal agent of severe epidemic outbreaks of the yellow leaf curl disease in tomato crops (Peteschmitt et al., 1999). Molecular analyses reveal the introduction of the so called "Mild" strain of TYLCV (TYLCV-Mld; Delatte *et al.*, 2005a)). In 2004, very severe symptoms of TYLCD, never seen before, have been recorded in the west of the tomato cropping region. Molecular diagnostic have revealed a novel introduction of the "Israel" strain of TYLCV (TYLCV-IL; Delatte et al., 2005a). Although the initial description of TYLCV-Mld by Antignus and Cohen (1994) suggested that it induced milder symptoms than TYLCV-IL, this question remains controversial in the absence of precise work on virulence. The spread and the molecular evolution of TYLCV-Mld in Réunion Island were well documented from 1998 to 2004 (Reynaud et al., 2003; Delatte *et al.*, 2007). The recent introduction of the IL strain of TYLCV into the same area provides an ideal field experiment and model to analyse the successive invasions and competition of two strains of one of the most emergent virus species in a new ecological environment.

In this study, we present the monitoring of the viral dynamic over a 5 year-period, since the presence of both strains in Réunion Island. In this time period, the Mild strain was completely displaced by the IL strain and was only recovered in mixed infection. To assess for the possibility of a selective advantage, the pathogenicity has been studied by evaluating fitness and virulence of both strains. We demonstrate for the first time the higher virulence on a susceptible tomato variety of the IL strain over the Mild strain of TYLCV without a difference in their fitness, a factor that could explain the observed viral dynamics in Réunion Island.

Materiel and methods

Field Sampling

Plants exhibiting TYLCD symptoms were sampled in 9 locations of the major tomato growing areas in the western and southern part of Réunion Island. Usually sampled fields were highly contaminated. So we choose to sample all range of symptom from low to high severity on each tomato field in order to sample the highest possible viral diversity. From 2004 to 2008, surveys were realized twice a year during the highest periods of whitefly and TYLCD epidemics (February–March and September–October, which correspond to the end and the beginning of the hot and wet season, respectively). Following collection, leaf samples were dehydrated with anhydrous calcium chloride and stored (Bos, 1977).

TYLCV strain characterisation

Total DNA was extracted using the DNeasy Plant miniprep kit (Qiagen, Germany) according to the manufacturer's instructions. In order to differenciate between TYLCV-IL and TYLCV-Mld, we used multiplex polymerase chain reaction (PCR) as described in Lefeuvre et al. (2007).

Virulence measure

To evaluate and compare the virulence of TYLCV-IL and TYLCV-Mld, symptomatology of agroinoculated plants was scored in a trial repeated three times. Each repetition consisted of three time twenty-eight plants agroinoculated with TYLCV-IL[Re4] (AM409201), TYLCV-Mld[Re] (AJ865337) and control, with symptom scoring every 3.5 days during 8 weeks post agroinoculation as described in Delatte et al. (2006). Basically, the notation scale was 1, no symptom, to 10, dead plant, with numbers 1–9 corresponding to the 0–4 scale of Lapidot and Friedmann (2006). The area under disease progress curve (AUDPC) was calculated from symptoms value and were compared using non parametric Kruskal-Wallis and Wilcoxon tests.

Quantification of viral DNA accumulation

Fitness is described for a virus as the aptitude to multiply in its host, and is mainly measured with the quantity of viral DNA copy (Elena et al., 1996). To determine and compare the fitness of TYLCV-IL and TYLCV-Mld, agroinoculated plants were sampled at 35 DPI with 20 plants infected by TYLCV-IL, 20 plants infected by TYLCV-Mld, 30 plants infected by both strain and 10 controlled inoculated plants. The sampling consisted in the three first leaves of every plant were collected and stored at -80°C. Plant samples were then lyophilised (Alpha 1-2 LD plus, Christ), ground and total nucleic acid extracts were prepared as reported above in the section "TYLCV strain characterisation".

Evaluation of viral copy numbers was achieved by real-time quantitative PCR kinetics using ABI PRISM 7000 Sequence Detection System (Applied Biosystems). Real-time PCR was performed with 2µl of appropriate diluted DNA, 40 nM of forward and reverse specific primers for TYLCV and 150 nM of TaqMan® MGB probe (see supplementary figure 1) and 1x TaqMan® Universal PCR Master Mix (Applied Biosystems). The PCR protocol used a degradation step of previously amplified products at 50°C during 2 min, an initial denaturating step at 95°C for 10 min, followed by 40 cycles of 95°C for 15 s and 62°C for 30 s. The specificity of the desired product was

validated with direct sequencing of the PCR products (Macrogen sequencing service, Korea). Quantification viral DNA was based on a standard curve with known amounts of recombinant plasmid DNA and was included in each real-time PCR experiment (Figure 1). Use of pGEM-T-TYLCV-Mld or pGEM-T-TYLCV-Il gave similar standard curves (data not shown). Description of pGEM-T-TYLCV-Mld or pGEM-T-TYLCV-Il was achieved for 10^3 to 10^8 viral copies number.

To avoid possible viral DNA content variation due to extraction, viral copy numbers were compared in relative to total DNA extract, using the NanoDrop 8000 Spectrophotometer (Thermo Fisher Scientific). Fitness results were compared using the nonparametric Kruskal-Wallis and Wilcoxon tests.

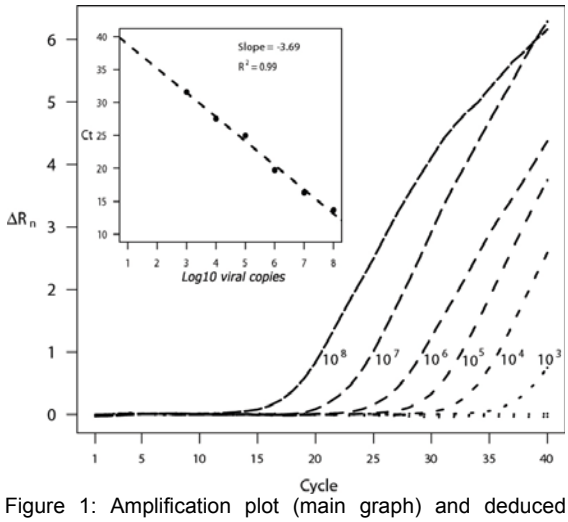


Figure 1: Amplification plot (main graph) and deduced slope curve (upper left graph) of real-time PCR assay performed over pGEM-T-TYLCV-Mld plasmid standards ranging from 10^3 to 10^8 copies. ΔRn : delta normalized reporter (magnitude of the signal).

Recombination detection

To account for the possible emergence of a new TYLCV variant resulting from the recombination between TYLCV-IL and TYLCV-Mld, the replication associated protein open reading frame (Rep ORF) of virus of both field and laboratory coinfected plants were amplified with two specifically designed primers (primer sequences, TYLCV1940F: 5'-ACAACGAAATCCGTGAACAG-3'; TYLCV200R: 5'-TTTTGCCTGTTCTGCTATCAC-3', 30 PCR cycle of 94°C-30s, 55°C-30s and 72°C-60s with an initial denaturation step of 94°C-5min and a final amplification step of 72°C-5min). Amplified fragment products of 1100bp were then cloned into pGem-T and sequenced using Macrogen Inc sequencing service (Korea). Seven samples collected in 2008 and nine agroinoculated plants aged of 8 weeks were treated this way. Sixteen sequences where then aligned using Clustal-W subalignment tool (Thompson *et al.*, 1994) included in MEGA 4 (Tamura *et al.*, 2007) and checked by eye for recombination.

Results

Réunion Island viral dynamics

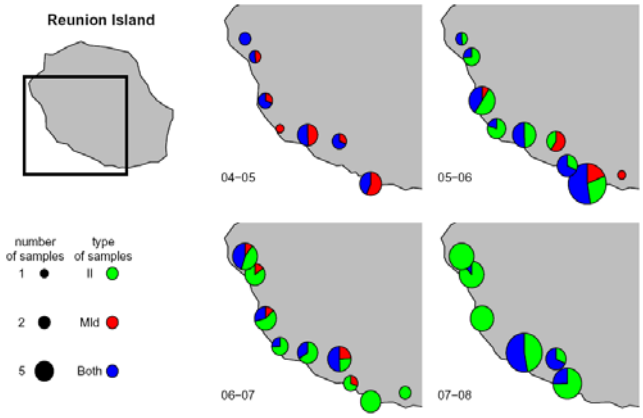


Figure 3: TYLCV-IL and TYLCV-Mld dynamics in Réunion Island from 2004 to 2008. Pie graph size depends of the number of samples tested. Pie graph portions represent the proportion of samples of each type with red for TYLCV-Mld, green for TYLCV-IL and blue for coinfection.

Results of the analysis of tomato samples collected during the 2004 to 2008 growing seasons of the main tomato-growing areas of Réunion Island are summarized in Figure 3. The data showed a clear displacement of TYLCV-Mld by TYLCV-IL in this short period. TYLCV-Mld was dominant in 2004 and progressively decreased in prevalence, with an increasing proportion of samples presenting TYLCV-IL in coinfection or not. TYLCV-Mld was recovered in 2008 only in mixed infection with TYLCV-IL (29%) whereas the majority of collected samples only presented TYLCV-IL infections (71%). To assess for the possible emergence of a new recombinant that could have displaced both TYLCV-Mld and TYLCV-IL, the Rep ORF of seven samples collected in 2008 were sequenced. This portion of the genome is the only that allow the differentiation of both strains, and so that could display visible recombinant pattern. None of the seven sequenced samples, collected throughout the island presented any recombination event and all the sequences were of TYLCV type. Instead of showing that no TYLCV-IL/TYLCV-Mld recombinant could emerge and outperform the parent in the future, these results show that at least, the

Table 1: Fitness and virulence results and associated statistics

		TYLCV-IL[Re4]	TYLCV-Mld[Re]	Coinfection
Number of viruses[$]	mean +/- sd	1.6e7 +/- 6.7e6	1.2e7 +/- 4.2e6	2.2e7 +/- 7.6e6*
	sample size	20	20	30
AUDPC	mean +/- sd	141.6 +/- 30.8*	98.5 +/- 27.2	NA
	sample size	55	33	

[$]number of viral copies per ng of total plant DNA extract
*wilcoxon test p-value < 0.05 for higher number of viruses/AUDPC in comparison to the other modalities
NA: not available

predominant circulating viruses in Réunion Island are of TYLCV-IL type.

Pathogenicity: fitness and virulence

Experiments were performed to determine if the displacement between the TYLCV species observed in epidemics on tomato fields was due to a differential fitness of both viral species in this host. Mean and standard deviation of viral load is reported in table 2. The numbers of viruses are in relative to the total DNA extracts and represented in mean 3.6, 4.8 and 6.6% of total DNA for TYLCV-Mld, TYLCV and mixed infection respectively. The results indicated that no statistical fitness differences at 0.05 statistical level were measured for any strain at 35 days post inoculation (PI) in tomato plants (*P*-value of 0.076). Viral load in mixed infected plants was also measured and compared to single infected one. Strong statistical support for higher viral load in coinfection was determined (P-value of 0.010 in comparison to TYLCV-IL and P-value of 1.3e-6 in comparison to TYLCV-Mld). For nine of the 30 mixed infected plants, Rep ORF was sequenced to determine if this higher fitness was driven by synergism or apparition of recombinant variant. All the sequences were parental-like with one of the “Mld” type and eight of the IL type. Furthermore, both strains were detected by PCR in every mixed infected plant.

Whereas no statistical difference between the two strains was showed for fitness, virulence was measured as being statistically superior for the "Severe" strain. Final symptoms levels were ranging from notes 3 to 8 for TYLCV-Mld and 4 to 9 for TYLCV-IL depending on the plant. A *P*-value of 8.0e-8 was calculated when comparing AUDPC of both infection dynamics (Figure 4). AUDPC of 95.5 (+/- 27.2) and 141.6 (+/-30.8) were measured for TYLCV-Mld and TYLCV-IL respectively.

Discussion

The recent introduction of the IL strain of TYLCV into a tropical insular area where TYLCD was widely present provided an ideal field experiment and model to analyse evolution of viral complexes. In this study, we presented the rapid displacement of the Mild strain by the IL strain of TYLCV over a five year-period monitoring. This displacement result to a situation where TYLCV-Mld is only found in mixed infected plants.

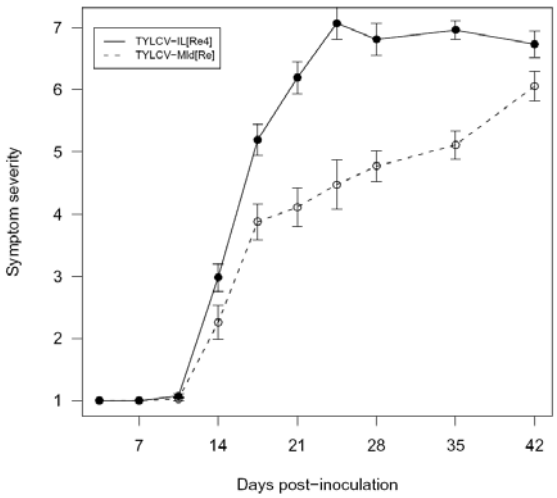


Figure 3: Mean values of disease severity following agro-inoculation of TYLCV-Mld[Re] (dashed line) and TYLCV-IL[Re4] (solid line). Vertical bars represent standard deviation.

Knowledge of the epidemiological factors underlying the selective advantage of TYLCV severe strain is essential for prediction of future progression of TYLCV epidemics and, therefore, for correct control strategies to be implemented. We have investigated the possibility of (1) the possible emergence of new recombinant variants that had outperformed both parents or (2) a higher pathogenicity of the severe strain.

Recombination contributes to the genetic diversification of geminivirus populations, the adaptation to new hosts and changing environmental conditions and has been related to the emergence of some serious diseases these ten last years with new pathogenetic phenotypes (Monci *et al.,* 2002; Zhou *et al.*, 1997). Therefore, we investigate the possibility of the presence of a new recombinant variant between the two TYLCV strains. Even if a small number of plants were tested, the results show that a new "dominant" recombinant viral strain is not likely to be circulating in

Réunion Island on tomato crops and did not emerged in laboratory mixed infected plants. Those results confirm that TYLCV-IL strain alone has the capacity to outperform in the field the TYLCV-Mld strain.

We demonstrated for the first time in controlled conditions on a susceptible tomato variety the higher virulence of the IL strain over the Mild strain without difference in their fitness level. TYLCV-IL presenting a higher selective value in Réunion Island environment than TYLCV-Mld in the ecological context of Réunion Island, one could expect that the symptoms level difference to explain a part of the displacement pattern observed. Symptom severity was positively correlated with the importance of yellowing. *B. tabaci*, as the majority of phytophagous insects as aphids, is attracted by the yellow part of the light spectrum (Mound, 1962; Moericke, 1969). The attractiveness of severely infected plants for insect vector could therefore promote the spread of the most severe viral strain and so explain in part the quick spread of TYLCV-IL. In addition, yellow radiations were reported to support a vegetative behaviour of *B. tabaci*, which should also act as an important factor in the incidence of the disease (Mound, 1962; Rojas *et al.*, 2005).

Another stricking epidemiological characteristic emerging from this epidemiological survey is the persistence of the Mild strain in mixed infection. Fitness results showed a significant higher viral load in mixed infection, suggesting synergism and possible selective advantage for coinfections. Mutualism was demonstrated between the invasive biotype B (dominant in Reunion Island; Delatte et al., 2006) and Asian monopartite begomoviruses (Jiu et al., 2007). This result might explain why TYLCV-Mld, and so coinfections presenting higher viral load, is still recovered in naturally infected plants.

Virulence is a key property of pathogens, and understanding the evolution of virulence has been a major goal for plant pathologist (Sacristan and Garcia-Arenal, 2008). The evolution of virulence may determine important phenomena such as emergence and re-emergence of pathogens, host switch and/or host range expansion and the over coming of host resistance (see Holmes and Rambaut, 2004 and Fargette et al., 2006 for review). However, little is known about the molecular basis of virulence. Experimental studies using Asian begomovirus tomato leaf curl pathological systems (ToLCV and TYLCCNV) have shown that the C4 protein was involved in the severity of symptoms (Gafni and Epel, 2002; Jupin et al., 1994; Rigden et al., 1994) and that the expression of the sole C4 gene produced virus like symptoms in transgenic plant (Dry *et al.,* 2000; Selth *et al.*, 2004; Van Wezel *et al.*, 2002). Considering the recombinant nature of TYLCV-IL, sharing a portion of genome comprising the C4 ORF with a ToLCV-Asian-like ancestors (Navas-Castillo *et al.*, 2000), we can reasonably hypothesize that the C4 protein of TYLCV-IL is implicated in the observed higher virulence. Despite the lack of knowledge about how genome rearrangement could produce new virus variant presenting altered host ranges or different pathogenicities, there are actually very few well supported examples of this having occurred in nature (Fondong *et al.,* 2000; Garcia-Andres *et al.*, 2007b; Monci *et al.*, 2002; Pita *et al.*, 2001; Varsani et al, 2008). This example of TYLCV-IL being more severe and with an emerging character is an important example of how recombination can provide advantage to a virus and drive emergence.

The only known mechanism for spread of begomovirus in nature is through their vector, *B. tabaci* which is a key element of the emergence of this virus genus. Therefore, the differential interaction between TYLCV strains and *B. tabaci* could be an important trait in the preferential dispersion of one virus species or strain (Monci *et al.,* 2002; Sanchez-Campos *et al.*, 1999). Our observations in the laboratory with the use of mass rearing population of biotype B do not suggest such a difference in transmission capacity of the two strains of TYLCV individually (J.M. Lett, unpublished data). These observations seem consistent with the high similarity between coat proteins (CP) of the two strains, the CP being probably the only protein involved in the specificity of transmission (Briddon *et al.,* 1998; Hofer *et al.*, 1997). Hence, the possible differential interaction between TYLCV strains and *B. tabaci* doesn't seem to be an important trait in the virus population composition and dynamics.

Maintenance of viruses between epidemics in alternate cultivated hosts or weeds present in the area provides a survival way throughout the seasonal or intercropping cycle and may play an important role in the emergence and dynamics of plant virus epidemics (Harper *et al.,* 2002; Hull *et al.*, 2000). Implication of some native and cultivated host plant species, acting as reservoirs, has been investigated in the dynamics of TYLCV and TYLCSV in south of Spain (Garcia-Andres *et al.,* 2006; Sanchez-Campos *et al.*, 1999). The possible existence of a differential host range between the two TYLCV strains could be an important aspect in TYLCV epidemiology that must be further studied.

Conclusion

The successive introduction in Réunion Island of two TYLCV strains, with increasing virulence, led to increasingly

losses of tomato crops in open fields (Delatte et al., 2005a). In less than ten years, this TYLCD epidemic has consequences in the mode of production of tomato on the island, with an increasing production in insect proof greenhouses and the progressive abandonment of open field tomato production. In this study, we demonstrate for the first time the higher virulence and the epidemiological advantage of the IL strain of TYLCV in comparison to the Mild one. This information is essential in order to better understand and predict viral epidemics on tomato, to adapt methods of epidemiological surveillance and to develop resistant varieties in breeding programs.

Acknowledgment

The authors thank M. Grondin and D. Fontaine for their excellent technical assistance and F. Perefarres for R coding help. This study was funded by the CIRAD and the Conseil Régional de la Réunion. P. Lefeuvre is a recipient of a PhD fellowship from the Ministère de la Recherche et de l'Enseignement Supérieur.

References

Abhary M, Patil L, Fauquet CM (2007) Molecular biodiversity, taxonomy, and nomenclature of Tomato yellow leaf curl-like viruses. Springer, Tomato Yellow Leaf Curl Disease 85-118.

Antignus EY, Cohen S. (1994) Cloning of Tomato yellow leaf curl virus (TYLCV) and the complete nucleotide sequence of a mild infectious clone. Phytopathology 84: 707-12.

Bos L (1977) Persistance of infectivity of three viruses in plant material dried over $CaCl_2$ and stored under different conditions. Neth J Plant Pathol 83: 217-20.

Briddon RW, Liu S, Pinner MS, Markham PG (1998) Infectivity of African cassava mosaic virus clones to cassava by biolistic inoculation. Arch Virol 143: 2487-92.

Davino S, Davino M, Accotto G (2008) A single-tube PCR assay for detecting viruses and their recombinants that cause tomato yellow leaf curl disease in the Mediterranean basin. J Virol Meth 147: 93-8.

Davino S, Napoli C, Davino M, Accotto G (2006) Spread of Tomato yellow leaf curl virus in Sicily: partial displacement of another geminivirus originally present. Eur J Plant Pathol 114: 293-9.

Delatte H, Holota H, Moury B, Reynaud B, Lett JM (2007) Evidence for a founder effect after introduction of Tomato yellow leaf curl virus-mild in an insular environment. J Mol Evol 65: 112-8.

Delatte H, Holota H, Reynaud B, Dintinger J (2006) Characterisation of a quantitative resistance to vector transmission of Tomato yellow leaf curl virus in Lycopersicon pimpinellifolium. Eur J Plant Pathol 114: 245-53.

Delatte H, Holota H, Naze F, Peterschmitt M, Reynaud B, Lett JM (2005a) The presence of both recombinant and non recombinant strains of *Tomato yellow leaf curl virus* on tomato in Réunion Island. Plant Pathol 54: 262.

Delatte H, Martin DP, Naze F, Goldbach R, Reynaud B et al. (2005b) South West Indian Ocean islands tomato begomovirus populations represent a new major monopartite begomovirus group. J Gen Virol 86: 1533-42.

Dry I, Krake L, Mullineaux P, Rezaian A (2000) Regulation of tomato leaf curl viral gene expression in host tissues. Mol Plant Microbe Interact 13: 529-37.

Elena SF, Gonzalezcandelas F, Novella IS, Duarte EA, Clarke DK, Domingo E, Holland JJ, Moya A (1996) Evolution of fitness in experimental populations of vesicular stomatitis virus. Genetics 142: 673-9.

Fargette D, Konate G, Fauquet C, Muller E, Peterschmitt M et al. (2006) Molecular ecology and emergence of tropical plant viruses. Annu Rev Phytopathol 44: 235-60.

Fondong VN, Pita JS, Rey ME, de Kochko A, Beachy RN et al. (2000) Evidence of synergism between African cassava mosaic virus and a new double-recombinant geminivirus infecting cassava in Cameroon. J Gen Virol 81: 287-97.

Gafni Y, Epel B L (2002) The role of host and viral proteins in intra- and inter-cellular trafficking of geminiviruses. Physiol Mol Plant Pathol 60: 231-41.

Garcia-Andres S, Accotto G, Navas-Castillo J, Moriones E (2007) Founder effect, plant host, and recombination shape the emergent population of begomoviruses that cause the tomato yellow leaf curl disease in the Mediterranean basin. Virology 359: 302-12.

Garcia-Andres S, Monci F, Navas-Castillo J, Moriones E (2006) Begomovirus genetic diversity in the native plant reservoir Solanum nigrum: Evidence for the presence of a new virus species of recombinant nature. Virology 350: 433-42.

Harper G, Hull R, Lockhart B, Olszewski N (2002) Viral sequences integrated into plant genomes. Annu Rev Phytopathol 40: 119-36.

Jiu M, Zhou XP, Tong L, Xu J, Yang X, Wan FH, Liu SS (2007) Vector-virus mutualism accelerates population increase of an invasive whitefly. PLoS ONE 2: e182.

Jupin I, De Kouchkovshy F, Jouanneau F, Gronenborn B (1994) Movement of tomato yellow leaf curl geminivirus (TYLCV) : involvement of the protein encoded by ORF C4. Virology 204: 82-90.

Hofer P, Bedford ID, Markham PG, Jeske H, Frischmuth T (1997) Coat protein gene replacement results in whitefly transmission of an insect nontransmissible geminivirus isolate. Virology 236: 288-95.

Holmes, E. C. & Rambaut, A. (2004). Viral evolution and the emergence of SARS coronavirus. Philos Trans R Soc Lond B Biol Sci 359: 1059-65.

Hull R, Harper G, Lockhart B (2000) Viral sequences integrated into plant genomes. Trends Plant Sci 5: 362-5.

Lapidot M, Ben-Joseph R, Cohen L, Machbash Z, Levy D (2006) Development of a scale for evaluation of Tomato yellow leaf curl virus resistance level in tomato plants. Phytopathology 96: 1404-8.

Lefeuvre P, Hoareau M, Delatte H, Reynaud B, Lett J (2007) A multiplex PCR method discriminating between the TYLCV and TYLCV-Mld clades of Tomato yellow leaf curl virus. J Virol Meth 144: 165-8.

Moericke V (1969) Hostplant specific colour behaviour by Hyalopterus pruni (Aphididae). Entomol Exp Appl 12: 524-34.

Monci F, Sanchez-Campos S, Navas-Castillo J, Moriones E (2002) A natural recombinant between the geminiviruses Tomato yellow leaf curl Sardinia virus and Tomato yellow leaf curl virus exhibits a novel pathogenic phenotype and is becoming prevalent in Spanish populations. Virology 303: 317-26.

Mound LA (1962) Studies on the olfaction and colour sensitivity of Bemisia tabaci (GENN.) (Homoptera, Aleurodidae). Entomol Exp Appl 5: 99–104.

Navas-Castillo J, Sanchez-Campos S, Noris E, Louro D, Accotto GP et al. (2000) Natural recombination between Tomato yellow leaf curl virus-is and Tomato leaf curl virus. J Gen Virol 81: 2797-801.

Pita JS, Fondong VN, Sangare A, Otim-Nape GW, Ogwal S et al. (2001) Recombination, pseudorecombination and synergism of geminiviruses are determinant keys to the epidemic of severe cassava mosaic disease in Uganda. J Gen Virol 82: 655-65.

Peterschmitt, M., Granier, M., Mekdoud, R., Dalmon, A., Gambin, O., Vayssières, J. F. & Reynaud, B. (1999). First report of tomato yellow leaf curl virus in Réunion Island. Plant Disease 83: 303.

Reynaud, B., Wuster, G., Delatte, H., Soustrade, I., Lett, J. M., Gambin, O. and Peterschmitt, M. (2003). Les maladies à bégomovirus chez la tomate dans les départements français d'Outre-Mer. Phytoma 562: 13-7.

Rigden J E, Krake LR, Rezaian MA, Dry IB (1994) ORF C4 of tomato leaf curl geminivirus is a determinant of symptom severity. Virology 204: 847-50.

Rojas MR, Hagen C, Lucas WJ, Gilbertson RL (2005) Exploiting chinks in the plant's armor: evolution and emergence of geminiviruses. Annu Rev Phytopathol 43: 361-94.

Sacristan S, Garcia-Arenal F (2008) The evolution of virulence and pathogenicity in plant pathogen populations. Mol Plant Pathol 9: 369-84.

Sanchez-Campos S, Navas-Castillo J, Camero R, Soria C, Diaz J et al. (1999) Displacement of tomato yellow leaf curl virus (TYLCV)-Sr by TYLCV-Is in tomato epidemics in spain. Phytopathology 89: 1038-43.

Seal SE, Jeger MJ, Van den Bosch F (2006) Begomovirus evolution and disease management. Adv Virus Res 67: 297-316.

Selth L, Randles J, Rezaian M (2004) Host responses to transient expression of individual genes encoded by Tomato leaf curl virus. Mol Plant Microbe interact 17: 27-33.

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596-9.

Thompson, JD, Higgins, DG, Gibson, TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-80.

Ueda, S., Kimura, T., Onuki, M., Hanada, K., Iwanami, T., 2004. Three distinct groups of isolates of *Tomato yellow leaf curl virus* in Japan and construction of an infectious clone. J Gen Plant Pathol 70, 232-8.

Van Wezel R, Dong X, Blake P, Stanley J, Hong Y (2002) Differential roles of geminivirus Rep and AC4 (C4) in the induction of necrosis in Nicotiana benthamiana. Molecular Plant Pathology 3: 461-71.

Varsani, A., Shepherd, D.N., Monjane, A.L., Owor, B.E., Erdmann, J.B., Rybicki, E.P., Peterschmitt, M., Briddon, R.W., Markham, P.G., Oluwafemi, S., Windram, O.P., Lefeuvre, P., Lett, J.M., and Martin, D.P., (2008). Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. J Gen Virol 89: 2063 – 74.

Zhou X, Liu Y, Calvert L, Munoz C, Otim-Nape GW et al. (1997) Evidence that DNA-A of a geminivirus associated with severe cassava mosaic disease in Uganda has arisen by interspecific recombination. J Gen Virol 78: 2101-11.

Chapitre II : Les bégomovirus du Sud-Ouest de l'océan Indien : un *Melting Pot* viral

Les premières études concernant la diversité des begomovirus dans les îles du Sud-Ouest de l'océan Indien (South West Indian Ocean ou SWIO) ont permis la mise en évidence de nouvelles espèces virales tout en montrant que la recombinaison avait largement participé à la création de la diversité observée (Delatte et al., 2005b). Cette étude ainsi que de nombreuses autres ont mis en évidence une diversité virale jusqu'à lors inattendue, ce qui a même permis au *Begomovirus* de détrôner les *Potyvirus* au rang de genre de phytovirus présentant le plus d'espèce (Fauquet et al., 2008). Si on ne peu comparer directement le nombre d'espèce de genres viraux différents, sachant que les seuils de distinction taxonomique et les contraintes virales ne sont pas les mêmes, cet exemple illustre néanmoins les avancées importantes réalisées ces dernières années dans la découverte de la diversité des bégomovirus. Bien évidemment, ces nouvelles descriptions virales sont imputables au regain d'intérêt pour ce genre viral du fait de l'implication des bégomovirus dans la plupart des nouvelles maladies virales émergentes dans le monde. Un nombre très important d'études a ainsi permis de décrire un peu plus les contours de la diversité en bégomovirus, en identifiant très souvent la présence de tel ou tel évènement de recombinaison au sein des nouvelles séquences virales obtenues. Il était dès lors indispensable de décrire ces complexes viraux à une plus large échelle et d'avoir une meilleure compréhension de la distribution des évènements de recombinaison.

A partir d'échantillons collectés entre 2002 et 2005 sur des plantes cultivées dans les différentes îles du Sud-Ouest de l'océan Indien (archipel des Comores, Madagascar et archipel des Seychelles), nous avons obtenus 14 nouvelles séquences virales de bégomovirus, constituant une diversité jusqu'à lors non décrite, puisque 7 nouvelles espèces ont pu être décrites à partir de ces échantillons.

La reconstruction phylogénétique réalisée à partir des nouvelles séquences de bégomovirus SWIO obtenues ainsi que celles disponibles dans les bases de données mondiales (GenBank-EMBL-DDJP) montre que ces virus sont associés au groupe « Africain-Méditerranéen » des bégomovirus monopartites et bipartites et qu'ils présentent une origine polyphylétique. Néanmoins, la majorité des espèces virales SWIO sont associées au sein d'un même groupe phylogénétique au

sein duquel se retrouve également des virus bipartites. Ce regroupement suggère que la plupart de ces virus sont apparentés et qu'ils ont vraisemblablement étés isolés des autres populations de bégomovirus durant une longue période. D'autre part, le regroupement de ces virus monopartites avec des virus bipartites suggère une origine commune à tous ces virus, avec soit la perte de l'ADN B par les virus monopartites, soit l'acquisition de l'ADN B par les virus bipartites. Le regroupement de ces virus suggère que plutôt qu'une structuration génétique liée à l'hôte, les premières radiations de virus se sont effectuées suivant l'origine géographique, avec pour résultat l'émergence de virus capables d'infecter différentes espèces de plantes cultivées à partir d'un probable ancêtre commun. L'importante divergence entre séquences virales identifiées à partir de plantes hôtes cultivées (ici originaires du Nouveau Monde, hôtes secondaires) suggère l'existence d'une diversité virale encore plus grande au sein des plantes indigènes réservoirs. Ces résultats soulignent la nécessité de réaliser un échantillonnage de grande envergure sur les plantes hôtes primaires dans le but d'appréhender toute la diversité des bégomovirus de la sous région et d'estimer ainsi un le risque potentiel pour l'agriculture régionale. La question de l'existence actuelle de ces populations virales en sympatrie ou en allopatrie au sein des plantes indigènes réservoirs reste ouverte.

L'étude de la recombinaison réalisée à partir de séquences décrites ici, ainsi qu'un jeu de donnée représentant l'ensemble de la diversité disponible des bégomovirus (hormis ceux du nouveau monde plus divergents) a souligné une fois de plus la présence d'un grand nombre de recombinaison. Au-delà de décrire une nouvelle fois la présence de séquences recombinantes, nous avons à l'aide de tests statistiques démontré la distribution non-aléatoire des évènements de recombinaison sur le génôme des bégomovirus avec la présence de points chauds (*hot spot*) et de points froids (*cold spot*) de recombinaison. Les *cold spots* de recombinaison se retrouvent principalement au sein du gène de la protéine de capside alors que les *hot spots* de recombinaison sont localisés au niveau des zones non codantes ainsi que des ORF de sens complémentaire. Ce travail, ainsi que d'autres études publiées à la même époque (Garcia-Andres et al., 2007b; Owor et al., 2007) ont permis d'avancer des hypothèses sur une possible origine mécanistique des évènements de recombinaison. Des conflits entre des complexes de réplication de l'ADN viral et de transcription des ORFs complémentaires (Jeske et al., 2001), qui évoluent en sens contraire, entraîneraient la formation de réplicons partiels. Ceux-ci interviendraient dans la réplication dépendante de la recombinaison (RDR). En présence de virus

d'origines différentes, il y aurait alors création de recombinants. Ces prédispositions biochimiques liées à la réplication et à la transcription du virus entraîneraient une augmentation de la recombinaison des ORF de sens complémentaire et expliquerait en partie la très grande capacité de recombinaison des bégomovirus. L'association de ces raisons mécanistiques et de la sélection opérant sur les recombinants créés expliquerait les profils de recombinaison observés.

En soulignant l'extraordinaire diversité des virus des îles de l'océan Indien, des détails des histoires évolutives ont été mis en avant. La diversité observée des virus de la région sur des plantes hôtes secondaires originaires du Nouveau Monde (Varsani et al., 2008) souligne la nécessité de réaliser un échantillonnage plus large sur les îles SWIO ainsi que sur la côte Est Africaine. Malgré tout, notre étude a permis (1) de démontrer le rôle essentiel de la recombinaison dans l'évolution de ce genre viral et (2) d'avancer des hypothèses fortes sur la manière dont ce phénomène semble se produire.

