UNIVERSITE MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

ECOLE DOCTORALE : **SIBAGHE**

THESE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER II

DISCIPLINE : **Biologie Intégrative des Plantes**

présentée et soutenue publiquement

par

Pierre-Olivier DUROY

le 19 décembre 2012

Quels sont les enjeux au cours de l'évolution du bananier qui ont conduit au maintien de séquences virales de *Banana streak virus* dans son génome ?

Thèse dirigée par Mr Jean-Loup Nottéghem (Professeur – Montpellier SupAgro)
co-dirigée par Mme Marie-Line Caruana (Chercheur – CIRAD – Montpellier)

JURY :
Président : Mr Christophe Brugidou (Directeur de recherche – IRD – Montpellier)
Rapporteur : Mr Daniel Prat (Professeur – Université Lyon 1)
Rapporteur : Mr Salah-Eddine Bouzoubaa (Maitre de conférence – IBMP – Strasbourg)
Examinateur : Mr Pierre Capy (Professeur - Université de Paris-Sud 11 – Orsay)
Examinateur : Mr Jean-Michel Drezen (Directeur de recherche – UMR CNRS IRBI – Tours)
Directeur de thèse : Mr Jean-Loup Nottéghem (Professeur – Montpellier SupAgro - Montpellier)
Co-directrice de thèse : Mme Marie-Line Caruana (Chercheur – UMR BGPI CIRAD – Montpellier)

UMR - BGPI
Biologie et Génétique
des Interactions Plante-Parasite

Laboratoire d'accueil

Cette thèse a été réalisée dans l'équipe **2B2E**
'**B**iodiversité des **B**adnavirus **E**ndogènes et **E**xogènes'

UMR BGPI
'Biologie & Génétique des Intéractions Plante-parasite'

UMR 54 - CIRAD-INRA-SupAgro
TA A-54 / K
Campus international de Baillarguet
34398 Montpellier Cedex 5
France

Sous la direction de Marie-Line Caruana et Jean-Loup Nottéghem

et

Grâce au soutien financier du CIRAD

cirad

Résumé :

Le génome du bananier (*Musa sp.*) est envahi par un nombre important de séquences de *Banana streak virus* (BSV), virus à ADN double brin de la famille *Caulimoviridae* qui n'a aucune étape d'intégration au génome hôte au cours de son cycle de multiplication. La majorité de ces intégrations eBSV (endogenous BSV) est défective mais certaines sont restées fonctionnelles et peuvent être à l'origine de particules virales suite à des stress. L'objectif de ce travail de thèse est de préciser si les eBSV sont maintenus ou non dans le génome *Musa balbisiana* des bananiers et d'étudier les conséquences évolutives que cela engendre. Nous avons tout d'abord caractérisé les eBSV fonctionnelles pour trois espèces BSV (*Banana streak goldfinger virus* (BSGFV), *Banana streak obino l'ewai virus* (BSOLV), *Banana streak imove virus* (BSImV) présentes dans le génome du bananier modèle *M. balbisiana cv* Pisang Klutuk Wulung (PKW). Nous avons montré que les intégrations eBSGFV et eBSOLV étaient di-alléliques avec un seul allèle fonctionnel à chaque fois, contrairement à eBSImV qui est mono-allélique et pour lequel nous n'avons pas pu identifier l'allèle à l'origine de l'infection. Leur contexte génomique d'intégration diffère avec une co-localisation d'eBSGFV et d'eBSOLV sur le chromosome 1 et d'eBSImV sur le chromosome 2. Ces résultats nous ont permis de développer les outils moléculaires nécessaires à la caractérisation de ces trois eBSV dans la diversité de *M. balbisiana*. Cette caractérisation a révélé la diversité de structures des eBSV et éclairé une partie encore inconnue de la phylogénie de l'espèce *M. balbisiana*. Dans un second temps nous avons abordé les mécanismes de régulation des eBSV. Ce travail a porté sur les mécanismes d'ARN interférent pouvant expliquer le maintien des eBSV dans le génome des bananiers. Cette analyse révèle que les eBSV sont effectivement sous contrôle d'un mécanisme de type ARNi et la forte production de petits ARNs de 24nt ciblant les eBSV suggère qu'il s'agit d'un silencing au niveau transcriptionnel (TGS). En parallèle, nous avons aussi recherché les mécanismes mis en place par les bananiers non-porteurs d'eBSV en cas d'infection afin de connaître les défenses constitutives des bananiers face à une attaque virale BSV. Nous avons, sur la base de ces résultats, proposé un modèle de régulation des eBSV et des BSV et discuté de l'impact que ces mécanismes auraient pu avoir sur l'évolution des eBSV. L'ensemble des données de ce travail ont permis de préciser les étapes évolutives qu'ont connues les eBSV dans le génome du bananier, expliquant le maintien que l'on observe aujourd'hui.

Mots-clés : Bananier (*Musa sp.*), *Banana streak virus* (BSV), Silencing, Phylogénie, Endogenous Pararetrovirus (EPRV).

ABSTRACT:

The nuclear genome of banana plants is invaded by numerous viral sequences of *banana streak virus* (BSV), a DNA virus belonging to the family *Caulimoviridae,* which does not require integration for its replication. These endogenous BSV (eBSV) are mostly defective; however, some can release a functional viral genome following activating stresses. The objectives of this work were to identify whether the eBSV are maintained or not in the *M. balbisiana* genome and to study the impacts of this on the evolution of banana plants. First, we characterized three functional eBSV sequences present within the *Musa balbisiana* cv PKW genome: (*Banana streak goldfinger virus* (BSGFV); *Banana streak obino l'ewai virus* (BSOLV) ; and, *Banana streak imove virus* (BSImV). We show that eBSOLV and eBSGFV are di-allelic with just one functional allele contrary to eBSImV which are mono-allelic and for which we cannot identified the functional allele. Their genomic areas of integration are different and we also observe that eBSOLV and eBSVGFV are both on chromosome 2 whereas eBSImV is on chromosome 1. These results allowed us to develop the molecular tools required for the characterization of these 3 functional eBSVs within the diversity of *M. balbisiana*. This characterization has revealed the structural diversity of eBSV and has thus clarified previously unresolved details of *M. balbisiana* phylogeny. Secondly, we studied the regulatory mechanism of eBSV expression. This work investigated if RNA interference (RNAi) mechanisms could explain the maintenance of eBSV in the *Musa* genome. Our analyses have shown that, as expected, eBSV was under the control of RNAi mechanisms and the strong production of 24nt small RNAs that target eBSV suggests that Transcriptional Gene Silencing (TGS) was involved in this control. In parallel, we investigated the mechanisms implicated in the anti-viral defense during a BSV infection on a banana plant without eBSV in order to understand the constitutive defense of banana plants. On the basis of these results we have proposed a regulation model of eBSV and BSV and we discuss the impact of silencing regulation on eBSV evolution. Data accumulated during this work have clarified several steps in the co-evolutionary history of *Musa sp*. and eBSV and explain the maintenance of eBSVs in *Musa* genomes that we observe today.

Key words: Banana Plant (*Musa* sp.), *Banana streak virus* (BSV), Silencing, Phylogeny, Endogenous Pararetrovirus (EPRV).

Remerciements

Je tiens tout d'abord à remercier les membres de mon jury de thèse Christophe Brigidou, Jean-Michel Drezen et Pierre Capy, ainsi que mes rapporteurs Daniel Prat et Salah-Eddine Bouzoubaa, d'avoir accepté de consacrer du temps à l'évaluation de mon travail.

Je remercie le CIRAD pour son soutien financier.

Je remercie aussi Jean-Loup Nottéghem d'avoir accepté d'être mon directeur de thèse et d'avoir participé à mes comités de thèse.

Un immense merci à Marie-Line Caruana qui m'a tout d'abord donné l'opportunité de réaliser cette thèse et qui ensuite a toujours été présente et de bons conseils dans les bons et les moments plus difficiles. Merci pour ton écoute, ton soutien sans faille tout au long de ces trois ans et surtout pour ta gentillesse et ton ouverture d'esprit qui m'ont énormément apporté.

Un très grand merci à Matthieu Chabannes qui a aussi largement contribué à ce travail de thèse par ces conseils toujours judicieux, sa réflexion scientifique de qualité et surtout pour sa bonne humeur et la bonne ambiance que tu as su mettre dans mon travail et les quelques superbes excursions que nous avons faites ensemble.

Je remercie tout particulièrement Nathalie Laboureau qui a su m'accueillir dans le labo de l'équipe, qui a toujours été présente et prête à m'aider dans les moments de galères et surtout qui, elle aussi, est d'une gentillesse sans faille.

Grand merci à Xavier Perrier qui a su trouver les mots pour que je comprenne la phylogénie et m'a permis de découvrir la grande, très grande diversité des bananiers. Il a lui aussi été très présent pour m'aider.

Je remercie aussi Marie Mirouze qui m'a expliqué le silencing de manière très pédagogique et m'a conseillé de façon très pertinente.

Merci aussi à Marie Umber pour son aide, son soutien à Montpellier et en Guadeloupe, avec ou sans cocktail. Et bien entendu à toute l'équipe de Pierre-Yves au Cirad Guadeloupe pour leur accueil et leur disponibilité qui m'ont aidé durant ma thèse.

Je remercie aussi ma fabuleuse stagiaire, Clémence Médina, pour qui les northern blot n'ont plus de secret, pour avoir contribué à ce travail de manière passionnée. Ce fut et de loin ma meilleure stagiaire (peut-être n'y en a-t-il pas eu d'autre ?).

Merci aussi aux membres de l'équipe 1 de BGPI avec qui on a passé de de très bons moments en particulier en Guadeloupe. Serge Galzi qui pris soin de ma collection très précieuse de bananiers et qui a réalisé des dessins superbes et ce malgré le fait qu'on l'ait un peu maltraité dans les montagnes. Emmanuelle Muller, elle aussi, pour ces bons conseils et la gentillesse dont elle a fait preuve à mon égard. Guy Noumbissié, l'autre thésard de l'équipe, avec qui on a passé de très bons moments en Guadeloupe. Et bien entendu les autres stagiaires qui sont passés parmi nous, Marc, Elisa et Abderrahman.

Je remercie particulièrement Rajeswaran Rajendran et Jonathan Seguin de l'équipe de Thomas Hohn et Mikail Pooggin à Bâle, pour m'avoir accueilli, initié et perfectionné aux northern et à la bioinformatique, et aussi à quelques spécialités indiennes.

Je remercie aussi toutes les personnes de l'UMR BGPI qui ont pu m'aider durant cette thèse. En particulier Dominique et Marie-Carmen pour avoir été toujours très efficaces face à mes demandes pas toujours simples. A Daniel aussi qui a « presque » réussi à me faire arrêter de fumer. Et Dave pour tous ces grands moments d'escalades. Et bien sur tous les autres avec qui j'ai partagé de bons moments.

Un Grand merci aussi à ceux avec qui j'ai partagé le fait de faire une thèse à l'UMR BGPI et qui ont grandement participé à la bonne ambiance de cette UMR. En particulier ceux qui sont devenus bien plus que des collègues (une thèse ça rapproche).

Tout d'abord Stella qui a toujours été présente pour me soutenir, m'aider et partager, merci pour tout, et amuse toi bien au pays des kangourous. Grand merci à mes collègues de bureau historique et qui le resteront pour longtemps encore, Mélanie et Audrey. Elles ont su m'accueillir, m'aider et surtout me soutenir pour la faire cette thèse. On a passé des très bons moments dans le bureau 124.