Journal of General Virology (2007), **88**, 3458–3468 **DOI** 10.1099/vir.0.83252-0

Begomovirus 'melting pot' in the south-west Indian Ocean islands: molecular diversity and evolution through recombination

P. Lefeuvre,[1] D. P. Martin,[2] M. Hoareau,[1] F. Naze,[1] H. Delatte,[1] M. Thierry,[1] A. Varsani,[3] N. Becker,[4] B. Reynaud[1] and J.-M. Lett[1]

Correspondence
J.-M. Lett
lett@cirad.fr

[1]CIRAD, UMR 53 PVBMT CIRAD-Université de la Réunion, Pôle de Protection des Plantes, 7 Chemin de l'IRAT, 97410 Saint Pierre, La Réunion, France

[2]Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory 7925, South Africa

[3]Electron Microscopy Unit, University of Cape Town, Rondebosch 7701, South Africa

[4]Museum National d'Histoire Naturelle, Dept RDDM, USM 501, CNRS UMR 5166, Evolution des Régulations Endocriniennes, 57 rue Cuvier, CP 32, 75005 Paris, France

During the last few decades, many virus species have emerged, often forming dynamic complexes within which viruses share common hosts and rampantly exchange genetic material through recombination. Begomovirus species complexes are common and represent serious agricultural threats. Characterization of species complex diversity has substantially contributed to our understanding of both begomovirus evolution, and the ecological and epidemiological processes involved in the emergence of new viral pathogens. To date, the only extensively studied emergent African begomovirus species complex is that responsible for cassava mosaic disease. Here we present a study of another emerging begomovirus species complex which is associated with serious disease outbreaks in bean, tobacco and tomato on the south-west Indian Ocean (SWIO) islands off the coast of Africa. On the basis of 14 new complete DNA-A sequences, we describe seven new island monopartite begomovirus species, suggesting the presence of an extraordinary diversity of begomovirus in the SWIO islands. Phylogenetic analyses of these sequences reveal a close relationship between monopartite and bipartite African begomoviruses, supporting the hypothesis that either bipartite African begomoviruses have captured B components from other bipartite viruses, or there have been multiple B-component losses amongst SWIO virus progenitors. Moreover, we present evidence that detectable recombination events amongst African, Mediterranean and SWIO begomoviruses, while substantially contributing to their diversity, have not occurred randomly throughout their genomes. We provide the first statistical support for three recombination hot-spots (V1/C3 interface, C1 centre and the entire IR) and two recombination cold-spots (the V2 and the third quarter of V1) in the genomes of begomoviruses.

Received 20 June 2007
Accepted 9 August 2007

INTRODUCTION

The genus *Begomovirus* (family *Geminiviridae)* is characterized by dicotyledonous plant-infecting, whitefly-transmitted viruses. Begomoviruses have either monopartite or bipartite genomes that are encapsidated as circular single-stranded DNA (ssDNA) molecules within twin icosahedral particles. During the last two decades, new begomovirus species have emerged worldwide, probably as a consequence of the spread of one or more highly polyphagous biotypes of their insect vector, *Bemisia tabaci* (Rybicki & Pietersen, 1999). Usually multiple begomovirus species have emerged simultaneously in a given region, with the ensuing species complexes causing diseases in a wide variety of plant species, including many of great agricultural importance.

The rate at which new species are emerging is perhaps best exemplified by the diversity of the almost 700 full begomovirus DNA-A sequences currently deposited in public sequence databases. Given the 89 % identity threshold of the International Committee on Taxonomy

Supplementary material is available with the online version of this paper.

of Viruses (ICTV), these genomes represent more than 200 species (Fauquet *et al.*, 2007).

While providing a major contribution to the richness of currently observed begomovirus species diversity, recombination continues both to fuel begomovirus diversification and complicate the classification of new species. The important contribution of recombination to geminivirus evolution is now well established (Umaharan *et al.*, 1998; Padidam *et al.*, 1999) and it is suspected that it is directly responsible for the emergence of many of the most agriculturally damaging begomovirus species complexes (Zhou *et al.*, 1997; Monci *et al.*, 2002; Garcia-Andres *et al.*, 2006). Despite this, very little is actually known either about why recombination seems to contribute to the emergence of species complexes, or how recombinants with enhanced pathogenicity arise and proliferate. Furthermore, both the biochemical processes that determine the kinds of recombinant genomes produced, and the evolutionary processes that determine which of these survive, remain a complete mystery. However, some studies have indicated that recombination hot-spots may exist within begomovirus genomes (Stanley, 1995; Ndunguru *et al.*, 2005; Fauquet *et al.*, 2005; Garcia-Andres *et al.*, 2007). Identifying the locations of any recombination hot- and cold-spots within begomovirus genomes sampled from nature would certainly be a valuable first step towards understanding the underlying processes controlling the generation and spread of recombinants within species complexes.

We decided to quantitatively evaluate the importance of recombination in the genetic diversification of begomoviruses within a newly discovered monopartite begomovirus species complex indigenous to the south-west Indian Ocean (SWIO) islands off the coast of Africa. Despite the pace at which begomovirus diversity has been explored in the past few years, very few full-length African begomovirus DNA-A sequences other than those of the African cassava mosaic disease (CMD) pathosystem are presently available. To increase the richness of the available African begomovirus genome sequence data, we therefore extended previous preliminary surveys of monopartite begomovirus species on the islands of Madagascar, Comoros and Seychelles archipelagos (Lefeuvre *et al.*, 2007; Delatte *et al.*, 2005b). We describe the molecular diversity and taxonomic relationships of 14 SWIO island begomovirus isolates, including seven new species, causing recent plant disease epidemics in the SWIO islands. Importantly, when analysed together with African and Mediterranean begomovirus sequences, we find solid statistical evidence of recombination hot- and cold-spots within the DNA-A components of these viruses. This result may indicate how and why recombination makes such a substantial contribution to begomovirus diversity in general.

METHODS

Sampling and DNA extraction. Tomato (*Solanum lycopersicon*), tobacco (*Nicotiana tabaci*) and bean (*Phaseolus vulgaris*) leaves presenting leaf curling symptoms were collected from individual plants on the islands of the Comoros archipelago (Anjouan, Grande Comore, Mayotte and Moheli), the Seychelles archipelago (Mahé) and Madagascar (Table 1) and stored dried (Bos, 1977). Total DNA was extracted using a DNeasy Plant miniprep kit (Qiagen) according to the manufacturer's instructions.

PCR detection. Polymerase chain reaction (PCR) was used to amplify two fragments from the extracted DNA of all samples using two degenerate primer sets: AV494-AC1048 (Wyatt & Brown, 1996), and VD360-CD1266 (Delatte *et al.*, 2005b). PCR reactions were carried out as described in Delatte *et al.* (2005b). The presence/absence of a DNA-B genome component and DNA-β molecules were also assessed for each of the isolates using, respectively, the PCR primer sets PBL1v2040-PCRc1 (Rojas *et al.*, 1993) and Beta 1-Beta 2 (Briddon *et al.*, 2002).

Cloning strategies. Circular viral DNA molecules were amplified using a TempliPhi kit (GE Healthcare) as described by Inoue-Nagata *et al.* (2004). Full genomes were cloned into the vector pBC-KS in the *Hin*dIII restriction site for AM701758, AM701759, AM701766, AM701767 and AM491778 and in the *Bam*HI restriction site for all others. A complete DNA-A-like component for each isolate was sequenced by gene walking using the Macrogen sequencing service (Korea).

Phylogenetic analysis. Full DNA-A-like sequences from 14 isolates (this study) were arranged so that the first nucleotide in the sequence corresponded to the first base (adenine) of virion strand replication (Laufs *et al.*, 1995). Forty-one other full DNA-A and DNA-A-like sequences of related viruses were obtained from public sequence databases using TaxBrowser (http://www.ncbi.nlm.nih.gov/) on May 2006. Multiple sequence alignments were constructed using partial order graphs (POA) (Lee *et al.*, 2002), the CLUSTAL W (Thompson *et al.*, 1994) based subalignment tool available in MEGA 3.1 (Kumar *et al.*, 2004) and manual editing.

The optimal model of sequence evolution defined by ModelTest (Posada, 2006) was used for phylogenetic reconstruction (GTR+I+G). The maximum-likelihood (ML) tree was determined from a preliminary neighbour-joining (NJ) analysis using PAUP* with the heuristic search algorithm. In addition to these analyses, we performed Bayesian phylogenetic reconstruction on the full dataset using MrBayes 3.1.2 (Ronquist & Huelsenbeck, 2003). Four runs with six Markov chains were conducted simultaneously for 1 000 000 generations starting from random initial trees, and sampled every 100 generations. Variation in the ML scores in this sample was examined graphically with Tracer (Rambaut & Drummond, 2004). The trees generated prior to stabilization of ML scores were discarded with the consensus phylogeny and posterior probability of their nodes being determined with a burn-in of 25 %. The method of Shimodaira & Hasegawa (1999) implemented in PAUP* was used to test whether the ML scores of the NJ, ML and Bayesian phylogenetic reconstructions fell within the same confidence limits.

Recombination analyses. Detection of potential recombinant sequences, identification of likely parental sequences and localization of possible recombination breakpoints was carried out on a 178-sequence alignment (170 begomovirus, seven curtovirus and one topocuvirus sequences) using the RDP (Martin & Rybicki, 2000), GENECONV (Padidam *et al.*, 1999), BOOTSCAN (Martin *et al.*, 2005a), MAXIMUM CHI SQUARE (Smith, 1992), CHIMAERA (Martin *et al.*, 2005b) and SISTER SCAN (Gibbs *et al.*, 2000) recombination detection methods as implemented in RDP3 (Martin *et al.*, 2005b), available from http://darwin.uvigo.es/rdp/rdp.html (see the RDP project file submitted as supplementary material for full details of program settings). The analysis was performed with default settings for the different

Table 1. Geographical origin and characterization of SWIO begomovirus isolates

ToLCAnjV-[Anj:Oua3:04], Tomato leaf curl Anjouan virus - [Anjouan:Ouani3:2004] (GenBank accession no. AM701758); ToLCKMJV-[Anj:Bam5:04], Tomato leaf curl Comoros virus - [Anjouan:Bambas5:2004] (AM701759); TbLCKMV-[GC:Sim18:04], Tobacco leaf curl Comoros virus - [Grande Comore:Simboussa18:2004] (AM701760); ToLCAntV-[GC:Dim44:04], Tomato leaf curl Antsiranana virus - [Grande Comore:Dimadjou44:2004] (AM701761); TbLCZV-[GC:Fbz95:05], Tobacco leaf curl Zimbabwe virus - [Grande Comore:Foumboudziouni95:2005] (AM701756); TbLCKMV-[GC:Fou99:05], Tobacco leaf curl Comoros virus - [Grande Comore:Foubouni99:2005] (AM701762); ToLCKMV-[YT:Dem:03], Tomato leaf curl Comoros virus - [Mayotte:Dembeni:2003] (AJ865341); ToLCYTV-[YT:Kah:03], Tomato leaf curl Mayotte virus - [Mayotte:Kahani:2003] (AJ865340); ToLCMohV-[KM:Fom163:05], Tomato leaf curl Moheli virus - [Comoros:Fomboni163:2005] (AM701763); ToLCAntV-[MG:Nam3:01], Tomato leaf curl Antsiranana virus - [Madagascar:Namakely3:2001] (AM701764); ToLCDiaV-[MG:Nam5:01], Tomato leaf curl Diana virus - [Madagascar:Namakely5:2001] (AM701765); ToLCAntV-[MG:Ant6:01], Tomato leaf curl Antsiranana virus - [Madagascar:Antsalaka6:2001] (AM701766); ToLCAntV-[MG:Mia1:01], Tomato leaf curl Antsiranana virus - [Madagascar:Miandrivazo1:2001] (AM701767); ToLCTolV-[MG:Mia2:01], Tomato leaf curl Toliara virus - [Madagascar:Miandrivazo2:2001] (AM701768); CLCuGV-Be[MG:For:01], Cotton leaf curl Gezira virus - Bean[Madagascar:Fort Dauphin:2001] (AM701757); ToLCMGV-Men[MG:Mor:01], Tomato leaf curl Madagascar virus - Menabe[Madagascar:Morondava:2001] (AJ865338); ToLCMGV-Ats[MG:Tol:01], Tomato leaf curl Madagascar virus - Atsimo[Madagascar:Toliary:2001] (AJ865339); ToLCSCV-[Mah:VE77:04], Tomato leaf curl Seychelles virus - [Mahé:Val d'Endor77:2004] (AM491778).

Region	Island or province/ district	Village	Host plant	Year	Acronym	Closest available virus (% identity)	DNA length	Predicted coding capacity (amino acid)						First description
								V2	V1	C1	C2	C3	C4	
Comoros archipelago	Anjouan	Ouani	Tomato	2004	ToLCAnjV-[Anj:Oua3:04]*	ToLCYTV-[Dem] (96 %)	2781	116	258	389	135	134	77†	This study
		Bambas	Tomato	2004	ToLCKMJV-[Anj:Bam5:04]	ToLCYTV-[Dem] (82 %)	2774	98‡	258	359	135	134	100	This study
	Grande Comore	Simboussa	Tobacco	2004	TbLCKMV-[GC:Sim18:04]*	ToLCYTV-[Dem] (83 %)	2755	116	258	358	135	134	100	This study
		Dimadjou	Tomato	2004	ToLCAntV-[GC:Dim44:04]*	ToLCMGV-[Mor] (86 %)	2772	98‡	258	215†	135	109‡	§	This study
		Foumboudziouni	Tobacco	2005	TbLCZV-[GC:Fbz95:05]	TbLCZV-[ZW] (96 %)	2764	116	258	371‖	135	134	85†	This study
		Foubouni	Tobacco	2005	TbLCKMV-[GC:Fou99:05]*	ToLCYTV-[Dem] (83 %)	2758	133‖	258	358	135	134	100	This study
	Mayotte	Dembeni	Tomato	2003	ToLCKMV-[YT:Dem:03]	ToLCYTV-[Dem] (88 %)	2765	116	258	358	135	134	100	Delatte *et al.* (2005b)
		Kahani	Tomato	2003	ToLCYTV-[YT:Kah:03]	ToLCKMV-[Kah] (88 %)	2768	116	258	379	135	134	143	Delatte *et al.* (2005b)
	Mohéli	Fomboni	Tomato	2005	ToLCMohV-[KM:Fom163:05]*	ToLCYTV-[Dem] (88 %)	2756	118	258	235†	135	134	100	This study
Madagascar	Antsiranana/Diana	Namakely	Tomato	2001	ToLCAntV-[MG:Nam3:01]*	ToLCYTV-[Dem] (86 %)	2769	116	258	359	135	134	100	This study
	Antsiranana/Diana	Namakely	Tomato	2001	ToLCDiaV-[MG:Nam5:01]*	ClCuGV-[Sha] (82 %)	2754	122	258	359	135	134	85†	This study
	Antsiranana/Diana	Antsalaka	Tomato	2001	ToLCAntV-[MG:Ant6:01]*	ToLCUGV-[Iga] (86 %)	2775	116	258	402‖	135	134	100	This study
	Toliara/Menabe	Miandrivazo	Tomato	2001	ToLCAntV-[MG:Mia1:01]*	ToLCUGV-[Iga] (86 %)	2774	116	258	358	135	134	100	This study
	Toliara/Menabe	Miandrivazo	Tomato	2001	ToLCTolV-[MG:Mia2:01]*	ToLCTZV-[Aru] (83 %)	2764	116	258	376	135	134	85†	This study
	Toliara/Anosy	Fort Dauphin	Bean	2001	CLCuGV-Be[MG:For:01]	ClCuGV-[Ok:Sha] (89 %)	2754	122	258	362	134	133	100	This study
	Toliara/Menabe	Morondava	Tomato	2001	ToLCMGV-Men[MG:Mor:01]	ToLCMGV-[Tol] (94 %)	2777	116	258	359	135	134	100	Delatte *et al.* (2005b)
	Toliara/Atsimo	Toliary	Tomato	2001	ToLCMGV-Ats[MG:Tol:01]	ToLCMGV-[Mor] (94 %)	2775	116	258	359	135	134	100	Delatte *et al.* (2005b)
Seychelles	Mahé	Val d'Endor	Tomato	2004	ToLCSCV-[Mah:VE77:04]*	ToLCYTV-[Dem] (81 %)	2742	116	258	375‖	183	§	85†	This study

*New species proposal.
†ORF containing a premature stop codon.
‡ORF containing a frame-shift mutation.
§No predicted ORF identified.
‖ORF containing an in-frame ATG codon upstream of the putative initiation codon.

detection methods and a Bonferroni corrected *P*-value cut-off of 0.05. The breakpoint positions and recombinant sequence(s) inferred for every detected potential recombination event were manually checked and adjusted where necessary using the extensive phylogenetic and recombination signal analysis features available in RDP3. Once a set of unique potential recombination events was identified, we compiled a breakpoint map by plotting the positions of all clearly identifiable breakpoints. A breakpoint density plot was then constructed from this map and the statistical significance of potential breakpoint hot- and cold-spots was tested as described in Heath *et al.* (2006). Briefly, the statistical analysis used takes the observed distribution of polymorphic sites in an alignment and randomly maps all the observed recombination events to this distribution, such that the real and randomly mapped events all involve exchanges of sequence tracts containing the same numbers of polymorphic sites. Doing this accounts for the fact that uneven distribution of polymorphic sites along the length of an alignment makes the identification of breakpoints in certain alignment regions more difficult than in others. This random mapping process is then repeated 1000 times and the actual distribution of breakpoints is compared to that of the 1000 permuted mappings using two tests. The first is a 'global' test which determines whether there are breakpoint clusters in the real distribution with more breakpoints than generally occur in the distributions determined from the permuted datasets. This analysis is highly conservative as it ignores the fact that it will be far harder to detect a genuinely significant breakpoint cluster in regions of conserved sequence than it will be to detect one in regions of more diverse sequence (as mentioned above, breakpoints are most easily and accurately detectable where diversity is high). Therefore, a second, less conservative, 'local' test compares corresponding portions of the real and permuted breakpoint distributions and determines whether local regions of the real distribution contain significantly more breakpoints than generally occur in corresponding regions of the permuted datasets. The *P*-values associated with both the global and local tests are simply the proportions of permuted datasets with greater breakpoint clusters. Whilst we judged *P*-values <0.05 to be significant for the conservative global test, to guard against false positives, we judged *P*-values <0.01 as being significant for the less conservative local test.

Species distinction analysis. Sequence identity was computed from the precedent multiple sequence alignments without curtovirus and topocuvirus sequences (170 sequences) using the dna.dist function available in the R package, APE (Paradis *et al.*, 2004). We identified genotypes belonging to different species using the ICTV-recommended 89 % complete DNA-A/DNA-A-like sequence identity threshold for species demarcation. To take into account possible influences of discovery order on species number estimates, we repeated the species identification operation 1000 times using the sequences in a random order. The mean and standard deviation of identified species numbers were calculated from the results of these permutations.

RESULTS

Cloning and sequencing

The complete nucleotide sequences of 14 DNA-A-like components were determined from dried leaf extracts originating from five different SWIO islands (Table 1; Fig. 1). While PCR amplification and cloning of apparently full-length DNA-A-like components was possible from all symptomatic leaf samples, DNA-B and DNA-β specific PCRs yielded no amplification products. This implied that the 14 viruses were most likely all monopartite, as it has been shown previously for four SWIO species with agroinfectious clones (Delatte *et al.*, 2005b). These DNA-A-like sequences were all of typical monopartite begomovirus size, ranging from 2742 to 2781 nt. Most of the sequences had predicted genes typical of monopartite begomoviruses in terms of both size and position. However, for some sequences, deduced protein sequences contained potential translation errors, as indicated in Table 1. All sequences are available in GenBank/EMBL/DDJP database under the accession numbers given in Table 1.

Species distinction

On the basis of nucleotide identity to their closest known relatives (Table 1), 7 of the 14 new sequences represent new species. Species names for these viruses, based on host plant and region of origin, are proposed in Table 1. The remaining seven sequences share >89 % identity with DNA-A-like sequences of previously described species such as *Cotton leaf curl Gezira virus* (CLCuGV), *Tomato leaf curl Comoros virus* (ToLCKMV), *Tomato leaf curl Mayotte virus* (ToLCYTV), *Tomato leaf curl Madagascar virus*

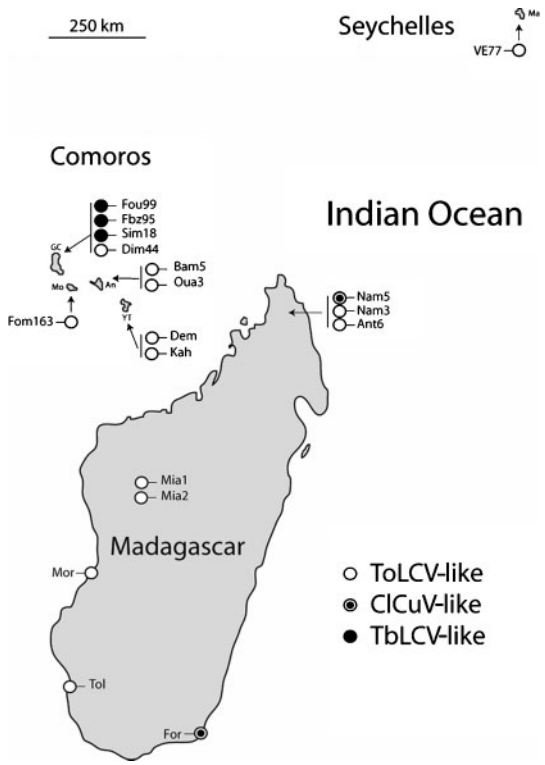


Fig. 1. Map of south-west Indian Ocean islands showing the distribution of begomovirus isolates examined in this study. Colours are used to designate related species. Isolate acronyms: Ant, Antsiranana; Bam, Bambas; Dem, Dembeni; Dim, Dimadjou; Fbz, Foumboudziouni; Fom, Fomboni; Fou, Foubouni; For, Fort Dauphin; Kah, Kahani; Mia, Miandrivazo; Mor, Morondava; Nam, Namakely; Oua, Ouani; Tol, Toliary. Islands acronyms: An, Anjouan; GC, Grande Comore; Ma, Mahé; Mo, Moheli; YT, Mayotte.

(ToLCMGV) and *Tobacco leaf curl Zimbabwe virus* (TbLCZV).

We assembled a dataset of 170 African, Mediterranean and SWIO begomovirus DNA-A-like sequences containing 51 ICTV-designated species and the seven tentative new species described here. We applied our species-counting algorithm test to this dataset and determined that, depending on the discovery order, 49 species ($\pm$1.3) should be defined as such. While this number is not substantially different from 51 classified by the ICTV, our dataset contains an additional seven sequences that are almost certainly legitimate novel species. We further attempted to determine which parts of the virus genomes mostly contribute to pairwise distance scores that are currently the primary taxonomic measure used for species demarcation. Using alignments of individual ORFs indicated that 32$\pm$1.1 (V1 ORF), 25$\pm$1.3 (V2 ORF), 43.5$\pm$1.1 (C1 ORF), 29.5$\pm$1.3 (C2 ORF), 29.5$\pm$1.1 (C3 ORF) and 43.5$\pm$1.3 (C4 ORF) groups of sequences might be defined if we consider only these subgenomes and a 89% identity threshold. This clearly indicated that sequences of the C1 and C4 ORFs are the primary source of taxonomic signal in begomoviruses.

Phylogenetic analysis

Phylogenetic reconstruction was achieved under the sequence evolutionary model GTR+I+G. The SH-test performed on NJ, ML and Bayesian trees concluded that both the ML and Bayesian phylogenetic reconstructions were congruent and had the greatest likelihood. Most of the nodes of the Bayesian phylogenetic tree had probabilities values greater than or equal to 0.95, indicating that branches are relatively stable (Fig. 2).

The Bayesian phylogenetic tree clearly indicated that the African, Mediterranean and SWIO sequences separate into four major clades or phylogroups (G1, G2, G3 and G4; Fig. 2). The SWIO isolates are found in three of these phylogroups (G1, G3 and G4). Viruses widely sampled from various host species (chayote, cotton, hollyhock, pepper, tobacco and tomato) throughout Africa are found in G1, which also contains four SWIO isolates [of which, *Tomato leaf curl Diana virus* (ToLCDiaV) and *Tomato leaf curl Toliara virus* (ToLCTolV) are new species; Table 1]. Phylogroup G2 contains Mediterranean tomato yellow leaf curl virus (TYLCV) isolates and closely related tomato infecting virus species, including African ToLCVs from Sudan and Mali. However, none of the currently described indigenous SWIO isolates fall into phylogroup G2. Phylogroup G3 contains *East African cassava mosaic virus* (EACMV) and other closely related species, such as *South African cassava mosaic virus* (ACMV), *East African cassava mosaic Zanzibar virus* (EACMZV) and *East African cassava mosaic Kenya virus* (EACMKV). G3 also contains ToLCMGV isolates from the west coast of Madagascar. Finally, the forth phylogroup, G4, contains ACMV, *Tomato leaf curl Uganda virus*-[Iganga] (ToLCUGV-[Iga]) and twelve SWIO begomovirus isolates including five new species: *Tomato leaf curl Moheli virus* (ToLCMohV), *Tobacco leaf curl Comoros virus* (TbLCKMV), *Tomato leaf curl Seychelles virus* (ToLSCV), *Tomato leaf curl Anjouan virus* (ToLCAnjV) and *Tomato leaf curl Antsiranana virus* (ToLCAntV). Importantly, G3 and G4 contain both monopartite and bipartite begomoviruses.

Analysis of recombination

We analysed evidence of recombination in a 178-sequence alignment containing 170 full-length SWIO, African and Mediterranean begomovirus DNA-A and DNA-A-like sequences, and eight curtovirus and one topocuvirus full genome sequences. It was apparent from this analysis that collectively the SWIO isolates bear detectable evidence of at least 22 past recombination events (Fig. 3). Only CLCuGV-Be[An:For:01] was not detectably recombinant. Among the recombination events that were detected, many were between different species: the TbLCKMV-[GC:Fou99:05] and TbLCKMV-[GC:Sim18:04] isolates have apparently obtained almost their entire CP ORF from a virus resembling TbLCZV-[ZW] (event 'p' in Fig. 3), whereas the rest of their genome resembles that of the tomato infecting virus ToLCAntV-[MG:Mia1:01] (AM701767). Another very striking recombination event was detected in the ToLCSCV-[Mah:VE77:04] sequence from the Seychelles archipelago (event 'v' in Fig. 3). We were surprised to find that part of the Rep ORF of this virus was apparently derived from a divergent begomovirus resembling *Sweet potato leaf curl virus* (SPLCGV). However, upon closer analysis, it is probably more feasible that both SPLCGV and ToLCSCV-[Mah:VE77:04] have obtained large portions of their C1 ORFs from a curtovirus-like source.

Eighteen out of a total of 22 unique events detected in the SWIO sequences were within the *rep* gene, indicating that the *rep* gene in general, and the sequences encoding the Rep N-terminal region in particular, might be a recombination hot-spot. To test this hypothesis, we plotted all unambiguously detectable breakpoint positions on a breakpoint density map and used a permutation test to determine whether the breakpoint distribution was significantly non-random (Fig. 4). This analysis revealed one large 'globally' significant recombination hot-spot (global P-values <0.05 across its length) and two smaller 'locally' significant hot-spots (local P-values <0.01). Whereas the large global hot-spot encompasses almost the entire intergenic region (IR) between the C1 start codon and approximately 50 nucleotides 5′ of the V2 ORF start codon, the locally significant hot-spots occur at the V1–C3 interface and in the centre of the C1 ORF. In addition to these hot-spots, the analysis also revealed two locally significant recombination cold-spots (local P-value <0.01). These occurred in the V2 ORF and in the third quarter of the V1 ORF (Table 2).

DISCUSSION

We have demonstrated that the SWIO islands harbour an extraordinarily diverse begomovirus population. On the

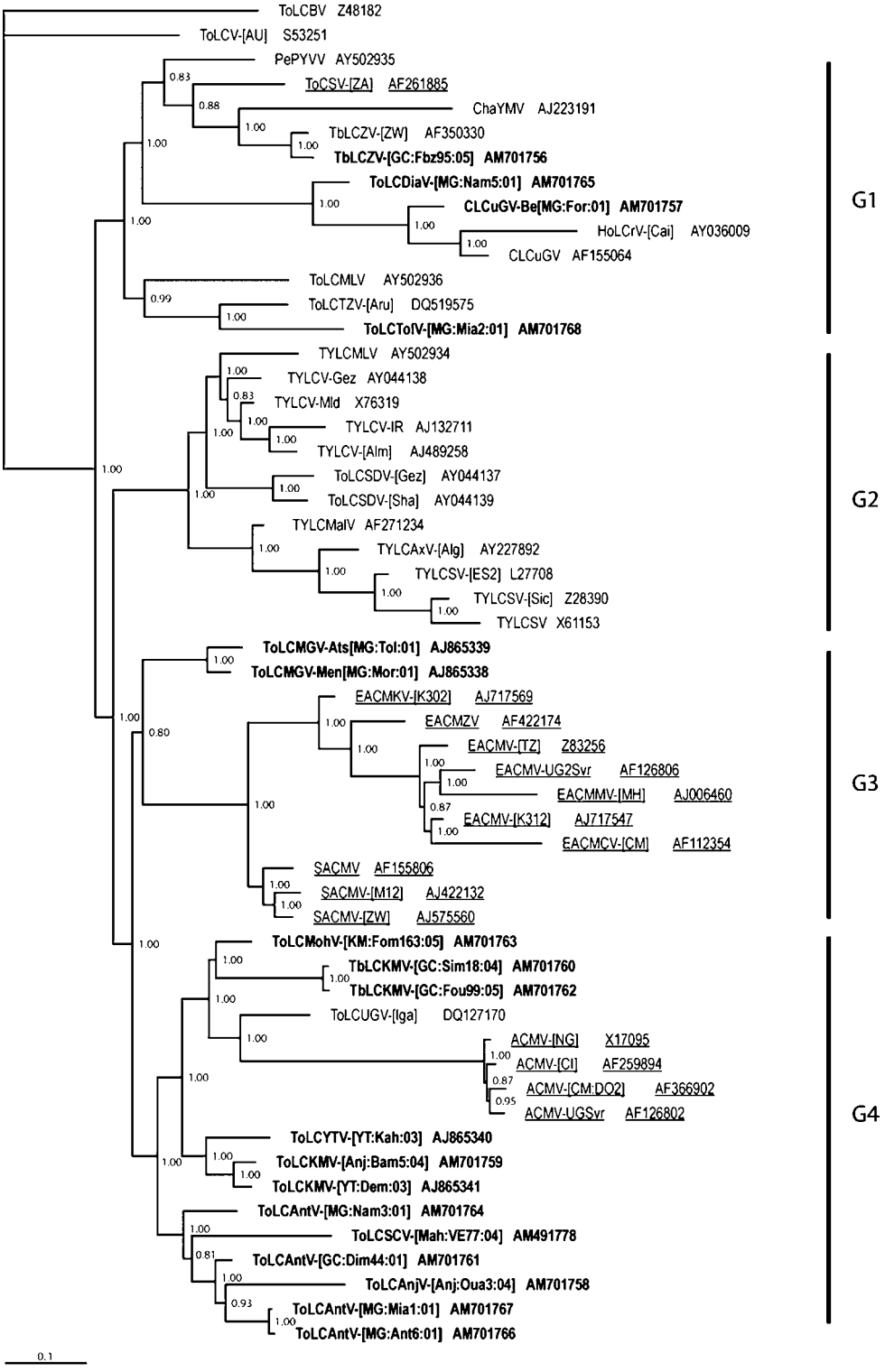
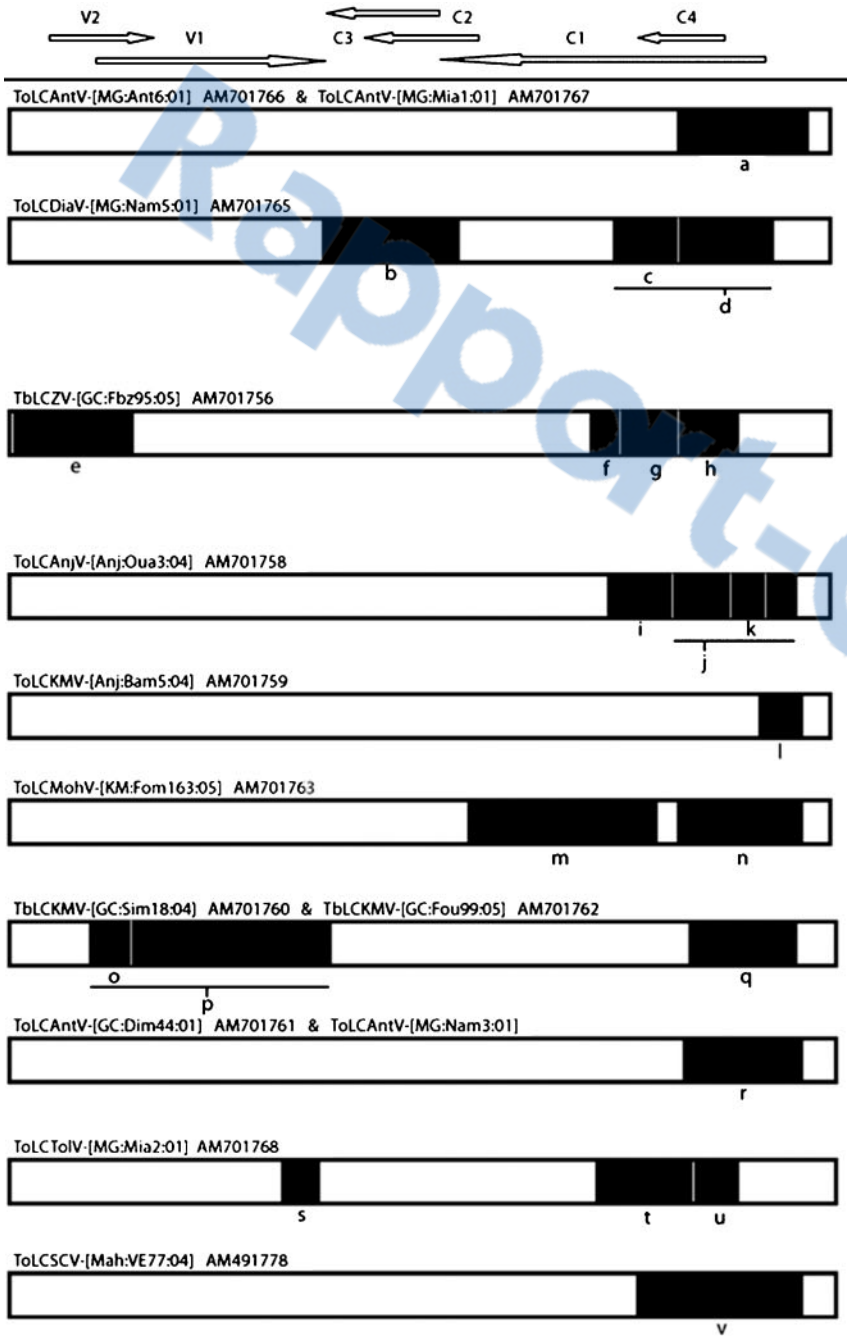


Fig. 2. Phylogenetic tree indicating the relationships between the DNA sequences of SWIO begomoviruses and those of a representative sampling of publicly available African and Mediterranean begomovirus sequences. Major clades or phylogroups are labelled G1 through G4. The tree was constructed using MrBayes and rooted using ToLCV-[AU] and ToLCBV as outliers. Numbers associated with nodes indicate the posterior probability for those nodes. Whereas horizontal bars represent genetic distances as indicated by the scale bar, vertical distances are arbitrary. SWIO sequences are in bold and bipartite begomoviruses are underlined. Four phylogenetic groups (G1 to G4) have been defined and are represented by vertical lines.



Event	Region	Parental sequences: Minor	Parental sequences: Major	Detected by	*P*-value
a	2293-2721	**ToLCMohV-[KM:Fom163:05] AM701763**	**ToLCAntV-[GC:Dim44:01] AM701761**	RGBMCS	5,6.10 $^{-31}$
b	1080-1528	**ToLCAntV-[MG:Mia1:01] AM701767**	**CLCuGV-Be[MG:For:01] AM701757**	RGBMCS	7,8.10 $^{-31}$
c	2081-2598	BYVMV-[301] AJ002453	Unknown	RGBMC	8,9.10 $^{-8}$
d	2085-2285	AF350330	**CLCuGV-Be[MG:For:01] AM701757**	RGBMCS	1,8.10 ^{18}
e	16-401	PepYVV AY502935	Unknown	RGBMC	4,8.10 $^{-5}$
f	2995-2083	AEV AJ437618	TbLCZV-[ZW] AF350330	RGBm	2,9.10 $^{-13}$
g	2111-2282	Unknown	BYVMV-[301] AJ002453	RGBMC	8,9.10 $^{-8}$
h	2321-2484	CLCuMV-[26] AJ002458	EACMV-[TZ10] AY795988	rmS	1,8.10 $^{-4}$
i	2065-2269	Unknown	**ToLCAntV-[MG:Mia1:01] AM701767**	RGBMC	5,8.10 $^{-30}$
j	2274-2670	Unknown	**ToLCAntV-[GC:Dim44:01] AM701761**	RGBMCS	8,9.10 $^{-27}$
k	2488-2590	Unknown	TYLCSV X61153	RGc	3,6.10 $^{-5}$
l	2578-2698	EACMV-[TZ] Z83256	**ToLCYTV-[YT:Kah:03] AJ865340**	RGBMCS	7,0.10 $^{-21}$
m	1586-2197	**ToLCAntV-[GC:Dim44:01] AM701761**	**ToLCKMV-[Anj:Bam5:04] AM701759**	RGBMCS	4,3.10 $^{-12}$
n	2303-2695	**TbLCKMV-[GC:Fou99:05] AM701762**	**ToLCYTV-[YT:Kah:03] AJ865340**	RGBMC	3,0.10 ^{41}
o	287-403	PePYVV AY502935	SiLCV AM050730	RGBMCS	5,0.10 ^{48}
p	286-1082	**TbLCZV-[GC:Fbz95:05] AM701756**	**ToLCAntV-[MG:Mia1:01] AM701767**	RGBMCS	4,8.10 $^{-5}$
q	2347-2678	ACMV-[KE] J02057	**ToLCYTV-[YT:Kah:03] AJ865340**	RGBMCS	9,2.10 $^{-8}$
r	2326-2709	SiLCV AM050730	EACMV-KE2[K48] AJ717542	RGBmC	1,2.10 $^{-9}$
s	936-1040	**ToLCKMV-[Anj:Bam5:04] AM701759**	Unknown	rgB	2,4.10 $^{-3}$
t	2023-2329	**TbLCZV-[GC:Fbz95:05] AM701756**	EACMV-[K6] AJ717546	RGBMC	7,0.10 $^{-26}$
u	2364-2485	CLCuMV-[26] AJ002458	EACMV-[TZ10] AY795988	rbS	1,8.10 $^{-4}$
v	2149-2703	SPLCGV-[16] AF326775	**ToLCAntV-[MG:Mia1:01] AM701767**	RGBMC	1,0.10 ^{65}

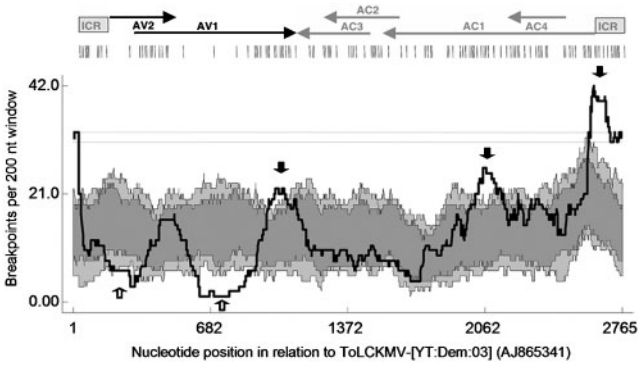


Fig. 4. The distribution of recombination breakpoints within SWIO, African and Mediterranean begomovirus full genome sequences. All detectable breakpoint positions are indicated by small vertical lines at the top of the graph. A 200 nucleotide window was moved along the alignment one nucleotide at a time and the number of breakpoints detected within the window region was counted and plotted (solid line). The upper and lower broken lines, respectively, indicate 99 and 95 % confidence thresholds for globally significant breakpoint clusters. Light and dark grey areas, respectively, indicate local 99 and 95 % breakpoint clustering thresholds taking into account local regional differences in sequence diversity that influence the ability of different methods to detect recombination breakpoints. Vertical black arrows indicate recombination hot-spots, while vertical white arrows represent recombination cold-spots. Horizontal black arrows represent virion strand ORFs (V1 and V2) and horizontal grey arrows represent complementary strand ORFs (C1, C2, C3 and C4) and the boxes represent the intergenic common region (ICR). The virion strand *ori* is at nucleotide position 1.

Table 2. Recombination hot-spots and cold-spots

Type	Position*	Region	Significance
Hot-spot	1–50	IR	Globally
	920–1030	V1–C3 interface	Locally
	2070–2210	C1 centre	Locally
Cold-spot	120–190	V2	Locally
	510–860	V1 third quarter	Locally

*Relative to ToLCKMV-[YT:Dem:03] (GenBank accession no. AJ865341).

basis of the 14 new complete DNA-A sequences and in accordance with the ICTV guidelines, we describe seven new island begomovirus species. Taken together, ten of the 18 complete DNA-A-like sequences so far determined for SWIO begomoviruses (this study and Delatte *et al.*, 2005b) represent new species. Interestingly, the ten viral species described in the SWIO islands are distributed amongst three of the four major phylogenetic groups identified within the African/Mediterranean begomovirus cluster. The presence of SWIO monopartite begomoviruses in two of the groups containing the bipartite cassava mosaic begomoviruses (G3 and G4) supports the hypothesis that either there have been multiple DNA-B component losses to produce the three different African monopartite virus lineages, or there have been multiple acquisitions of DNA-B components to produce the bipartite virus lineages (Saunders *et al.*, 2002; Mansoor *et al.*, 2003).

Our attempts at provisional classification of the novel virus genotypes described in this study led us to examine begomovirus species demarcation criteria. Given the prevailing demarcation criteria, we determined that, amongst the 170 sequences examined, the 89 % ICTV begomovirus species demarcation criterion implies that only 49 of these should be classified as species. This analysis revealed something quite interesting about the begomovirus genomes examined. Whereas the genomes contained approximately 43.5 ± 1.1 distinct *rep* genes, they contained far fewer different kinds of other genes (ranging from 25 ± 1.3 distinct V2 ORFs to 32 ± 1.1 distinct *cp* genes). There is the equivalent of 36 % more *rep* genes in circulation than all other SWIO/African/Mediterranean begomovirus genes. Our results clearly indicate that recombination is almost certainly the driving force behind this apparent proliferation of *rep* genes. This simple result clearly illustrates how recombination confounds the definition of useful taxonomic criteria (Seal *et al.*, 2006; Fauquet *et al.*, 2003).

The phylogenetic analysis performed in this study clearly demonstrates that the breadth of begomovirus diversity found on the SWIO islands is qualitatively similar to that

Fig. 3. Recombinant regions detected within SWIO virus sequences. The genome at the top of the figure corresponds to the schematic representation of sequences below. Region coordinates are nucleotide positions of detected recombination breakpoints in the multiple sequence alignment used to detect recombination. Wherever possible, parental sequences are identified. 'Major' and 'Minor' parents are sequences that were used, along with the indicated recombinant sequence, to identify recombination. Whereas for each identified event the minor parent is apparently the contributor of the sequence within the indicated region, the major parent is the apparent contributor of the rest of the sequence. Note that the identified 'parental sequences' are not the actual parents but are simply those sequences most similar to the actual parents in the dataset analysed. Recombinant regions and parental viruses were identified using the RDP (R), GENECONV (G), BOOTSCAN (B), MAXIMUM CHI SQUARE (M), CHIMAERA (C) and SISTER SCAN (S) methods. The reported *P*-value is for the method in bold type and is the best *P*-value calculated for the region in question. Whereas upper-case letters imply that a method detected recombination with a multiple comparison corrected $P<0.01$, lower-case letters imply that the method detected recombination with a multiple comparison corrected $P<0.05$ but $\geqslant 0.01$. New virus isolates presented in this study are in bold.

identifiable across the entire African continent. Also, besides six isolates classified as belonging to the G1 and G3 groups, the island isolates are all most closely related to one another and only share a distant common ancestor with the mainland viruses. This probably indicates that the SWIO islands have, with the exception of infrequent transmission events from mainland Africa, been epidemiologically isolated for a long time.

Phylogeographically the results are also intriguing: for the G4 group, there is a well supported cluster of SWIO isolates with ACMV. There is evidence here that either (i) ACMV, a lonely outlier amongst the other African viruses, originated on the SWIO islands, or (ii) the SWIO isolates are an extant and thriving population of an ancestral lineage that, besides ACMV and ToLCUGV-[Iga], has largely disappeared on the African mainland. As with cassava, the plants sampled in this study are exotic introduced species. One would expect original (and possibly still the natural) hosts of these viruses to be indigenous uncultivated plants. Further studies should aim to characterize begomovirus diversity in these hosts.

Our recombination analysis clearly indicates the presence of breakpoint hot- and cold-spots within SWIO/African/Mediterranean begomovirus genomes. This indicates either that DNA breakage and repair do not occur randomly in begomoviruses or that, if breakpoints do occur randomly, selection has preferentially culled recombinants with breakpoints in certain positions while permitting the survival of recombinants with breakpoints in other positions. That all recombinants are not created equal has been clearly demonstrated with laboratory constructed geminivirus recombinants (Liu *et al.*, 2001; Martin & Rybicki, 2002) and one would expect that many, if not most, natural recombinants would experience serious fitness deficits. There are in fact two well-supported explanations as to why recombinants are generally less fit than their parents. First, protein engineers have discovered that hybrid genes with bits of sequence from distantly related sources tend to encode proteins that do not fold properly – probably due to disruptions of co-evolved amino acid contacts within their structures (Voigt *et al.*, 2002; Saraf & Maranas, 2003). Second, when genes are transferred wholesale into distantly related genetic backgrounds, they appear only to function well either when they do not interact with a lot of other genes, or when they are co-transferred with the other genes with which they do interact (Jain *et al.*, 1999; Martin *et al.*, 2005c; Escriu *et al.*, 2007). It is therefore likely that, whereas the recombination hot-spots we have detected represent genomic regions where breakpoints are both biochemically permissive and highly 'survivable', the cold-spots represent regions where breakpoints are either particularly deleterious or are biochemically very unlikely to occur.

Although we detected a large number of unique recombination breakpoints (i.e. breakpoints that occurred during different recombination events) across the entire 3′ portion of *rep* spanning the C4 ORF, this region is bounded by two recombination hot-spots. This pattern of recombination is almost certainly due to both a biochemical predisposition to recombination in these sequences, and a high tolerance for recombination in the proteins encoded in this region. Importantly, experimental analyses of recombination in geminiviruses (Schnippenkoetter *et al.*, 2001; Stenger *et al.*, 1991; Garcia-Andres *et al.*, 2007) and the replicational release mechanisms put into practice during agroinoculation of geminiviruses, have indicated that the origin of virion strand replication is a biochemically predisposed recombination hot-spot. While there is a clear breakpoint distribution peak detected at the virion strand *ori*, the highest breakpoint distribution peak is 5′ of the *ori*, close to the *rep* start codon. This region corresponds to the most variable region of begomovirus genomes. It is probable that at least part of the reason why so many breakpoints are detected here is that this is the genome region where breakpoints are easiest to detect. Nevertheless, the statistical test used to detect hot spots takes this increased variability into account and has still identified that there are an improbably large number of breakpoints in this region. We propose first, that the IR-wide breakpoint hot-spot is a consequence of recombinants with breakpoints outside of genes generally being fitter than those with breakpoints within genes. This possibility is supported by the fact that the V1–C3 interface, the only other genome region where breakpoints are possible outside of genes, is also a recombination hot-spot.

Importantly, there exists direct experimental support for our observation that the V1–C3 interface is a recombination hot-spot because recombination at this point does not incur a significant fitness cost. In experimental recombination in controlled mixed TYLCSV and TYLCV-Mld infection, the most prevalent (and hence probably the most fit) emergent recombinant had one breakpoint within 100 nucleotides of the V1–C3 ORF interface and another at precisely the virion strand *ori* (Garcia Andres *et al.*, 2007). That this particular recombinant genotype is highly fit is further evidenced by its close resemblance to the widespread natural TYLCSV–TYLCV-Mld recombinant, *Tomato yellow leaf curl Malaga virus* (Monci *et al.*, 2002). The problem remains, however, to explain the recombination hot-spot in the middle of the *rep* gene. Our second proposal is therefore that the N-terminal portion of Rep and any protein expressed from the C4 ORF are exceptionally tolerant of recombination, with the most tolerable breakpoint positions (i.e. those that disrupt Rep folding the least) occurring near the centre of the gene around the recombination hot-spot.

The presence of recombination cold-spots within the V2 ORF and the third quarter of the V1 ORF is consistent with our first proposal that recombination breakpoints within coding regions are generally more damaging than those outside of coding regions. However, the fact that the detectable breakpoint cold-spots are within the virion sense ORFs, whereas the greatest number of breakpoints are

within the complementary sense ORFs, leads us to a third proposal: the uneven distribution of recombination breakpoints is possibly due, at least in part, to clashes between virion strand replication and gene transcription. Whereas replication and virion strand transcription proceed in the same direction and are therefore unlikely to interfere with one another, transcription of the complementary strand ORFs tends to disrupt replication forks moving in the opposite direction. Analysis of replicating begomoviral DNA intermediates has revealed a wide distribution of so-called heterogeneous length linear dsDNA forms (hDNA), possibly created during such clashes. The ends of these hDNA molecules tend to map most frequently to the V-*ori* and either the C2/C3 transcription promoter near the hot-spot we detected in the centre of *rep*, or the C2/C3 terminator near the hot-spot we detected at the V1–C3 ORF interface (Jeske *et al.*, 2001). Completion of replication from displaced, partially replicated virion strands would then proceed via the recombination-dependent replication pathway (Preiss & Jeske, 2003), which in the presence of potential template DNAs with different sequences could result in detectable recombination events. Completion of replication would result in a recombinant virion strand with one breakpoint at the point where replication was initially disrupted and the other at the virion sense *ori* where replication was completed.

Novel environments, such as the new host species offered to begomoviruses by invasive polyphagous vector biotypes, are possibly the defining force driving begomovirus evolution worldwide. For example, introduction into Reunion of the polyphagous *B. tabaci* biotype B is believed to be responsible for severe TYLC disease epidemics on the island in the late 1990s (Peterschmitt *et al.*, 1999; Delatte *et al.*, 2005a). Spread of this biotype to other SWIO islands may (i) facilitate host switching into cultivated crops of uncharacterized begomoviruses that currently only infect weeds and (ii) induce an overlap of exotic TYLCV and indigenous begomovirus distributions. Given the propensity of begomoviruses to recombine, emergence of new recombinants with increased virulence and/or modified host ranges are to be expected. An emergent TYLCV–*Tomato yellow leaf curl Sardinia virus* (TYLCSV) recombinant lineage in Spain (Monci *et al.*, 2002; Garcia-Andres *et al.*, 2006) demonstrates that the probability of such an occurrence is high, especially as the genetic distance between TYLCV and the SWIO indigenous ToLCVs is similar to the distance between TYLCV and TYLCSV.

By highlighting the extraordinary diversity of begomoviruses on the SWIO islands, we have provided a detailed description of their phylogenetic and recombinant histories. The phylogenetic association between the monopartite SWIO isolates and both monopartite and bipartite mainland African isolates indicate that they are probably indigenous to the islands. The large number of unique recombination events that we have detected amongst the SWIO isolates and their nearest mainland relatives reiterates the pivotal role of this process in begomovirus evolution. It is, however, apparent from our breakpoint distribution analysis that purifying selection and/or varying biochemical predispositions to recombination in different parts of begomovirus genomes place substantial constraints on the degree of evolutionary innovation that is possible by recombination.

ACKNOWLEDGEMENTS

This work was funded by the Conseil Régional de la Réunion, the Ministère de l'Outre-Mer, CIRAD and the MRES. D. P. M. is funded by the Harry Oppenheimer Foundation, a Sydney Brenner Fellowship and the South African Bioinformatics Network. A.V. is supported by the Carnegic Corporation of New York.

REFERENCES

Bos, L. (1977). Persistence of infectivity of three viruses in plant material dried over $CaCl_2$ and stored under different conditions. *Eur J Plant Pathol* **83**, 217–220.

Briddon, R. W., Bull, S. E., Mansoor, S., Amin, I. & Markham, P. G. (2002). Universal primers for the PCR-mediated amplification of DNA beta: a molecule associated with some monopartite begomoviruses. *Mol Biotechnol* **20**, 315–318.

Delatte, H., Holota, H., Naze, F., Peterschmitt, M., Reynaud, B. & Lett, J. M. (2005a). The presence of both recombinant and nonrecombinant strains of Tomato yellow leaf curl virus on tomato in Reunion Island. *Plant Pathol* **54**, 262.

Delatte, H., Martin, D. P., Naze, F., Golbach, R. W., Reynaud, B., Peterschmitt, M. & Lett, J. M. (2005b). South West Indian Ocean islands tomato begomovirus populations represent a new major monopartite begomovirus group. *J Gen Virol* **86**, 1533–1542.

Escriu, F., Fraile, A. & Garcia-Arenal, F. (2007). Constraints to genetic exchange support gene coadaptation in a tripartite RNA virus. *PLoS Pathog* **3**, e8.

Fauquet, C. M., Bisaro, D. M., Briddon, R. W., Brown, J. K., Harrison, B. D., Rybicki, E. P., Stenger, D. C. & Stanley, J. (2003). Revision of taxonomic criteria for species demarcation in the family *Geminiviridae*, and an updated list of begomovirus species. *Arch Virol* **148**, 405–421.

Fauquet, C. M., Sawyer, S., Idris, A. M. & Brown, J. K. (2005). Sequence analysis and classification of apparent recombinant begomoviruses infecting tomato in the Nile and Mediterranean Basins. *Phytopathology* **95**, 549–555.

Fauquet, C. M., Briddon, R. W., Brown, J. K., Moriones, E., Stanley, J., Zerbini, M. & Zhou, X. (2007). Geminivirus strain demarcation and nomenclature. *Arch Virol* (in press).

Garcia-Andres, S., Monci, F., Navas-Castillo, J. & Moriones, E. (2006). Begomovirus genetic diversity in the native plant reservoir *Solanum nigrum*: evidence for the presence of a new virus species of recombinant nature. *Virology* **350**, 433–442.

Garcia-Andres, S., Tomas, D. M., Sanchez-Campos, S., Navas-Castillo, J. & Moriones, E. (2007). Frequent occurrence of recombinants in mixed infections of tomato yellow leaf curl disease-associated begomoviruses. *Virology* **359**, 302–312.

Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. (2000). Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573–582.

Heath, L., van der Walt, E., Varsani, A. & Martin, D. P. (2006). Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *J Virol* **80**, 11827–11832.

Inoue-Nagata, A. K., Albuquerque, L. C., Rocha, W. B. & Nagata, T. (2004). A simple method for cloning the complete begomovirus genome using the bacteriophage φ29 DNA polymerase. *J Virol Methods* **116**, 209–211.

Jain, R., Rivera, M. C. & Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* **96**, 3801–3806.

Jeske, H., Lutgemeier, M. & Preiss, W. (2001). DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. *EMBO J* **20**, 6158–6167.

Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5**, 150–163.

Laufs, J., Traut, W., Heyraud, F., Matzeit, V., Rogers, S. G., Schell, J. & Gronenborn, B. (1995). In vitro cleavage and joining at the viral origin of replication by the replication initiator protein of tomato yellow leaf curl virus. *Proc Natl Acad Sci U S A* **92**, 3879–3883.

Lee, C., Grasso, C. & Sharlow, M. F. (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**, 452–464.

Lefeuvre, P., Delatte, H., Naze, F., Dogley, W., Reynaud, B. & Lett, J. M. (2007). A new tomato leaf curl virus from the Seychelles archipelago. *Plant Pathol* **56**, 342.

Liu, H., Lucy, A. P., Davies, J. W. & Boulton, M. I. (2001). A single amino acid change in the coat protein of *Maize streak virus* abolishes systemic infection, but not interaction with viral DNA or movement protein. *Mol Plant Pathol* **2**, 223–228.

Mansoor, S., Briddon, R. W., Zafar, Y. & Stanley, J. (2003). Geminivirus disease complexes: an emerging threat. *Trends Plant Sci* **8**, 128–134.

Martin, D. & Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562–563.

Martin, D. P. & Rybicki, E. P. (2002). Investigation of *Maize streak virus* pathogenicity determinants using chimaeric genomes. *Virology* **300**, 180–188.

Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. (2005a). A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses* **21**, 98–102.

Martin, D. P., Williamson, C. & Posada, D. (2005b). RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* **21**, 260–262.

Martin, D. P., van der Walt, E., Posada, D. & Rybicki, E. P. (2005c). The evolutionary value of recombination is constrained by genome modularity. *PLoS Genet* **1**, e51.

Monci, F., Sanchez-Campos, S., Navas-Castillo, J. & Moriones, E. (2002). A natural recombinant between the geminiviruses *Tomato yellow leaf curl Sardinia virus* and *Tomato yellow leaf curl virus* exhibits a novel pathogenic phenotype and is becoming prevalent in Spanish populations. *Virology* **303**, 317–326.

Ndunguru, J., Legg, J. P., Aveling, T. A., Thompson, G. & Fauquet, C. M. (2005). Molecular biodiversity of cassava begomoviruses in Tanzania: evolution of cassava geminiviruses in Africa and evidence for East Africa being a center of diversity of cassava geminiviruses. *Virol J* **2**, 21.

Padidam, M., Sawyer, S. & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218–225.

Paradis, E., Claude, J. & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290.

Peterschmitt, M., Granier, M., Mekdoud, R., Dalmon, A., Gambin, O., Vayssières, J. F. & Reynaud, B. (1999). First report of tomato yellow leaf curl virus in Réunion Island. *Plant Dis* **83**, 303.

Posada, D. (2006). ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. *Nucleic Acids Res* **34**, W700–W703.

Preiss, W. & Jeske, H. (2003). Multitasking in replication is common among geminiviruses. *J Virol* **77**, 2972–2980.

Rambaut, A. & Drummond, A. J. (2004). Tracer v1.3, Available from http://evolve.zoo.ox.ac.uk/software.html

Rojas, M. R., Gilbertson, R. L., Russel, D. R. & Maxwell, D. P. (1993). Use of degenerate primers in the polymerase chain reaction to detect whitefly-transmitted geminivirus. *Plant Dis* **77**, 340–347.

Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.

Rybicki, E. P. & Pietersen, G. (1999). Plant virus disease problems in the developing world. In *Advances in Virus Research*, vol. 53, pp. 127–178. Edited by K. Maramorosch, F. A. Murphy & A. J. Shatkin. San Diego, CA: Academic Press.

Saraf, M. C. & Maranas, C. D. (2003). Using a residue clash map to functionally characterize protein recombination hybrids. *Protein Eng* **16**, 1025–1034.

Saunders, K., Salim, N., Mali, V. R., Malathi, V. G., Briddon, R., Markham, P. G. & Stanley, J. (2002). Characterisation of Sri Lankan cassava mosaic virus and Indian cassava mosaic virus: evidence for acquisition of a DNA B component by a monopartite begomovirus. *Virology* **293**, 63–74.

Schnippenkoetter, W. H., Martin, D. P., Willment, J. A. & Rybicki, E. P. (2001). Forced recombination between distinct strains of *Maize streak virus*. *J Gen Virol* **82**, 3081–3090.

Seal, S. E., vandenBosch, F. & Jeger, M. J. (2006). Factors influencing begomovirus evolution and their increasing global significance: implications for sustainable control. *Crit Rev Plant Sci* **25**, 23–46.

Shimodaira, H. & Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* **16**, 1114–1116.

Smith, J. M. (1992). Analyzing the mosaic structure of genes. *J Mol Evol* **34**, 126–129.

Stanley, J. (1995). Analysis of African cassava mosaic virus recombinants suggests strand nicking occurs within the conserved nonanucleotide motif during the initiation of rolling circle DNA replication. *Virology* **206**, 707–712.

Stenger, D. C., Revington, G. N., Stevenson, M. C. & Bisaro, D. M. (1991). Replicational release of geminivirus genomes from tandemly repeated copies: evidence for rolling-circle replication of a plant viral DNA. *Proc Natl Acad Sci U S A* **88**, 8029–8033.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680.

Umaharan, P., Padidam, M., Phelps, R. H., Beachy, R. N. & Fauquet, C. (1998). Distribution and diversity of geminiviruses in Trinidad and Tobago. *Phytopathology* **88**, 1262–1268.

Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nat Struct Biol* **9**, 553–558.

Wyatt, S. D. & Brown, J. K. (1996). Detection of subgroup III geminivirus isolates in leaf extracts by degenerate primers and polymerase chain reaction. *Phytopathology* **86**, 1288–1293.

Zhou, X., Liu, Y., Calvert, L., Munoz, C., Otim-Nape, G. W., Robinson, D. J. & Harrison, B. D. (1997). Evidence that DNA-A of a geminivirus associated with severe cassava mosaic disease in Uganda has arisen by interspecific recombination. *J Gen Virol* **78**, 2101–2111.

Chapitre III : La recombinaison et l'évolution des bégomovirus : forces et contraintes

La recombinaison permet la création de nouveaux arrangements génomiques et ainsi l'obtention d'une diversité nouvelle. Pour de nombreux genres viraux, ce phénomène joue un rôle important dans l'évolution. Dans le cas des bégomovirus, il apparaît que ce rôle est même essentiel, à la vue du nombre extrêmement important d'évènements de recombinaison détectés au sein des séquences. Néanmoins, ces évènements de recombinaison ne sont pas distribués aléatoirement sur le génome (Chapitre II). Si des raisons mécanistiques sont avancées pour expliquer certains déséquilibres au sein de ces profils de recombinaison, nous savons peu de choses sur les facteurs responsables de la sélection des recombinants créés mécanistiquement. Quelques auteurs se sont penchées sur le sujet et ont émis l'hypothèse qu'un des critères majeurs influant sur la *fitness* d'organismes recombinants ou la fonctionnalité de protéines recombinantes était le maintien des réseaux d'interactions formés au sein du génome ou entre les acides-aminés des protéines (Martin et al., 2005; Jain et al., 1999; Voigt et al., 2002).

En enzymologie, une équipe a mis au point une méthode de mesure de la perturbation des protéines recombinantes (Voigt et al., 2002). En se basant sur la structure tridimensionnelle d'une protéine donnée, il est possible de déterminer quels acides aminés sont en contact du fait de leur proximité. Pour une protéine recombinante, le degré de fonctionnalité de la protéine a été associé au nombre de couples d'acides aminés en contact qui ne sont pas de type parental. Les parents étant supposés raisonnablement « *fit* », tout changement dans l'architecture des contacts entre acides aminés est susceptible de perturber la structure tridimensionnelle de la protéine et donc d'altérer sa fonction.

Dans notre étude, nous avons appliquée cette méthode de mesure appelée SCHEMA à des jeux de donnée de bégomovirus. Sur la base des structures tridimensionnelles disponibles pour la protéine de capside et la partie terminale de la protéine associée à la réplication (Bottcher et al., 2004; Campos-Olivas et al., 2002), nous avons pu analyser un tiers du génome en utilisant la méthode SCHEMA.

L'étude a consisté en un premier temps à déterminer les distributions des points de recombinaison en utilisant une batterie de méthodes de détection de la recombinaison accompagnée d'une analyse manuelle rigoureuse. Pour tous les

évènements de recombinaison présentant un point de recombinaison au sein des portions du génome codant pour les protéines dont la structure était connue, les niveaux de mutation (m) et de perturbation (E) ont été déterminés. Par la suite, en se basant sur ces évènements de recombinaison (appelé ici "évènements réels") une simulation de l'ensemble des évènements de recombinaison dérivés a été réalisée, ceux-ci présentant soit (1) un nombre variable de mutations, mais une taille identique en terme de polymorphisme le long du génome, soit (2) un nombre identique de mutations non synonymes dans les zones analysées. Pour ces jeux "d'évènements simulés", les valeurs de m et E ont aussi été déterminées puis comparées aux évènements réels par un test de permutation.

Cette analyse statistique a permis de montrer que les évènements naturels de recombinaison ont tendance à impliquer un échange de polymorphisme moindre que les évènements simulés. De plus, quand du polymorphisme non synonyme (codant pour des acides aminés différents) est échangé, des niveaux moindres de perturbation de la structure protéique sont mesurés. En clair, les évènements réels de recombinaison détectés sur des séquences « *in natura* » échangent moins de polymorphisme et, si c'est le cas, ces évènements sont moins perturbateurs qu'ils ne le seraient dans le cas où les recombinaisons se feraient au hasard. Il apparaît ici de manière hautement significative qu'une sélection purificatrice s'opère sur l'ensemble des recombinants créés de manière mécanistique pour ne conserver que ceux qui présentent le niveau de *fitness* le plus important. En considérant que les virus parentaux présentent un niveau d'adaptation élevé, on peut supposer que la grande majorité des virus recombinants à l'origine d'émergence (ou raisonnablement adaptés pour tout simplement faire partie des virus circulants dans la population) ne représente qu'une faible portion de l'ensemble des recombinants originellement créés. Ainsi, la plupart des génomes recombinants, qui semblent supporter des réarrangements délétères, seraient rapidement éliminés par sélection naturelle. Ces résultats suggèrent que la sélection purificatrice qui agit sur la conservation de la structure et de la fonction des protéines, est un des facteurs majeurs façonnant la distribution des recombinaisons au sein des populations naturelles de bégomovirus. D'autres paramètres, non étudiés ici, interviennent probablement aussi dans cette sélection comme par exemple les interactions entre protéines. De nouvelles méthodes de détection de ces interactions ont ainsi récemment vu le jour, mais leur robustesse et leur acuité à détecter des interactions au sein de jeux de données de séquences recombinantes peuvent être mises en doute. Toute utilisation de méthodes phylogénétiques sur des

séquences recombinantes doit se faire avec prudence, dans la mesure où la recombinaison fausse la phylogénie. De nombreux faux positifs/négatifs directement générés par recombinaison, pourraient ainsi être détectés. L'avantage principal de la méthode du SCHEMA est de se baser sur des structures tridimensionnelles dont la détermination n'est pas faussée par recombinaison.

Dans cette étude nous avons pu montrer que le maintien de l'intégrité des réseaux d'interactions formés au sein d'un génome est un critère majeur de sélection des bégomovirus recombinants. Cette même règle doit prévaloir au sein du vivant en général pour le très grand nombre d'interactions présent au sein d'un génome, d'une cellule, d'un individu ou d'un système écologique.

Avoidance of Protein Fold Disruption in Natural Virus Recombinants

Pierre Lefeuvre[1], Jean-Michel Lett[1], Bernard Reynaud[1], Darren P. Martin[2*]

1 CIRAD, UMR 53 PVBMT CIRAD-Université de la Réunion, Pôle de Protection des Plantes, Ligne Paradis, Saint Pierre, La Réunion, France, **2** Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory, South Africa

With the development of reliable recombination detection tools and an increasing number of available genome sequences, many studies have reported evidence of recombination in a wide range of virus genera. Recombination is apparently a major mechanism in virus evolution, allowing viruses to evolve more quickly by providing immediate direct access to many more areas of a sequence space than are accessible by mutation alone. Recombination has been widely described amongst the insect-transmitted plant viruses in the genus *Begomovirus* (family Geminiviridae), with potential recombination hot- and cold-spots also having been identified. Nevertheless, because very little is understood about either the biochemical predispositions of different genomic regions to recombine or what makes some recombinants more viable than others, the sources of the evolutionary and biochemical forces shaping distinctive recombination patterns observed in nature remain obscure. Here we present a detailed analysis of unique recombination events detectable in the DNA-A and DNA-A-like genome components of bipartite and monopartite begomoviruses. We demonstrate both that recombination breakpoint hot- and cold-spots are conserved between the two groups of viruses, and that patterns of sequence exchange amongst the genomes are obviously non-random. Using a computational technique designed to predict structural perturbations in chimaeric proteins, we demonstrate that observed recombination events tend to be less disruptive than sets of simulated ones. Purifying selection acting against natural recombinants expressing improperly folded chimaeric proteins is therefore a major determinant of natural recombination patterns in begomoviruses.

Citation: Lefeuvre P, Lett JM, Reynaud B, Martin DP (2007) Avoidance of protein fold disruption in natural virus recombinants. PLoS Pathog 3(11): e181. doi:10.1371/journal.ppat.0030181

Introduction

Besides its vital cellular role in maintaining and repairing broken DNA molecules [1,2], recombination is also evolutionarily significant in that it defends genomes against the otherwise unavoidable accumulation of deleterious mutations [3–5]. However, by enabling the creation of novel genetic combinations from existing genomes, recombination has the potential to do more than just reverse the mutational decay of genomes: it can also provide organisms with vastly more evolutionary options than are available through mutation alone [6,7].