Je remercie aussi très très chaleureusement les autres thésards/amies pour tout les supers moments passés ensemble à BGPI mais aussi en dehors, autour d'une bière, d'un téléski ou bien sur une place de Barcelone. Merci beaucoup donc à Juliette et Jean-Phi bien entendu, Dounia, Steph, Enrique, Béranger, Flo et la petite dernière Emilie.

Grand merci aussi à mes collègues de master surtout toi Johann qui, malgré ton expatriation au pays de la saucisse, a su être présent et à la super Nono avec qui on a passé des moments magiques quand il le fallait. Tout comme Guillaume qui malgré son expatriation chez les hockeyeurs a pu revenir pour qu'on puisse fêter nos anniversaires de la meilleure des manières.

Je remercie toutes les personnes extraordinaires avec qui j'ai habité, sans qui la vie aurait été beaucoup moins drôle durant ces trois ans. Tout d'abord celui qui est là depuis le début Yvou, qui m'a fait découvrir tant de choses incroyables comme les chiroptères certes mais aussi des amis comme Vincent, Blandine ou Thomas ou 'Anne'. Il y a eu aussi Julien avec qui ont a partagé des moments au top, que ce soit au ski, à Padern ou ailleurs et qui, lui aussi, m'a fait découvrir des personnes formidables. Et enfin je remercie Blandine, Sophie, Marylène d'avoir été à mes côtés quotidiennement pendant ma thèse.

Un immense et inconditionnel merci à mes amis du brouillard et de la moutarde, de Dijon quoi. Ces personnes incroyables qui ont été présentes depuis le tout début de mon cursus scolaire (c'est le moins que l'on puisse dire pour certains) et qui sont restés à mes côtés pour que l'on continue à vivre tous ces moments extraordinaires que j'ai partagés avec vous. Et donc par ordre d'apparition dans ma vie (pas de jaloux comme ça) Françous, Oliv, Nico, Gogo (le big four en quelques sortes), Tonton, Matthias, Vaness, (sans elle je ne serais pas venu à Montpellier), Alex, Anne et Bérangère (le crew de l'Université de Bourgogne on peut dire) et Fab (il est le groupe du Jura à lui tout seul).

Merci donc à mes quatre amis d'enfance avec qui, je pense, on a tout vécu et avec qui j'espère on vivra encore plus. Et merci aussi à mes amis du 37 (c'est mieux avec …) pour toutes ces nuits de discussions, ces escapades aux bout de la France et de la Bourgogne et surtout merci d'avoir été présents.

Sans oublier celui qui est le cadeau de mon directeur de thèse et qui est né avec ma thèse, il a su être présent à sa manière, mon chat Gaspard.

Je remercie Clothilde.

Et bien sûr je remercie tout particulièrement mes parents pour leur soutien total tout au long de ce cursus scolaire un peu particulier. Et pour leur présence, leur amour et le bonheur qu'ils ont su m'apporter tout au long de ma vie.

Table des matières

Liste des figures et tableaux

Abréviations et acronymes

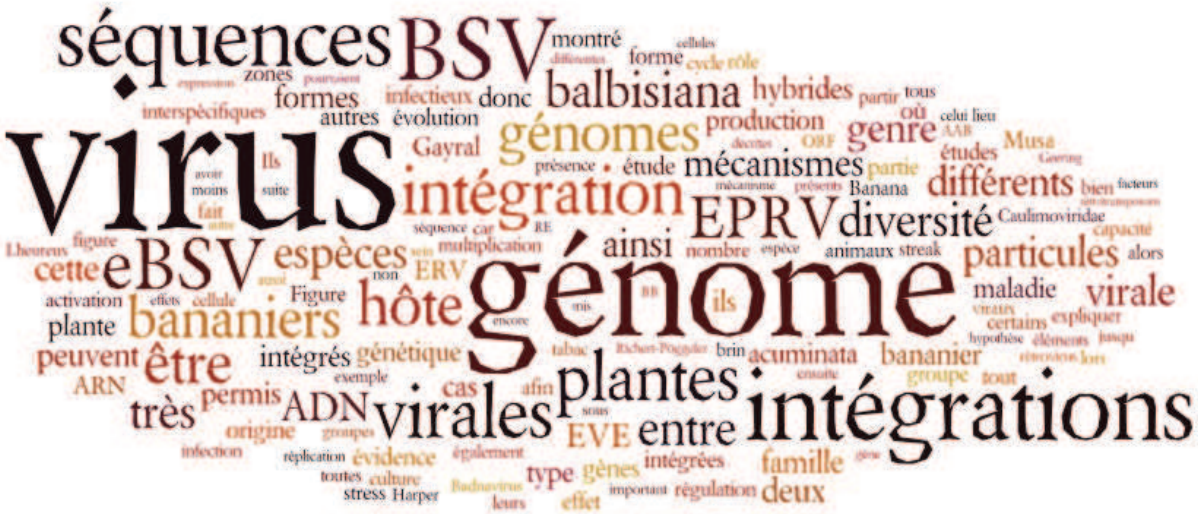
ADN :	Acide Désoxiribonucléique
AFLP	Amplified Fragment Length Polymorphism
AP :	Aspartate Protease
ARN :	Acide Ribonucléique
ARNdb	ARN double brin
ARNi :	ARN interférent
ARNm :	ARN Messager
BAC :	Bacterial Artificial Chromosom
BEL :	BSV Expressed Locus
BLAST :	Basic Local Alignement Search Tool
BSD :	Banana Streak Disease
BWA :	Burrows-Wheeler Aligner
ChIP :	Chromatin Immuno P
CIV :	Culture In Vitro
CP :	Capside Protein
CRISPR :	Clustered Regularly Interspaced Short Palindromic Repeats
dCAPS :	derived Cleaved Amplified Polymorphic Sequences
DCL1 :	Dicer-Like protein 1
DSBR :	Double Strand Break Repair
eBSGFV :	endogenous BSGFV
eBSImV :	endogenous BSImV
eBSMyV :	endogenous BSMyV
eBSOLV :	endogenous BSOLV
eBSV :	endogenous BSV
EPRV :	Endogenous Pararetrovirus
ePVCV :	endogenous PVCV
ERTBV :	Endogenous RTBV
ERV :	Endogenous Retrovirus
ET :	Element Transposable
EVE :	Endogenous Viral Element
FISH :	Fluorescent In Situ Hybridization
Flaps :	Single stranded overhanging sequences
GRD :	Geminivirus-related DNA
GVCP :	Geminiviral Coat Protein
HERV :	Human Endogenous Retrovirus
IC-PCR :	Immunocapture Polymerase Chain Reaction
ICTV:	International Committee On Taxonomy Of Viruses
IG :	region InterGenique
IN :	Integrase
ITC :	International Transit Center
LTR :	Long Terminal Repeat
LUCA :	Last Universal Cellular Ancestor
LycEPRV :	EPRV Chez Solanum lycopersicum
miRNA :	microARN
NHEJ :	Non-homologous End-joining

NsEPRV :	EPRV Chez Nicotiana sylvestris
NtoEPRV :	EPRV Chez Nicotiana tomentosiformis
ORF :	Open Reading Frame
PBS :	Primer Binding Site
PCR :	Polymerase Chain Reaction
PKW :	Pisang Klutuk Wulung
Poly (A) :	Poly Adénilation
PRV :	Pararetrovirus
PTGS :	Post-transcriptional Gene Silencing
q-PCR :	quantitatice-PCR
RdDM :	RNA-directed DNA methylation
RdRp :	RNA-dependant RNA polymerase
RE :	RétroElément
RIN :	ARN Integrity Number
RISC :	RNA induced silencing complex
RNaseH :	Ribonuclease H
RPKM :	Reads Par Kilobase par Million
RPM :	Read Par Million
RT :	Reverse Transcriptase
siARN :	Small Interfering RNA
SotuEPRV :	EPRV Chez Solanum tuberosum
TGS :	Transcriptional Gene Silencing
TSD :	Target-sites Duplication
vsARN :	petit ARN viraux

Liste des virus

BSCaV :	*Banana streak cavendish virus*
BSGFV :	*Banana streak goldfinger virus*
BSImV :	*Banana streak Imové virus*
BSMyV :	*Banana streak Mysore virus*
BSOLV :	*Banana streak obino l'ewai virus*
BSPeV :	*Banana streak perou virus*
BSV :	*Banana streak virus*
BSVNV :	*Banana streak vietnam virus*
BSYnV	*Banana streak Yunnan virus*
CaMV :	*Cauliflower mosaic virus*
CMBV :	*Citrus mosaic virus*
CMBV :	*Citrus mosaic bacilliform virus*
CMV :	*Cucumber mosaic virus*
ComYMV :	*Commelina yellow mottle virus*
CSSV :	*Cacao swollen shoot virus*
CVMV :	*Cassava vein mosaic virus*
DENV :	*Dengue Virus*
DMV-D10 :	*Dahlia mosaic virus D-10*
HTDV :	*Human teratocarcinoma-derived virus*
KRV :	*Kamiti River virus*
KTSV :	*Kalanchoe top-spotting virus*
MuLV:	*Murine leukemia virus*
PSTV :	*Potato splindle tuber viroid*
PVCV :	*Petunia vein clearing virus*
PVY :	*Potato virus Y*
RTBV :	*Rice tungro bacilliform virus*
SCBMV :	*Sugarcane bacilliform Mor virus*
ScBV :	*Sugarcane bacilliform virus*
TGMV :	*Tomato golden mosaic virus*
TMV :	*Turnip mosaic virus*
TVCV :	*Tobacco vein clearing virus*
TyMV :	*Turnip yellow mosaic virus*
WNV :	*West-Nile Virus Yellow Fiever Virus*
YFV :	*Yellow Fiever Virus*

INTRODUCTION



Les virus sont les organismes biologiques les plus abondants et génétiquement les plus divers existant sur terre (Bergh et al., 1989). Ce sont des parasites qui sont présents dans tous les milieux et écosystèmes et qui peuvent infecter tous les types d'organismes vivants ainsi que d'autres virus (Braitbart et al., 2005). L'étendue réelle de cette diversité n'a été appréhendée que récemment, car durant très longtemps, l'identification de virus n'a été abordée en grande partie qu'au travers des effets qu'ils provoquaient sur les êtres vivants et plus particulièrement les effets délétères. Les études récentes basées sur des approches de métagénomiques virales, que ce soit en milieu marin (Suttle et al., 2005) ou terrestre (Kristensen et al., 2010), ont permis d'accéder sans a priori à l'identification d'une large gamme de nouveaux virus et d'entre apercevoir l'étendue de la diversité qu'il reste encore à découvrir (Roossinck, 2011).