In virology, two recombinational processes can be distinguished: genome reassortment and true recombination. Genome reassortment, also called pseudo-recombination, involves the exchange of intact genome components between viruses with multipartite genomes to yield viruses whose genomes are comprised of new combinations of components. True recombination, on the other hand, involves the exchange of genetic material between individual genomic molecules. The rearrangement of genetic information mediated by both true recombination and pseudo-recombination must yield fully functional and reasonably fit genomes for these processes to be easily detectable in nature. However, analysis of the functionality of recombinant genes [8,9] and the viability of recombinant genomes [10–12] has indicated that a large proportion (and possibly the vast majority) of recombination events between genomes sharing less than 90% nucleotide sequence identity yield progeny with decreased viability. Bacterial recombination [13] and DNA shuffling studies [8,9,14,15] have indicated that the evolutionary value of recombination can vary depending on both the specific genes and sub-gene modules that are exchanged. A key factor determining the survival of recombinants is the degree to which recombination disrupts coevolved intra-genome interactions. At the whole genome scale, potentially disrupted interactions could include sequence-specific interactions between viral proteins, DNA, and RNA. At the scale of individual viral proteins, interactions include those occurring between amino acids required for proper folding.

While full accounts of experimentally verified intra-genome interactions are currently unavailable for any virus species, potential amino acid interactions within folded proteins can be inferred with reasonable accuracy given high resolution protein structural data. In the past five years, protein engineers have made substantial progress in the development of computational methods capable of accurately inferring degrees of recombination-induced fold disruption in experimentally generated chimaeras of proteins with known structures [8,14,15]. Although these methods

Editor: Edward C. Holmes, The Pennsylvania State University, United States of America

Received August 21, 2007; **Accepted** October 12, 2007; **Published** November 30, 2007

Abbreviations: CP, coat protein; hDNA, heterogeneous length linear dsDNA; IR, intergenic region; ORF, open reading frame; Rep, replication-associated protein; v-*ori*; virion strand origin of replication

* To whom correspondence should be addressed. E-mail: Darrin.Martin@uct.ac.za

Author Summary

The exchange of genetic material between different virus species, called inter-species recombination, has the potential to generate, within a single genome replication cycle, an almost unimaginable number of genetically distinct virus strains, including many that might cause deadly new human, animal, or plant diseases. Many fear that inter-species recombination could provide viruses with quick access to evolutionary innovations such as broader host ranges, altered tissue tropisms, or increased severities. However, mounting evidence suggests that recombination is not an unconstrained process and that most inter-species recombinants that occur in nature are probably defective. It is suspected that networks of coevolved interactions between different parts of virus genomes and their encoded proteins must be kept intact for newly formed inter-species recombinants to have any chance of out-competing their parents. One category of coevolved interaction is that between contacting amino acids within the 3-D structures of folded proteins. Here we examine the distributions of recombination events across the genomes of a group of rampantly recombining plant viruses and find very good evidence that this class of interaction tends to be preserved amongst recombinant sequences sampled from nature. This indicates that selection against misfolded proteins strongly influences the survival of natural recombinants.

have, to our knowledge, never been used to analyse any naturally generated chimaeric proteins, we realised they should also be useful for understanding breakpoint distribution patterns found within coding regions of recombining virus genomes.

Recently, Lefeuvre et al. [16] reported the first statistically supported evidence of recombination hot- and cold-spots in the genomes of begomoviruses, members of a highly recombinagenic family of single-stranded DNA viruses called the Geminiviridae. Importantly, they detected a substantial number of recombination events within a portion of the begomovirus replication–associated protein (*rep*) gene encoding a protein for which a high resolution crystal structure is available. In this paper we describe an expanded analysis of recombination amongst begomoviruses. We identify sets of unambiguously unique recombination events detectable in publicly available monopartite begomovirus DNA-A-like sequences and bipartite begomovirus DNA-A sequences. We then determine the distribution of recombination breakpoints across the analysed sequences and confirm the recombination hot- and cold-spots identified previously. We use a method called SCHEMA [8] to predict degrees of fold disruption in chimaeric begomovirus Rep and coat protein (CP) molecules (for which a reasonably high resolution structural model exists) expressed by viruses determined to have recombinant *rep* and *cp* genes. We then compare these predictions with those for an exhaustive set of all possible recombination breakpoint pairs within these genes and provide the first statistical evidence to our knowledge that avoidance of protein fold disruption is a major factor shaping the patterns of recombination that are detectable in natural virus populations.

Results/Discussion

We anticipated that general rules governing the evolutionary advancement of viruses through recombination should be most manifest in virus groups in which distinctive conserved patterns of recombination have emerged [16–21]. Given that begomoviruses are both highly recombinogenic [22] and display some evidence of recombination breakpoint hot- and cold-spots [16], we undertook a detailed analysis of recombination in this group.

Are Patterns of Recombination Conserved amongst All Begomoviruses?

We began by precisely mapping the distributions of recombination events across begomovirus DNA-A and DNA-A-like sequences sampled throughout the world. Using a battery of recombination signal detection tools and rigorous manual and automated evaluation of recombination signals, we identified sets of 120 and 164 non-ambiguous unique recombination events in the bipartite begomovirus DNA-A and monopartite DNA-A-like sequences, respectively (see Datasets S1 and S2, Figure S1, and Tables S1 and S2 for detailed descriptions of all the detected events).

These events were mapped onto "recombination count matrices" (Figure 1). These matrices represent the number of times that recombinational movement of sequence tracts within the analysed genomes has separated pairs of nucleotide sites. This representation of the characterised recombination events highlights the differential "exchangeability" of sequence tracts within begomovirus DNA-A and DNA-A-like sequences. Whereas highly exchangeable genome regions (i.e., those separated many times by recombination from their original genetic background) are represented by red/purple shades, the least exchangeable regions (i.e., those separated the fewest times by recombination) are represented by yellow/green shades. As can be seen from Figure 1, the region of the *rep* gene encoding the N-terminal portion of Rep and the adjacent intergenic region sequences up to the virion strand origin of replication (v-*ori*) are the regions of both monopartite and bipartite begomovirus genomes most frequently exchanged during recombination. As a result of this, the 5′ and 3′ portions of *rep* are very frequently inherited from different parents. This implies that *rep* must be comprised of highly modular subregions capable of proper functioning in diverse foreign genetic backgrounds. Conversely, the small numbers of detectable recombination events that separate fragments of the *cp* gene indicate that naturally occurring monopartite and bipartite begomovirus recombinants tend to inherit the portions of this gene encoding CP amino acids 75 through 220 from a single source.

To visualise the distribution of recombination breakpoints in monopartite and bipartite begomovirus genomes, all approximated recombination breakpoint locations were plotted on density maps and a permutation test was used to determine whether there were any statistically significant hot- or cold-spots in the breakpoint distribution. This test indicated that the distribution of breakpoints was significantly non-random, with clear recombination hot- and cold-spots being detectable (Figure 2). It is apparent that, as for the recombinant region count matrices (Figure 1), the recombination breakpoint distributions detected in the monopartite and bipartite datasets are very similar. The clusters of inferred breakpoint positions in the two datasets do not, however, have identical significance levels, probably due to differences in datasets with respect to both their sequence diversity, and the number of detectable recombination events they contain.

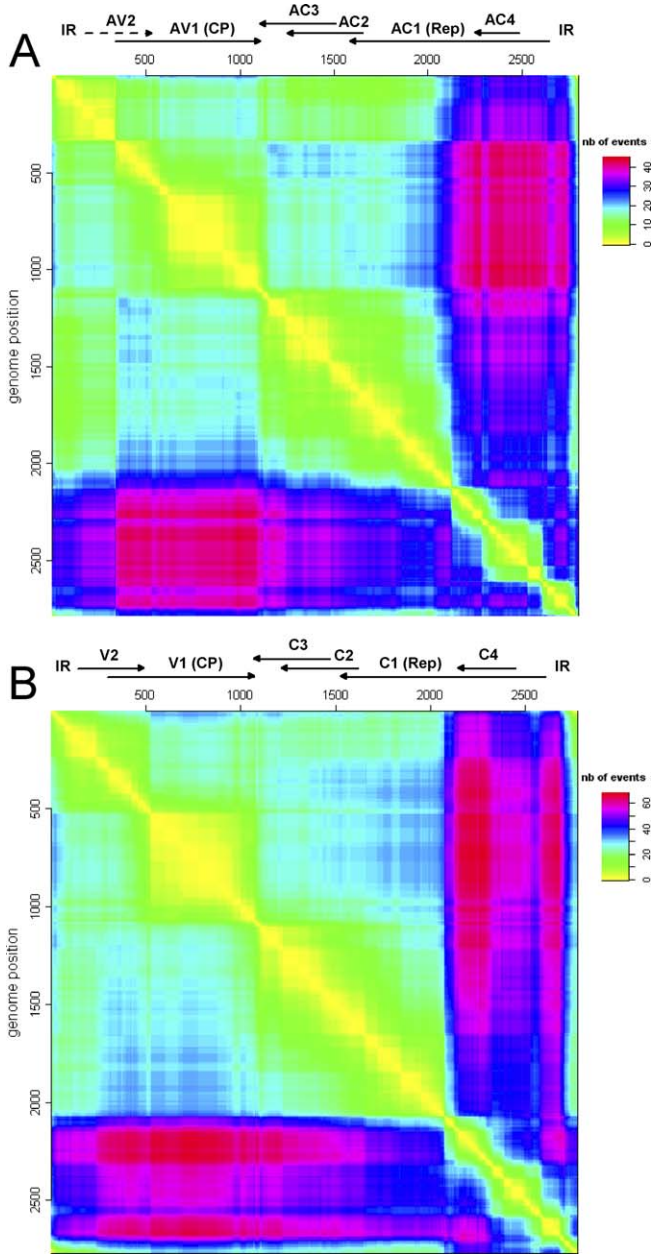


Figure 1. Recombination Region Count Matrix of Unique Recombination Events Detected amongst (A) DNA-A Sequences of Bipartite Begomoviruses and (B) DNA-A-Like Sequences of Monopartite Begomoviruses

Unique recombination events have been mapped onto the matrix based on their estimated breakpoint positions. The shades displayed are a function of the number of times pairs of nucleotides (plotted on the *x*- and *y*-axis) are separated during the observed set of unique recombination events. Diagrams indicating the positions of landmarks in begomovirus DNA-A/DNA-A-like sequences are shown on the top of the matrices. Positions were drawn in relative to EACMCV-[TZ] (AY795983) for bipartite sequences and ToLCYT-[Dem] (AJ865341) for monopartite sequences.

doi:10.1371/journal.ppat.0030181.g001

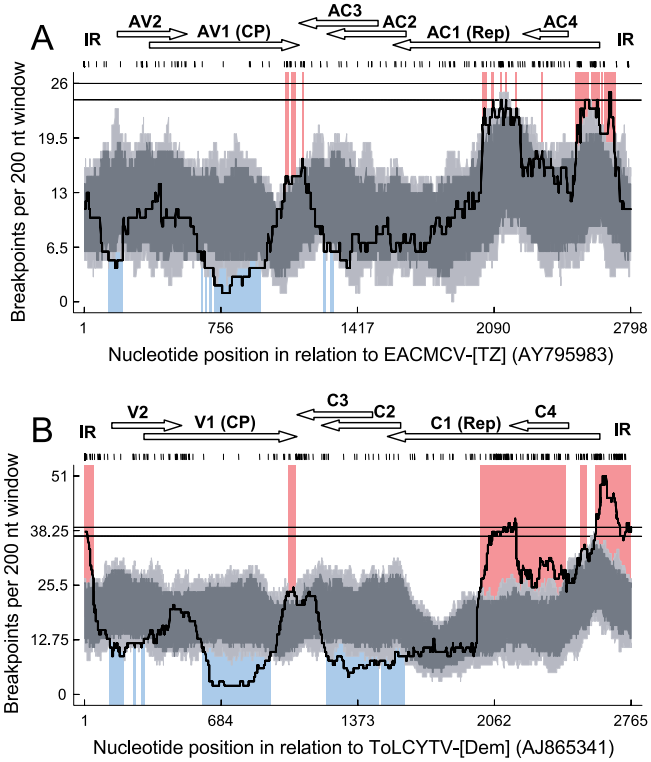


Figure 2. The Distribution of Recombination Breakpoints Detected within (A) DNA-A Sequences of Bipartite Begomovirus and (B) DNA-A-Like Sequences of Monopartite Begomoviruses

All estimated breakpoint positions are indicated by small vertical lines at the top of the graph. A 200-nucleotide window was moved along the alignment one nucleotide at a time and the number of breakpoints detected within the window region was counted and plotted (solid line). The horizontal lines at the top of each graph indicate 99% and 95% confidence thresholds for globally significant breakpoint clusters. Light and dark grey areas respectively indicate local 99% and 95% breakpoint clustering thresholds, taking into account local regional differences in sequence diversity that influence the ability of different recombination detection methods to identify recombination breakpoints. Red areas indicate recombination hot-spots, while blue areas represent recombination cold-spots. ORFs (horizontal arrows) and IR are represented on the top of the graph.

doi:10.1371/journal.ppat.0030181.g002

In both the monopartite and bipartite datasets, larger "globally" significant (global p-values <0.05 across its length) and smaller "locally" significant (local p-values <0.01) recombination hot-spots are apparent in the intergenic region (IR) and complementary strand ORFs. In the monopartite dataset one large globally significant ($p < 0.01$) hot-spot encompasses the entire IR 5′ of the v-*ori*, and another ($p < 0.01$) occurs near the centre of *rep*. Globally significant hotspots ($p < 0.05$ and $p = 0.05$) are also detected in these positions in the bipartite dataset, but the extent of the IR hotspot is not as great. This is probably due in large part to the particularly low quality of nucleotide sequence alignment achievable amongst the highly divergent bipartite begomovirus IR sequences. Locally significant hot-spots occur at the interface of *cp* and the C3 ORFs in both bipartite and monopartite sequences. In addition to hot-spots, the analysis also revealed locally significant recombination cold-spots. These occurred in the first half of *cp* and in the third quarter of the V1 ORF for both datasets and in the overlapping region of the C2 and C3 ORFs of the bipartite dataset.

These results clearly indicate that recombination hot- and cold-spots previously identified amongst African and Mediterranean begomoviruses [16] are conserved amongst both monopartite and bipartite begomoviruses found worldwide.

Importantly, the distribution of these recombination hot- and cold-spots is largely consistent with observations made

during experimental analyses of geminivirus recombination [23–25] in which the V1/C3 ORF interface and the v-*ori* have been identified as potential recombination hot-spots. Also, analysis of replicating begomoviral DNA intermediates has revealed a wide distribution of so-called heterogeneous length linear dsDNA forms (hDNA). The ends of these hDNA molecules tend to map most frequently to the v-*ori* and either the AC2/AC3 transcription promoter at the hot-spot we detect in the centre of *rep*, or the C2/C3 terminator at the hot-spot we detect at the V1/C3 ORF interface. It has been convincingly demonstrated that these "broken" replicative intermediates are diverted into the recombination-dependent replication pathway of begomoviruses, which would neatly explain the recombination hot-spots detected in these regions [26].

Furthermore, population genetic analysis of recombination rates in large full genome datasets of very closely related groups of maize streak viruses (a geminivirus species in the genus *Mastrevirus*) and cassava-infecting geminiviruses has indicated that the base biochemical recombination rates in sequences encoding complementary sense genes are probably five to 12 times higher than they are in sequences encoding the virion sense genes [27]. Importantly, these studies also show that greatest changes in recombination rates occur near the V1/C3 ORF interface and the v-*ori*.

That all of these lines of evidence indicate the complementary sense ORFs of geminiviruses are biochemically more predisposed to recombination than their virion sense ORFs strongly suggests that something about the direction of transcription of these ORFs may be responsible for the recombination rate imbalance. For example, it has been proposed that complementary sense gene transcription, which occurs in the opposite direction to virion strand synthesis during rolling circle replication of geminivirus genomes, may be responsible for an increased rate of replication complex displacement during replication of the complementary sense ORFs [16,26]. Completion of replication from partially replicated virion strands would then proceed via the recombination-dependent replication pathway [26] which, in the presence of potential template DNAs with dissimilar sequences, could result in an increased prevalence of detectable recombination events across the complementary sense ORFs.

Do Recombination Breakpoints "Avoid" Disruption of Protein Folding?

While these mechanistic processes might account for both a general imbalance between recombination rates in the virion and complementary sense ORFs and hot-spots at the V1/C3 ORF interface, the centre of *rep* and the v-*ori*, they cannot completely explain, for example, the apparently conserved breakpoint clusters (albeit not hot-spots) at the 3′ end of the V2 ORF and breakpoint cool/cold-spots in the C2 ORF. We have noted previously [16] that, besides the hot-spot in *rep*, peaks in breakpoint density tend to occur either outside or near the ends of genes, whereas cold-spots tend to occur well within genes. As has been previously suggested in analyses of both virus recombination [28] and DNA shuffling experiments [8], this indicated to us that selection might preferentially favour the survival of recombinants that either do not express chimaeric proteins, or express chimaeric proteins in which recombination has not damaged amino acid interactions required for proper protein folding. We therefore decided to test whether selection against disruption of protein folding might not also be at least partially responsible for some of the conserved breakpoint density peaks and troughs observable in Figure 2.

Recombination events with inferred breakpoint positions within the portions of *rep* and *cp* genes that encode protein fragments with available 3-D structural data were identified. These included 12 and five events in the *cp* genes of the monopartite and bipartite begomoviruses, respectively, and 29 and 19 events in the *rep* genes of the monopartite and bipartite begomoviruses, respectively.

We used the SCHEMA method to predict degrees of fold disruption in the chimaeric Rep and CP molecules expressed by the recombinant viruses we identified. This analysis indicated that the average degree of potential fold disruption in the chimaeric Rep molecules was higher than that in the chimaeric CP molecules (E-values in Table 1). Rather than indicating that recombination should be more tolerable in *cp* than it is in *rep*, this result simply reflects that CP molecules are much more conserved than Rep molecules and that there are consequently fewer potentially disruptive combinations of amino acids. It should also be pointed out that both CP and Rep have overlapping ORFs: V2/AV2 in the case of CP and C4/AC4 in the case of Rep. Whereas in the case of CP the overlap with V2/AV2 involves only approximately 3% (six codons) of the analysed CP region, the overlap of Rep and C4/AC4 involves approximately 58% (67 codons) of the analysed Rep region. It is possible that the evolutionary constraints of overlapping coding regions have made Rep more robust than CP with respect to potentially clashing amino acid interactions within its folded structure. Importantly, by increasing the number of tolerable Rep mutations, the only influence increased folding robustness might have had on our analyses would have been to decrease the power of our tests for preservation of intra-protein interactions.

We performed SCHEMA analyses on the two simulated datasets (described in Materials and Methods) to determine whether the degrees of Rep and CP fold disruption predicted for the observed recombinants were significantly lower than would be expected if selection did not act against recombinants expressing chimaeric proteins with high degrees of predicted fold disruption. Analysis of the exhaustive genome event dataset indicates that predicted degrees of fold disruption in real CP and Rep chimaeras were significantly lower (Table 1; p-values $= 5.1 \times 10^{-2}$ and $< 1.0 \times 10^{-4}$ for the bipartite and monopartite CP datasets, respectively, and 1.7×10^{-3} and 6.6×10^{-3} for the monopartite and bipartite Rep datasets, respectively) than would be expected in the absence of selection against fold disruption. However, the amino acid mutation levels (m-values in Table 1) of the real datasets indicate that the real recombination events also tended to involve transfers of significantly fewer non-synonymous mutations than the simulated datasets (Table1; p-values $= 1.0 \times 10^{-4}$ and 1.7×10^{-2} for the mono and bipartite CP datasets, respectively, and 1.0×10^{-4} and 2.4×10^{-2} for the monopartite and bipartite Rep datasets, respectively). This indicated that the significantly reduced degrees of predicted fold disruption in the real datasets relative to the simulated datasets are at least in part due to the real recombination events involving transfers of significantly fewer non-synonymous mutations than the simulated events. This implies that

Table 1. SCHEMA Derived Estimates of Recombination Induced Effective Mutation (m) and Protein Fold Disruption (E) for Monopartite and Bipartite Begomovirus Datasets Comparing Real and Simulated Recombination Events

Genomic Portion Analysed		V1 / AV1			C1 / AC1		
Position in Monopartite and Bipartite Sequences[a]		468-1055 / 502-1089			2249–2601 / 2282–2635		
Recombination Type		Real	Exhaustive Genome	Exhaustive Non-Synonymous	Real	Exhaustive Genome	Exhaustive Non-Synonymous
Monopartite	Number of events	12 (11[b])	3080	274	29 (28[b])	5281	835
	Mutation (m) mean ± sd (*p*-value[c])	1.7 ± 1.1	6.0 ± 3.2 (1.0e-4)	1.8 ± 1.0 (1.0)	6.5 ± 4.4	8.12 ± 3.5 (1.0e-4)	6.75 ± 4.3 (1.0)
	Disruption (E) mean ± sd (*p*-value[c])	0.33 ± 0.89	2.3 ± 2.4 (<1.0e-4)	1.3 ± 1.1 (3.0e-4)	6.0 ± 4.6	8.9 ± 6.6 (1.7e-3)	8.8 ± 6.9 (<1.0e-4)
Bipartite	Number of events	5 (5[b])	1543	188	19 (18[b])	3942	750
	Mutation (m) mean ± sd (*p*-value[c])	2.8 ± 1.6	9.1 ± 3.2 (1.7e-2)	2.8 ± 1.6 (1.0)	8.3 ± 5.0	10.8 ± 4.5 (2.4e-2)	8.7 ± 4.7 (1,0)
	Disruption (E) mean ± sd (*p*-value[c])	0.80 ± 0.84	4.0 ± 2.8 (5.1e-2)	2.2 ± 1.7 (2.9e-2)	8.7 ± 8.8	12.7 ± 8.7 (6.6e-3)	12.3 ± 8.3 (6.0e-4)

[a]Nucleotide positions within the ORFs analysed in relation to ToLCYTV-[Dem] (AJ865341) for monopartite sequences and EACMV-[TZ] (Z83256) for bipartite sequences.
[b]Number of events involving the transfer of non-synonymous polymorphisms.
[c]Probability that real events are not less mutative/disruptive than the simulated events (exhaustive genome events and exhaustive non-synonymous events). $p < 0.05$ was considered significant.
doi:10.1371/journal.ppat.0030181.t001

the real breakpoints tend to occur closer to the edges of the analysed regions than one would expect if breakpoints occurred randomly throughout the regions. Given the breakpoint density peaks on either side of the CP encoding V1 ORF (Figure 2), we had anticipated this result for the CP dataset. However, a similar result obtained for the Rep dataset implies that, despite the high density of breakpoints throughout the C1 region encoding the fragment of Rep that was analysed, there is still a significant tendency for breakpoints to occur closer to the edges of the analysed region than would be expected by chance.

We therefore decided to test whether avoidance of protein fold disruption is achieved only through avoidance of non-synonymous mutation mixing, or whether, controlling for unequal degrees of non-synonymous mutation mixing, it is also achieved through preferential mixing of non-disruptive non-synonymous mutations. Analysis of the simulated non-synonymous event dataset indicated that in the natural monopartite and bipartite *cp* and *rep* recombinants, there does indeed appear to have been preferential mixing of non-disruptive non-synonymous mutations (Table 1). This demonstrates, therefore, that whenever an interaction between a pair of polymorphic amino acid residues is predicted to be important for Rep or CP folding, there is a tendency for nucleotide sequences encoding these amino acids to be inherited from the same parental virus significantly more often than those encoding non-interacting pairs of polymorphic amino acids.

Despite the common conception that recombination is a highly efficient mechanism used by both microorganisms and protein engineers in the discovery of phenotypic novelty and/or improved fitness, rules constraining the evolutionary utility of recombination are beginning to emerge. It has been experimentally demonstrated that the viability of recombinant viruses and the activities of chimaeric proteins are strongly influenced by both the relatedness of their parents and the inherent "modularity" of the sequence tracts they inherit from them [8,10]. Put simply, fragments of sequence that do not interact extensively with other sequence fragments tend to function well when transferred into even highly divergent foreign genetic backgrounds, whereas those that interact extensively with other sequence fragments tend to only work properly when transferred into foreign genetic backgrounds that are not very different from those in which they evolved [10,11]. It is therefore probable that high profile natural recombinant viruses, such as those that are responsible for disease outbreaks [29] or those that have novel host ranges and phenotypes [30,31], or even those that have simply emerged as prominent circulating members of virus populations [32–35], represent the exceptional, reasonably fit subset of a vastly greater but vastly more ephemeral group of defective "hopeful monsters" culled by purifying selection.

Our results provide clear supporting evidence for this notion that purifying selection is a major factor shaping at least part of the distinctive patterns of natural recombination found in begomoviruses. Within the genome regions analysed, we find strong statistical evidence that natural recombination events have tended to involve sequence exchanges that avoid the transfer of non-synonymous nucleotide polymorphisms (i.e., those encoding different amino acids in the different parents) between genomes. We also show that, when non-synonymous polymorphisms are transferred between genomes, there is a statistically significant tendency to avoid transfer of those non-synonymous polymorphisms that are predicted to disrupt the folding of expressed chimaeric proteins. While these results indicate that interaction networks required for proper protein folding are preserved in the natural recombinants, they imply that strong selective forces must operate against any novel recombinant in which these interaction networks are not preserved.

A major omission in our analysis of intra-protein amino acid interactions is our failure to consider all the other potential interactions that most likely occur within the genomes analysed. These include many sequence-specific inter-protein and protein–DNA interactions that might also constrain the viability of recombinants [10]. While experimental work towards obtaining high resolution genome-wide interaction maps has only begun for most virus taxa (including the begomoviruses), exciting new analysis methods

are being developed to identify both coevolving (or covarying) amino acids within protein sequence alignments [36–38] and epistatically interacting nucleotide sites within DNA sequence alignments [39,40]. Although these methods promise the mapping of interaction networks directly from naturally sampled viral genome sequences, it is currently unknown how well they will fare given datasets containing (1) large numbers of recombinant sequences or (2) obvious recombination hot-spots. It is very likely, for example, that many non-synonymous polymorphisms on sequence tracts between recombination hot-spots will be detectably "covariant" if they are frequently transferred amongst genomes (i.e., on an imposed phylogenetic tree it will appear as though the same sets of sites change simultaneously on multiple branches of the tree). If some of these or future related analytical methods prove robust to the influences of recombination, an obvious application of these would be to determine whether there is also a significant tendency for recombination to avoid disrupting these genome-wide protein–protein, protein–nucleic acid, and nucleic acid–nucleic acid interactions.

In fields as diverse as microbial evolution [41,42], protein engineering [8,9], and computer science [43], maintenance of interaction networks is emerging as a common theme unifying studies aimed at delimiting recombination's potential as an exploratory strategy. Many and complex interactions is a defining feature of living systems. When these interactions are encoded within genome sequences they form an epistatic architecture. It is really just common sense that for productive recombination to occur it must happen without damaging the integrity of these largely intangible network-like structures. Maintenance of these networks might in fact be directly responsible for the evolution of differential biochemical predispositions for recombination across genomes: If recombination events mostly occur in genome regions with low connectivity, a greater proportion of recombinants would be viable than if recombination events were randomly scattered across genome regions with low and high connectivity. Other evolutionary strategies to ensure maintenance of interaction networks in the face of continual recombination might be the evolution of network robustness, or an increased capacity to mutationally compensate for deleterious recombination events. Conversely, however, the network architectures themselves might also evolve over time to accommodate biochemically predisposed recombination hot-spots that have some biological importance. We have shown here that in the case of the begomoviruses at least, various biochemical and selective processes working in tandem most likely combine to produce the distinctive patterns of recombination seen in nature.

Materials and Methods

Sequence data. All available monopartite and bipartite begomovirus DNA-A and DNA-A-like sequences were obtained from public sequence databases using TaxBrowser (http://www.ncbi.nlm.nih.gov/) in May 2006. Multiple sequence alignments were constructed separately for monopartite and bipartite sequences using POA [44], the ClustalW [45] based sub-alignment tool available in MEGA 3.1 [46], and manual editing. While great care was taken to ensure the most accurate alignment possible, during subsequent recombination analyses additional alignment checks were performed in RDP3 (also using the ClustalW method) for every recombination signal detected to ensure that they were not misalignment artefacts [45]. To minimise the number of tests performed during recombination analyses (and therefore increase the statistical power of the analyses) all but one sequence within groups of sequences sharing more than 98% nucleotide identity were discarded. The resulting monopartite DNA-A-like and bipartite DNA-A sequence alignments contained 123 and 116 sequences, respectively.

Structural data. The Rep protein structure (catalytic domain; residues 4–121) of TYLCSV has been determined by NMR spectroscopy. This Rep structure (PDB ID 1L2M) comprises five anti-parallel β sheets in the centre with a two-stranded β sheet, a β hairpin, and two α helices on the periphery [47]. The begomovirus capsid structures (196-aa core CP) has been modeled based on the crystal structure of *Satellite Tobacco Necrotic Virus* and fitted into an approximately 20-Å density map generated from cryoelectron microscopy reconstructions of *African cassava mosaic virus* particles [48]. The PDB file of this structure was kindly provided by B. Böttcher.

Recombination analysis. Identification of potential recombinants, parental sequences, and approximation of possible recombination breakpoint positions was carried out using the RDP [49], Geneconv [22], RecScan [50], Maximum Chi Square [51], Chimaera [52], and SisterScan [53] methods as implemented in RDP3 [52], which is available from http://darwin.uvigo.es/rdp/rdp.html (for full details of program settings, see Datasets S1 and S2). The analyses were performed with default settings for the detection methods, a Bonferroni-corrected *p*-value cutoff of 0.05, and a requirement that any potential event be detectable by two or more methods. It is important to point out that implementations of all these recombination detection methods in RDP3 were not severely constrained by the initial window size settings specified at the onset of the analyses. All of the methods used include an algorithm for dynamically optimising window sizes for the detection of recombination signals during an initial exploratory phase of recombination detection. Following this exploratory phase, RDP3 rechecks every detected recombination signal with all six methods with a starting window size seeded with that used by whatever method initially detected the signal.

The approximate breakpoint positions and recombinant sequence(s) inferred for every detected potential recombination event were manually checked and adjusted where necessary using the extensive phylogenetic and recombination signal analysis features available in RDP3. This process further reduced any possible influence that initial window size settings had on the final estimates of breakpoint positions. Once a set of unique recombination events was identified, a breakpoint map containing the positions of all clearly identifiable breakpoints was compiled. A breakpoint density plot was then constructed from this map as described in Heath et al. [17].

SCHEMA analysis. SCHEMA takes as input a PDB protein structure file and parental amino acid sequence files. It uses the protein structural information to properly fold the parental amino acid sequences and then identifies potentially interacting amino acid pairs based on their proximity (in this case within 4.5 Å) within the resulting folds. The amino acid contact map yielded by this process can be used to determine the degree of fold disruption expected in any conceivable chimaera of the parental amino acid sequences. The way this is done is relatively simple: For all the amino acid residues that are potentially interacting within a folded chimaeric protein, SCHEMA counts the number of instances where the interacting pairs are non-parental. Non-parental interacting amino acid pairs arise when the parental molecules differ from one another at two potentially interacting amino acid residues and the chimaera inherits one-half of the potentially interacting pair from one parent and the other half from the other parent. Counts of these non-interacting pairs in chimaeric proteins, called "E" values, have been shown to correlate directly with degrees of fold disruption experienced by the proteins. The value of E therefore corresponds with expected degree of fold disruption. SCHEMA also counts the number of amino acid substitutions that would be required to convert a chimaera into the parental sequence that it most closely resembles—this value is referred to as "m" [8].

We selected recombination events in the monopartite and bipartite sequence datasets for which (1) sequences closely resembling inferred parental sequences were identifiable and (2) recombination breakpoints occurred in genome regions encoding the portions of Rep and CP with known/approximated 3-D structures. These events constituted a "real event" dataset that we analysed using the SCHEMA method.

We devised a permutation test to determine whether predicted CP and Rep fold disruptions incurred by real events were less severe than those incurred by random recombination events with the same

parental sequences simulated throughout the *rep* and *cp* regions under consideration. The permutation test involved two different sets of simulated recombination events. The first set was derived from each real event by moving the breakpoints observed in the real event backwards and forwards along the entire nucleotide sequence alignment one polymorphic alignment position at a time until every possible unique recombination event involving the "exchange" of exactly the same number of polymorphic nucleotides as the real event were simulated within the parental sequences. We called the complete set of simulated events constructed from the entire real event dataset the "exhaustive genome event" dataset. Note that although these events involved exchanges of the same number of total polymorphic sites as was observed for the real events, they can involve a different number of polymorphic sites in the particular genomic regions analysed (i.e., those encoding portions of CP and Rep with available 3-D structure information). This set of simulated events was used to determine whether there was a significant tendency for the observed recombination breakpoints to occur on the edges of these analysed regions. The second set of simulated events was generated by considering only non-synonymous polymorphisms within the alignment regions encoding CP or Rep fragments with available structural data. A window containing the same number of non-synonymous mutations as a corresponding real event was moved along the analysed region, and all possible recombination events were simulated. The subsequent events share exchanges of exactly the same numbers of non-synonymous mutations as the real events and consequently all have the same SCHEMA m values. This set of events was called the "exhaustive non-synonymous event" dataset (see Figure S2 for simulation details). This set of simulated events was used to determine whether, given "exchanges" of the same numbers of polymorphic amino acids as were observed for real events, there was a significant tendency for the real recombination events to exchange less disruptive amino acid polymorphisms.

Quantification of potential fold disruption in real and simulated chimaeric CP and Rep molecules, respectively, expressed by real and simulated recombinants, was carried out using SCHEMA. For each of these chimaeras, amino acid sequences of inferred parents and chimeras were aligned with MUSCLE using default settings [54]. The python scripts SCHEMACONTACTS and SCHEMAENERGY [9] were used to compute m (mutational distance between chimaeras and their most closely related parent) and E (predicted fold disruption) scores for all simulated and real chimaeras. The analysis procedure was automated using some of the extensive functionalities available in the R package [55], APE [56], and seqinR (available on http://www.cran.r-project.org/). R scripts for these analyses are available on request. We grouped the E and m scores determined for the observed and simulated chimaeras and determined the sum of ranks for the observed chimaeras. We then repeated the entire process 10,000 times but with "real" events randomly chosen from amongst every subset of corresponding simulated events. We propose that the proportion of simulated events with a sum of ranks score lower than or equal to that of the observed event is equivalent to the probability that the breakpoint distributions observed in the real dataset have not tended to avoid disruption of protein folding. Put another way, we estimate a p-value from the proportion of permuted recombination event datasets that on the whole are predicted to be less disruptive to protein folding than the set of actual observed recombination events.

Supporting Information

Dataset S1. RDP3 Bipartite Sequences Analysis

RDP3 project file.

Found at doi:10.1371/journal.ppat.0030181.sd001 (379 KB ZIP).

Dataset S2. RDP3 Monopartite Sequences Analysis

RDP3 project file.

Found at doi:10.1371/journal.ppat.0030181.sd002 (412 KB ZIP).

Figure S1. Recombination Event Density and Parental Sequences Relatedness

Found at doi:10.1371/journal.ppat.0030181.sg001 (25 KB PDF).

Figure S2. Recombination Event Simulation Process

Figures describing how the simulated recombination events are created.

Found at doi:10.1371/journal.ppat.0030181.sg002 (108 KB PDF).

Table S1. Bipartite Sequences Recombination Events

List of recombination events detected with RDP3 in bipartite sequences.

Found at doi:10.1371/journal.ppat.0030181.st001 (132 KB XLS).

Table S2. Monopartite Sequences Recombination Events

List of recombination events detected with RDP3 in monopartite sequences.

Found at doi:10.1371/journal.ppat.0030181.st002 (140 KB XLS).

Accession Numbers

The National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) accession numbers for the sequences discussed in this paper are EACMCV-[TZ] (AY795983), ToLCYTV-[Dem] (AJ865341), and TYLCSV (X61153). The Protein Data Bank (http://www.pdb.org/) ID number for TYLCSV N-terminal region structure is 1L2M.

Acknowledgments

The authors want to thank Frédéric Chiroleu and Caroline Domerg for R programming assistance, Cathal Seoighe and David Posada for critical readings, and Arvind Varsani for helpful comments on structural data.

Author contributions. PL, JML, BR, and DPM conceived and designed the experiments. PL performed the experiments. PL and DPM analyzed the data. PL and DPM contributed analysis tools. PL, JML, BR, and DPM wrote the paper.

Funding. PL is funded by CIRAD and the French Ministère de la Recherche et de l'Enseignement Supérieur. JML and BR are funded with the regional council of Reunion Island. DPM is funded by the South African National Bioinformatics Network, the Wellcome Trust, and the Harry Oppenheimer Trust, and holds a Sydney Brenner Fellowship.

Competing interests. The authors have declared that no competing interests exist.

Reference

1. Michel B, Flores MJ, Viguera E, Grompone G, Seigneur M, et al. (2001) Rescue of arrested replication forks by homologous recombination. Proc Natl Acad Sci U S A 98: 8181–8188.
2. Cromie GA, Connelly JC, Leach DR (2001) Recombination at double-strand breaks and DNA ends: conserved mechanisms from phage to humans. Mol Cell 8: 1163–1174.
3. Felsenstein J (1974) The evolutionary advantage of recombination. Genetics 78: 737–756.
4. Keightley PD, Otto SP (2006) Interference among deleterious mutations favours sex and recombination in finite populations. Nature 443: 89–92.
5. Martin G, Otto SP, Lenormand T (2006) Selection for recombination in structured populations. Genetics 172: 593–609.
6. Crameri A, Raillard SA, Bermudez E, Stemmer WP (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature 391: 288–291.
7. Stemmer WP (1994) Rapid evolution of a protein in vitro by DNA shuffling. Nature 370: 389–391.
8. Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH (2002) Protein building blocks preserved by recombination. Nat Struct Biol 9: 553–558.
9. Meyer MM, Silberg JJ, Voigt CA, Endelman JB, Mayo SL, et al. (2003) Library analysis of SCHEMA-guided protein recombination. Protein Sci 12: 1686–1693.
10. Martin DP, van der Walt E, Posada D, Rybicki EP (2005) The evolutionary value of recombination is constrained by genome modularity. PLoS Genet 1: e51. doi:10.1371/journal.pgen.0010051
11. Escriu F, Fraile A, Garcia-Arenal F (2007) Constraints to genetic exchange support gene coadaptation in a tripartite RNA virus. PLoS Pathog 3: e8. doi:10.1371/journal.ppat.0030008
12. Moreno IM, Malpica JM, Diaz-Pendon JA, Moriones E, Fraile A, et al. (2004) Variability and genetic structure of the population of watermelon mosaic virus infecting melon in Spain. Virology 318: 451–460.
13. Jain R, Rivera MC, Moore JE, Lake JA (2003) Horizontal gene transfer accelerates genome innovation and evolution. Mol Biol Evol 20: 1598–1602.
14. Hayes RJ, Bentzien J, Ary ML, Hwang MY, Jacinto JM, et al. (2002) Combining computational and experimental screening for rapid optimization of protein properties. Proc Natl Acad Sci U S A 99: 15926–15931.
15. Saraf MC, Maranas CD (2003) Using a residue clash map to functionally characterize protein recombination hybrids. Protein Eng 16: 1025–1034.
16. Lefeuvre P, Martin DP, Hoareau M, Naze F, Delatte H, et al. (2007)

Begomovirus "melting pot" in the South West Indian Ocean Islands: molecular diversity and evolution through recombination. J Gen Virol 88: 3458–3468.
17. Heath L, van der Walt E, Varsani A, Martin DP (2006) Recombination patterns in aphthoviruses mirror those found in other picornaviruses. J Virol 80: 11827–11832.
18. Shapka N, Nagy PD (2004) The AU-rich RNA recombination hot spot sequence of Brome mosaic virus is functional in tombusviruses: implications for the mechanism of RNA recombination. J Virol 78: 2288–2300.
19. Ohshima K, Tomitaka Y, Wood JT, Minematsu Y, Kajiyama H, et al. (2007) Patterns of recombination in turnip mosaic virus genomic sequences indicate hotspots of recombination. J Gen Virol 88: 298–315.
20. Magiorkinis G, Paraskevis D, Vandamme AM, Magiorkinis E, Sypsa V, et al. (2003) In vivo characteristics of human immunodeficiency virus type 1 intersubtype recombination: determination of hot spots and correlation with sequence similarity. J Gen Virol 84: 2715–2722.
21. Chin MP, Rhodes TD, Chen J, Fu W, Hu WS (2005) Identification of a major restriction in HIV-1 intersubtype recombination. Proc Natl Acad Sci U S A 102: 9002–9007.
22. Padidam M, Sawyer S, Fauquet CM (1999) Possible emergence of new geminiviruses by frequent recombination. Virology 265: 218–225.
23. Stenger DC, Revington GN, Stevenson MC, Bisaro DM (1991) Replicational release of geminivirus genomes from tandemly repeated copies: evidence for rolling-circle replication of a plant viral DNA. Proc Natl Acad Sci U S A 88: 8029–8033.
24. Schnippenkoetter WH, Martin DP, Willment JA, Rybicki EP (2001) Forced recombination between distinct strains of Maize streak virus. J Gen Virol 82: 3081–3090.
25. Garcia-Andres S, Tomas DM, Sanchez-Campos S, Navas-Castillo J, Moriones E (2007) Frequent occurrence of recombinants in mixed infections of tomato yellow leaf curl disease-associated begomoviruses. Virology 365: 210–219.
26. Jeske H, Lutgemeier M, Preiss W (2001) DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. EMBO J 20: 6158–6167.
27. Owor BE, Martin DP, Shepherd DN, Edema RE, Monjane AL, et al. (2007) Genetic analysis of maize streak virus (MSV) isolates from Uganda reveals widespread distribution of a recombinant MSV variant in Uganda. J Gen Virol 88: 3154–3165.
28. Bonnet J, Fraile A, Sacristan S, Malpica JM, Garcia-Arenal F (2005) Role of recombination in the evolution of natural populations of Cucumber mosaic virus, a tripartite RNA plant virus. Virology 332: 359–368.
29. Legg JP, Thresh JM (2000) Cassava mosaic virus disease in East Africa: a dynamic disease in a changing environment. Virus Res 71: 135–149.
30. Russell CJ, Webster RG (2005) The genesis of a pandemic influenza virus. Cell 123: 368–371.
31. Gibbs MJ, Weiller GF (1999) Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus. Proc Natl Acad Sci U S A 96: 8022–8027.
32. Chare ER, Holmes EC (2006) A phylogenetic survey of recombination frequency in plant RNA viruses. Arch Virol 151: 933–946.
33. Rousseau CM, Learn GH, Bhattacharya T, Nickle DC, Heckerman D, et al. (2007) Extensive intrasubtype recombination in South african human immunodeficiency virus type 1 subtype C infections. J Virol 81: 4492–4500.
34. Varsani A, van der Walt E, Heath L, Rybicki EP, Williamson AL, et al. (2006) Evidence of ancient papillomavirus recombination. J Gen Virol 87: 2527–2531.
35. Tyler SD, Severini A (2006) The complete genome sequence of herpesvirus papio 2 (Cercopithecine herpesvirus 16) shows evidence of recombination events among various progenitor herpesviruses. J Virol 80: 1214–1221.
36. Fares MA, McNally D (2006) CAPS: coevolution analysis using protein sequences. Bioinformatics 22: 2821–2822.
37. Gloor GB, Martin LC, Wahl LM, Dunn SD (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 44: 7156–7165.
38. Dutheil J, Pupko T, Jean-Marie A, Galtier N (2005) A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol 22: 1919–1928.
39. Shapiro B, Rambaut A, Pybus OG, Holmes EC (2006) A phylogenetic method for detecting positive epistasis in gene sequences and its application to RNA virus evolution. Mol Biol Evol 23: 1724–1730.
40. Zuker M, Jacobson AB (1998) Using reliability information to annotate RNA secondary structures. RNA 4: 669–679.
41. Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 96: 3801–3806.
42. Fraser C, Hanage WP, Spratt BG (2007) Recombination and the nature of bacterial speciation. Science 315: 476–480.
43. Holland J (1975) Adaptation in natural and artificial systems. Ann Arbor (Michigan): University of Michigan Press.
44. Lee C, Grasso C, Sharlow MF (2002) Multiple sequence alignment using partial order graphs. Bioinformatics 18: 452–464.
45. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
46. Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5: 150–163.
47. Campos-Olivas R, Louis JM, Clerot D, Gronenborn B, Gronenborn AM (2002) The structure of a replication initiator unites diverse aspects of nucleic acid metabolism. Proc Natl Acad Sci U S A 99: 10310–10315.
48. Bottcher B, Unseld S, Ceulemans H, Russell RB, Jeske H (2004) Geminate structures of African cassava mosaic virus. J Virol 78: 6758–6765.
49. Martin D, Rybicki E (2000) RDP: detection of recombination amongst aligned sequences. Bioinformatics 16: 562–563.
50. Martin DP, Posada D, Crandall KA, Williamson C (2005) A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses 21: 98–102.
51. Maynard SJ (1992) Analyzing the mosaic structure of genes. J Mol Evol 34: 126–129.
52. Martin DP, Williamson C, Posada D (2005) RDP2: recombination detection and analysis from sequence alignments. Bioinformatics 21: 260–262.
53. Gibbs MJ, Armstrong JS, Gibbs AJ (2000) Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16: 573–582.
54. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
55. R Development Core Team (2006) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07–0. Available: http://www.R-project.org/. Accessed 23 October 2007.
56. Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–290.

Chapitre IV : Elargissement aux virus à ADN simple brin circulaire

Si la recombinaison est un phénomène qui a depuis longtemps été décrit chez les bégomovirus, avec notamment des avancés significatives dans la compréhension du phénomène (chapitre II et III), il n'en est pas de même pour les autres virus apparentés. Ainsi, au regard du nombre de familles et de genres viraux à ADN simple brin circulaire peu d'études concernent la recombinaison (Amin et al., 2006; Csagola et al., 2006; Gao et al., 2003; Heath et al., 2006; Hughes, 2004; Lukashov & Goudsmit, 2001; Olvera et al., 2007; Padidam et al., 1999; Rokyta et al., 2006; Shackelton et al., 2007).

Les virus à ADN simple brin circulaire, malgré des gammes d'hôtes très variés (Animaux pour les *Circoviridae*, les *Anellovirus* et les *Parvoviridae*, Plantes pour les *Geminiviridae* et les *Nanoviridae* et Bactéries pour les *Microviridae*), présentent certaines particularités communes avec notamment la réplication du génome par le mécanisme de réplication en cercle roulant. Une origine commune a par ailleurs été proposée pour ces virus en se basant sur l'homologie de certaines séquences protéiques (Gibbs et al., 2006; Koonin and Ilyina, 1992).

Afin de déterminer s'il existe une certaine conservation des profils et règles de recombinaison liés par exemple à l'organisation génomique, à la réplication en cercle roulant ou encore a la nature des gènes, une analyse comparative a été réalisée sur ces virus. Nous nous sommes attachés à décrire et comparer les profils de recombinaison et la capacité de recombinaison de ces différents virus.

Après constitution des jeux de données et analyse des séquences, nous avons montré que la recombinaison était pour la majorité de ces virus un moteur majeur d'évolution. L'analyse des profils de recombinaison et des structures d'échange de gènes, a révélé que certaines particularités d'échange de matériel génétique précédemment décrites pour les géminivirus semblent également prévaloir pour la grande majorité des virus à ADN circulaire simple brin. La distribution du nombre de points de recombinaison a été analysée statistiquement laissant apparaître que (1) les *hot spot* de recombinaison sont majoritairement présents dans les sections non codantes des génomes, que (2) les *cold spot* de recombinaison sont principalement présents au niveau des zones codantes, qui lorsqu'elles sont échangées le sont

généralement en un seul bloc, et que (3) la protéine de capside est particulièrement épargnée par la recombinaison tandis que le gène codant pour la protéine associée à la réplication est celui qui supporte le plus de recombinaison.

Ces résultats soulignent encore une fois l'importance de la sélection sur le « façonnage » des profils de recombinaison avec des niveaux différents de capacité de recombinaison entre régions codantes et non codantes mais aussi entre les régions codantes elles-mêmes. Les possibilités de création par recombinaison de génomes viraux peu adaptés avec des protéines dont la structure et la fonction sont altérées semblent élevées. Les virus moins adaptés portant ces re-arrangements délétères seront probablement éliminés par sélection purificatrice. Par ailleurs, la présence de *cold spot* ou de *hot spots* au sein de certains gènes semble refléter le faible ou le haut niveau de plasticité, respectivement, de certaines familles de gènes. Ainsi, la faible capacité de recombinaison de la protéine de capside est très probablement liée à son caractère multifonctionnel et notamment son implication dans la structure des particules virales. Concernant, la protéine associée à la réplication (absente chez les microvirus, qui possèdent leur propre protéine de réplication), sa grande capacité de recombinaison avait été préalablement démontrée chez les bégomovirus. Pour une majorité de virus à ADN circulaire simple brin, la partie N-terminale de cette protéine semble particulièrement « échangeable » et modulable. Il semblerait que cette plasticité par recombinaison puisse jouer un rôle important dans l'adaptation des virus à de nouveaux hôtes (Prasanna and Rai, 2007).

La recombinaison joue un rôle majeur dans l'évolution des virus à ADN circulaire simple brin en général. L'aptitude des virus à ADN circulaire simple brin à échanger du matériel génétique par recombinaison et le fait qu'ils semblent présenter un taux de mutation élevé similaire aux virus à ARN (Duffy et al., 2008 pour revue) sont de solides atouts leurs permettant d'évoluer rapidement et de s'adapter à de nouvelles niches écologiques. Concernant les géminivirus, leur distribution mondiale à la fois sur des plantes monocotylédones et dicotylédones semble le prouver. L'ensemble de leurs capacités évolutives fait de ces virus des candidats potentiels à l'émergence, domaine dans lequel ils ont rencontré ces dernières années un franc succès (Carman et al., 2008; Duffy and Holmes, 2008; Shackelton and Holmes, 2006; Shackelton et al., 2005; Zhou et al., 1997).

Widely conserved recombination patterns amongst single stranded DNA viruses and their satellites

P. Lefeuvre[a], J.-M. Lett[a], A. Varsani[b] and D.P. Martin[c]*

[a]CIRAD, UMR 53 PVBMT CIRAD-Université de la Réunion, Pôle de Protection des Plantes, Ligne Paradis, 97410, Saint Pierre, La Réunion, France

[b]Electron Microscope Unit, University of Cape Town, Private Bag, Rondebosch 7701, South Africa

[c]Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory 7925, South Africa

*Author for correspondence

Abstract

Recombination, by enabling the creation of new genomic combinations from existing diversity, provides to organism the potential to explore a large diversity. More and more studies are reporting the recombinant nature of virus from a wide genus range. However, despites those progress in describing recombination, a few is known about the rules that constrain recombination and shape recombination patterns observable in nature. It has been hypothesis that networks of coevolving interactions, that could define function itself, must be preserved for a virus to be viable. This concept was successfully applied to the high recombinagenic *Geminiviridae* family. Here, we extend this model to the majority of circular single strand DNA viruses. By analysing sequences available in public database, we show that recombination was not randomly distributed along the genome. While hypothesis on possible mechanistic origin of those imbalance recombination patterns were proposed, we statistically show the varying recombinability of the different gene classes, with strong support for preservation of structural genes. Those finding clearly highlight the constraint on protein functionality and so selection that recombinant viruses experiment.

Introduction

Genetic recombination is a ubiquitous biological process that is both a central feature in DNA repair pathways and an important evolutionary mechanism. By generation of novel combinations of pre-existing genetic variation, recombination can paradoxically accelerate evolution by increasing the genetic diversity upon which adaptive selection relies, while at the same time decreasing the rate at which harmful mutations accumulate within populations (Keightley and Otto, 2006; Martin et al., 2006). Whereas its ability to defend high fitness genomes from mutational decay most likely underlies the evolutionary value of sexuality in higher organisms, in many microbial species where pseudo-sexual genetic exchange is permissible even amongst highly divergent genomes, recombination provides immediate access to evolutionary innovations that would be unachievable by mutation alone (Crameri et al., 1998).

Such inter-species recombination appears to be fairly common in many virus families (Baird et al., 2006; Chare et al., 2003; Heath et al., 2006; Magiorkinis et al., 2007; Varsani et al., 2006). It is, however, becoming clear that as with mutations, most recombination events between distantly related genomes will be maladaptive. As genetic distances between parental genomes increases so too does the probability of fitness defects in their recombinant offspring (Escriu et al., 2007; Martin et al., 2005a). The viability of recombinants is apparently largely dependent on how severely recombination disrupts co-evolved intra-genome interaction networks. These interactions might include tracts of nucleotide sequence that form secondary structures, sequence specific protein-DNA binding, inter-protein binding, and intra-protein amino acid-amino acid interactions within 3D folds.

One virus family where within genome interaction networks appear to have a large impact on patterns of natural inter-species recombination are the single stranded DNA (ssDNA) geminiviruses. As with other ssDNA viruses inter-species recombination is very common amongst the species of this family. The partially conserved recombination hot and cold-spots in different geminivirus genera are apparently the products of both differential mechanistic

predispositions genome regions to recombination, and natural selection disfavouring the survival of recombinants with disrupted intra-genome interactions.

Genome organisation and rolling circle replication (RCR; see Gutierrez, 1999 for a review) – the mechanism by which geminiviruses and many other ssDNA viruses replicate – seem to have a large influence on base recombination rates in different parts of geminivirus genomes (Garcia-Andres et al., 2007; Jeske et al., 2001; Lefeuvre et al., 2007b; Owor et al., 2007; van der Walt et al., 2008; Varsani *et a*l., 2008). To initiate RCR, virion strand ssDNA molecules are converted by host mediated pathways into a double stranded "replicative form" (RF). Initiated by a virus encoded replication associated protein (Rep) at a well defined virion strand replication origin (v-*ori*), new virion strands are synthesised on the complementary strand of RF DNAs. Virion strand replication is concomitant with the displacement of the old virion strands ultimately yielding covalently closed ssDNA geminivirus genomes which are then either encapsidated or converted into additional RFs.

Geminivirus genomes are organised with genes expressed from both the complementary and virion strands. This arrangement poses a potential problem in that during RCR virion strand replication proceeds on the same strand but in the opposite direction to complementary sense gene transcription. Clashes between replication and transcription complexes within the complementary sense genes is implied by the fact that the ends of linear heterogeneous length DNA fragments commonly occurring during geminvirus replication occur most frequently within the complementary sense ORFs (Jeske et al., 2001). That these clashes result in base recombination rates within the complementary sense genes being higher than those in the virion sense ORFs is in turn suggested by the complementary sense genes of viruses sampled from nature both containing more detectable recombination breakpoints (Lefeuvre et al., 2007b) and yielding higher population scaled recombination rate estimates (Owor et al., 2007).

Besides RCR playing a role in the apparent imbalance in recombination rates between geminivirus complementary and virion sense genes, it is probably responsible for two discrete recombination hotspots at both the v-*ori* and complementary sense gene transcript termination site. The v-*ori* hotspot is likely caused by interrupted RCR resulting in new virion strands frequently being synthesised from two or more separate template molecules – a situation that results in one recombination breakpoint at the point where the interruption occurred and another at the v-ori where the 5' and 3' virion strand ends replicated from different templates get joined following completion of virion strand replication. The hotspot at the transcription termination site is possibly caused by an increased probability of RCR interruption due to complementary sense gene transcription complexes spending a disproportionately large amount of time there.

Whereas all members of most ssDNA virus families replicate via either a rolling circle mechanism (the Nanoviridae, Microviridae and Geminiviridae) or a related rolling hairpin mechanism (the Parvoridae), amongst the circoviruses only the circovirus genus–ie the genus that PCV and BFDV belong to is known to use RCR. Although the Gyrovirus, the other member of circovirus genera, and the Annelovirus, might also use RCR, it currently unknown whether they do. Additionally, some members of the geminivirus genus Begomovirus have either a second genome component, called DNA-B, or are associated with satellite ssDNA molecules called DNA-1 and DNA-Beta, all of which also replicate by rolling circle mechanisms.

Recombination is known to occur in the parvoviruses (Gao et al., 2003; Lukashov and Goudsmit, 2001; Shackelton et al., 2007), microviruses (Rokyta et al., 2006), circoviruses (Csagola et al., 2006; Heath et al., 2004; Olvera et al., 2007), geminivirus DNA-B components (Padidam et al., 1999), DNA-Beta and DNA-1 satellite molecules (Amin et al., 2006) and nanoviruses (Hughes, 2004). Given that most if not all of these ssDNA replicons are evolutionarily related to (Koonin and Ilyina, 1992) and share many biological features with the geminiviruses it would be interesting to determine whether conserved recombination patterns observed in the geminivruses are evident in these other groups. To date no comparative analyses have ever been performed with different ssDNA virus families to determine, for example, possible influences of genome organisation on recombination breakpoint distributions observable in viruses sampled from nature.

In this study we compare recombination breakpoint distributions and patterns of sequence exchange

amongst most currently described ssDNA viruses and satellite molecules. Here we present evidence that recombination features prominently in the evolution of most if not all ssDNA replicons. While we find that detectable recombination events between ssDNA virus species have tended to involve the transfer of entire genes, detectable recombination breakpoints tend to occur within the non-coding regions of ssDNA replicons. The v-*ori* recombination hotspot is a common but not universal feature of ssDNA virus genomes that are replicated by a rolling circle mechanism. Different gene recombinabilities were shown with in particular low and high recombinabilities for structural gene and Rep gene respectively. Those results highlight the differential modularities of the different gene classes.

Materiel and methods

Sequence datasets

All publicly available full-length circovirus, microvirus and parvovirus sequences, annelovirus sequences that were greater than 50% full genome size, nanovirus DNA-1 sequences and geminivirus DNA B and DNA-1 sequences were obtained from public sequence databases using TaxBrowser (http://www.ncbi.nlm.nih.gov/) between October and December 2007. An alignment of geminivirus DNA-Beta sequences was obtained from Robert Briddon (Plant Biotechnology Division, National Institute for Biotechnology and Genetic Engineering, Jhang Road, Faisalabad, Pakistan). Sequence alignments were constructed using POA (Lee et al., 2002) and edited both by eye and using the ClustalW-based (Thompson et al., 1994) sub-sequence realignment tool implemented in MEGA 4 (TAMURA ET AL., 2007). All but one sequence within groups of sequences sharing more than 98% nucleotide identity were discarded. Sequences found to be highly divergent to the rest of the align one were also discarded. Finally, to ensure an efficient detection of recombination, dataset were split in groups of sequences sharing at least 60% identity. Begomovirus and Mastreviruses datasets are the one described in (Lefeuvre et al., 2007a) and (Owor et al., 2007) respectively. Details about all the datasets are given in supplementary table 1.

Detection of individual recombination events

Detection of potential recombinant sequences, identification of likely parental sequences, and localization of possible recombination breakpoints was carried out with the RDP (Martin, 2000), GENECONV (Padidam et al., 1999), BOOTSCAN (Martin, 2005), MAXCHI (Maynard, 1992), CHIMAERA (Martin et al., 2005b), SISCAN (Gibbs, 2000), LARD (Holmes et al., 1999) and 3Seq (Boni et al., 2007) methods implemented in RDP3 (Martin et al., 2005b) (see the RDP project files submitted as supplementary material for full details of program settings). Default settings were used throughout and only potential recombination events detected by two or more of the above methods coupled with phylogenetic evidence of recombination were considered significant.

Recombination breakpoint density plots and "recombinant region count matrices" were constructed using RDP3 as described in Heath et al. (2006) and Lefeuvre et al. (2007a) respectively. The matrices represent the numbers of times recombinational movement of sequence tracts between genomes separates pairs of nucleotide sites. This representation of detectable recombination events highlights the differential "exchangeability" of sequence tracts between genomes. Whereas highly exchangeable genome regions (i.e. those represented by warm colours in the matrices due to their frequent movement into foreign genetic backgrounds) are expected to be most modular, the less exchangeable regions (i.e. represented by cooler colours due to their infrequent movement into foreign genetic backgrounds) are expected to be the least modular.

Recombination hot- and cold-spots were identified from breakpoint distribution plots using the permutation based linear "local" and "global" tests described in Heath et al. (2006). The statistical significance of potential recombination region hot- and cold-spots in recombinant region count matrices was tested using a two dimensional version of the linear local recombination hot/cold-spot permutation test. Briefly this involved: (i) Starting with the first recombination signal, determining the positions of all variable nucleotide positions (VNPs) between the triplet of sequences used to detect the recombination event; (ii) Counting the VNPs between the breakpoints; (iii) randomly placing the 5' breakpoint position along the alignment midway between two VNPs and then placing the 3' breakpoint at a VNP exactly the same number of VNPs away from the randomised 5' breakpoint as the actual 3' breakpoint was from the actual 5' breakpoint; (iv) In cases where sequences were linear (such as in the parvovirus datasets), wherever 5' breakpoints either overlapped the end of the sequence or were within 3 variable nucleotide positions of the end of the sequence, repeating step (iii) until it was located in a suitable

position; (iv) Recording recombination events on a linear map of the alignment; (v) In cases where more than one sequence contained evidence of the same recombination events breakpoints were also recorded on the linear maps representing these sequences making sure to preserve their positions relative to the randomised breakpoint positions; (vi) Wherever the newly mapped breakpoint positions either bounded any previously mapped breakpoint positions then the newly mapped positions were erased and the process repeated from step (iii); (vii) Starting with the next recombination signal identified, determining the positions of all VNPs between the triplet of sequences used to detect the recombination event and repeating the process from steps (ii) through (vi) until all unique events were analysed; (viii) Generating and storing a recombination region count matrix of the permuted dataset; (ix) Repeating the process from steps (i) through (viii) a further 999 times. (x) Ranking the score of each cell in the recombinant region count matrix relative to corresponding cells recorded in the 1000 permuted matrices and recording the cells in the actual matrix were either higher or lower than 95% or 99% of corresponding cells in the permuted matrices.

To specifically test for clustering of recombination breakpoints in different genome regions we also used a modification of the local permutation test decribed in Heath et al. (2006). In this test rather than partitioning the alignment with a moving window of set length the alignment was partitioned in various other ways. So, for example, to test for significant clustering of breakpoints in the intergenic regions, alignments were partitioned into coding and on-coding regions and individually tested to determine whether more/fewer breakpoints were detectable in the intergenic regions than could be accounted for by chance. Other similar tests included (1) discounting breakpoints falling outside coding regions and determining whether individual genes contained significantly more/fewer detectable breakpoints than the remainder of the coding regions, and (2), again discounting breakpoints falling outside coding regions, determining whether the middle 50%, 75% and 87.5% of all genes collectively contained significantly more/fewer detectable breakpoints than the 25%, 12.5% and 6.25% ends of the genes.

Results and Discussion

Previous analyses of geminivirus DNA-A and DNA-A-like sequences have indicated that both the genome regions exchanged between geminiviruses, and the recombination breakpoints occurring within detectable recombinants sampled from nature are not randomly distributed (Lefeuvre et al., 2007a; Lefeuvre et al., 2007b; Owor et al., 2007; Padidam et al., 1999). It is apparent that a combination of varying mechanistic predispositions to recombination in different genome regions and natural selection against defective recombinants has resulted in recombination hot- and cold-spots within geminivirus genomes.

Table 1: recombination analysis on ssDNA viruses dataset

Host	Family	Genus	Dataset name	Nb. of event	Prop<90(<80)
Animal		Anellovirus	Merged	-	-
			TTV	67	99 (87)
			TTMV	21	100 (100)
	Circoviridae	PCV/BF	Merged	-	-
		Bird Circovirus	Merged	-	-
		Circovirus/PCV	PCV	16	0 (0)
		Circovirus/BF	BF	17	45 (18)
		Circovirus/Goose	Goose	1	0 (0)
		Gyrvovirus	Gyro	1	0 (0)
	Parvoviridae	Dependo/Parvo	Merged	-	-
		Dependovirus	Dependo	17	78 (33)
		Erythrovirus	Erythro	2	100 (0)
		Parvovirus	Parvo	13	100 (33)
Bacteria	Microviridae	Microvirus	Micro	38	45 (27)
Plant	Geminiviridae	Mastrevirus	Mastre	37	66 (11)
		Begomovirus	Begomo	245	100 (84)
		DNA B	Merged	-	-
			Indian ToLCV	17	100 (11)
			EACMV	10	14 (14)
			Asia	9	71 (57)
			New world	41	100 (100)
		DNA Beta	Merged	-	-
			betaA	25	100 (90)
			betaC	20	80 (50)
			betaD	11	100 (67)
		DNA 1	dna1	38	100 (75)
	Nanoviridae	Babuvirus	Merged babu nano	-	-
			other	4	100 (100)
			other	3	-
			CP	1	-
			Rep	0	-
			other	2	100 (0)
		Nanovirus	nano	7	100 (100)

Distribution of discrete recombination events in natural virus populations

To determine the net effect of natural selection and varying mechanistic predispositions to recombination on the distribution of individual recombination events across ssDNA virus genomes, we analysed 24 datasets using a battery of recombination detection methods and rigorous manual and automated evaluation of recombination signals. These datasets included: Four geminivirus DNA-B, one geminivirus associated DNA-1, three geminivirus associated DNA-Beta and one DNA-1, five babuvirus, one nanovirus, two annelovirus, three circovirus, one gyrovirus, one microvirus, and 3 parvoviridae datasets.

We identified and characterised 381 individual recombination events in these 24 datasets all of which

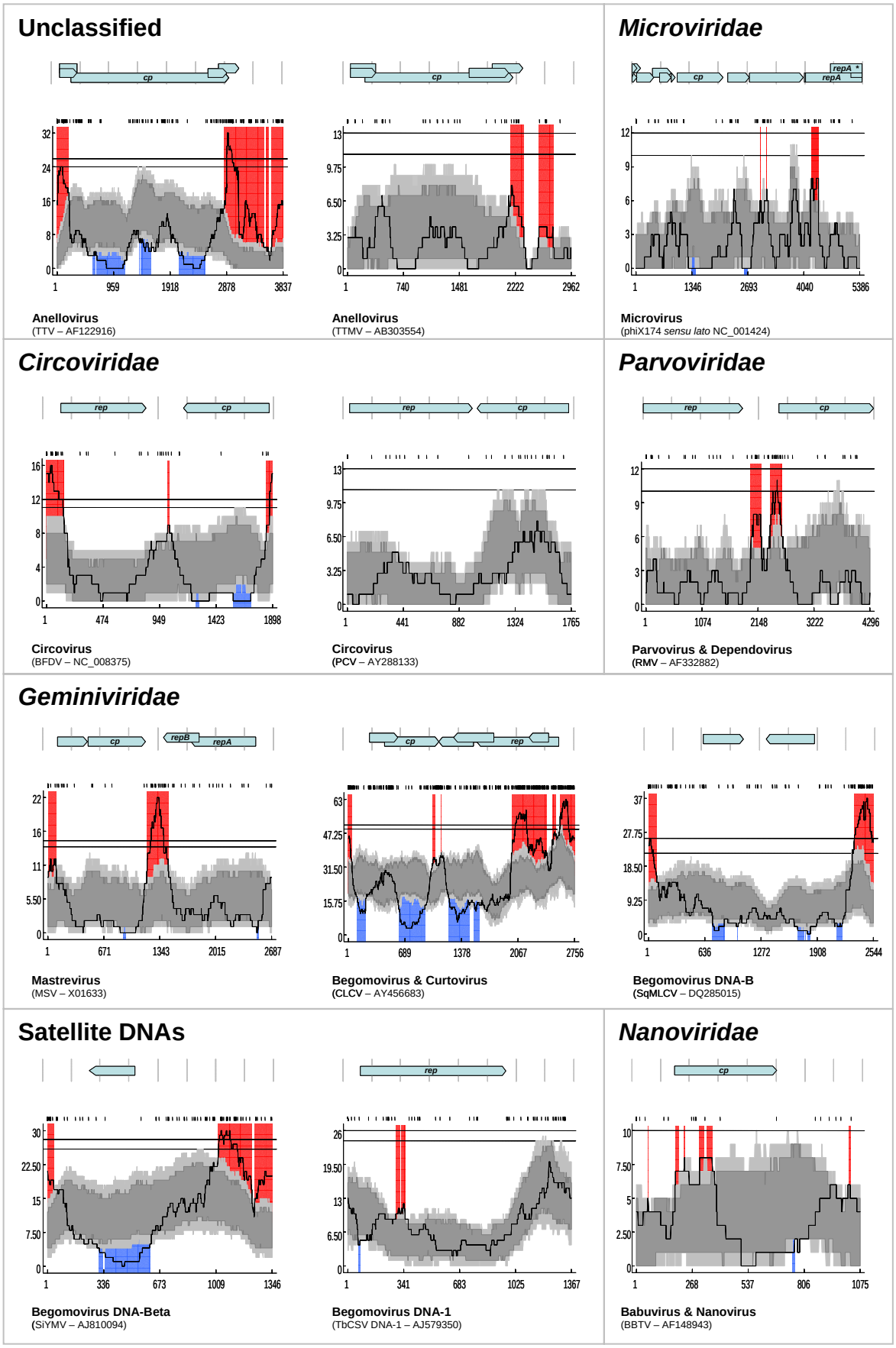


Figure 1: The distribution of recombination breakpoints detected within the different ssDNA virus datasets. All detectable breakpoint positions are indicated by small vertical lines at the top of the graph. A 200 nucleotide window was moved along the alignment one nucleotide at a time and the number of breakpoints detected within the window region was counted and plotted (solid line). The upper and lower broken lines respectively indicate 99% and 95% confidence thresholds for globally significant breakpoint clusters. Light and dark grey areas respectively indicate local 99% and 95% breakpoint clustering thresholds taking into account local regional differences in sequence diversity that influence the ability of different recombination detection methods to identify recombination breakpoints. Red areas indicate recombination hot spots while blue areas represent recombination cold spots. ORFs (horizontal arrows) are represented on the top of the graph.