Les virus, de par leurs interactions avec le monde vivant en tant que parasites, ont participé et participent encore grandement à la diversité génétique existante. Ils apparaissent ainsi comme un des facteurs d'évolution des génomes hôtes de par leur mode d'interactions avec l'individu qu'ils infectent. En effet, afin d'améliorer et/ou du moins de maintenir, tant à l'hôte qu'au virus, la capacité de se multiplier une optimisation de l'interaction sous forme de course continuelle, nommée « course à l'armement »,a lieu. Les virus semblent s'être adaptés en ayant développé toutes sortes de mécanismes leur permettant d'être toujours plus efficaces dans cette quête. La survie du virus étant essentielle, les populations virales se doivent d'être transmises à la descendance de l'individu infecté (transmission dite « verticale ») et/ou contaminer d'autres individus de la même espèce (transmission dite « horizontale »). La grande majorité des virus se multiplie en contaminant de nouvelles cellules et est transmise d'un individu à un autre de manière horizontale par un partenaire vecteur extérieur tel que des insectes, acariens, nématodes... La transmission à l'ensemble de la descendance de génération en génération est rare car elle oblige une contamination des organes de multiplication. C'est par exemple le cas des potyvirus, du pois que l'on retrouve dans la graine cette plante ce qui lui permet d'infecter l'embryon (Wang et Maule, 1994). Quelques virus cependant ont développé, en interaction avec leur hôte, des échanges de matériel génétique pouvant assurer ce transfert à la descendance. Ainsi, les avancées récentes des méthodes de séquençage et d'analyses bio-informatiques ont permis de mettre en évidence l'ubiquité de tels échanges. Des résidus de séquences d'origine virale ont pu être identifiés pour tous ses êtres vivants dont le génome a été séquencé, allant des séquences les plus ancestrales des éléments transposables jusqu'aux séquences de virus complets et fonctionnels comme ceux étudiés dans cette thèse. L'identification de

séquences virales présentes dans le génome de leur hôte est récente et par la même, la compréhension de leur rôle potentiel pour l'hôte ainsi que leur évolution n'en sont qu'à leurs balbutiements. Une question préalable essentielle est de reconnaître s'il s'agit d'un cas de mutualisme entre l'hôte et le virus ou bien de simples insertions accidentelles non corrigées n'ayant pas d'intérêt de part et d'autre.

1-Qu'est ce qu'un virus et d'où vient-il ?

Un virus est défini comme un parasite cellulaire obligatoire n'ayant aucun métabolisme propre, qui utilise les structures de l'hôte pour sa réplication et son assemblage en particules virales ou virions. Ces particules sont constituées d'acide(s) nucléique(s) (une ou plusieurs molécules simple brin ou double brin, d'ADN ou d'ARN, circulaire ou linéaire) porteur(s) d'informations génétiques, protégé(s) par une structure protéique : la capside, parfois entourée d'une membrane lipidique : l'enveloppe. Les particules virales constituent le moyen utilisé par le virus de se transmettre d'une cellule à une autre et infecter l'organisme dans sa totalité, puis d'un organisme à l'autre pour infecter d'autres individus de la population hôte. En dehors de la cellule hôte, les particules virales sont inertes et incapables de se multiplier. Cette incapacité du virus à se multiplier par lui-même, clause explicite du statut du vivant, amène certains à ne pas reconnaître les virus comme des êtres vivants. Par ailleurs la classification des êtres vivants utilise en majorité les séquences ribosomales, d'où l'impossibilité d'y positionner les virus qui n'en possèdent pas.

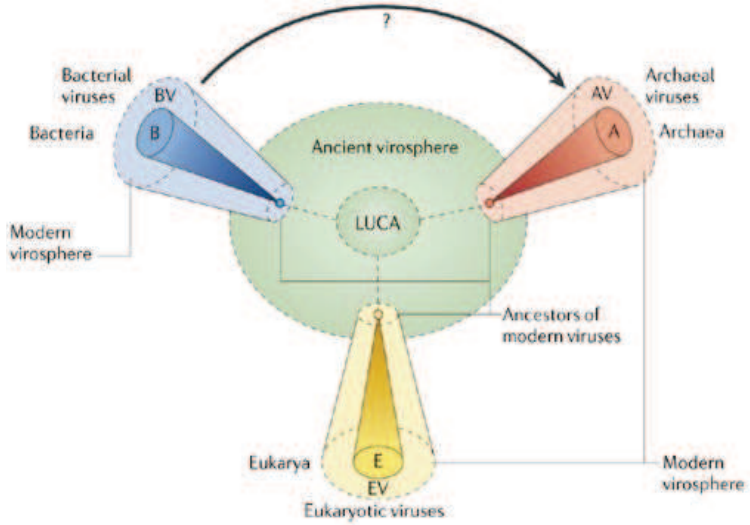


Figure 1-1 : Hypothèse de l'origine pré-LUCA des virus

Les virus peuvent avoir existés avant la vie cellulaire et auraient formé une virosphère importante expliquant la diversité actuelle. Les virus, une fois la vie cellulaire apparue, auraient continué d'évoluer avec leurs hôtes dans les trois domaines de la vie.
LUCA (Last Universal Cellular Ancestror). *D'après Prangishvili et al., (2006)*

Malgré ces conclusions, des études phylogénétiques ont permis de proposer des hypothèses sur l'origine des virus. L'hypothèse la plus communément admise est celle de « l'escape gene theory », c'est à dire que des fragments de matériels génétiques se seraient échappés de cellules hôtes ancestrales et seraient devenus par la suite des parasites (Forterre, 2006). Cette théorie aurait eu lieu dans un contexte pré-LUCA (Last Universal Cellular Ancestor) et les virus auraient évolué dans ce contexte pour former une virosphère complexe à l'origine et au-delà de la diversité que nous connaissons aujourd'hui (Pranggishvili et al., 2006) (figure 1-1). Les virus auraient évolué génétiquement de façon indépendante selon le règne auquel appartenait leur hôte tout en continuant leur adaptation à ce même hôte. L'hypothèse de Forterre (2006) est l'hypothèse la plus adaptée pour expliquer toute la diversité présente au sein de la virosphère comme, par exemple, le fait que les virus utilisent différents types de molécules pour coder leur information génétique. Les virus se seraient ensuite spécialisés une fois LUCA apparu en développant différentes méthodes de multiplication, en s'adaptant à différents hôtes et en utilisant différents moyens pour se propager, la course à « l'armement »contribuant de façon très importante au développement de toute cette diversité. Cette capacité évolutive réside également dans la petite taille de leur génome qui est plus petite que celle des êtres vivants à l'exception des mimivirus, et surtout qui présenterait des taux d'erreurs plus importants lors de la réplication (Gago et al., 2009). Néanmoins, l'étude des relations phylogénétiques comprenant toute la diversité virale connue reste très complexe car les virus très diversifiés, ne possèdent pas de structures génétiques ubiquitaires permettant d'entreprendre ces études de manières classique, ce qui complexifie grandement la compréhension de l'origine des virus.

Cette abondance de diversité a mené les virus jusqu'à des cas de parasitisme extrêmes où, tout comme un retour à la mère par rapport à « l'escape gene theory », la multiplication virale oblige à l'intégration du patrimoine génétique dans celui de l'hôte. L'étude des virus intégrés dans le génome hôte apparaît dès lors comme une donnée essentielle pour aborder et tenter de comprendre la diversité virale. Les différents virus découverts intégrés sont présentés dans la figure 1-2. L'intégration obligatoire représente pour certains virus une étape à part entière pour se multiplier et se propager de façon plus efficace (Gifford et Tristem, 2003). Cette intégration dans le génome hôte peut se traduire par des marques ou jalons qui témoignent comme autant de traces des virus passés. L'étude de ces traces virales fossiles peut nous permettre d'augmenter la connaissance de la biodiversité virale existante en retraçant son histoire évolutive.

Group/type	Family or genus	Taxa	Number per haploid genome
Group I/dsDNA	Baculovirus	Insects	Unknown (hybridization data, no sequencing)
Group I/dsDNA	Herpesviridae	Humans	1
Group I/dsDNA	Nudivirus	Parasitic wasps	Several
Group I/dsDNA	Phycodnaviridae	Brown algae	1
Group II/ssDNA	Circoviridae	Mammals	1 to 2
Group II/ssDNA	Geminiviridae	Tomentosae (tobacco and three other species)	5 to 120
Group II/ssDNA	Parvoviridae	Mammals; shrimp	1 to 3
Group III/dsRNA	Partitiviridae	Plants; arthropods; Protozoa	1 to 4
Group III/dsRNA	Reovirus	*Aedes* spp. mosquitoes	1
Group III/dsRNA	Totiviridae	Fungi; plants; ticks	1 to 6
Group IV/+ssRNA	Dicistroviridae	Honeybees	1
Group IV/+ssRNA	Flaviviridae	Medaka fish; mosquitoes	1 to 4
Group IV/+ssRNA	Potyviridae	Grapes	Several
Group V/–ssRNA	Bornaviridae	Vertebrates	1 to 17
Group V/–ssRNA	Bunyaviridae	Ticks	14
Group V/–ssRNA	Filoviridae	Mammals	1 to 13
Group V/–ssRNA	Nyavirus	Zebrafish	6
Group V/–ssRNA	Orthomyxoviridae	Ticks	1
Group V/–ssRNA	Rhabdoviridae	Insects (ticks and mosquitoes)	1 to 28
Group VI/ssRNA-RT	Retroviridae	Vertebrates	Several hundreds to several hundreds of thousands
Group VII/dsDNA-RT	Hepadnavirus	Passerine birds	15
Group VII/dsDNA-RT	Pararetrovirus	Plants	A dozen to a thousand

+, positive sense; –, negative sense; RT, reverse transcriptase.

Figure 1-2 : Virus décrits comme intégrés dans le génome de leur hôte en dehors des rétroéléments

(+) : virus avec un acide nucléique de sens positif ; (-) : virus avec un acide nucléique de sens négatif ; RT : reverse transcriptase. *D'après Feschotte et Gilbert, (2012)*

2-Les Endogenous Virals Elements : rôle et évolution chez les animaux

2-1 Les éléments transposables (ET) premières pièces du puzzle

Les premières séquences virales retrouvées dans le génome des êtres vivants sont celles d'éléments transposables et plus particulièrement celles de rétroéléments (RE) avec une Longue région Terminale Répétée (LTR). Ils ont été découverts sans savoir qu'il s'agissait de séquences virales primitives et le lien avec les rétrovirus a été démontré par la comparaison de séquences des gènes de la Reverse Transcriptase (RT) (Xiong et Eichbush, 1990) et de la ribonucléase H (Malik, 2001) (Figure 1-3). Ces éléments utilisent la même stratégie pour se multiplier que les rétrovirus et produisent des particules pseudo-virales. Cependant ces pseudo-particules ont un cycle unique de réplication et sont incapables d'infecter les cellules voisines comme le font habituellement les virus. Toutefois, de par le grand nombre de gènes en commun à l'exception de celui de l'enveloppe, et dernièrement l'identification d'un ancêtre commun avec les rétrovirus (Grandbastien, 2008), les RE à LTR sont considérés aujourd'hui comme des virus. Ils font partie depuis 2005 de la classification des virus sous forme de deux familles : les RE à LTR *Ty3-gypsy* dans la famille des *Metaviridae* et les RE à LTR *Ty1-copia* dans la famille des *Pseudoviridae* (Fauquet, 2005) (Figure 1-4). De manière surprenante, bien que ne pouvant pas sortir de la cellule infectée, les RE à LTR sont présents chez un très grand nombre d'espèces de toutes les familles du vivant (Levin et Moran, 2011). Et beaucoup d'exemples de l'influence des ET sur les génomes sont maintenant connus et permettent d'expliquer leur conservation. Tout d'abord, ils peuvent avoir une influence sur la taille des génomes, l'exemple le plus connu étant celui du maïs où les ET représentent 80% du génome total de la plante (San Miguel et al., 1998). Chez l'espèce de riz *Oriza australiensis* ils peuvent provoquer un doublement de la taille du génome (Piegu et al., 2006). Nous savons ensuite, que les ET de par leur capacité à transposer, peuvent avoir des effets sur le fonctionnement des génomes. Ces effets peuvent être délétères lorsque les ET se retrouvent intégrés au sein ou à proximité de gènes ou lorsque que leur régulation ne se fait plus correctement et qu'ils envahissent le génome. Malgré tout, la sélection naturelle les élimine ou peut bloquer leur transmission. Cependant, leur ubiquité au sein des génomes peut s'explique par des effets bénéfiques pour ces derniers. Tout d'abord, nous savons que ces éléments peuvent accélérer la réponse à la sélection en produisant de la diversité génétique (Chao et al., 1983), Ils peuvent être recrutés par des espèces afin de créer de nouveaux gènes fonctionnels (Lin et al., 2007 ; Hudson et al., 2003), ou peuvent faciliter la création de nouveaux gènes par l'intermédiaire de la RT qui en faisant plus d'erreurs que celle