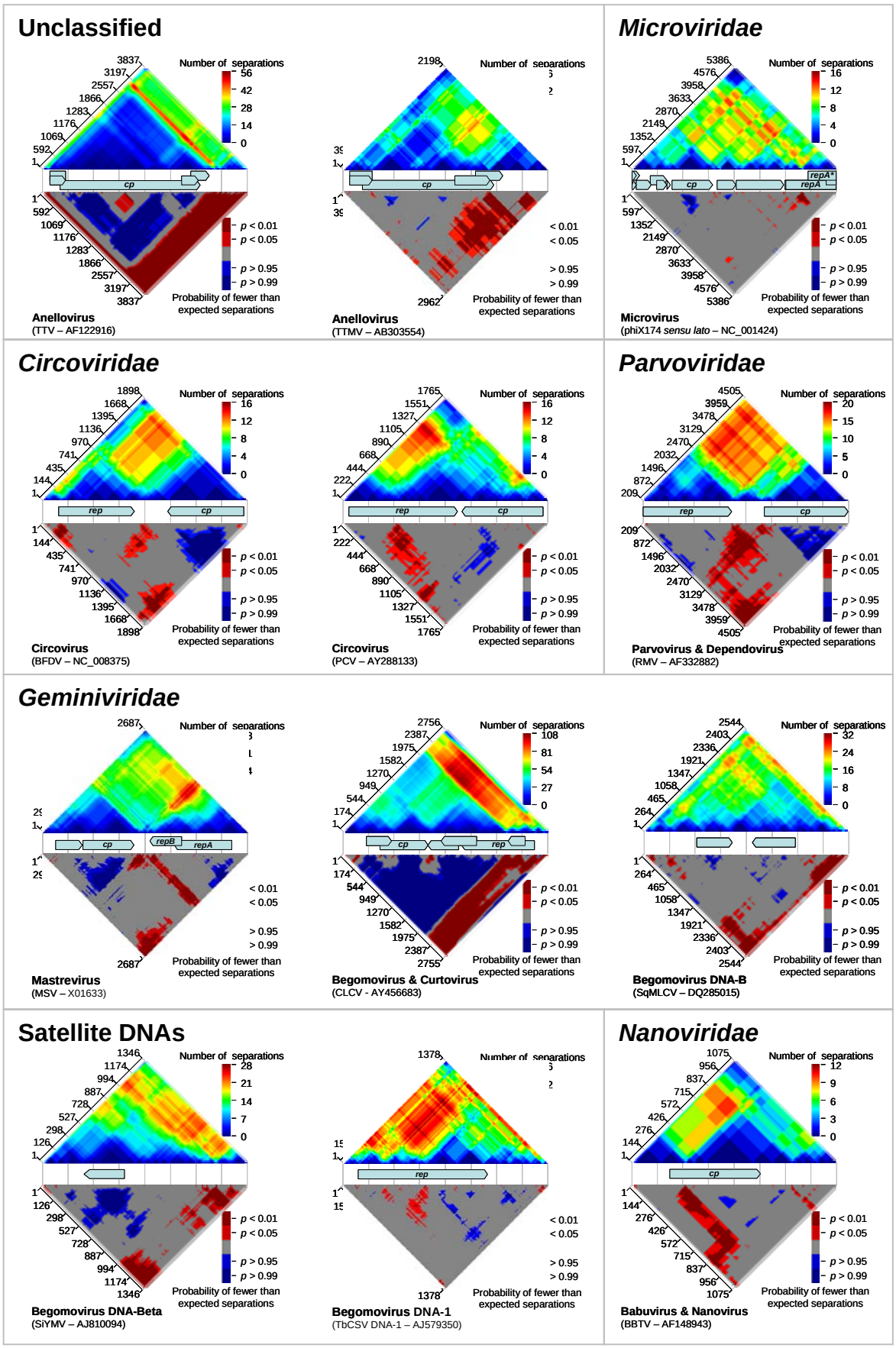


Figure 2: Recombination region count matrix (upper hemi-matrices) and recombination hot/cold spot matrix (lower hemi-matrices) of the different ssDNA datasets. Unique recombination events have been mapped onto the matrix based on their estimated breakpoint positions. In the upper matrices, the shades displayed are a function of the number of times pairs of nucleotides (plotted on the x and y axis) are separated during the observed set of unique recombination events. In the lower matrices, color represents recombination hot and cold-spots for sequences tracks. Horizontal arrows indicate ORF.

Table 2: Probabilities of breakpoints occuring in the different coding/non coding region

Host	Family	Genus	Dataset name	Coding vs Noncoding	Middle 50% vs end 50% of genes	CP vs rest of genes
Animal	Unasigned	Anellovirus	TTV	**<0.0001**	0.356	**0.0002**
			TTMV	**0.001**	**0.0082**	0.9243
	Circoviridae	Circovirus	PCV	1	0.878	0.7366
		Circovirus	BFDV	**0.0011**	**0.0029**	**0.0013**
	Parvoviridae	Parvo + Dependovirus	Merged	**<0.0001**	**0.0499**	0.977
		Dependovirus	Dependo	1	0.0988	0.826
		Parvovirus	Parvo	**0.0197**	0.2001	**0.0932**
Bacteria	Microviridae	Microvirus	Micro	1	**0.0104**	**0.0047**
Plant	Geminiviridae	Mastrevirus	African streak virus	**<0.0001**	**0.0286**	**0.0133**
		Begomo + Curtovirus	DNA-A	0.084	**<0.0001**	**0.0172**
		Begomovirus	DNA-B	**<0.0001**	**0.0314**	-
		Satelite	Beta	**0.0001**	**0.0016**	-
		Satelite	DNA-1	0.5669	0.7779	-
	Nanoviridae	Nano + Babuvirus	CP component	0.4827	**0.001**	-

are summarised in table 1 (each of these events is detailed in the supplementary RDP3 project and spreadsheet files). Most of the sequences in all of the datasets were recombinant, with for example 38 events detectable in 37 microvirus sequences and 38 events detectable in 35 begomovirus satellites DNA 1 sequences. Between 0 (PCV, Goose circovirus and Gyrovirus datasets) and 100% (TTMV, Erythrovirus, Parvovirus, Begomovirus, Nanovirus and Babuvirus datasets) of the detected recombination events involved sequence exchanges between parental viruses sharing less than 90% sequence identity. Although for some groups such as the gyroviruses, the goose circoviruses and the erythroviruses, only few recombination events were detected, this is probably just a reflection of both the low diversity and small size of these datasets (see supplementary table 1).

For datasets where more than 5 individual recombination events were detectable we tested for the presence of both recombination breakpoint and recombinant region hot- and cold-spots. Whereas breakpoint distribution maps were generated and tested (Figure 1), the tracts of sequences exchanged during events were mapped onto “recombinant region count matrices” (Figure 2). These matrices describe the relative frequencies with which different parts of the analysed ssDNA replicons are separated during recombination. Recombinant region hot- and cold-spots, indicating genome regions that are respectively more or less frequently exchanged during recombination than can be accounted for by chance (with the null hypothesis that tracts of sequence get randomly exchanged). Importantly the permutation tests used for both the recombination breakpoint distribution plots and the recombinant region count matrices take into account the fact that, due to variations in diversity across the lengths of analysed nucleotide sequence alignments, recombination is invariably easier to detect in certain genome regions than it is in others.

Importantly, recombination hotspots are detectable at the origins of virion strand replication (Figure 1 and 2, position zero of the different genome representations) in TTV, BFDV, Mastrevirus, Begomovirus, DNA A, DNA B and satellite genomes all known to replicate via a rolling circle mechanism. The main exceptions were the microvirus, and geminivirus DNA-1 datasets. Although technically no v-*ori* hotspot was detected for the nanovirus datasets either, the number of detectable recombination breakpoints in these three datasets was very low and as a result the hotspot test lacked sufficient power. Also, whereas in the BFDV circovirus dataset a clear recombination hotspot is detectable at the v-ori, none is detectable in the PCV circovirus dataset.

Interruption of replication by polymerase uncoupling will result in a partially replicate virion strand with one end corresponding to v-ori. If one imagine that recombination dependent replication could occur in all those viruses as what was shown for begomovius (Preiss & Jeske, 2003), completion of replication from

partially replicated virion strands and in the presence of potential template DNAs from different sequences could result in detectable recombination events. Those events would then display breakpoint at the point where replication was initially disrupted and the other at the virion sense ori where replication was started. The conservation of v-ori hotspot in some viral genus (anellovirus, gyrovirus, mastrevirus, begomovirus and satellites) and the absence in others (nanovirus, parvovirus, dependovirus, circovirus) is controversial. Factors underlying those differences are not known, possible variations in aptitude to replicate through recombination-dependent replication could be one explanation.

Another potentially conserved recombination pattern related to genome organisation is observed in the multipartite geminiviruses (begomovirus DNA-A and DNA-B components) and nanoviruses (Babuvirus DNA-1 and DNA-2 and Nanovirus DNA-1). Although too few nanovirus sequences are currently available to effectively detect recombination hot and cold-spots, recombination breakpoints tend to occur most frequently within the segment of the intergenic region that is conserved amongst different nanovirus genome components. Likewise, in the geminiviruses a similar pattern is observed either in DNA A or DNA B (Figure 1 and 2). Every component of those viruses shares a highly conserved common region (Figure 1 and 2, correspond mainly to the intergenic region). As can be seen either on recombination density plot or matrices (Figure 1 and 2), hotspot of recombination is found over CR in nanovirus and begomovirus DNA B. Those results support the hypothesis that repeated events of genetic exchange have occurred between non-coding regions of genome components encoding non-homologous proteins, as it was previously suspected for nanovirus component (Hughes, 2004).
While recombination should be the reason of the CR to be "common", it's also possible that certain of the regions that show evidence of genetic exchange between components are conserved because they play important roles in the biology of these viruses and thus are subject to functional constraint.
Nevertheless, the phenomenon of concerted evolution, whereby recombination serves to homogenize the members of a multi-gene family over evolutionary time (Arnheim et al., 1980), should be of great importance in multipartite viruses family. A constant update of CR from one component to another would ensure the co-evolution of each element for multipartite viruses.

We could find no obviously conserved recombination patterns between either the parvoviruses or microviruses and the other smaller ssDNA virus replicons. With the exception of a small breakpoint hotspot at the abutting region of NS (Rep) and VP (CP) ORF, we detected no conserved recombination breakpoint hot/cold-spots between the two analysed parvovirus datasets. It should, however, be noted that given the large size of these genomes and the relatively small numbers of analysed recombination breakpoints, our analysis lacked sufficient power to detect any but the most obvious recombination hot- and cold-spots. With bigger datasets and additional detectable recombination events, it is possible that evidence of conserved recombination breakpoint patterns will emerge.

Recombination matrices indicated three potential recombination patterns shared by many of the analysed viruses: (i) That intergenic regions tended to be moved between genomes more frequently than individual genes (note red/orange/light green diagonals in the upper matrices and red diagonals in the lower matrices originating on intergenic regions in the anellovirus, geminivirus, DNA-Beta and nanovirus datsets); (ii) That certain genes, particularly those encoding coat proteins, tended to be moved by recombination either as complete or mostly complete (>50% of the middle regions) units (Note light/dark blue triangles in the upper matrices and blue patches in the lower matrices associated in particular with coat protein genes in the anellovirus, microvirus, circovirus-BFDV, parvovirus and geminivirus datasets); (iii) That coat protein genes tended to contain fewer recombination breakpoints than other genes.

Intragenic recombination is negatively selected in natural virus population

It was immediately noticeable when comparing breakpoint patterns in different groups of ssDNA replicons, that whereas recombination hotspots tended to occur within intergenic regions (52% of 27 detected hotspots), recombination cold-spots tended to occur within coding regions (41% of 17 detected coldspots).

It was also quite clearly evident from visual inspection of the recombinant region count matrices that recombination tended to transfer complete, or almost complete genes between genomes. Exceptions to this pattern were for circovirus – PCV Rep,

begomovirus DNA-A Rep and Microviridae, in which breakpoints into ORF, even if it's not a rule are frequently observed. Whereas for the Microviridae, the small size of the non-coding regions might explain this tendency, in the PCV and Begomovirus datasets the recombination hotspot in the Rep gene may indicate that the replication associated proteins encoded by these genes are particularly tolerant of recombination

Those hypothesis (i) that recombination breakpoints tend to occur within intergenic regions (ii) that breakpoints that do occur within genes tend to occur towards the end of genes (iii) That breakpoints tend to not occur in CP genes and (iv) that breakpoints tend to occur in Rep more than in other genes were statistically tested.
The results, summarised in table 2 and supplementary table 2, indicate for all but three of the analysed datasets (Circovirus PCV, Dependovirus and Satellites DNA-1) there is indeed a significant statistical tendency for breakpoints to occur within IR sequences or at least that when breakpoint occur, they tend to map at the edges and not the central part of the genes (begomovirus DNA A and DNA A-like, microvirus and nanovirus capsid component). These findings are consistent with the hypothesis that breakpoints are less tolerable when they occur within genes (Bonnet et al., 2005; Lefeuvre et al., 2007a), possibly because there is a relatively high probability that they will disrupt proper protein folding (see Carbone and Arnold, 2007 for review).

An important result is also the tendency of breakpoint to avoid CP gene in comparison to the other ORF in five of the nine datasets analysable this way (begomovirus, circovirus except PCV dataset, microvirus, mastrevirus and TTV datasets). The Rep ORF was found to be more recombinagenic then the other ORF in three of nine datasets (begomovirus, mastrevirus and the other circovirus then PCV). While those results highlight the previous observations of recombination count matrices, they clearly show again the existence of variations in recombinability of the different genes families (Jain et al., 1999; Martin et al., 2005a). The tendency of capsid protein to avoid recombination is probably explain by the oligomerisation role of this protein. At contrario, the rep protein seems to be more recombinagenic then any other genes. Possible adaptation of rep to host factor was proposed (Prasanna and Rai, 2007) and should be involved in this "forced" adaptability.

While a great number of recombination events were detected in publicly available sequence database for most of the circular ssDNA replicons, it seems that, as has been reported previously for the geminiviruses, recombination patterns are most likely determined by the interplay between selective and mechanistic processes. We have, however, shown that many of these undrelying mechanistic and selective processes are probably quite widely conserved amongst the ssDNA replicons. While it is very probably that the v-ori or rolling circle replicons is a mechanistic recombination hotspot in the circoviruses, and geminivirus DNA-A (or DNA-A-like), DNA-B and DNA-beta genome constituents, it is likely that both geminivirus and circovirus genes transcribed in the opposite direction to RCR experience increased recombination rates. If they are to survive, newly produced recombinants must be able to productively compete with their parents. Our results suggest that amongst the ssDNA replicons we have analysed natural selection in general tends to: (i) Penalise breakpoints within coding regions more harshly than it does breakpoints in intergenic regions (ii) favour recombinants with breakpoints on the edges of genes more than recombinants with breakpoints within the centres of genes, (iii) disfavour recombination events with breakpoints within CP genes and (iv) either favour or is neutral with respect to recombination breakpoints that occur in the Rep genes.

The notion that all ssDNA replicons might be applying the same basic evolutionary strategy is further supported by the recent finding that circoviruses, parvoviruses and geminiviruses (and probably other ssDNA viruses and other replicons too) are unusual amongst DNA viruses in that they most likely experience nucleotide substitution rates that are as rapid as those of some RNA viruses (Duffy et al., 2008). Although high mechanistic mutation and recombination rates do not necessarily translate into rapid evolution rates, both our recombination results and estimates of mutation driven evolution rates in ssDNA viruses (ie the rates at which mutations become fixed within populations) emphasise the huge evolutionary potential of ssDNA replicons. Whereas in the past this potential most likely enabled adaptation of these evolutionarily related viruses to bacterial, plant and animal hosts, in recent times it has been most manifest in the emergence of various plant and animal pathogens (Carman et al., 2008; Duffy and

Holmes, 2008; Shackelton et al., 2005; Zhou et al., 1997).

Acknowledgments

P.L. is supported by the French Ministère de la Recherche et de l'Enseignement Supérieur. J.M.L. is supported by CIRAD. A.V. is supported by the Carnegie Corporation of New York; D.P.M. is supported by the Wellcome trust.

Reference

Amin, I., Mansoor, S., Amrao, L., Hussain, M., Irum, S., Zafar, Y., Bull, S.E. and Briddon, R.W., 2006. Mobilisation into cotton and spread of a recombinant cotton leaf curl disease satellite. Arch Virol 151, 2055-65.

Arnheim, N., Krystal, M., Schmickel, R., Wilson, G., Ryder, O. and Zimmer, E., 1980. Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. Proc Natl Acad Sci U S A 77, 7323-7.

Baird, H.A., Galetto, R., Gao, Y., Simon-Loriere, E., Abreha, M., Archer, J., Fan, J., Robertson, D.L., Arts, E.J. and Negroni, M., 2006. Sequence determinants of breakpoint location during HIV-1 intersubtype recombination. Nucleic Acids Res 34, 5203-16.

Boni, M.F., Posada, D. and Feldman, M.W., 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics.

Bonnet, J., Fraile, A., Sacristan, S., Malpica, J.M. and Garcia-Arenal, F., 2005. Role of recombination in the evolution of natural populations of Cucumber mosaic virus, a tripartite RNA plant virus. Virology 332, 359-68.

Carbone, M.N. and Arnold, F.H., 2007. Engineering by homologous recombination: exploring sequence and function within a conserved fold. Curr Opin Struct Biol 17, 454-9.

Carman, S., Cai, H.Y., DeLay, J., Youssef, S.A., McEwen, B.J., Gagnon, C.A., Tremblay, D., Hazlett, M., Lusis, P., Fairles, J., Alexander, H.S. and van Dreumel, T., 2008. The emergence of a new strain of porcine circovirus-2 in Ontario and Quebec swine and its association with severe porcine circovirus associated disease--2004-2006. Can J Vet Res 72, 259-68.

Chare, E.R., Gould, E.A. and Holmes, E.C., 2003. Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. J Gen Virol 84, 2691-703.

Crameri, A., Raillard, S.A., Bermudez, E. and Stemmer, W.P., 1998. DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature 391, 288-91.

Csagola, A., Kecskemeti, S., Kardos, G., Kiss, I. and Tuboly, T., 2006. Genetic characterization of type 2 porcine circoviruses detected in Hungarian wild boars. Arch Virol 151, 495-507.

Duffy, S. and Holmes, E.C., 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. J Virol 82, 957-65.

Duffy, S., Shackelton, L.A. and Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9, 267-76.

Escriu, F., Fraile, A. and Garcia-Arenal, F., 2007. Constraints to genetic exchange support gene coadaptation in a tripartite RNA virus. PLoS Pathog 3, e8.

Gao, G.P., Alvira, M.R., Somanathan, S., Lu, Y., Vandenberghe, L.H., Rux, J.J., Calcedo, R., Sanmiguel, J., Abbas, Z. and Wilson, J.M., 2003. Adeno-associated viruses undergo substantial evolution in primates during natural infections. Proceedings of the National Academy of Sciences of the United States of America 100, 6081-6086.

Garcia-Andres, S., Tomas, D.M., Sanchez-Campos, S., Navas-Castillo, J. and Moriones, E., 2007. Frequent occurrence of recombinants in mixed infections of tomato yellow leaf curl disease-associated begomoviruses. Virology.

Gibbs, M.J., Armstrong, J.S. & Gibbs, A.J., 2000. Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16, 573-582.

Gutierrez, C., 1999. Geminivirus DNA replication. Cell Mol Life Sci 56, 313-29.

Heath, L., Martin, D.P., Warburton, L., Perrin, M., Horsfield, W., Kingsley, C., Rybicki, E.P. and Williamson, A.L., 2004. Evidence of unique genotypes of beak and feather disease virus in southern Africa. J Virol 78, 9277-84.

Heath, L., van der Walt, E., Varsani, A. and Martin, D.P., 2006. Recombination patterns in aphthoviruses mirror those found in other picornaviruses. J Virol 80, 11827-32.

Holmes, E.C., Worobey, M. and Rambaut, A., 1999. Phylogenetic evidence for recombination in dengue virus. Mol Biol Evol 16, 405-9.

Hughes, A.L., 2004. Birth-and-death evolution of protein-coding regions and concerted evolution of non-coding regions in the multi-component genomes of nanoviruses. Mol Phylogenet Evol 30, 287-94.

Jain, R., Rivera, M.C. and Lake, J.A., 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 96, 3801-6.

Jeske, H., Lutgemeier, M. and Preiss, W., 2001. DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. EMBO J 20, 6158-67.

Keightley, P.D. and Otto, S.P., 2006. Interference among deleterious mutations favours sex and recombination in finite populations. Nature 443, 89-92.

Koonin, E.V. and Ilyina, T.V., 1992. Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. J Gen Virol 73 (Pt 10), 2763-6.

Lee, C., Grasso, C. and Sharlow, M.F., 2002. Multiple sequence alignment using partial order graphs. Bioinformatics 18, 452-64.

Lefeuvre, P., Lett, J.M., Reynaud, B. and Martin, D.P., 2007a. Avoidance of protein fold disruption in natural virus recombinants. PLoS Pathog 3, e181.

Lefeuvre, P., Martin, D.P., Hoareau, M., Naze, F., Delatte, H., Thierry, M., Varsani, A., Becker, N., Reynaud, B. and Lett, J.M., 2007b. Begomovirus 'melting pot' in the south-west Indian Ocean islands: molecular diversity and evolution through recombination. J Gen Virol 88, 3458-68.

Lukashov, V.V. and Goudsmit, J., 2001. Evolutionary relationships among parvoviruses: virus-host coevolution among autonomous primate parvoviruses and links between adeno-associated and avian parvoviruses. J Virol 75, 2729-40.

Magiorkinis, G., Ntziora, F., Paraskevis, D., Magiorkinis, E. and Hatzakis, A., 2007. Analysing the evolutionary history of HCV: Puzzle of ancient phylogenetic discordance. Infect Genet Evol 7, 354-60.

Martin, D., & Rybicki, E., 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16, 562-563.

Martin, D.P., Posada, D, Crandall, K. A., & Williamson, C., 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS and Human Retroviruses In Press.

Martin, D.P., van der Walt, E., Posada, D. and Rybicki, E.P., 2005a. The evolutionary value of recombination is constrained by genome modularity. PLoS Genet 1, e51.

Martin, D.P., Williamson, C. and Posada, D., 2005b. RDP2: recombination detection and analysis from sequence alignments. Bioinformatics 21, 260-2.

Martin, G., Otto, S.P. and Lenormand, T., 2006. Selection for recombination in structured populations. Genetics 172, 593-609.

Maynard, S.J., 1992. Analyzing the mosaic structure of genes. J. Mol Evol. 34, 126-129.

Olvera, A., Cortey, M. and Segales, J., 2007. Molecular evolution of porcine circovirus type 2 genomes: phylogeny and clonality. Virology 357, 175-85.

Owor, B.E., Martin, D.P., Shepherd, D.N., Edema, R., Monjane, A.L., Rybicki, E.P., Thomson, J.A. and Varsani, A., 2007. Genetic analysis of maize streak virus isolates from Uganda reveals widespread distribution of a recombinant variant. J Gen Virol 88, 3154-65.

Padidam, M., Sawyer, S. and Fauquet, C.M., 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265, 218-25.

Prasanna, H.C. and Rai, M., 2007. Detection and frequency of recombination in tomato-infecting begomoviruses of South and Southeast Asia. Virol J 4, 111.

Rokyta, D.R., Burch, C.L., Caudle, S.B. and Wichman, H.A., 2006. Horizontal gene transfer and the evolution of microvirid coliphage genomes. J Bacteriol 188, 1134-42.

Shackelton, L.A., Hoelzer, K., Parrish, C.R. and Holmes, E.C., 2007. Comparative analysis reveals frequent recombination in the parvoviruses. J Gen Virol 88, 3294-301.

Shackelton, L.A., Parrish, C.R., Truyen, U. and Holmes, E.C., 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc Natl Acad Sci U S A 102, 379-84.

Tamura, K., Dudley, J., Nei, M. and Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24, 1596-9.

Thompson, J.D., Higgins, D.G. and Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22, 4673-80.

van der Walt, E., Palmer, K.E., Martin, D.P. and Rybicki, E.P., 2008. Viable chimaeric viruses confirm the biological importance of sequence specific maize streak virus movement protein and coat protein interactions. Virol J 5, 61.

Varsani, A., van der Walt, E., Heath, L., Rybicki, E.P., Williamson, A.L. and Martin, D.P., 2006. Evidence of ancient papillomavirus recombination. J Gen Virol 87, 2527-31.

Varsani, A., Shepherd, D.N., Monjane, A.L., Owor, B.E., Erdmann, J.B., Rybicki, E.P., Peterschmitt, M., Briddon, R.W., Markham, P.G., Oluwafemi, S., Windram, O.P., Lefeuvre, P., Lett, J.M., and Martin, D.P., 2008. Recombination, decreased host specificity and increased mobility may have driven the emergence of maize streak virus as an agricultural pathogen. J Gen Virol 89, 2063 – 2074.

Zhou, X., Liu, Y., Calvert, L., Munoz, C. and Otim-Nape, G.W., 1997. Evidence that DNA-A of a geminivirus associated with severe cassava mosaic disease in Uganda has arisen has arised by interspecific recombination. J Gen Virol 78, 2101-2111.

Discussion Générale

Les bégomovirus des îles du Sud Ouest de l'océan Indien et de l'Afrique : chroniques d'une diversité encore cachée

L'obtention de nouvelles séquences complètes de bégomovirus provenant de la zone du Sud Ouest de l'océan Indien (SWIO) a permis de décrire une diversité virale très importante, mais probablement encore largement sous estimée. La mise en relation du nombre de nouvelles espèces définies (10) et du nombre de génomes complets séquencés (18, Delatte et al., 2005b, Chapitre II) suggère que seuls les contours de la diversité en bégomovirus ont été définis pour le moment dans les îles SWIO, d'autant plus que cette diversité n'a été décrite qu'à partir de plantes cultivées originaires du Nouveau Monde (Delatte et al., 2005c; Chapitre II). Cette sous-estimation de la diversité réelle des bégomovirus semble être encore plus importante sur le continent Africain où les nombreuses études réalisées jusqu'à présent se sont principalement penchées sur les virus infectant le manioc, et très peu sur les plantes légumières (Chen et al., 2007; Idris and Brown, 2005; Zhou et al., 2008).

Toutefois, notre analyse phylogénétique (Chapitre II) a montré que les virus des îles SWIO étaient apparentés aux virus monopartites et bipartites africains, exception faite du clade qui regroupe les TYLCVs et ToLCVs méditerranéens. La majorité des espèces virales SWIO sont associées au sein d'un même groupe phylogénétique dans lequel se trouvent également des virus bipartites. Cette proximité phylogénétique des virus monopartites et bipartites suggère une origine commune et un isolement ancien des virus Africains par rapport aux autres groupes de bégomovirus. La question de la perte ou de l'acquisition d'un ADN B par les virus monopartites ou bipartites respectivement reste à être élucidée. L'éloignement de certains virus SWIO du groupe de virus Africain est aussi intéressant. En effet, la très grande majorité des bégomovirus décrits actuellement en Afrique l'ont été à partir de plantes hôtes cultivées originaires du Nouveau Monde (tomate, haricot, tabac, piment, poivron, manioc et coton). Savoir si les phylogroupes que nous avons mis en évidence sont la résultante d'une forte adaptation à l'hôte, d'évènements d'émergence distincts entre plantes hôtes originelles et plantes cultivées ou d'une origine géographique différente reste une énigme.

L'ensemble de nos interrogations confirme la nécessité de poursuivre la caractérisation de la diversité génétique des bégomovirus des îles SWIO et de l'Afrique avec un échantillonnage plus important. Cette description exhaustive devrait

permettre d'avancer de nouvelles hypothèses et de mieux comprendre l'origine et la dissémination locale et mondiale de ces virus. L'analyse de la diversité génétique au niveau populationnel entre les différentes îles volcaniques d'âges connues et de l'île continent Madagascar devrait permettre d'étudier les flux migratoires et l'importance des effets de fondation et de dérive génétique lors de la colonisation virale de ces îles.

Quelles origines à ces virus ?

Comme souligné plus haut, l'immense majorité des séquences de bégomovirus déterminée en Afrique ainsi que dans le monde en général provient de plantes cultivées, souvent apportées par l'agriculture moderne. La distinction entre virus du Nouveau Monde (zone d'origine de ces hôtes) et de l'Ancien Monde, ainsi que l'association forte entre des virus isolés à partir d'hôtes différents en Afrique (exemple ACMV et ToLCVs) suggère non pas l'importation de plantes hôtes virosées, mais plutôt des phénomènes de saut d'espèce avec l'adaptation de virus préalablement présents dans la zone d'introduction à de nouvelles plantes hôtes. La réalisation d'une étude globale de la diversité génétique avec un échantillonnage plus important des bégomovirus et mastrévirus en Afrique et dans les îles de l'océan Indien représente une opportunité majeure de (1) réaliser des études comparatives d'émergence indépendante de plusieurs lignées africaines au niveau du genre et de l'espèce et (2) mieux comprendre à l'échelle continentale la dynamique d'adaptation et d'expansion des virus émergents, comme d'importantes pestes agricoles.

Quelques études récentes se sont penchées sur cette question de la diversité moléculaire au sein des hôtes non cultivés (Hernandez-Zepeda et al., 2007; Varsani et al., 2008; Shepherd et al., 2008b). Les travaux de Varsani et al. (2008) ont notamment montré une association phylogénétique forte entre les mastrévirus des poacées indigènes d'Afrique et les mastrévirus du maïs ou de la canne à sucre, suggérant une origine commune. De nouveaux isolats de *Sugarcane streak virus* (SSV) ont aussi été découverts sur des poacées non cultivées africaines, indiquant que ces plantes sont probablement les hôtes naturels des SSV africains isolés depuis sur la canne à sucre. L'analyse de la diversité actuelle des mastrévirus africains a d'ores et déjà permis d'identifier une importante diversité génétique inter et intra espèces avec des isolats de *Sugarcane streak Reunion virus* (SSRV) au Zanzibar et de SSV à La Réunion (Shepherd et al., 2008b). De manière intéressante, ces nouveaux isolats de SSV ont été isolés à partir de quatre genres différents de poacées indiquant qu'ils ont très probablement une large gamme d'hôtes à la fois chez les poacées pérennes ou annuelles d'Afrique.

Dissémination des complexes viraux

Chaque groupe de *Geminivirus* infectant les plantes cultivées a ainsi sa spécificité et la connaissance de la phylogéographie de ces virus et des dynamiques d'échanges entre plantes hôtes cultivées et / ou non cultivées devraient permettre d'identifier et de caractériser les paramètres épidémiologiques impliqués dans les pandémies et ainsi de mieux les contrôler. L'absence de bégomovirus indigènes et la présence d'une grande diversité de mastrévirus dans l'archipel des Mascareignes (Varsani et al, 2008) reste une énigme et souligne bien la spécificité de dissémination de chacun de ces genres viraux. Pour les plantes comme la canne à sucre, disséminée par l'intermédiaire de boutures, la propagation à courte et longue distance des virus semble être principalement associée à la distribution du matériel végétal contaminé. Une dissémination via la canne à sucre a ainsi été proposée pour le MSV qui a été récemment associée à de nouvelles épidémies sur canne à sucre dans le sud de l'Afrique (Shepherd et al., 2008b). Comme les infections sont généralement accompagnées de faibles symptômes et que la canne à sucre est propagée de manière végétative, il n'est pas impossible que du matériel infecté par du MSV ait occasionnellement été transporté dans différentes régions du continent. Malgré tout, les analyses phylogénétiques montrent une diversification des virus du MSV en fonction de la distance géographique. Ce résultat suggère que la dissémination du MSV par le transport de matériel contaminé ne semble pas associée à des mouvements à longues distances. Cet exemple souligne toutefois la nécessité de la connaissance de la diversité virale et de la gamme d'hôtes de ces mêmes virus pour la compréhension des épidémies.

La recombinaison chez les géminivirus

Si la contribution de la recombinaison à l'émergence est difficile à mesurer directement, conceptuellement, il est clair que ce phénomène doit permettre au virus une adaptation rapide, que ce soit par augmentation de fitness ou adaptation à de nouveaux environnements et de nouveaux hôtes. L'avantage évolutif lié à la recombinaison de part la possibilité de créer de nouveaux arrangements à partir de la diversité préexistante semble être directement associée à la capacité d'évolution et au caractère émergent des bégomovirus. Bien que ce phénomène soit très largement décrit chez ce genre viral, seules les premières pièces du puzzle sont assemblées concernant la compréhension des facteurs qui contraignent et façonnent la distribution des recombinaisons au sein d'un génome.

En se basant sur les séquences disponibles dans les bases de données et sur les connaissances du cycle viral des bégomovirus dans la plante (Jeske et al., 2001), il a été (1) montré que la distribution des évènements de recombinaison au sein des génômes de ces virus était non aléatoire et (2) proposé que le mécanisme principal impliqué dans la recombinaison soit lié au conflit entre le complexe de transcription des ORFs du brin complémentaire et le complexe de réplication. La transcription des ORFs du brin complémentaire dans le sens inverse de celui de la réplication par cercle roulant entraînerait une augmentation du taux de recombinaison par un déplacement du complexe de réplication (Jeske et al., 2001). L'achèvement de la réplication à partir de génomes partiellement répliqués se réaliserait par la stratégie de réplication dépendante de la recombinaison, qui en présence de matrices d'ADN de différentes origines devrait entraîner une augmentation du nombre d'évènements de recombinaison détectables le long des ORFs de type complémentaire (Chapitre II et III).