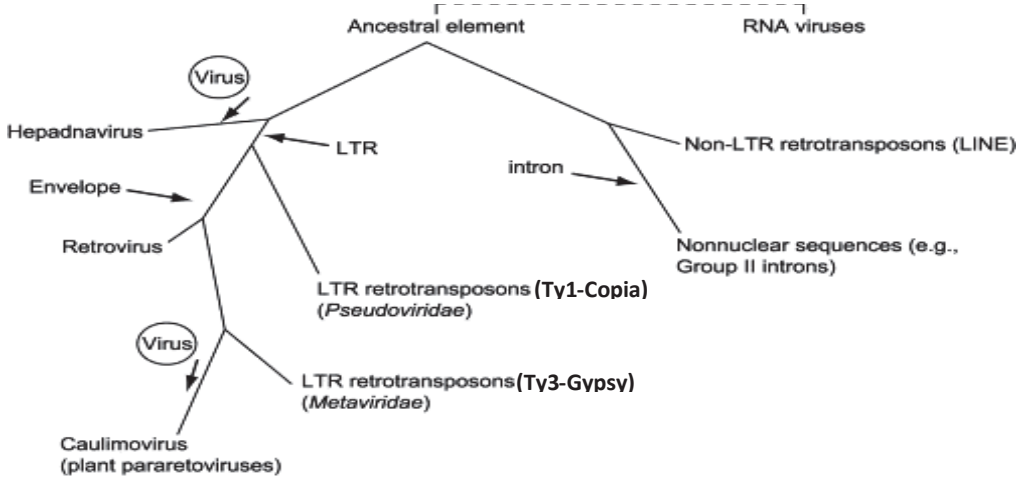


Figure 1-3 : Relations et origines des éléments porteurs de reverse transcriptase (RT)

Basée sur les alignements de séquences RT de Xiong et Eichbush, (1990).

D'après Hansen et Heslop-Harrison, (2004)

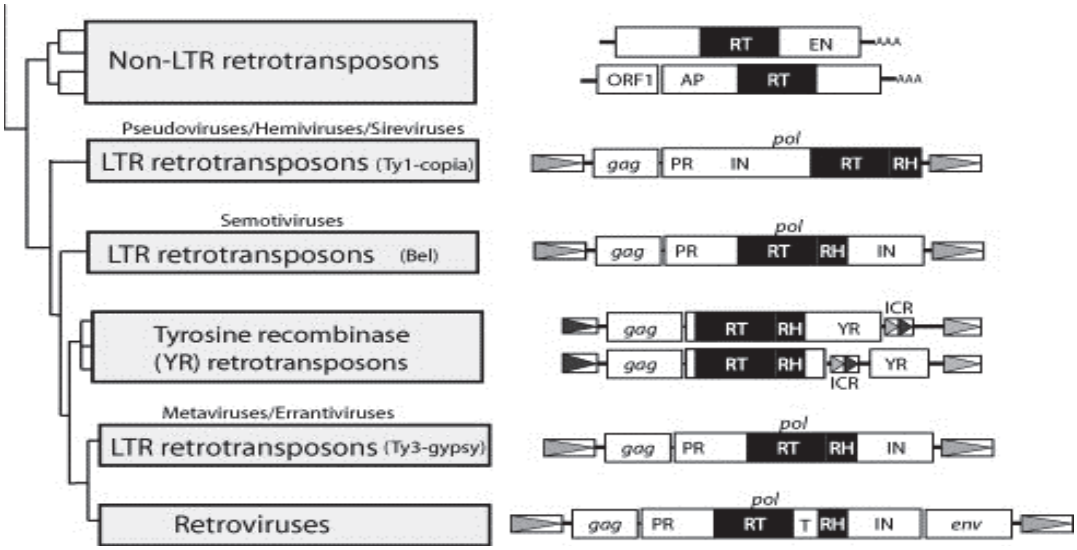


Figure 1-4 : Relation entre la structure et phylogénétique des différents groupes de rétrotransposons

Côté gauche: relation phylogénétique des rétrotransposons basée sur la séquence de leurs domaines de la transcriptase inverse.

Côté droit: structures communes pour les éléments de chaque groupe. Si d'importantes variations de structure se produisent au sein d'un groupe deux structures sont schématisés pour représenter les types de variation les plus fréquemment observées. Les cadres de lecture ouverts (ORF) de chaque élément sont représentés par des cases horizontales. Les ORF multiples dans le même élément peuvent être dans différents cadres de lecture ou séparés par des codons de terminaison. Les ORF présentant des similitudes avec les gènes gag, pol et env du rétrovirus de vertébrés sont étiquetés comme tels. Abréviations pour les domaines codant pour des protéines: RT: domaine reverse transcriptase ; HR : domaine RNase H ; PR: protéinase ; IN: intégrase ; T : binding site ; APE: endonucléase apurinique ; AP : aspartate protease ;FR : endonucléase ; YR, domaine présentant une similitude avec les tyrosines recombinantes. Les boites avec les pointes de flèches grisées : les régions LTR ou ICR ; lignes fines, les régions non traduites des éléments; AAA, le poly (A) des queues.

Adaptée Eickbush et al., (2008)

présente chez les êtres vivants va induire de la diversité (Wang et Dooner, 2006). Ils contribuent aussi au maintien de l'intégrité du centromère (Biemont, 2009 ; Weber et Schmidt, 2009) ; ils servent comme source d'éléments cis-régulateurs et de petits ARN (Feschotte, 2008). Ils pourraient jouer un rôle dans l'effet hétérosis (Springer et Stupar, 2007).

2-2 EVE de virus ayant une étape d'intégration dans leur cycle de multiplication

Pour certains virus d'animaux ou de bactéries, intégrer le génome de l'hôte fait partie du cycle de réplication virale. Les plus connus sont les « temperate phages » mais d'autres exemples existent en particulier chez les virus d'animaux comme par exemple les virus à ADN doubles brins polydnavirus, les Adeno associated virus, et les rétrovirus qui sont des virus ARN simple brin.

2-2-1 Les « temperate phage »

Les phages bactériens peuvent être virulents et lytiques lors d'infection bactérienne ou quiescents en intégrant le génome de la bactérie sous forme de provirus, ces derniers sont appelés phages tempérés. Lors d'une infection, certaines bactéries vont être colonisées par des virus lytiques qui une fois multipliés vont produire la lyse de la cellule et la libération de virions néoformés. Chez d'autres cellules bactériennes pour lesquelles les conditions environnementales sont moins intéressantes pour le virus, les phages infectant la cellule vont intégrer le génome bactérien sous forme de provirus ou prophage et entrer dans une phase de latence où leurs gènes vont être réprimés. Le génome bactérien devient alors lysogène, et va pouvoir être transmis à sa descendance. Le prophage va redevenir lytique dès que les conditions environnementales lui seront favorables (Brussow et al., 2004) (Figure 1-5).

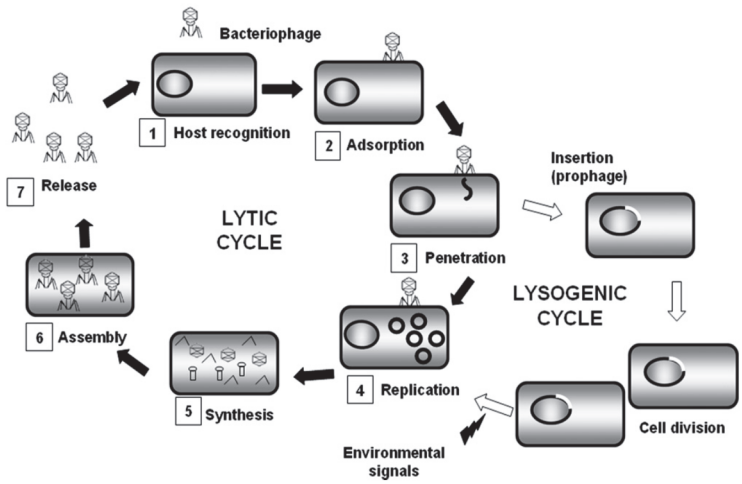


Figure 1-5 : Cycles de multiplication des phages tempérés
Sont représentés sur le schéma les cycles de vie lytiques ou lysogéniques que peuvent suivre les phages tempérés. *D'après Garcia et al., (2010)*

Ces étapes d'intégration au génome hôte peuvent être l'occasion de transferts verticaux de gènes qui vont permettre l'acquisition de nouvelles fonctions aux cellules bactériennes et participer à l'évolution des bactéries (Brussow et al., 2004). Ainsi, il a été montré que la *staphylococcal enterotoxin* A produite par la bactérie *Staphylococcus aureus* provient du génome d'un temperate phage infectant cette bactérie (Betley et Mekalanos, 1985). L'interaction bactérie/prophage peut même être de type symbiotique lorsque le prophage intégré est complet et fonctionnel (Wagner et Waldor, 2002). Enfin, sur des temps d'évolution plus longs, les bactéries peuvent « domestiquer » les prophages en utilisant les fragments viraux qui leurs sont nécessaires (Banks et al., 2002) et réprimer leur fonction lytique. C'est le cas des séquences de type « Clustered Regularly Interspaced Short Palymdromic Repeats » (CRISPR) présentes chez un grand nombre de bactéries, qui dérivent de prophages et sont impliquées dans le phénotype de résistance aux phages par l'intermédiaire de mécanismes épigénétiques (Barrangou et al., 2007 ; Brouns et al., 2008).

2-2-2 Les Endogenous Retrovirus (ERV)

Les rétrovirus sont des virus à ARN simple brin, protégés par une capside et une enveloppe lipoglycoprotéique. L'intégration au génome de l'hôte sous forme de provirus fait partie de leur cycle de multiplication virale. Elle se réalise grâce à la reverse transcriptase-RT qui permet la réplication de l'ARN en ADN pour pouvoir intégrer le génome hôte grâce à une intégrase. Les rétrovirus sont retrouvés chez tous les mammifères et chez un grand nombre de vertébrés. Lorsqu'une cellule germinale est infectée par le virus, celui-ci se retrouve intégré dans le génome de la cellule et va ensuite être transmis à la descendance comme n'importe quel gène de la cellule. Ce phénomène de transfert vertical résulte d'une endogénisation du provirus qui devient alors un Endogenous Retrovirus (ERV) (Dewannieux et al., 2006). Il a été montré par exemple que le génome humain est composé à 7-8% d'ERV (Horie et Tomonaga, 2011). Ces intégrations sont le plus souvent le résultat d'infections répétées et d'endogénisations anciennes pouvant dater de plusieurs millions d'années (Johnson et Coffin, 1999). Tarlington et al. (2006) ont suggéré que le processus d'infection et d'endogénisation du rétrovirus du koala, initié il y a une centaine d'années dans le nord de l'Australie, était toujours en cours, ce qui en fait un sujet d'étude unique des facteurs influençant l'endogénisation rétrovirale.

Les ERV découverts chez les animaux en dehors du laboratoire ont tous été découverts comme ne pouvant pas produire de particules virales (Jern et Coffin 2008 ; Feschotte et Gilbert, 2012). En effet, différents mécanismes de régulations sont mis en place par l'hôte

afin de réguler leur expression et, par là même, la multiplication des ERV dans les génomes. Ces mécanismes font appel au « Post-Insertional Gene Rearrangement » (Hughes et Coffin 2001), ou bien aux mécanismes d'ARN interférant (Thomas et Schneider, 2011 ; Rowe et Trono, 2011). L'invasion des génomes par les ERV peut alors résulter de mécanismes habituels de duplication ou de translocation des rétroéléments (RE) qui se retrouvent intégrés en un nombre important de copies. Il a d'ailleurs été très récemment montré le lien entre l'absence du gène d'enveloppe chez les ERV et leur potentialité à envahir les génomes comme des éléments transposables (Magiorkinis et al., 2012). Ce qui tend à montrer qu'une fois leur potentiel d'infection virale systémique perdu, ces intégrations ne sont pas éliminées et peuvent rester dans le génome. C'est par exemple ce qui a été décrit durant l'évolution des primates et qui explique le nombre important d'ERV présents actuellement dans le génome humain (Gifford et Tristem., 2003).

Tout comme pour les « temperate phage » ou les rétroéléments, l'importance des ERV dans l'évolution du génome de leur hôte a pu être démontré car ils sont une source importante de diversité génétique pour l'adaptabilité de l'hôte. Ils ont ainsi apporté un grand nombre de promoteurs forts dans le génome humain, ainsi que de nombreux sites d'accroches de facteurs de transcription utilisés ensuite par les gènes situés à proximité (Jern et Coffin 2008). Il a été mis en évidence que chez les bactéries, les ERV sont utilisés pour la course à « l'armement » afin de lutter contre les virus libres. L'étude récente de Aswad et Katzourakis (2012) recense les différents gènes viraux utilisés par les animaux dans la défense anti-virale et tend à montrer l'impact de ces éléments dans cette course à « l'armement » entre les virus et leurs hôtes. C'est le cas par exemple chez la souris où le « Friend virus susceptibility 1 gene » permet la répression du *Murine Leukaemia Virus* (Mura et al., 2004). Mais, l'exemple le plus intéressant d'utilisation de gènes viraux est la participation du gène de la syncitine au développement et à la physiologie du placenta des mammifères. Ce gène dérive de façon certaine de gènes d'enveloppe de différents ERV chez les animaux placentaires (Mi et al., 2000) et chez l'homme, provient du rétrovirus endogène HERV-W. Cet ERV a été désarmé pour éviter la production de particules virales mais a conservé la capacité de produire la protéine d'enveloppe qui a, par la suite, été domestiquée et participe au développement du placenta chez la femme enceinte. Cet exemple illustre le détournement des ERV au profil de leur hôte.

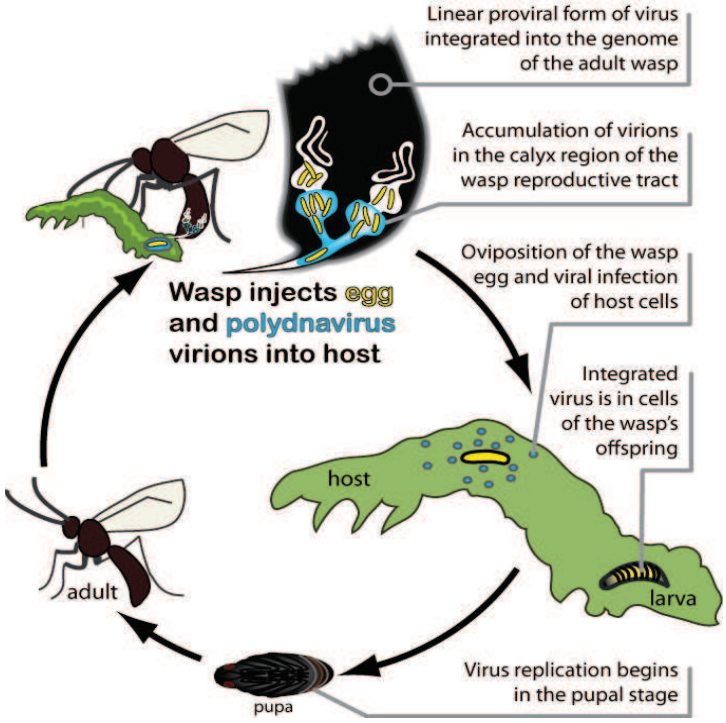


Figure 1-6 : Cycle de vie d'une guêpe parasitoïde et des polydnavirus (PDV) intégrés dans son génome lors du parasitage d'une larve de lépidoptère

D'après Burke and Strand, (2012)

Family (genus)	Replication	Exogenous host range	EVE host classes
DNA viruses			
Parvoviridae*	Nuclear	Mammals, birds	Mammals
Dependovirus	-	Mammals, birds	-
Parvovirus	-	Mammals	-
Amdovirus	-	Mammals	-
Circoviridae*	Nuclear	Mammals, birds	Mammals
*Hepadnaviridae**	Nuclear	Mammals, birds	Birds
RNA viruses			
Bornaviridae*	Nuclear	Mammals, birds	Mammals
Filoviridae	Cytoplasmic	Mammals	Mammals
Bunyaviridae	Cytoplasmic	Vertebrates, insects	Insects
Nairovirus	-	Vertebrates, insects	-
Phlebovirus	-	Mammals, insects	-
Rhabdoviridae*	Cytoplasmic	Mammals, birds, insects	Insects
Orthomyxoviridae	Nuclear	Mammals, birds, insects	Insects
Reoviridae	Cytoplasmic	Mammals, birds, insects	Insects
Flaviviridae*	Cytoplasmic	Mammals, birds, insects	Insects
Unclassifiable	N/A	N/A	Mammals

Figure 1-7 : Endogenous Viral Element identifiés dans le génome des animaux par Katzourakis et Gilford (2011)

* : indiquent les familles de virus avec des représentants connus pour être capables d'établir une infection persistante/latente. Les EVE pris en compte possèdent des insertions d'un génome viral complet ou de gènes ayant plus 90% d'identité avec ceux des virus connus de l'espèce considérée. Les insertions orthologues chez des espèces distinctes n'ont pas été comptabilisées. *D'après Katzourakis et Gilford, (2011)*

2-2-3 Les polydnavirus

Les polydnavirus sont des virus symbiontes obligatoires et dérivent de nudivirus ancestraux (groupe frère des baculovirus) qui aurait été intégrés dans le génome des guêpes parasitoïdes il y a 100 millions d'années (Bézier et al., 2009). Il s'agit là d'un cas extrême d'utilisation d'intégration virale puisque les génomes des ichnovirus et des bracovirus, les deux genres de la famille des *polydnaviridae*, n'existent plus que sous leur forme intégrée dans le génome de leurs hôtes hyménoptères (Drezen et al., 2003 ; Dupuy et al., 2006). L'étude des séquences virales intégrées dans le génome des différentes guêpes parasitoïdes indique une origine polyphylétique des polydnavirus qui appartiendraient à l'origine à des familles de virus distinctes (Thézé et al., 2011).

Les polydnavirus sont des symbiotes nécessaires au succès reproducteur des hyménoptères parasitoïdes du genre *Braconidae* et *Ichneumonidae*. Ils se répliquent uniquement dans les ovaires des guêpes, et sont injectés dans leur proies : des larves de lépidoptères. Des facteurs viraux dérèglent alors le système immunitaire de la chenille afin que les larves de guêpes puissent se développer (Espagne et al., 2005) (figure 1-6). Ces différents éléments tendent à montrer que des polydnavirus ont été totalement domestiqués par la guêpe en tant que porteurs d'ADN pour aller infecter la chenille. Ces virus ont perdu leur capacité à se multiplier en dehors de la guêpe, puisque les gènes nécessaires à la multiplication virale ne sont pas intégrés dans la particule virale. On peut donc les considérer plus comme des sécrétions géniques d'un organelle de guêpe que comme réellement des virus (Burke et Strand, 2012).

2-3 EVE n'ayant pas d'étape d'intégration dans leur cycle de multiplication

Un troisième type d'intégration virale EVE est celui qui correspond à des virus qui n'ont pas d'étape d'intégration obligatoire dans leur cycle de multiplication.

Une étude de Katzourakis et Gifford (2011) a permis de mettre en évidence un grand nombre d'EVE intégrés dans le génome d'animaux (mammifères, oiseaux et insectes) correspondant à 10 familles non rétrovirales. L'étude a été menée *in silico* sur tous les génomes séquencés disponibles afin d'obtenir une vision exhaustive des intégrations virales chez les animaux (figure 1-7). Elle reste néanmoins certainement en dessous de la réalité puisque seules les séquences de virus connus référencés ont été étudiées. Elle a permis néanmoins de mettre à jour que tout matériel génétique issu de tout type de génome viraux connu et de tout mode de réplication pouvait être intégré au génome hôte. La plupart des