Différentes études avaient montré que la viabilité des virus recombinants et l'activité des protéines chimères sont fortement influencées à la fois par la similarité nucléotidique des parents et la "modularité" des fragments de séquences échangées (Martin et al., 2005; Voigt et al., 2002). Ainsi, l'association de fragments de séquences nucléotidiques, qui ne présentent pas d'interaction l'un avec l'autre, génère des unions compatibles, mêmes si ils sont transférés dans des fonds génétiques très divergents. Au contraire, des fragments de séquences interagissant fortement les uns avec les autres ont tendance a n'être fonctionnel que si ils sont transférés dans un environnement génétique peu différent de celui dans lequel ils ont évolué (Martin et al., 2005; Escriu et al., 2007). Il est dès lors probable que les recombinants naturels, comme ceux responsables d'épidémies (Owor et al., 2007) ou ceux présentant une nouvelle gamme d'hôtes ou un nouvel avantage sélectif (Bonnet et al., 2005; Zhou et al., 1997) ou tout simplement ceux qui font partie des virus circulants dans la population (Chare and Holmes, 2006; Gibbs and Weiller, 1999; Rousseau et al., 2007; Russell and Webster, 2005), représentent la part exceptionnelle, raisonnablement adaptée, d'une frange beaucoup plus large de variants délétères rapidement éliminés par la sélection purificatrice.

Les résultats du Chapitre III indiquent que les interactions responsables d'un repliement correct des protéines sont préservées par recombinaison, avec en parallèle de fortes contraintes sélectives agissant contre les réarrangements au sein desquels ces interactions ne sont pas préservées. La sélection purificatrice serait un facteur majeur façonnant, en partie au moins, les profils de génotypes recombinants

observés pour les populations naturelles de bégomovirus. Le maintien des réseaux d'interactions apparaît être un thème commun unifiant les études visant à définir les limites du potentiel de recombinaison comme stratégie d'exploration de la diversité. Le nombre, la complexité et le maintien des interactions sont des caractéristiques majeures des systèmes vivants. Par conséquent, il apparaît que pour assurer la viabilité d'une descendance recombinante, il ne faut pas que les réseaux d'interactions soient endommagés. Ce modèle a été étendu par la suite aux autres virus à ADN simple brin circulaire, montrant que ces phénomènes biochimiques associées à la sélection purificatrice participent ensemble à la création des recombinants et au « façonnage » des profils de recombinaison observés dans la nature (Chapitre III).

Adaptation aux contraintes de la recombinaison

S'il est difficile de quantifier réellement les taux de recombinaison dans une population virale (Froissart et al., 2005) ou de les interpréter (Owor et al., 2007), un nombre croissant d'études ainsi que nos travaux (Chapitre III) ont mis en avant la présence de nombreux recombinants et de génomes partiels (défectifs, absence de certains gènes; Liu et al., 1998; Patil et al., 2007; Zhou et al., 2001), suggérant que le coût associé à la recombinaison n'est pas anodin.

Dans un tel contexte comment expliquer la possibilité pour un virus de présenter d'importants taux de recombinaison ? Le maintien des réseaux épistatiques devrait être en réalité directement responsable de l'évolution des différentes prédispositions biochimiques pour la recombinaison tout au long des génomes. On peut penser que l'architecture même des génomes se soit adaptée à ces forts taux de recombinaison. Par conséquent, si les évènements de recombinaison ont lieu majoritairement dans des régions de faible connectivité, une plus grande proportion de recombinants devrait être fonctionnelle que si la recombinaison avait lieu aléatoirement dans des régions de fortes connectivités.

D'autres stratégies évolutives comme l'évolution de la robustesse des réseaux d'interactions ou l'augmentation de la capacité à compenser par mutation les réarrangements délétères liés à la recombinaison pourrait être envisagées. Un virus présentant des taux de recombinaison élevés serait aussi un virus avec un fort taux de mutation. Cette dernière hypothèse est corroborée par les taux élevés de mutation et de substituion observés chez les virus à ADN simples brins circulaires (Duffy et al., 2008; Ge et al., 2007; Isnard et al., 1998).

Evolution de la fitness et création de fonctions nouvelles

Malgré les spéculations sur les mécanismes par lesquels la recombinaison produirait de nouvelles espèces ou souches virales émergentes, avec des gammes d'hôtes élargies ou des pouvoirs pathogènes accentués par exemple, il y a en réalité très peu d'exemples de ce phénomène décrits dans la nature (Fondong et al., 2000; Garcia-Andres et al., 2007a; Monci et al., 2002; Pita et al., 2001; Varsani et al, 2008). L'un des challenges actuel est d'identifier dans quelle mesure la recombinaison peut influer sur la virulence ou la *fitness* d'un virus et de mieux comprendre comment de nouveaux virus aux caractéristiques biologiques nouvelles peuvent apparaître au sein d'une population. Dans le cas des mastrévirus, l'analyse de la diversité des populations de mastrévirus sur les plantes cultivées et non cultivées a permis de montrer que les variants du MSV appartenant au groupe A (MSV-A), le plus répandu et dommageable pour la culture de maïs en Afrique subsaharienne, est issu d'une recombinaison entre les ancêtres des MSVs appartenant actuellement aux groupes B et G/F identifiés sur plantes non cultivées (Varsani et al., 2008). Ce résultat représente une preuve directe de la contribution d'une souche non adaptée au maïs à l'évolution des variants du MSV-A et ce par un phénomène de recombinaison. Nul doute qu'une meilleure description de la diversité virale présente dans les hôtes non cultivés devrait permettre de multiplier ce type d'exemple. Toutefois, l'obtention de virus recombinants avec la réalisation d'infections mixtes de plantes (Garcia-Andres et al., 2007b), la création de librairies de virus recombinants ou encore la synthèse directe de génomes viraux recombinants puis leur inoculation semblent des méthodes plus prometteuses pour la compréhension de l'évolution de la *fitness* et de la virulence par recombinaison. D'ores et déjà, les premiers travaux réalisés à partir de virus recombinants de synthèse ont souligné la possibilité offerte aux virus de recombiner rapidement dans l'hôte, de rétablir un *fitness* élevé et d'établir des infections stables (Monjane et al., 2007).

Le point de vue développé précédemment (Chapitre III) se focalise sur le rétablissement de l'architecture des réseaux d'interactions parentaux au sein du virus recombinant pour que celui-ci soit viable. Si l'avantage de la recombinaison est lié à l'adaptation à de nouveaux hôtes et l'exploration de nouvelles combinaisons, il paraît aussi vraisemblable que l'adaptation passe par la création d'interactions non parentales propres à créer de nouvelles fonctions. Les limites de la création de ces fonctions nouvelles sont liées aux interactions possibles avec l'hôte. Des recherches réalisées au niveau des interactions entre protéines virales et protéines hôtes, des structures tridimensionnelles de ces mêmes protéines et de la manière dont elles interagissent en général devraient permettre de mieux cerner le potentiel d'évolution

de ces complexes d'interactions. La connaissance des niveaux de plasticité de chaque protéine suivant les fonctions qu'elles doivent obligatoirement remplir permettrait de prédire le potentiel d'évolution de chaque virus et les possibilités d'adaptation à de nouveaux hôtes. Pour cette raison, la protéine de capside, qui est une protéine multifonctionnelle et qui intervient dans plusieurs étapes du cycle viral (particule virale, transmission par insecte vecteur et mouvement du virus dans la plante), semble posséder une plasticité faible contrairement à la protéine d'initiation de la réplication. Cette dernière pourrait présenter une malléabilité plus importante liée à la multitude d'hôtes rencontrés par ces virus (chapitre IV).

La révolution du séquençage de génomes viraux à haut débit : du niveau moléculaire à un niveau populationnel

Enormément de travail reste à être réalisé pour la compréhension de l'évolution et de l'émergence des bégomovirus. La diversité globale et la gamme d'hôte des virus, l'interaction avec le vecteur, l'interaction avec la plante, l'évolution de la *fitness* et de la virulence présentent encore de nombreuses inconnues.

Concernant la description de la diversité globale et de la gamme d'hôtes des bégomovirus, des campagnes d'échantillonnage et de séquençage très importantes doivent seront initiées (projet *Emerge* financé par le CRVOI, plusieurs centaines de séquences attendues) et devront s'intéresser à l'ensemble des plantes sur lesquelles *Bemisia tabaci* s'alimente, et non plus se limiter aux plantes cultivées. Dans le cas de l'évolution de la *fitness* et de la virulence, des expériences d'inoculation en conditions contrôlées et de suivi de la descendance virale devraient permettre de suivre l'évolution des traits viraux impliqués dans le pouvoir pathogène.

L'avènement de la polymérase phi29 du phage *phi29* de *Bacillus subtilis* pour l'amplification de l'ADN circulaire double brin (Inoue-Nagata et al., 2004) des bégomovirus, associée à la baisse des coûts de séquençage et la diffusion de nouveaux protocoles d'extraction (Shepherd et al., 2008a), ont grandement facilité les tâches du clonage et du séquençage des génomes complets de géminivirus. Ces nouvelles méthodologies rendent aujourd'hui envisageables des approches de séquençage à haut débit. Le nombre de séquences complètes de géminivirus disponibles dans les bases de données a ainsi considérablement augmenté ces dernières années (Fauquet et al., 2008).

La prochaine révolution méthodologique liée à ce genre viral pourrait d'ailleurs dépendre des progrès et de la baisse des coûts associés aux "nouvelles techniques de séquençage" (Mardis, 2008; Strausberg et al., 2008). Ces techniques permettent un grand nombre de lecture de matrices apparentées et pourraient ainsi permettre de "déchiffrer" en une seule fois la diversité présente au sein d'une population de virus

telle que celle infectant une plante. A partir d'amplificons d'ADN viral obtenus en utilisant la polymérase phi 29 (ne requiert aucune connaissance à priori des séquences amplifiées), une librairie de la population de virus infectant une plante pourrait être créée puis entièrement séquencée. Cette stratégie permettrait d'obtenir non pas la séquence virale la plus représentée dans la population, ni la séquence consensus des principaux variants constituant la population virale, mais les séquences des virus majoritaires. Cette approche populationnelle permettra sans nul doute une meilleure compréhension de la diversité, des mécanismes d'évolution et d'adaptation à l'hôte des géminivirus.

Références bibliographiques

Abo, M. E., Sy, A. A. & Alegbejo, M. D. (1998). Rice yellow mottle virus (RYMV) in Africa: Evolution, distribution, economic significance on sustainable rice production and management strategies. 11, 85-111.

Abhary, M., Patil, L. & Fauquet C.M. (2007). Molecular biodiversity, taxonomy, and nomenclature of Tomato yellow leaf curl-like viruses. Springer, Tomato Yellow Leaf Curl Disease, 85-118.

Accotto, G. P., Bragaloni, M., Luison, D., Davino, S. & Davino, M. (2003). First report of Tomato yellow leaf curl virus (TYLCV) in Italy. Plant Pathology 52, 799.

Amin, I., Mansoor, S., Amrao, L., Hussain, M., Irum, S., Zafar, Y., Bull, S. E. & Briddon, R. W. (2006). Mobilisation into cotton and spread of a recombinant cotton leaf curl disease satellite. Arch Virol 151, 2055-65.

Anderson, P. K., Cunningham, A. A., Patel, N. G., Morales, F. J., Epstein, P. R. & Daszak, P. (2004). Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends in Ecology & Evolution 19, 535-544.

Antignus, Y. & Cohen, S. (1994). Complete nucleotide sequence of an infectious clone of a mild isolate of tomato yellow leaf curl virus (TYLCV). Molecular plant pathology 84, 707-712.

Baranowski, E., RuizJarabo, C. M. & Domingo, E. (2001). Evolution of cell recognition by viruses. Science 292, 1102-1105.

Betancourt, A. J. & Bollback, J. P. (2006). Fitness effects of beneficial mutations: the mutational landscape model in experimental evolution. Curr Opin Genet Dev 16, 618-23.

Bethke, J.A., Paine, T.D. & Nuessly, G.S. (1991). Comparative biology, morphometrics, and development of two populations of *Bemisia tabaci* (Homoptera : Aleyrodidae) on cotton and poinsettia. Annals of Entomo. Soc. Am., 84, 407-11.

Bird, J., Idris, A. M., Rogan, D. & Brown, J. K. (2001). Introduction of the exotic tomato yellow leaf curl virus -Israel in tomato to Porto Rico. Plant Disease 85, 1028.

Bisaro, D. M. (2006). Silencing suppression by geminivirus proteins. Virology 344, 158-68.

Blackmer, J. L., Byrne, D. N. & Tu, Z. (1995). Behavioral, morphological, and physiological traits associated with migratory Bemisia tabaci (Homoptera: Aleyrodidae). J. Insect Behav. 8, 251-267.

Bonnet, J., Fraile, A., Sacristan, S., Malpica, J. M. & Garcia-Arenal, F. (2005). Role of recombination in the evolution of natural populations of Cucumber mosaic virus, a tripartite RNA plant virus. Virology 332, 359-68.

Bosque-Perez, N. A. (2000). Eight decades of maize streak virus research. Virus Research 71, 107-121.

Bottcher, B., Unseld, S., Ceulemans, H., Russell, R. B. & Jeske, H. (2004). Geminate structures of African cassava mosaic virus. J Virol 78, 6758-65.

Boykin, L. M., Shatters, R. G., Rosell, R. C., McKenzie, C. L., Bagnall, R. A., De Barro, P. & Frohlich, D. R. (2007). Global relationships of Bemisia tabaci (Hemiptera : Aleyrodidae) revealed using Bayesian analysis of mitochondrial COI DNA sequences. Molecular Phylogenetics and Evolution 44, 1306.

Briddon, R. W., Bedford, I. D., Tsai, J. H. & Markham, P. G. (1996). Analysis of the nucleotide sequence of the treehopper-transmitted geminivirus, tomato pseudo-curly top virus, suggests a recombinant origin. Virology 219, 387-94.

Briddon, R. W., Pinner, M. S., Stanley, J. & Markham, P. G. (1990). Geminivirus coat protein gene replacement alters insect specificity. Virology 177, 85-94.

Briddon, R. W. & Stanley, J. (2006). Subviral agents associated with plant single-stranded DNA viruses. Virology 344, 198-210.

Brown, J. & Bird, J. (1995). Variability within the Bemisia tabaci species complex and its relation to new epidemics caused by Geminiviruses. CEIBA 36, 73-80.

Campos-Olivas, R., Louis, J. M., Clerot, D., Gronenborn, B. & Gronenborn, A. M. (2002). The structure of a replication initiator unites diverse aspects of nucleic acid metabolism. Proc Natl Acad Sci U S A 99, 10310-5.

Carman, S., Cai, H. Y., DeLay, J., Youssef, S. A., McEwen, B. J., Gagnon, C. A., Tremblay, D., Hazlett, M., Lusis, P., Fairles, J., Alexander, H. S. & van Dreumel, T. (2008). The emergence of a new strain of porcine circovirus-2 in Ontario and Quebec swine and its association with severe porcine circovirus associated disease--2004-2006. Can J Vet Res 72, 259-68.

Casado, C. G., Javier Ortiz, G., Padron, E., Bean, S. J., McKenna, R., Agbandje-McKenna, M. & Boulton, M. I. (2004). Isolation and characterization of subgenomic DNAs encapsidated in "single" T = 1 isometric particles of Maize streak virus. Virology 323, 164-71.

Castillo, A. G., Collinet, D., Deret, S., Kashoggi, A. & Bejarano, E. R. (2003). Dual interaction of plant PCNA with geminivirus replication accessory protein (Ren) and viral replication protein (Rep). Virology 312, 381-94.

Castillo, A. G., Kong, L. J., Hanley-Bowdoin, L. & Bejarano, E. R. (2004). Interaction between a geminivirus replication protein and the plant sumoylation system. J Virol 78, 2758-69.

Chare, E. R. & Holmes, E. C. (2006). A phylogenetic survey of recombination frequency in plant RNA viruses. Arch Virol 151, 933-46.

Chen, L. F., Hagen, C., Zhou, Y., Noussourou, M., Kon, T., Rojas, M. & Gilbertson, R. L. (2007). Emergence of tomato leaf curl and yellow leaf curl diseases in West Africa: Identification of a novel begomovirus-satellite DNA complex. Phytopathology 97, S154.

Chiel, E., Gottlieb, Y., Zchori-Fein, E., Mozes-Daube, N., Katzir, N., Inbar, M. & Ghanim, M. (2007). Biotype-dependent secondary symbiont communities in sympatric populations of Bemisia tabaci. Bulletin of Entomological Research 97, 407.

Chouchane, S. G., Gorsane, F., Nakhla, M. K., Maxwell, D. P., Marrakchi, M. & Fakhfakh, H. (2007). First report of tomato yellow leaf curl virus-israel species infecting tomato, pepper and bean in Tunisia. Journal of Phytopathology 155, 236.

Cohen, S. & Antignus, Y. (1994). Tomato yellow leaf curl virus, a whitefly-borne geminivirus of tomatoes. Advances in Disease Vector Research 10, 259-288.

Cohen, S., Harpaz, J., (1964). Periodic, rather than continual acquisition of a new tomato virus by its vector, the tobacco whitefly (Bemisia tabaci). Entomol. Exp. Appl. 7, 155-166.

Cohen, S. & Nitzany, F. E. (1966). Transmission and host range of the tomato yellow leaf curl virus. Phytopathology 56, 1127-1131.

Crameri, A., Raillard, S. A., Bermudez, E. & Stemmer, W. P. (1998). DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature 391, 288-91.

Csagola, A., Kecskemeti, S., Kardos, G., Kiss, I. & Tuboly, T. (2006). Genetic characterization of type 2 porcine circoviruses detected in Hungarian wild boars. Arch Virol 151, 495-507.

Czosnek, H., Ghanim, M., Morin, S., Rubinstein, G., Fridman, V. & Zeidan, M. (2001). Whiteflies: vectors, and victims (?), of geminiviruses. Adv Virus Res 57, 291-322.

Czosnek, H. (2007). Interactions of Tomato yellow leaf curl virus with its whitefly vector. Springer, Tomato Yellow Leaf Curl Disease, 157-70.

Czosnek, H. & Laterrot, H. (1997). A worldwide survey of tomato yellow leaf curl viruses. Arch Virol 142, 1391-406.

Dalmon A., Peterschmitt M., Cailly M., Dufour O., Jeay M. & Baguet A. (2000). Yellow leaf curl of tomatoes: a serious virus caused by TYLCV, accidentally introduced into France. Phytoma, 527, 10-13.

Davino, S., Napoli, C., Davino, M. & Accotto, G. P. (2006). Spread of Tomato yellow leaf curl virus in Sicily: partial displacement of another geminivirus originally present. European Journal of Plant Pathology 114, 293.

Delatte H, Duyck PF, Triboire A, David P, Becker N, Bonato O and Reynaud B (2008) Differential invasion success among biotypes: case of Bemisia tabaci. Biological Invasions 10.

Delatte, H., Dalmon, A., Rist, D., Soustrade, I., Wuster, G., Lett, J. M., Goldbach, R. W., Peterschmitt, M. & Reynaud, B. (2003). Tomato yellow leaf curl virus can be acquired and transmitted by Bemisia tabaci (Gennadius) from tomato fruit. Plant Dis. 87, 1297-1300.

Delatte, H., David, P., Granier, M., Lett, J. M., Goldbach, R., Peterschmitt, M. & Reynaud, B. (2006). Microsatellites reveal the coexistence and genetic relationships between invasive and indigenous whitefly biotypes in an insular environment. Genetical Research, 87, 109-24.

Delatte, H., Holota, H., Naze, F., Peterschmitt, M., Reynaud, B. & Lett J.M. (2005a). The presence of both recombinant and non recombinant strains of Tomato yellow leaf curl virus on tomato in Réunion Island. Plant Pathol 54, 262.

Delatte, H., Martin, D. P., Naze, F., Golbach, R. W., Reynaud, B., Peterschmitt, M. & Lett, J. M. (2005b). South West Indian Ocean islands tomato begomovirus populations represent a new major monopartite begomovirus group. J Gen Virol 86, 1533–1542.

Delatte, H., Reynaud, B., Lett, J. M., Peterschmitt, M., Granier, M., Ravololonandrianina, J. & Goldbach, R. (2002). First molecular identification of a begomovirus in Madagascar. Plant Disease 86, 1404.

Dewar, R. E. & Richard, A. F. (2007). Evolution in the hypervariable environment of Madagascar. Proc Natl Acad Sci U S A 104, 13723-7.

Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998). Rates of spontaneous mutation. Genetics 148, 1667-86.

Dry, I. B., Krake, L. R., Rigden, J. E. & Rezaian, M. A. (1997). A novel subviral agent associated with a geminivirus: the first report of a DNA satellite. Proc Natl Acad Sci U S A 94, 7088-93.

Duffy, S. & Holmes, E. C. (2007). Multiple introductions of the old world Begomovirus Tomato yellow leaf curl virus into the new world. Applied and Environmental Microbiology 73, 7114.

Duffy, S. & Holmes, E. C. (2008). Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. J Virol 82, 957-65.

Duffy, S., Shackelton, L. A. & Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9, 267-76.

Escriu, F., Fraile, A. & Garcia-Arenal, F. (2007). Constraints to genetic exchange support gene coadaptation in a tripartite RNA virus. PLoS Pathog 3, e8.

Fargette, D., Konate, G., Fauquet, C., Muller, E., Peterschmitt, M. & Thresh, J. M. (2006). Molecular ecology and emergence of tropical plant viruses. Annu Rev Phytopathol 44, 235-60.

Fargette, D., Pinel, A., Abubakar, Z., Traore, O., Brugidou, C., Fatogoma, S., Hebrard, E., Choisy, M., Sere, Y., Fauquet, C. & Konate, G. (2004). Inferring the evolutionary history of Rice yellow mottle virus from genomic, phylogenetic, and phylogeographic studies. Journal of Virology 78, 3252-3261.

Fauquet, C. M., Briddon, R. W., Brown, J. K., Moriones, E., Stanley, J., Zerbini, M. & Zhou, X. (2008). Geminivirus strain demarcation and nomenclature. Archives of Virology 153, 783.

Fondong, V. N., Pita, J. S., Rey, M. E., de Kochko, A., Beachy, R. N. & Fauquet, C. M. (2000). Evidence of synergism between African cassava mosaic virus and a new double-recombinant geminivirus infecting cassava in Cameroon. J Gen Virol 81, 287-97.

Frischmuth, T., Ringel, M. & Kocher, C. (2001). The size of encapsidated single-stranded DNA determines the multiplicity of African cassava mosaic virus particles. J Gen Virol 82, 673-6.

Frohlich, D. R., TorresJerez, I., Bedford, I. D., Markham, P. G. & Brown, J. K. (1999). A phylogeographical analysis of the Bemisia tabaci species complex based on mitochondrial DNA markers. Molecular Ecology 8, 1683-1691.

Froissart, R., Roze, D., Uzest, M., Galibert, L., Blanc, S. & Michalakis, Y. (2005). Recombination every day: abundant recombination in a virus during a single multi-cellular host infection. PLoS Biol 3, e89.

Gafni, Y. & Epel, B. L. (2002). The role of host and viral proteins in intra- and inter-cellular trafficking of geminiviruses. Physiological and Molecular Plant Pathology 60, 231-241.

Gao, G. P., Alvira, M. R., Somanathan, S., Lu, Y., Vandenberghe, L. H., Rux, J. J., Calcedo, R., Sanmiguel, J., Abbas, Z. & Wilson, J. M. (2003). Adeno-associated viruses undergo substantial evolution in primates during natural infections. Proceedings of the National Academy of Sciences of the United States of America 100, 6081-6086.

Garcia-Andres, S., Accotto, G. P., Navas-Castillo, J. & Moriones, E. (2007a). Founder effect, plant host, and recombination shape the emergent population of begomoviruses that cause the tomato yellow leaf curl disease in the Mediterranean basin. Virology 359, 302-12.

Garcia-Andres, S., Tomas, D. M., Sanchez-Campos, S., Navas-Castillo, J. & Moriones, E. (2007b). Frequent occurrence of recombinants in mixed infections of tomato yellow leaf curl disease-associated begomoviruses. Virology 365, 210-9.

Garcia-Andres, S., Monci, F., Navas-Castillo, J. & Moriones, E. (2006). Begomovirus genetic diversity in the native plant reservoir Solanum nigrum: Evidence for the presence of a new virus species of recombinant nature. Virology 350, 433-42.

Ge, L., Zhang, J., Zhou, X. & Li, H. (2007). Genetic Structure and Population Variability of Tomato Yellow Leaf Curl China Virus. J Virol. 81, 2902-7.

Gennadius, P. (1889). Disease of tobacco plantations in the Trikonia. The aleyrodid of tobacco. Ellenike Georgia 5, 1-13.

Gibbs, M. J., Smeianov, V. V., Steele, J. L., Upcroft, P. & Efimov, B. A. (2006). Two families of rep-like genes that probably originated by interspecies recombination are represented in viral, plasmid, bacterial, and parasitic protozoan genomes. Mol Biol Evol 23, 1097-100.

Gibbs, M. J. & Weiller, G. F. (1999). Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus. Proc Natl Acad Sci U S A 96, 8022-7.

Grubman, M. J. & Baxt, B. (2004). Foot-and-mouth disease. Clin Microbiol Rev 17, 465-93.

Gubler, D. J. (2007). The continuing spread of West Nile virus in the western hemisphere. Clin Infect Dis 45, 1039-46.

Haevermans, T., Hoffmann, P., Lowry, P. P., Labat, J. N. & Randrianjohany, E. (2004). Phylogenetic analysis of the Madagascan Euphorbia subgenus Lacanthis based on its sequence data. Annals of the Missouri Botanical Garden 91, 247-259.

Hanley-Bowdoin, L., Elmer, J. S. & Rogers, S. G. (1989). Functional expression of the leftward open reading frames of the A component of tomato golden mosaic virus in transgenic tobacco plants. Plant Cell 1, 1057-67.

Hanley-Bowdoin, L., Settlage, S. B., Orozco, B. M., Nagar, S. & Robertson, D. (2000). Geminiviruses: Models for plant DNA replication, transcription, and cell cycle regulation. Critical Reviews in Biochemistry and Molecular Biology 35, 105-140.

Heath, L., van der Walt, E., Varsani, A. & Martin, D. P. (2006). Recombination patterns in aphthoviruses mirror those found in other picornaviruses. J Virol 80, 11827-32.

Hernandez-Zepeda, C., Idris, A. M., Carnevali, G., Brown, J. K. & Moreno-Valenzuela, O. A. (2007). Preliminary identification and coat protein gene phylogenetic relationships of begomoviruses associated with native flora and cultivated plants from the Yucatan Peninsula of Mexico. Virus Genes 35, 825-33.

Holmes, E. C. & Drummond, A. J. (2007). The evolutionary genetics of viral emergence. Curr Top Microbiol Immunol 315, 51-66.

Holmes, E. C. & Rambaut, A. (2004). Viral evolution and the emergence of SARS coronavirus. Philos Trans R Soc Lond B Biol Sci 359, 1059-65.

Horowitz, A. R., Kontsedalov, S., Khasdan, V. & Ishaaya, I. (2005). Biotypes B and Q of Bemisia tabaci and their relevance to neonicotinoid and pyriproxyfen resistance. Archives of Insect Biochemistry and Physiology 58, 216-225.

Hughes, A. L. (2004). Birth-and-death evolution of protein-coding regions and concerted evolution of non-coding regions in the multi-component genomes of nanoviruses. Mol Phylogenet Evol 30, 287-94.

Idris, A. M., Guerrero, J.C., & Brown, J. K. (2007). Two Distinct Isolates of Tomato yellow leaf curl virus Threaten Tomato Production in Arizona and Sonora, Mexico. Plant disease, 91, 910.

Idris, A. M. & Brown, J. K. (2005). Evidence for interspecific-recombination for three monopartite begomoviral genomes associated with the tomato leaf curl disease from central Sudan. Arch Virol 150, 1003-12.

Inoue-Nagata, A. K., Albuquerque, L. C., Rocha, W. B. & Nagata, T. (2004). A simple method for cloning the complete begomovirus genome using the bacteriophage phi29 DNA polymerase. J Virol Methods 116, 209-11.

Isnard, M., Granier, M., Frutos, R., Reynaud, B. & Peterschmitt, M. (1998). Quasispecies nature of three maize streak virus isolates obtained through different modes of selection from a population used to assess response to infection of maize cultivars. J Gen Virol 79, 3091-9.

Jain, R., Rivera, M. C. & Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 96, 3801-6.

Jenkins, G. M., Rambaut, A., Pybus, O. G. & Holmes, E. C. (2002). Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. Journal of Molecular Evolution 54, 156-165.

Jeske, H., Lutgemeier, M. & Preiss, W. (2001). DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. EMBO J 20, 6158-67.

Jupin, I., De Kouchkovshy, F., Jouanneau, F. & Gronenborn, B. (1994). Movement of tomato yellow leaf curl geminivirus (TYLCV) : involvement of the protein encoded by ORF C4. Virology 204, 82-90.

Klute, K. A., Nadler, S. A. & Stenger, D. C. (1996). Horseradish curly top virus is a distinct subgroup II geminivirus species with rep and C4 genes derived from a subgroup III ancestor. J Gen Virol 77 (Pt 7), 1369-78.

Kong, L. J. & Hanley-Bowdoin, L. (2002). A geminivirus replication protein interacts with a protein kinase and a motor protein that display different expression patterns during plant development and infection. Plant Cell 14, 1817-32.

Koonin, E. V. & Ilyina, T. V. (1992). Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. J Gen Virol 73 (Pt 10), 2763-6.

Krake, L. R., Rezaian, M. A. & Dry, B. (1998). Expression of the tomato leaf curl Geminivirus C4 gene produces viruslike symptoms in transgenic plants. Mol Plant-Microbe Interac 11, 413-417.

Latham, J. R., Saunders, K., Pinner, M. S. & Stanley, J. (1997). Induction of plant cell division by curly top virus gene C4. The Plant Journal 11, 1273-1283.

Legg, J. P. & Fauquet, C. M. (2004). Cassava mosaic geminiviruses in Africa. Plant Mol Biol 56, 585-99.

Lett, J. M., Delatte, H., Naze, F., Reynaud, B., Abdoul-Karime, A. L. & Peterschmitt, M. (2004). A new tomato leaf curl Begomovirus from Mayotte. Plant Disease 88, 681.

Ling, K.-S., Simmons, A.M., Hassell, R.L., Keinath, A.P. & Polston, J.E. (2006). First report of Tomato yellow leaf curl virus in South Carolina. Plant Disease, 90, 379.

Liu, Y., Robinson, D. J. & Harrison, B. D. (1998). Defective forms of cotton leaf curl virus DNA-A that have different combinations of sequence deletion, duplication, inversion and rearrangement. J Gen Virol 79 (Pt 6), 1501-8.

Lukashov, V. V. & Goudsmit, J. (2001). Evolutionary relationships among parvoviruses: virus-host coevolution among autonomous primate parvoviruses and links between adeno-associated and avian parvoviruses. J Virol 75, 2729-40.

Luque, A., SanzBurgos, A., RamirezParra, E., Castellano, M. M. & Gutierrez, C. (2002). Interaction of geminivirus rep protein with replication factor C and its potential role during geminivirus DNA replication. Virology 302, 83-94.

Mardis, E. R. (2008). Next-Generation DNA Sequencing Methods. Annu Rev Genomics Hum Genet.

Martin, D. P., van der Walt, E., Posada, D. & Rybicki, E. P. (2005). The evolutionary value of recombination is constrained by genome modularity. PLoS Genet 1, e51.

Monci, F., SanchezCampos, S., NavasCastillo, J. & Moriones, E. (2002). A natural recombinant between the geminiviruses Tomato yellow leaf curl Sardinia virus and Tomato yellow leaf curl virus exhibits a novel pathogenic phenotype and is becoming prevalent in Spanish populations. Virology 303, 317-326.

Monjane, A.L., shepherd, D., van der Walt, E., Varsani, A. and Martin, D.P. (2007) Highly predictable restoration of MSV fitness through recombination. 5th Internationnal Geminivirus Symposium, 20-26 may 2007, Ouro Preto, Brazil.

Morilla, G., Krenz, B., Jeske, H., Bejarano, E. R. & Wege, C. (2004). Tete a tete of tomato yellow leaf curl virus and tomato yellow leaf curl sardinia virus in single nuclei. J Virol 78, 10715-23.

Morin, S., Ghanim, M., Sobol, I. & Czosnek, H. (2000). The GroEL protein of the whitefly Bemisia tabaci interacts with the coat protein of transmissible and nontransmissible begomoviruses in the yeast two-hybrid system. Virology 276, 404-16.

Morin, S., Ghanim, M., Zeidan, M., Czosnek, H., Verbeek, M. & van den Heuvel, J. (1999). A GroEL homologue from endosymbiotic bacteria of the whitefly Bemisia tabaci is implicated in the circulative transmission of tomato yellow leaf curl virus. Virology 256, 75.

Moriones, E. & NavasCastillo, J. (2000). Tomato yellow leaf curl virus, an emerging virus complex causing epidemics worldwide. Virus Research 71, 123-134.

Mosig, G. (1998). Recombination and recombination-dependent DNA replication in bacteriophage T4. Annu Rev Genet 32, 379-413.

Mosig, G., Gewin, J., Luder, A., Colowick, N. & Vo, D. (2001). Two recombination-dependent DNA replication pathways of bacteriophage T4, and their roles in mutagenesis and horizontal gene transfer. Proc Natl Acad Sci U S A 98, 8306-11.

Mound, L. A. (1962). Studies on the olfactory and colour sensitivity of Bemisia tabaci (Genn.) (Homoptera, Aleyrodidae). Entomol. exp. appl. 5, 99-104.

NavasCastillo, J., SanchezCampos, S., Noris, E., Louro, D., Accotto, G. P. & Moriones, E. (2000). Natural recombination between Tomato yellow leaf curl virus-Is and Tomato leaf curl virus. Journal of General Virology 81, 2797-2801.

Nosal, B. & Pellizzari, R. (2003). West Nile virus. CMAJ 168, 1443-4.

Novick, R. P. (1998). Contrasting lifestyles of rolling-circle phages and plasmids. Trends Biochem Sci 23, 434-8.

Olvera, A., Cortey, M. & Segales, J. (2007). Molecular evolution of porcine circovirus type 2 genomes: phylogeny and clonality. Virology 357, 175-85.