Host plant		Endogenous viral sequence				
Plant family	Plant species	Virus species	Virus genus	Virus family	Replication competent	References
Asteraceae	*Dahlia mirabilis*	Unassigned	*Caulimovirus*	*Caulimoviridae*	?	Pahalawatta *et al.* (2008)
Bromeliaceae	*Ananas comosus*	AcomV	Unassigned		?	Gambley *et al.* (2008), Geering and Teycheney (2010)
Musaceae	*Musa acuminata*	Unassigned	*Badnavirus*		Unlikely	Ndowora *et al.* (1999), Geering *et al.* (2001, 2005a)
	Musa balbisiana	BSOLV, BSGFV, BSMysV, BSImV and other unassigned			Yes	Ndowora *et al.* (1999), Harper *et al.* (1999), Geering *et al.* (2001, 2005a,b, 2008)
	Musa schizocarpa	Unassigned			Unlikely	Geering *et al.* (2005a)
Poaceae	*Oryza sativa*	OsatV	[2]Orendovirus		Unlikely	Kunii *et al.* (2004), Geering *et al.* (201[0])
Ruscaceae	*Dracaena sanderiana*	DrMV	*Badnavirus*	*Caulimoviridae*	Probably	Su *et al.* (2007)
Rutaceae	*Poncirus trifoliata*	Unassigned	Unassigned	*Caulimoviridae*	?	Yang *et al.* (2003b)
Solanaceae	*Datura* sp.	Unassigned	Solendovirus	*Caulimoviridae*	?	Jakowitsch *et al.* (1999)
	Nicotiana clevelandii	Unassigned	Solendovirus	*Caulimoviridae*	?	Matzke *et al.* (2004)
	Nicotiana edwardsonii	TVCV	Solendovirus	*Caulimoviridae*	Yes	Lockhart *et al.* (2000)
	Nicotiana glutinosa	Unassigned	Solendovirus	*Caulimoviridae*	?	Matzke *et al.* (2004)
	Nicotiana kawakamii	Unassigned	*Begomovirus*	*Geminiviridae*		Murad *et al.* (2004)
	Nicotiana otophora	Unassigned	Solendovirus	*Caulimoviridae*	?	Jakowitsch *et al.* (1999)
	Nicotiana sylvestris	TVCV	Solendovirus	*Caulimoviridae*	Unlikely	Jakowitsch *et al.* (1999), Gregor *et al.* (2004)
	Nicotiana tabacum	TVCV	Solendovirus	*Caulimoviridae*	Unlikely	Jakowitsch *et al.* (1999)
		Unassigned	*Begomovirus*	*Geminiviridae*	Unlikely	Bejarano *et al.* (1994), Murad *et al.* (2002, 2004)
	Nicotiana tomentosa	Unassigned	*Begomovirus*	*Geminiviridae*	Unlikely	Murad *et al.* (2004)
	Nicotiana tomentosiformis	TVCV	Solendovirus	*Caulimoviridae*	Unlikely	Gregor *et al.* (2004), Geering *et al.* (2010)
		Unassigned	*Begomovirus*	*Geminiviridae*	Unlikely	Murad *et al.* (2002, 2004)
	Petunia hybrida	PVCV	*Petuvirus*	*Caulimoviridae*	Yes	Richert-Pöggeler *et al.* (2003)
	Solanum cheesmaniae	Unassigned	Solendovirus	*Caulimoviridae*	Unlikely	Staginnus *et al.* (2007)
	Solanum habrochaites	TVCV	Solendovirus	*Caulimoviridae*	Unlikely	Staginnus *et al.* (2007), Geering *et al.* (2010)
	Solanum lycopersicum	TVCV	Solendovirus	*Caulimoviridae*	Unlikely	Jakowitsch *et al.* (1999), Staginnus *et al.* (2007), Geering *et al.* (2010)
	Solanum peruvianum	Unassigned	Solendovirus	*Caulimoviridae*	Unlikely	Staginnus *et al.* (2007)
	Solanum pimpinellifolium	Unassigned	Solendovirus	*Caulimoviridae*	Unlikely	Staginnus *et al.* (2007)
	Solanum tuberosum	Unassigned	Solendovirus	*Caulimoviridae*	Unlikely	Hansen *et al.* (2005)
Salicaceae	*Populus trichocarpa*	Unassigned	Unassigned	*Caulimoviridae*	?	Bertsch *et al.* (2009)
Vitaceae	*Vitis vinifera*	Unassigned	Unassigned	*Caulimoviridae*	?	Bertsch *et al.* (2009)

AcomV, *Ananas comosus* virus; BSOLV, banana streak Obino l'Ewai virus; BSGFV, banana streak GF virus; BSImV, banana streak Imové virus; BSMysV, banana streak Mysore virus; DrMV, *Dracaena* mottle virus; OsatV, *Oryza sativa* virus; PVCV: petunia vein clearing virus; TVCV, tobacco vein clearing virus.

Figure 1-8 : Endogenous Viral Element identifiés dans le génome des plantes

Les genre viraux qui ne sont pas en italique correspondent aux propositions de Teycheney et Geering pour renommer deux genres des *Caulimoviridae*. Ces nouveaux noms n'ont pas encore été acceptés par l'International Committee on Taxonomy of Virus, dans le texte ils sont donc appelés par leurs noms actuels. Les *Orendovirus* correspondent au genre des *Tungrovirus* et les *Solendovirus* à celui des *Cavemovirus.* Tous ces genres appartiennent à la famille *Caulimoviridae*. D'après Teycheney et Geering , (2011)

séquences sont hautement mutés et réarrangés et ne possèdent que très peu de séquences actives. Néanmoins, les auteurs ont identifié des séquences de virus à ARN et à ADN ayant des cadres de lecture ouverts fonctionnels ou retrouvées sous forme d'ARN messager. Ces données, en révélant de nouvelles informations sur l'histoire évolutive des virus, vont certainement alimenter et élargir le cadre des études actuelles de paléo-virologie. Elles démontrent également que les échanges de gènes entre virus et génomes animaux sont plus importants et anciens que ce qui avait été jusqu'alors envisagé. L'observation notamment chez les primates d'une séquence virale dérivée des bornavirus EBLN-1 et maintenue intacte chez plusieurs espèces au-delà de plusieurs millions d'années témoigne de la possibilité que ces intégrations virales puissent jouer un rôle plus important sur l'évolution du génome hôte que celui supposé précédemment (Horie et al., 2011 ; Kazourakis et Gilfford, 2011).

3- Les virus intégrés dans le génome des plantes

Les intégrations virales dans le génome des plantes sont nécessairement des intégrations accidentelles car aucun rétrovirus n'existe chez les plantes et aucun virus de plantes ne possède de cycle de réplication nécessitant une intégration dans le génome de l'hôte. Toutes les intégrations virales de plantes sont donc considérées comme étant des évènements illégitimes d'intégrations de virus présents dans le noyau car les seuls virus retrouvés à ce jour intégrés sont des virus à ADN. Ces différents virus sont présentés dans la figure 1-8.

3-1 Les virus à ADN de plante

Chez les plantes, il existe trois familles de virus à ADN : *Geminiviridae*, *Nanoviridae* et *Caulimoviridae* (figure 1-9). Les quatre genres de la famille *Geminiviridae* (Begomovirus, Curtovirus, Mastrevirus et Topocuvirus) sont classés selon leurs insectes vecteurs, leurs hôtes et l'organisation de leur génome (Fauquet, 2005). Le génome est à ADN simple brin et peut comprendre un ou deux composants. Il est converti en ADN double brin lorsqu'il entre dans le noyau et se multiplie par le mécanisme de cercle roulant ou « rolling circle ». La famille *Nanoviridae* est divisée en deux genres *(*Nanovirus et Babuvirus). Le génome de ces virus se compose de 12 ou plus, molécules d'ADN simple brin. Elles sont toutes en sens positif, sont transcrites dans une seule direction et conservent une structure de type tige boucle dans la région intergénique (Fauquet, 2005).

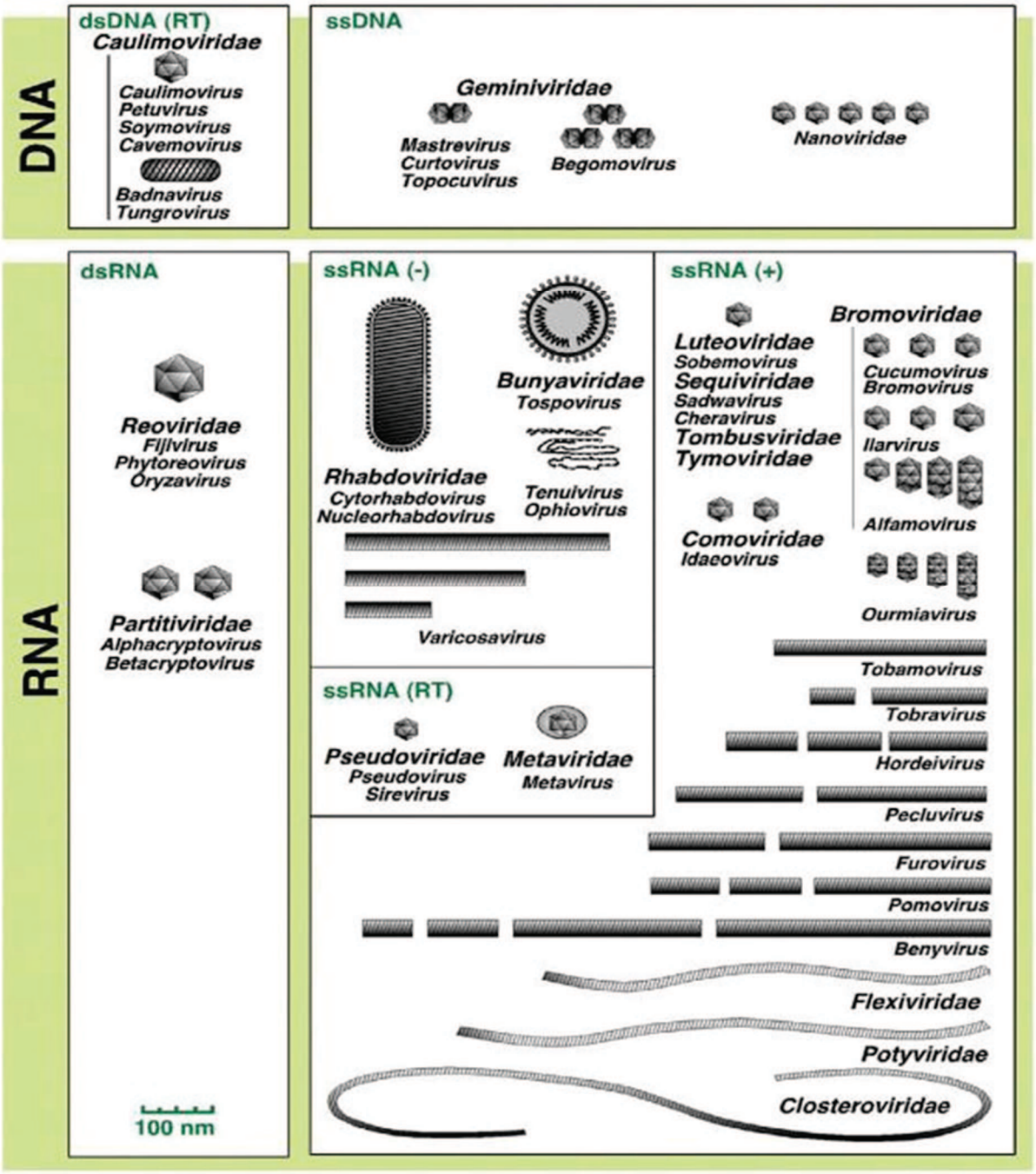


Figure 1-9 : Famille et genre des virus infectant les plantes

Adapté de Fauquet, (2005)

La famille *Caulimoviridae* appartient au sous-groupe des pararétrovirus de l'ordre des *Retrovirales* qui contient les sous-ordres *Orthoretrovirinea* (famille *Retroviridae*) et *Retrotransposineae* (familles *Pseudoviridae* et *Metaviridae).* Elle est composée de 6 genres : 4 de ces genres correspondent à des virus icosahédraux *Caulimovirus*, *Soymovirus*, *Cavemovirus* et *Petuvirus* ; et deux à des virus baciliformes *Badnavirus* et *Tungrovirus.* Ces genres sont définis selon le nombre et l'ordre des cadres ouverts de lecture (ORF) de leurs génomes (Fauquet, 2005) (Figure 1-10). Il a été mis en évidence dès le début des années 90 par Xiong et Eickbush (1990) l'existence de liens entre la famille *Caulimoviridae* et les rétrotransposons de type *Ty3-Gypsy* de la famille *Metaviridae* (Figure 1-3). En effet, ces virus et les rétrotransposons possèdent une organisation du domaine contenant le gène *pol* similaire : protease-reverse transcriptase (RT) ribonucléase H (RH) avec le ensuite le gène de l'intégrase (IN) seulement pour les rétrotransposons (Hansen et Heslop-Harrison, 2004). Les virus de la famille *Caulimoviridae,* tout comme les rétrovirus animaux, se multiplient via un mécanisme de transcription inverse. Après dissociation de la particule virale dans le cytoplasme, l'ADN viral pénètre dans le noyau cellulaire. Il est converti en ADN super-enroulé par réparation des interruptions de séquences (gaps) par les enzymes cellulaires, et associé à des histones pour former un minichromosome circulaire (Hull et al., 2001). Il est alors transcrit en ARN prégénomique redondant en partie terminale 5' et 3'.Ces redondances sont l'équivalent des séquences R des structures LTR (long Terminal Repeat) des rétrovirus. Cet ARN est exporté vers le cytoplasme de la cellule hôte où il est soumis d'une part à une transcription inverse pour produire de l'ADN et d'autre part à une traduction des gènes en protéines virales. L'ADN viral peut être encapsidé pour former une particule virale.