Orozco, B. M. & Hanley-Bowdoin, L. (1996). A DNA structure is required for geminivirus replication origin function. J Virol 70, 148-58.

Otto, S. P. & Gerstein, A. C. (2006). Why have sex? The population genetics of sex and recombination. Biochem Soc Trans 34, 519-22.

Owor, B. E., Martin, D. P., Shepherd, D. N., Edema, R., Monjane, A. L., Rybicki, E. P., Thomson, J. A. & Varsani, A. (2007). Genetic analysis of maize streak virus isolates from Uganda reveals widespread distribution of a recombinant variant. J Gen Virol 88, 3154-65.

Padidam, M., Sawyer, S. & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. Virology 265, 218-25.

Patil, B. L., Dutt, N., Briddon, R. W., Bull, S. E., Rothenstein, D., Borah, B. K., Dasgupta, I., Stanley, J. & Jeske, H. (2007). Deletion and recombination events between the DNA-A and DNA-B components of Indian cassava-infecting geminiviruses generate defective molecules in Nicotiana benthamiana. Virus Res 124, 59-67.

Perring, T. M. (2001). The Bemisia tabaci species complex. Crop Protection 20, 725-737.

Peterschmitt, M., Granier, M., Mekdoud, R., Dalmon, A., Gambin, O., Vayssières, J. F. & Reynaud, B. (1999). First report of tomato yellow leaf curl virus in Réunion Island. Plant Disease 83, 303.

Pico, B., Diez, M.-J. & Nuez, F. (1996). Viral diseases causing the greatest economic losses to the tomato crop. II. The tomato yellow leaf curl virus - a review. Scientia Horticulturae 67, 151-196.

Pilartz, M. & Jeske, H. (1992). Abutilon mosaic geminivirus double-stranded DNA is packed into minichromosomes. Virology 189, 800-2.

Pita, J. S., Fondong, V. N., Sangare, A., Otim-Nape, G. W., Ogwal, S. & Fauquet, C. M. (2001). Recombination, pseudorecombination and synergism of geminiviruses are determinant keys to the epidemic of severe cassava mosaic disease in Uganda. J Gen Virol 82, 655-65.

Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. (2007). Empirical fitness landscapes reveal accessible evolutionary paths. Nature 445, 383-6.

Polston, J. E., Bois, D., Serra, C. A. & Concepcion, S. (1994). First report of a tomato yellow leaf curl-like geminivirus in the Western Hemisphere. Plant Dis. 78, 831.

Prasanna, H. C. & Rai, M. (2007). Detection and frequency of recombination in tomato-infecting begomoviruses of South and Southeast Asia. Virol J 4, 111.

Rigden, J. E., Krake, L. R., Rezaian, M. A. & Dry, I. B. (1994). ORF C4 of tomato leaf curl geminivirus is a determinant of symptom severity. Virology 204, 847-50.

Rojas, M. R., Hagen, C., Lucas, W. J. & Gilbertson, R. L. (2005). Exploiting chinks in the plant's armor: evolution and emergence of geminiviruses. Annu Rev Phytopathol 43, 361-94.

Rojas, M. R., Jiang, H., Salati, R., XoconostleCazares, B., Sudarshana, M. R., Lucas, W. J. & Gilbertson, R. L. (2001). Functional analysis of proteins involved in movement of the monopartite begomovirus, tomato yellow leaf curl virus. Virology 291, 110-125.

Rokyta, D. R., Burch, C. L., Caudle, S. B. & Wichman, H. A. (2006). Horizontal gene transfer and the evolution of microvirid coliphage genomes. J Bacteriol 188, 1134-42.

Rousseau, C. M., Learn, G. H., Bhattacharya, T., Nickle, D. C., Heckerman, D., Chetty, S., Brander, C., Goulder, P. J., Walker, B. D., Kiepiela, P., Korber, B. T. & Mullins, J. I. (2007). Extensive intrasubtype recombination in South african human immunodeficiency virus type 1 subtype C infections. J Virol 81, 4492-500.

Russell, C. J. & Webster, R. G. (2005). The genesis of a pandemic influenza virus. Cell 123, 368-71.

Rybicki, E. P., Briddon, R., Brown, J. K., Fauquet, C., Maxwell, D. P., Stanley, J., Harrison, B. D., Markham, P., Bisaro, D. M. & Robinson, D. J. (2000). Family Geminiviridae. In Virus Taxonomy, pp. 285-297. Edited by M. H. V. Van Regemnmortel, Fauquet, C., Bishop, D.H.L., Carstens, E., Estes, M., Lemon, S., Maniloff, J., Mayo, M.A., McGeoch, D., Pringle, C., Wickner, R. New York: Academic Press.

Rybicki, E. P. & Pietersen, G. (1999). Plant virus disease problems in the developing world. In Advances in Virus Research, Vol 53, pp. 127+. Edited by K. Maramorosch, F. A. Murphy & A. J. Shatkin. 525 B Street/ Suite 1900/San Diego/CA 92101-4495/USA: Academic Press Inc.

SanchezCampos, S., NavasCastillo, J., Camero, R., Soria, C., Diaz, J. A. & Moriones, E. (1999). Displacement of tomato yellow leaf curl virus (TYLCV)-Sr by TYLCV-Is in tomato epidemics in spain. Phytopathology 89, 1038-1043.

Saunders, K., Lucy, A. & Stanley, J. (1992). RNA-primed complementary-sense DNA synthesis of the geminivirus African cassava mosaic virus. Nucleic Acids Res 20, 6311-5.

Saunders, K. & Stanley, J. (1999). A nanovirus-like DNA component associated with yellow vein disease of Ageratum conyzoides: evidence for interfamilial recombination between plant DNA viruses. Virology 264, 142-52.

Schuffenecker, I., Iteman, I., Michault, A., Murri, S., Frangeul, L., Vaney, M. C., Lavenir, R., Pardigon, N., Reynes, J. M., Pettinelli, F., Biscornet, L., Diancourt, L., Michel, S., Duquerroy, S., Guigon, G., Frenkiel, M. P., Brehin, A. C., Cubito, N., Despres, P., Kunst, F., Rey, F. A., Zeller, H. & Brisse, S. (2006). Genome microevolution of chikungunya viruses causing the Indian Ocean outbreak. PLoS Med 3, e263.

Seal, S. E., Jeger, M. J. & Van den Bosch, F. (2006). Begomovirus evolution and disease management. Adv Virus Res 67, 297-316.

Selth, L. A., Dogra, S. C., Rasheed, M. S., Healy, H., Randles, J. W. & Rezaian, M. A. (2005). A NAC domain protein interacts with tomato leaf curl virus replication accessory protein and enhances viral replication. Plant Cell 17, 311-25.

Selth, L. A., Randles, J. W. & Rezaian, M. A. (2004). Host responses to transient expression of individual genes encoded by Tomato leaf curl virus. Mol Plant-Microbe Interac 17, 27-33.

Servin, R., MartinezCarrillo, J. L. & Hirales, L. (1999). Weeds and cultivated hosts of the silverleaf whitefly Bemisia argentifolii Bellows and Perring in Baja California Sur, Mexico. Southwestern Entomologist 24, 31-36.

Shackelton, L. A., Hoelzer, K., Parrish, C. R. & Holmes, E. C. (2007). Comparative analysis reveals frequent recombination in the parvoviruses. J Gen Virol 88, 3294-301.

Shackelton, L. A. & Holmes, E. C. (2006). Phylogenetic evidence for the rapid evolution of human B19 erythrovirus. J Virol 80, 3666-9.

Shackelton, L. A., Parrish, C. R., Truyen, U. & Holmes, E. C. (2005). High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc Natl Acad Sci U S A 102, 379-84.

Shepherd, D. N., Martin, D. P., Lefeuvre, P., Monjane, A. L., Owor, B. E., Rybicki, E. P. & Varsani, A. (2008a). A protocol for the rapid isolation of full geminivirus genomes from dried plant tissue. J Virol Methods 149, 97-102.

Shepherd, D. N., Varsani, A., Windram, O. P., Lefeuvre, P., Monjane, A. L., Owor, B. E. & Martin, D. P. (2008b). Novel sugarcane streak and sugarcane streak Reunion mastreviruses from southern Africa and La Reunion. Arch Virol.

Shivaprasad, P. V., Akbergenov, R., Trinks, D., Rajeswaran, R., Veluthambi, K., Hohn, T. & Pooggin, M. M. (2005). Promoters, transcripts, and regulatory proteins of Mungbean yellow mosaic geminivirus. J Virol 79, 8149-63.

Stanley, J., Saunders, K., Pinner, M. S. & Man Wong, S. (1997). Novel defective interfering DNAs associated with ageratum yellow vein geminivirus infection of Ageratum conyzoides. Virology 239, 87-96.

Stemmer, W. P. (1994). Rapid evolution of a protein in vitro by DNA shuffling. Nature 370, 389-91.

Strausberg, R. L., Levy, S. & Rogers, Y. H. (2008). Emerging DNA sequencing technologies for human genomic medicine. Drug Discov Today 13, 569-77.

Sunter, G. & Bisaro, D. M. (1989). Transcription map of the B genome component of tomato golden mosaic virus and comparison with A component transcripts. Virology 173, 647-55.

Tahiri, A., Sekkat, A., Bennani, A., Granier, M., Delvare, G. & Peterschmitt, M. (2006). Distribution of tomato-infecting begomoviruses and Bemisia tabaci biotypes in Morocco. Annals of Applied Biology 149, 175.

Ueda, S., Kimura, T., Onuki, M., Hanada, K. & Iwanami, T. (2004). Three distinct groups of isolates of Tomato yellow leaf curl virus in Japan and construction of an infectious clone. J. Gen. Plant Pathol., 70, 232-8.

Urbino, C., Gerion, A. L., Poliakoff, F., Coranson, R., Dalmon, A., Tiego, G. & Babo, E. (2003). Begomovirus diseases on tomatoes in French atlantic overseas departments. Phytoma, 52-55.

van Regenmortel, M. H. V., Fauquet, C. M., Bishop, D. H. L., Carstens, E., Estes, M. K., Lemon, S., Maniloff, J., Mayo, M. A., McGeoch, D., Pringle, C. R. & Wickner, R. B. (2000). Virus Taxonomy. Seventh Report of the International Committee on Taxonomy of Viruses. Academic Press London, pp. 1044. San Diego: Academic Press London.

van Wezel, R., Liu, H., Tien, P., Stanley, J. & Hong, Y. (2001). Gene C2 of the monopartite geminivirus tomato yellow leaf curl virus-China encodes a pathogenicity determinant that is localized in the nucleus. Mol Plant Microbe Interact 14, 1125-8.

Vana, G. & Westover, K. M. (2008). Origin of the 1918 Spanish influenza virus: a comparative genomic analysis. Mol Phylogenet Evol 47, 1100-10.

Vanderschuren, H., Stupak, M., Futterer, J., Gruissem, W. & Zhang, P. (2007). Engineering resistance to geminiviruses--review and perspectives. Plant Biotechnol J 5, 207-20.

Vanitharani, R., Chellappan, P. & Fauquet, C. M. (2005). Geminiviruses and RNA silencing. Trends Plant Sci 10, 144-51.

Varsani, A., Shepherd, D.N., Monjane, A.L., Owor, B.E., Erdmann, J.B., Rybicki, E.P., Peterschmitt, M., Briddon, R.W., Markham, P.G., Oluwafemi, S., Windram, O.P., Lefeuvre, P., Lett, J.M. & Martin, D.P., (2008). Recombination, decreased host specificity and increased mobility may

have driven the emergence of maize streak virus as an agricultural pathogen. J Gen Virol, 89, 2063–74.

Vega-Rocha, S., Gronenborn, B., Gronenborn, A. M. & Campos-Olivas, R. (2007). Solution structure of the endonuclease domain from the master replication initiator protein of the nanovirus faba bean necrotic yellows virus and comparison with the corresponding geminivirus and circovirus structures. Biochemistry 46, 6201.

Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. (2002). Protein building blocks preserved by recombination. Nat Struct Biol 9, 553-8.

Woolhouse, M. E., Taylor, L. H. & Haydon, D. T. (2001). Population biology of multihost pathogens. Science 292, 1109-12.

Wu, J.B. (2006). First report of Tomato yellow leaf curl virus in China. Plant Disease, 90, 1359.

Xie, Q., Sanz-Burgos, A. P., Guo, H., Garcia, J. A. & Gutierrez, C. (1999). GRAB proteins, novel members of the NAC domain family, isolated by their interaction with a geminivirus protein. Plant Mol Biol 39, 647-56.

Yassin, A. M. & Nour, M. A. (1965). Tomato leaf curl disease in the Sudan and its relation to tobacco leaf curl. Annals of Applied Biology 56, 207-217.

Zhou, X., Liu, Y., Calvert, L., Munoz, C. & Otim-Nape, G. W. (1997). Evidence that DNA-A of a geminivirus associated with severe cassava mosaic disease in Uganda has arisen has arised by interspecific recombination. Journal of General Virology 78, 2101-2111.

Zhou, X. P., Xie, Y., Zhang, Z. K., Qi, Y. J. & Wu, J. J. (2001). Molecular characterization of a novel defective DNA isolated from tobacco tissues infected with tobacco leaf curl virus. Acta Virol 45, 45-50.

Zhou, Y. C., Noussourou, M., Kon, T., Rojas, M. R., Jiang, H., Chen, L. F., Gamby, K., Foster, R. & Gilbertson, R. L. (2008). Evidence of local evolution of tomato-infecting begomovirus species in West Africa: characterization of tomato leaf curl Mali virus and tomato yellow leaf crumple virus from Mali. Archives of Virology 153, 693.

Figure supplémentaire de l'article « Rapid displacement and mixed infection as a result of direct interaction between strains of TYLCV presenting differential severity in a tropical insular environment ».

Supplementary figure 1: Oligonucleotides and probe used for quantitative PCR of both Mild and IL *Tomato yellow leaf curl virus* strains. Numbers correspond to positions of the (+) strand of complete genome in relative to TYLCV-Mld[Re3] (Accession number AJ865337). Primers are indicated by arrows and blue boxes while Taqman MGB-Il/Mld 1470 probe is indicating by red box.

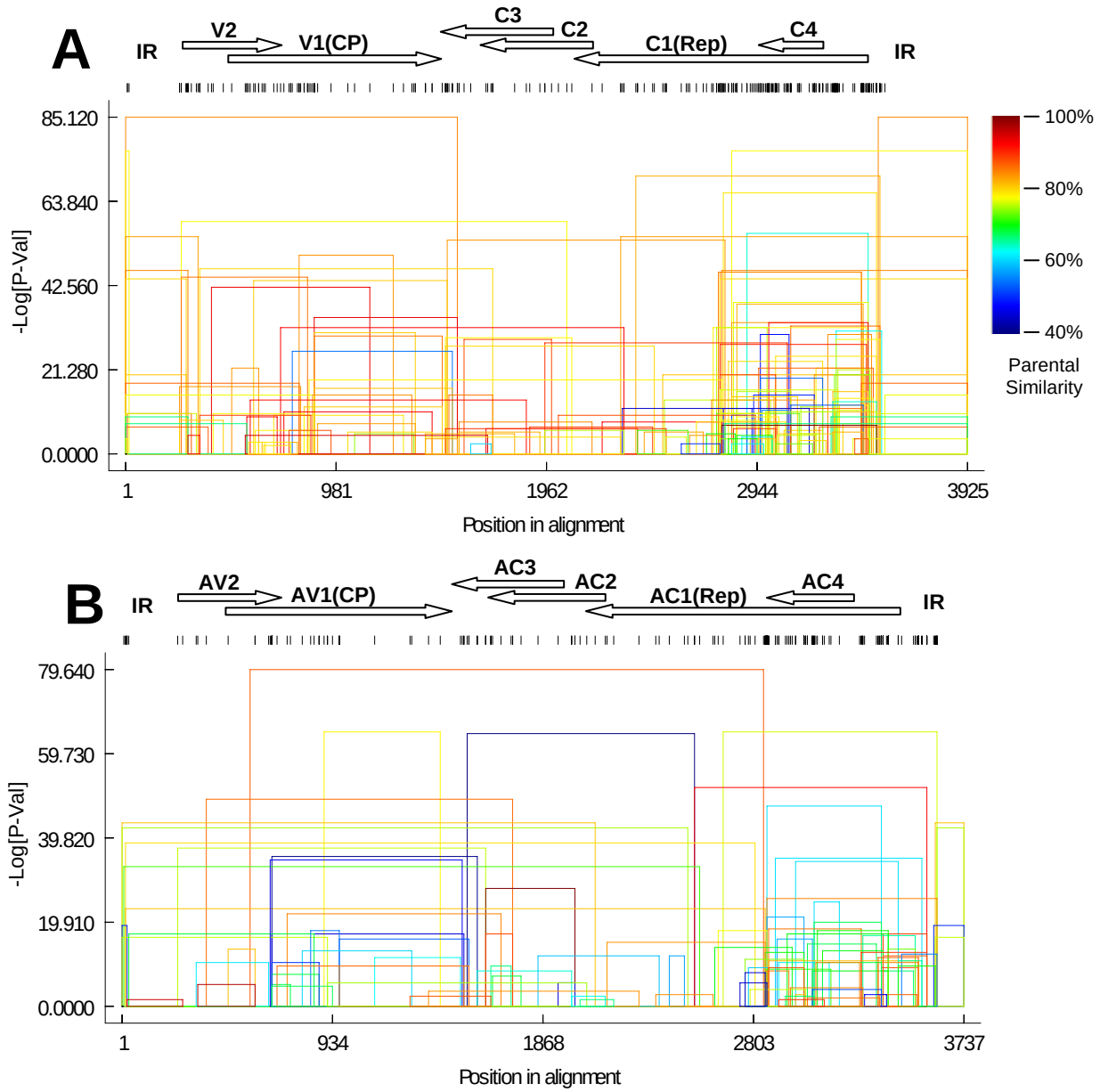


Figure supplémentaire 1 de l'article « Avoidance of Protein Fold Disruption in Natural Virus Recombinants ».

Supplementary figure 1. Schematic representations of individual recombination events detectable within the genomes of monopartite begomoviruses (A) and the DNA-A components of bipartite begomoviruses (B). Each rectangle represents one recombination event with vertical lines delimiting the tract of sequence transferred during the event. Whereas the heights of the vertical lines indicate the statistical certainty with which individual events were detected (in general greater statistical certainty correlates with greater accuracy of breakpoint estimation), the colours of boxes reflect the approximate degree of parental relatedness at the time when recombination events occured.

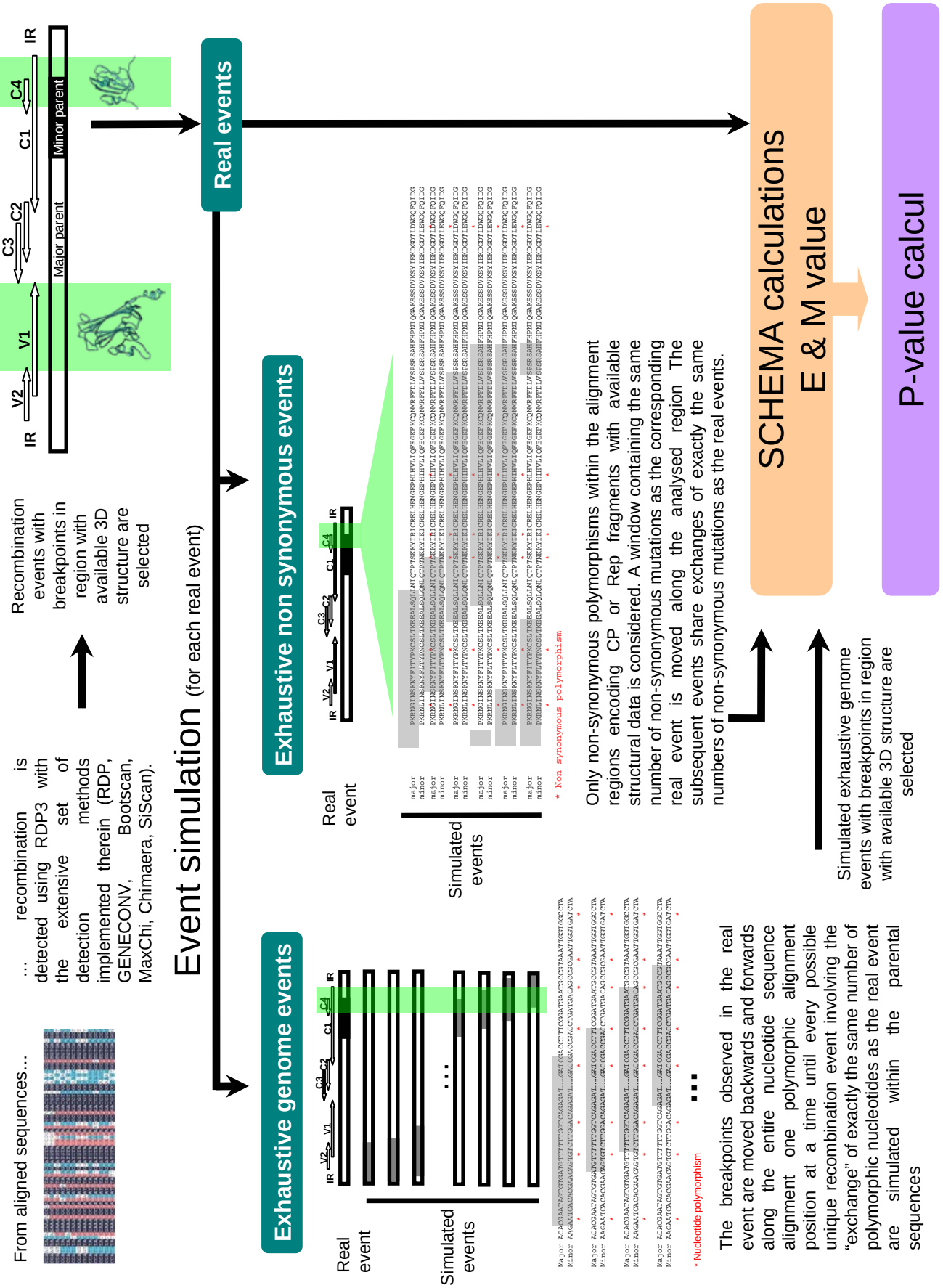


Figure supplémentaire 2 de l'article « Avoidance of Protein Fold Disruption in Natural Virus Recombinants » : méthode de simulation des évènements de recombinaison à partir d'évènements détectés.

Table supplémentaire de l'article : "Widely conserved recombination patterns amongst single stranded DNA viruses and their satellites".

Supplementary Table 1: Dataset description

Host	Family	Genus	Dataset name	mean seq. length (kb)	Pi	Nb. of seq
Animal	-	Anellovirus	Merged	3192*	-	198
			TTV	3319*	0.397	159
			TTMV	2674*	0.414	39
	Circoviridae	PCV/BF	Merged	1873	-	86
		Bird Circovirus	Merged	1964	-	56
		Circovirus/PCV	PCV	1767	0.069	48
		Circovirus/BF	BF	2007	0.284	38
		Circovirus/Goose	Goose	1872	0.17	18
		Gyrvovirus	Gyro	2299	0.024	44
	Parvoviridae	Dependo/Parvo	Merged	4251	-	61
		Dependovirus	Dependo	4100	0.239	22
		Erythrovirus	Erythro	4373	0.237	21
		Parvovirus	Parvo	4292	0.198	18
Bacteria	Microviridae	Microvirus	Micro	5600	0.275	37
Plant	Geminiviridae	Mastrevirus	Mastre	2691	0.14	195
		Begomovirus	Begomo	2728	0.35	328
		DNA B	Merged	2618	-	199
			ow1	2647	0.213	25
			ow2	2710	0.153	54
			ow3	2683	0.261	16
			nw7	2528	0.311	79
		DNA Beta	Merged	1348	-	126
			betaA	1354	0.388	66
			betaC	1342	0.318	42
			betaD	1340	0.399	18
		DNA 1	dna1	1372	0.229	39
	Nanoviridae	Babuvirus	Merged babu nano	1065	-	145
			babuB1	1054	0.292	18
			babuB2	1033	0.287	16
			babuB3	1059	0.102	24
			babuC	1100	0.1	51
			babu3	1060	0.065	23
		Nanovirus	nano	995	0.347	13

*partial sequences

Table supplémentaire de l'article "Widely conserved recombination patterns amongst single stranded DNA viruses and their satellites".

Supplementary Table 2: ORF recombinability

Annelovirus - TTV dataset

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	85	71	12.23099061	2.705675952	0	1
ORF4 vs rest of ORFs	31	40	2.404735479	0.837071987	0.0022	0.9985
ORF3 vs rest of ORFs	17	54	1.403435802	1.122648651	0.6044	0.4806
ORF2 vs rest of ORFs	13	58	2.199320703	1.068085908	0.114	0.9245
ORF1 vs rest of ORFs	47	24	0.886724764	3.328886348	0.9998	0.0002
End 50% vs Middle 50%	49	59	2.082386713	1.771691464	0.356	0.6816
End 25% vs Middle 75%	27	81	2.325795581	1.791113274	0.2497	0.793
End 10% vs Middle 90%	11	97	2.576529861	1.845408316	0.2632	0.8051

Anellovirus - TTMV dataset

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	8	29	2.427339508	2.017920797	0.001	0.9999
ORF4 vs rest of ORFs	12	17	0.901265238	0.61640588	0.0066	0.9975
ORF3 vs rest of ORFs	16	13	1.038649755	0.510019314	0.0025	0.999
ORF2 vs rest of ORFs	6	23	0.984475577	0.66093415	0.3346	0.8022
ORF1 vs rest of ORFs	28	1	0.798245614	0.171912193	0.2683	0.9243
End 50% vs Middle 50%	30	32	1.932399905	1.861091555	0.0082	0.9946
End 25% vs Middle 75%	22	40	2.659540625	1.636203148	0.0005	0.9999
End 10% vs Middle 90%	4	58	1.447067784	1.9362547	0.2212	0.8844

Geminiviridae - DNA-Beta dataset

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	79	4	13.6553974	1.908089223	0.0001	0.9999
End 50% vs Middle 50%	3	1	2.270143527	0.798949737	0.0016	1
End 25% vs Middle 75%	2	2	3.067396313	1.041055718	0.0129	0.9995
End 10% vs Middle 90%	1	3	4.404466501	1.278716081	0.1031	0.9979

Geminiviridae - DNA-1 dataset

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	26	27	8.731808732	5.199926632	0.5669	0.542
End 50% vs Middle 50%	17	10	3.757301479	3.973134046	0.7779	0.3778
End 25% vs Middle 75%	17	10	4.507291206	3.058326658	0.4445	0.7215
End 10% vs Middle 90%	13	14	3.922695596	3.755988502	0.7577	0.3859

Geminiviridae - DNA-B dataset

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	73	35	8.756347352	2.317559216	0	1
nsp vs rest of ORFs	18	17	1.325549451	1.034689372	0.401	0.7213
mp vs rest of ORFs	17	18	1.034689372	1.325549451	0.7213	0.401
End 50% vs Middle 50%	24	11	1.833847587	0.897219784	0.0314	0.9868
End 25% vs Middle 75%	13	22	1.788252556	1.216971217	0.1076	0.9495
End 10% vs Middle 90%	6	29	2.027483667	1.295336788	0.1444	0.9354

Geminiviridae - Begomovirus/Curtovirus dataset

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	51	302	35.24402568	24.06390996	0.084	0.9403
AV2 vs rest of ORFs	36	266	3.806984804	4.97428705	0.8452	0.1903
AV1 vs rest of ORFs	65	237	1.51E-02	2.22E-02	0.9861	0.0172
AC3 vs rest of ORFs	41	261	3.577778995	5.096823218	0.9433	0.0731
AC2 vs rest of ORFs	23	279	1.718853467	5.600630169	1	0
AC1 vs rest of ORFs	183	119	6.587290599	3.284907897	0	1
AC4 vs rest of ORFs	56	246	8.846409133	4.346209093	0.0009	0.9992
End 50% vs Middle 50%	352	52	27.83630863	10.30005623	0	1
End 25% vs Middle 75%	309	95	31.41541549	12.08968949	0	1
End 10% vs Middle 90%	193	211	39.113267	16.53671905	0.0936	0.9249

Geminiviridae - Mastrevirus dataset

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	29	34	10.19768105	2.628579123	0	1
MP vs rest of ORFs	5	29	0.728451224	0.836719072	0.3783	0.7844
CP vs rest of ORFs	6	28	0.452180734	0.991009049	0.9962	0.0133
Rep vs rest of ORFs	23	11	1.075261123	0.546369154	0.0353	0.9849
End 50% vs Middle 50%	31	3	2.753181775	0.959875476	0.0286	0.9927
End 25% vs Middle 75%	30	4	3.119301576	0.839002268	0.002	0.9992
End 10% vs Middle 90%	25	9	3.008108342	1.481668249	0.0166	0.9934

Circoviridae - Circovirus dataset (BFDV + PCV)

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	15	41	6.536054778	3.326561861	0.0088	0.9964
Rep vs rest of ORFs	20	21	1.541166349	1.649407494	0.0255	0.9898
Cap vs rest of ORFs	21	20	1.649407494	1.541166349	0.9898	0.0255
End 50% vs Middle 50%	31	10	2.736098853	2.445572227	0.0115	0.9967
End 25% vs Middle 75%	25	16	2.824282887	2.436343653	0.0018	0.9991
End 10% vs Middle 90%	9	32	4.435633291	2.389843167	0.0041	0.9987

Circoviridae - Circovirus dataset (Bird circovirus)

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	15	19	4.603345097	1.260950358	0.0011	0.9995
Rep vs rest of ORFs	14	5	1.397504456	0.496982606	0.0013	0.9999
Cap vs rest of ORFs	5	14	0.496982606	1.397504456	0.9999	0.0013
End 50% vs Middle 50%	15	4	1.610363584	0.489236791	0.0029	0.9995
End 25% vs Middle 75%	12	7	2.435064935	0.557206538	0.0005	0.9999
End 10% vs Middle 90%	7	12	3.112033195	0.787332692	0.0007	1

Circoviridae - Circovirus PCV dataset

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	0	26	0	2.334197383	1	0.2926
Rep vs rest of ORFs	16	10	1.523507241	0.710143212	0.7366	0.4463
CP vs rest of ORFs	10	16	0.710143212	1.523507241	0.4463	0.7366
End 50% vs Middle 50%	9	17	1.588603819	3.051678803	0.8732	0.2375
End 25% vs Middle 75%	3	23	1.061602871	2.734793676	0.878	0.2895
End 10% vs Middle 90%	2	24	1.792929293	2.371409486	0.7249	0.5448

Microviridae - Microvirus dataset

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	4	57	5.05160239	2.688351996	0.2136	0.9131
vs rest of ORFs	18	39	0.406199111	0.3457248	0.1298	0.9173
vs rest of ORFs	18	39	0.412992521	0.343505181	0.1144	0.9299
vs rest of ORFs	6	51	0.23243396	0.388405869	0.6699	0.4807
vs rest of ORFs	1	56	0.109299303	0.37845371	0.9117	0.31
vs rest of ORFs	0	57	0	0.374091378	1	0.3728
vs rest of ORFs	1	56	8.43E-02	0.385526387	0.8587	0.4149
vs rest of ORFs	6	51	0.457298233	0.354168661	0.231	0.8855
vs rest of ORFs	2	55	0.2494354	0.368875934	0.6315	0.6526
vs rest of ORFs	2	55	0.680084551	0.356728406	0.2491	0.9242
CP vs rest of ORFs	6	51	0.184047554	0.409574537	0.9986	0.0047
vs rest of ORFs	7	50	0.468846168	0.351643506	0.66	0.4784
vs rest of ORFs	17	40	0.447317517	0.335808789	0.1686	0.8899
End 50% vs Middle 50%	52	32	2.829062829	1.520461861	0.0104	0.9931
End 25% vs Middle 75%	22	62	2.153321546	2.122552953	0.3598	0.7146
End 10% vs Middle 90%	6	78	1.582484898	2.188836151	0.6719	0.4625

Nanoviridae - Nanovirus dataset

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	19	8	3.689416548	1.58388477	0.0344	0.9854
End 50% vs Middle 50%	7	1	1.716279146	0.251604188	0.001	0.9999
End 25% vs Middle 75%	1	7	0.469661151	1.181655451	0.7928	0.5773
End 10% vs Middle 90%	0	8	0	1.089420194	1	0.6288

Nanoviridae - Nanovirus encoding capsid protein dataset

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	12	9	3.152781597	2.581893388	0.4827	0.684
End 50% vs Middle 50%	8	1	3.784378438	0.504694836	0.0011	1
End 25% vs Middle 75%	2	7	1.990740741	2.264860798	0.5692	0.7464
End 10% vs Middle 90%	0	9	0	2.415730337	1	0.5302

Parvoviridae - Parvovirus and Dependovirus dataset

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	19	29	4.184303855	0.996292347	0	1
NS1 vs rest of ORFs	15	14	0.295159386	0.291725923	0.1151	0.9422
VP2 vs rest of ORFs	14	15	0.291608952	0.295271231	0.9422	0.1155
End 50% vs Middle 50%	13	16	1.274185336	0.971608833	0.0499	0.977
End 25% vs Middle 75%	10	19	1.917330677	0.88559322	0.0043	0.9987
End 10% vs Middle 90%	5	24	2.474293059	0.973656481	0.0492	0.9854

Parvoviridae - Dependovirus dataset

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	0	22	0	1.151909451	1	0.9467
nonstructural protein vs rest c	6	16	0.139104044	0.323095846	0.826	0.3233
structural protein vs rest of Ol	16	6	0.323095846	0.139104044	0.3233	0.826
End 50% vs Middle 50%	14	8	1.382120652	0.786512153	0.0988	0.9531
End 25% vs Middle 75%	5	17	1.008670272	1.107937212	0.4832	0.7122
End 10% vs Middle 90%	0	22	0	1.202046036	1	0.1011

Parvoviridae - Parvovirus dataset

Test	Breakpoints No in region	Breakpoint No outside region	breakpoints/100nts in region	breakpoints/100nts outside region	Probability of fewer than expected BPs in region	Probability of more than expected BPs in region
IR vs gene	5	13	1.817404319	0.652726091	0.0193	0.9962
NS1 vs rest of ORFs	7	6	0.182196773	0.175108632	0.096	0.9671
VP2 vs rest of ORFs	6	7	0.172728145	0.18446403	0.9684	0.0932
End 50% vs Middle 50%	6	7	0.756059595	0.556959656	0.2001	0.913
End 25% vs Middle 75%	6	7	1.462365591	0.426798651	0.0151	0.997
End 10% vs Middle 90%	3	10	1.696606786	0.533735205	0.0876	0.9783