3-2 Découvertes des intégrations virales chez les plantes

Les premières séquences virales découvertes intégrées dans le génome des plantes correspondent à des séquences de bégomovirus. Elles ont été découvertes de manière fortuite, lors d'études sur le tabac (*Nicotiana tabacum*) pour identifier des résistances au *Tomato golden mosaic virus* (TGMV). Les intégrations virales découvertes par la suite ont toutes correspondu à des virus de la famille *Caulimoviridae*. Elles ont également été découvertes de manière fortuite suite au séquençage des génomes de plantes (tabac-Jakowitsch et al., 1999 ; pétunia-Richert-Pöggeler et al., 2003 ; riz-Kunii et al., 2004 ; vigne-Murad et al., 2004 ; pomme de terre-Hansen et al., 2005 ; tomate-Stagginus et al., 2007 ; ananas-Gambley et al., 2008 ; peuplier-Bertsch et al., 2009), et/ou au développement de dia-

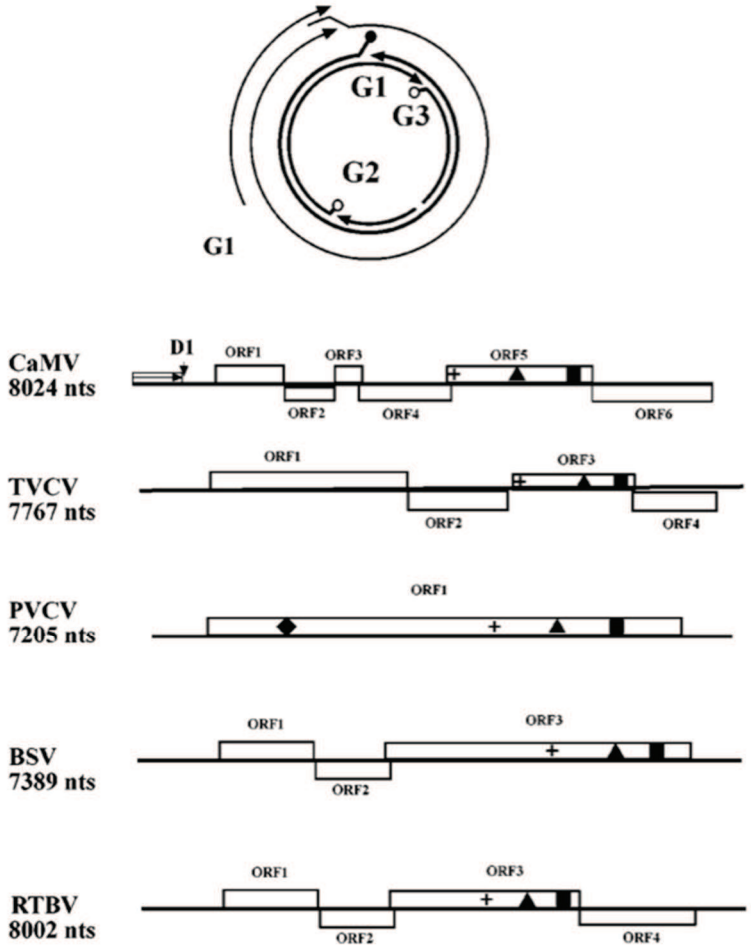


Figure 1-10 : Organisation génomique des *Caulimoviridae*

En haut : le génome circulaire des *Caulimovirus*, les transcrits sont présentés avec des lignes plus fines. On peut y voir aussi les discontinuitées G1 à G3 et les transcrits qui sont représentés par des lignes plus fines. Chez les *Badnavirus* seulement deux discontinuités sont présentes sur chacun des deux brins.

En bas : Organisation génomique des virus appartenant aux autres genres viraux des *Caulimoviridae*, ils sont représentés sous forme linéaire pour une meilleure lecture. Les rectangles représentent les ORF. Les symboles dans les ORF représentent les motifs typiques partagés entre les *Caulimoviridae* avec les *Rétrovirus* et les *Rétrotransposons*. L'aspartate protéase (+), Reverse transcriptase (▲), Ribonucléase H (■), et intégrase (♦).

CaMV : Cauliflower Mosaic Virus ; TVCV : Turnip Vein Clearing Virus ; PVCV : Petunia Vein Clearing virus ; BSV : Banana Streak Virus ; RTBV : Rice Tungro Baciliform Virus

D'après Harper et al., (2002)

-gnostic viraux par PCR utilisant des amorces spécifiques du génome viral et par southern blots (bananier-Lafleur et al., 1996 ; Kunii et al., 2004 ; Geering ., 2005 ; Bertsch ., 2009 ; Bousalem et al., 2009). Leur description a montré une répartition ubiquiste quant à l'hôte concerné allant de plantes monocotylédone à dicotylédone comme une diversité du nombre d'intégrations allant de plusieurs milliers de copies à quelques copies voire une copie par génome (Jakowitch et al., 1999 ; Lockhart et al., 2000).

3-3 *Geminiviridae* chez les solanacées et fabacées

Les *geminiviridae* sont des phytovirus ayant un petit génome à ADN simple brin de 2,5 à 3 kb (Fauquet, 2005). Les premières intégrations virales décrites par Bejarano et al. (1996) correspondent au gène codant pour la protéine associée à la réplication (ORF AC1, Rep protein) ainsi qu'à une séquence intergénique contenant les itérons (Rep protein-binding DNA motifs).Sur la base de l'observation de la conservation systématique des mêmes séquences, les auteurs ont avancé l'hypothèse d'une recombinaison illégitime à partir d'une forme réplicative subgénomique défective. Par la suite, ces séquences intégrées, ont été dénommées Geminivirus related DNA (GRD). Elles ont été décrites comme présentes en centaines de copies dans le génome du tabac (*Nicotiana tabacum*), et ayant subi de nombreux événements de duplication, délétion et réarrangement (Ashby et al., 1997). Ces séquences forment deux groupes polyphylétiques distincts : GRD3 et GRD5 qui dérivent des bégomovirus. Ces deux groupes sont issus de deux événements d'intégrations distincts. Les virus appartenant au groupe GRD5 sont les plus anciens et se sont intégrés chez l'ancêtre commun des espèces *Nicotiana kawakamii, N. tomentosa* et *N. tomentosiformis*. La lignée GRD3 provient d'une intégration plus récente, dans *N. tomentosiformis* seulement qui est l'ancêtre paternel du tabac (*Nicotiana tabacum*) (Murad et al., 2004).

3-4 Les *Caulimoviridae* intégrés dans le génome des plantes ou EPRV

Les intégrations de cette famille virale sont appelées EPRV pour endogenous pararetrovirus. Elles représentent la très grande majorité des intégrations décrites chez les plantes. Ceci tend à montrer qu'elles ne sont pas des phénomènes isolés et qu'elles ont eu lieu de façon récurrente pour les virus de cette famille. Des virus appartenant à 5 des 6 genres de la famille des *Caulimoviridae* ont été retrouvés intégrés dans un génome hôte. Seul les *Soymovirus* n'ont pour le moment jamais été décrits intégrés.

3-4-1 les EPRV non infectieux

La grande majorité des EPRV ont subi des mutations altérant la fonctionnalité de leurs gènes et/ou leur capacité à se multiplier et sont incapables de produire des particules virales restituant une infection. Aucune des 29 EPRV du *Rice tungro baciloform virus* (RTBV) du genre *Tungrovirus* dispersées dans le génome du riz *Oryza japonica* cv nipponbare ne présente un génome viral complet et tous sont fortement réarrangés. Ces EPRV ont pu être attribués à trois groupes phylogénétiques indiquant des évènements d'intégrations multiples et indépendants. De nombreuses intégrations de RTBV ont pu ainsi être mises en évidence dans le génome de quatre espèces de riz d'origine Sud-asiatique ou Australienne par la méthode de southern blot (Kunii et al., 2004).

Les EPRV identifiés dans le génome des *Solanacae* possèdent des caractéristiques communes, ils sont en général nombreux, dispersés dans le génome de la plante, non infectieux et proches du *Tabacco vein clearing virus* ou TVCV. Ils sont insérés préférentiellement dans l'hétérochromatine et dans les régions péricentromériques des chromosomes qui sont des régions naturellement non-transcrites des chromosomes (Hohn et al., 2008 ; Staginnus et Richert-Pöggeler, 2006). Ces intégrations ont été classées en cinq groupes, en fonction de l'hôte et de la séquence virale. Tout d'abord les NsEPRV qui ont été les premiers découverts chez *Nicotiana tabacum* par Jakowitsch et al. en 1999. Ils sont présents en une centaine de copies dans le génome de *N. tabacum* ainsi que de *N. sylvestris*. Les *Nto*EPRV ont été décrits dans le génome de *N. tomentosiformis* qui en abrite environ 600 copies (Gregor et al., 2006) et plus de mille dans le génome de *N. tabacum*. Ces intégrants ont été retrouvés très fréquemment dans l'environnement immédiat de rétroéléments de type Ty3-Gypsy qui pourraient jouer un rôle en facilitant leur intégration et leur duplication (Matzke et al., 2004). Le troisième groupe d'EPRV est celui des *Sotu*EPRV intégré sur 36 des 48 chromosomes de la pomme de terre (*Solanum tuberosum*) (Hansen et al., 2005). Le quatrième type correspond aux *Lyc*EPRV intégrés dans le génome de la tomate cultivée (*Solanum licopersicum*) et celui des espèces sauvages proches (Staginnus et al., 2007). Les études de diversité ont montré que ce dernier groupe d'intégrations est proche des EPRV de tabac et serait donc issu d'un événement d'intégration différent de celui qui a eu lieu chez la pomme de terre. Le cinquième groupe de *Caulimoviridae* intégrés dans le génome des solanacées est celui des TVCV EPRV qui est le seul à avoir gardé sa capacité à produire des particules virales, nous y reviendrons dans la partie suivante.

L'intégration de virus du genre *Caulimovirus* a été décrite pour la première fois dans le génome du Dahlia et correspond au *Dahlia mosaic virus* D-10 (Pahalawatta et al., 2008). Beaucoup plus récemment, P.-Y. Teycheney et A. Geering (com. personnelle) ont identifié la présence d'EPRV de la famille des *Caulimoviridae* dans le génome de nombreuses plantes. Ils ont mis en évidence des séquences virales et proposé un nouveau genre viral nommé *Dyonivirus* sur la base de la reconstruction d'un génome viral non décrit à ce jour sous une forme épisomale. Ces intégrations semblent ubiquitaires et ont été retrouvées dans le génome de presque toutes les plantes séquencées à ce jour (vigne, riz, tabac, …).

3-4-2 Les EPRV infectieux

Il existe seulement trois cas décrits d'EPRV dits« infectieux » c'est à dire responsables de la production de particules virales induisant la maladie chez la plantes hôte. Ces trois cas possèdent un certain nombre de similarités. Tout d'abord il est important de noter que ces trois insertions contiennent le génome viral complet possédant donc l'information génétique lui permettant d'être infectieux. De plus, on remarque que pour chacune de ces intégrations la production de particule virale est consécutive à des stress et ne concerne que des plantes hybrides interspécifiques.

Les EPRV de TVCV chez le tabac

Le virus de l'éclaircissement des nervures du tabac ou *Tobacco vein clearing virus* (TVCV) appartient au genre des *Cavemovirus*. Il a été mis en évidence à partir d'intégrations découvertes dans le tabac. Le TVCV n'infecte que l'hybride *Nicotiana edwardsonii,* hybride interspécifique hexaploides entre les espèces *N. clevelandii* et *N. glutinosa*. Il provoque des symptômes foliaires tardifs et est transmis uniquement par graines. Il possède un génome de 7,8kb qui comporte 4 ORFs. Des essais de transmission mécanique et par insectes vers les autres espèces de solanacées n'ont pas été suivis d'infection. Lockhart et al., (2000) ont fait l'hypothèse d'EPRV transmis par le génome parental de *N. glutinosa* et démontré leur existence par southern blot. Le parent incriminé ne développe jamais ni les symptômes de la maladie ni le virus et semble être un porteur sain des EPRV. Cette observation laisse à penser que l'hybridation interspécifique peut jouer un rôle dans la libération du virus.

Les EPRV de PVCV chez le pétunia

Le virus de l'éclaircissement des nervures du pétunia ou *Petunia vein clearing virus* appartient au genre *Petuvirus*. Ces virus ont un génome de 7,2kb, composé de deux longs ORFs (Fauquet, 2005). Cinq locus d'intégrations ont été identifiés chez le pétunia *Petunia*

hybrida grâce à la méthode d'hybridation in situ ou Fluorescente In-Situ Hybridization (FISH). Entre 100 et 200 copies de génome de PVCV complet, insérées en tandem ont été décrites. Le *P. hybrida* est un hybride entre deux espèces sauvages de pétunia où seul un des parents possède ces intégrations. Richert-Pöggeler et al., (2003) ont démontré le caractère infectieux de ces EPRV ou ePVCV (endogenous PVCV) en observant la présence de virions et le développement de la maladie dans des plantes *P. parodii* sans ePVCV bombardées avec des fragments du génome de plantes contenant l'ePVCV.

Les EPRV de BSV chez le bananier

Le dernier cas connu de séquences virales intégrées infectieuses concerne le genre *Badnavirus* et correspond au virus de la mosaïque en tirets des bananiers ou *Banana streak virus*. Ces intégrations, objet de ce travail, seront détaillées dans le paragraphe 7.

4-Evolution des EPRV dans le génome des plantes

Les études d'évolution des intégrations virales apportent des informations tant sur le virus que sur son hôte. Tout d'abord, les EVE représentent des fossiles de formes anciennes de virus suite à leur intégration dans le génome hôte et viennent donc très utilement enrichir les connaissances sur les formes ancestrales pour les études d'évolution des virus. Ensuite, les EVE informent sur les interactions qui existent entre l'hôte et son pathogène, et l'intérêt réciproque pour le virus et pour l'hôte en terme de coévolution. Enfin elles permettent de mieux comprendre les forces qui régissent l'évolution des génomes qu'ils soient viraux ou de plantes, grâce aux différents mécanismes qui sont liés aux intégrations.

4-1 Mécanismes d'intégration des EPRV

Différents mécanismes d'intégrations ont été envisagés pour expliquer la présence des ERV dans le génome de certaines plantes. Cependant, ces différents mécanismes restent des hypothèses qui ne sont pas démontrées ; il reste encore un grand champ de recherche afin de comprendre comment ont lieu ces intégrations. Il semblerait que les rétrotransposons soient impliqués dans ces mécanismes eu égard à leur présence quasi systématique aux abords des séquences intégrées. La co-localisation des rétrotransposons et des ERV peut s'expliquer aussi par l'évolution des génomes qui favorise les intégrations dans les zones génomiques où elles n'auront pas d'impact et donc ces différentes séquences se retrouvent dans les mêmes zones génomiques. Il est aussi important de souligner que les ERV détectés

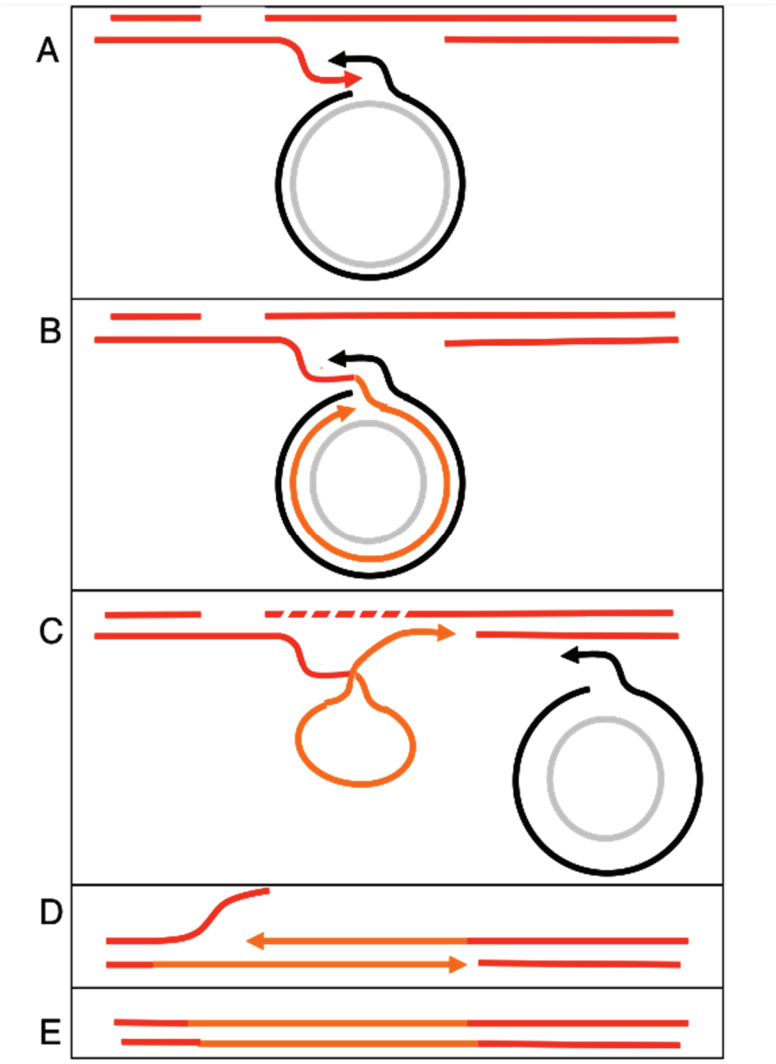


Figure 1-11 : Mécanisme d'intégration des EPRV par recombinaison non homologue

L'ADN viral est représenté en noir, celui de l'hôte en rouge. Les brins néosynthétisés sont indiqués en orange. (A) Une recombinaison non homologue peut s'initier par NHEJ (non-homologous end-joining) lorsque des fragments simple brin générés au niveau des cassures du génome hôte, s'apparient sur des micro-homologies à des séquences virales hétérologues, plutôt qu'à des séquences cellulaires homologues. Les matrices virales simple brin peuvent être soit les intermédiaires ssDNA des *Geminiviridae*, soit les courtes séquences simples brins présentes au niveau des interruptions de séquences du génome des *Caulimoviridae*. (B-D) Les enzymes de réparation de l'ADN synthétisent les brins complémentaires à la séquence virale, réparent les nucléotides manquants et les cassures par ligation. (E) La séquence virale devenue endogène est maintenant intégrée de manière stable dans le génome de la cellule hôte. *Adapté de Hohn et al., (2008)*

appartiennent à des groupes de virus qui ne nécessitent pas d'intégration dans leur cycle de multiplication, les intégrations correspondent donc à des évènements d'intégration accidentels. Le deuxième élément est que, ces EVE de plantes correspondent à des virus qui ont un cycle de multiplication qui se passe au moins pour partie dans le noyau. Et comme l'a démontré Paszlowski et al. (1984) tout ADN se trouvant dans le noyau a une chance, même si elle est très faible, de se retrouver intégré dans le génome, c'est le fondamental de la transformation des plantes par transfert direct de gène.

Un des mécanismes pouvant expliquer l'intégration est l'autostop génétique. Les pararétrovirus et les rétrotransposons possèdent un fort taux de recombinaison et les séquences de protéines permettant l'intégration pourraient être associées au virus lors de la production de particules virales dans le cytoplasme et ensuite être utilisées dans le noyau pour l'intégration. Des hybridations sont aussi possibles entre les séquences de rétrotransposons et de virus, ce qui favoriserait leurs intégrations lors d'événements de transposition des rétrotransposons. Ceci expliquerait également que l'on retrouve souvent les EVE dans des zones riches en rétroéléments voire au sein des rétroéléments (Richert-Pöggeler et al. 2003 ; Gregor et al. 2004 ; Staginnus et al., 2007 ; Gayral et al., 2008).

Le mécanisme généralement admis pour expliquer les intégrations d'EVE dans le génome des plantes et, plus particulièrement l'intégration des EPRV qui sont des virus à ADN se multipliant dans le noyau est celui de la réparation de cassures ADN double brin (Double strand break repair (DSBR)), qui mène à l'intégration d'ADN par jonction terminale non homologue (non-homologous end joining) (Putcha., 2004 ; Bill et al., 2004) (Figure 1-11). Parmi les deux mécanismes envisagés pour la DSBR, celui impliquant le single strand annealing est le plus probable. Car le premier mécanisme implique de la recombinaison homologue mitotique que l'on sait très rare chez les plantes. Plus fréquemment des micro-homologies peuvent être impliquées lors de l'initiation de l'intégration du simple brin. Les régions riches en Adénine et Uracile sont connues pour faciliter les changements de matrice entre le virus et l'ARN de l'hôte (White et Nagy, 2004). Les intégrations déjà présentes dans le génome des plantes peuvent servir d'amorces et de matrices pour initier le mécanisme de recombinaison. Tout comme les régions Flaps présentes à l'extrémité des ARN prégénomiques viraux qui sont des sites très utilisés pour les mécanismes d'intégration simple brin. Pour l'ePVCV (Richert-Pöggeler et al., 2003), les NsEPRV (Jakowitsch et al., 1999) et l'eRTBV (Kunii et al., 2004), des régions Flaps ont été mises en évidence au niveau des sites d'intégration.

4-2 Les sites d'intégrations

Il est maintenant clair que les intégrations des EVE ne se sont pas faites au hasard dans le génome des plantes mais dans des zones particulières des génomes, plusieurs hypothèses sont proposées afin de comprendre les forces évolutives qui favorisent les intégrations dans ces zones.

Tout d'abord, il a été montré que, très majoritairement, les EPRV se trouvaient intégrés dans des zones riches en rétransposons à LTR (Jakowitsch et al., 1999, Ndowora et al., 1999 ; Richert-Pöggeler et al., 2003 ; Gregor et al., 2004 ; Staginnus et al., 2007 ; Gayral et al., 2008). Les expériences de type FISH ont permis de mettre en évidence que les ePVCV du pétunia et les EPRV de la pomme de terre se trouvent situés dans les zones centromériques et péricentromériques des chromosomes (Richert-Pöggeler et al., 2003 ; Hansen et al., 2005). Les différentes observations montrent que la plupart du temps, les EPRV sont intégrés dans l'hétérochromatine tout comme les éléments transposables avec lesquels majoritairement ils co-localisent (Wang et al., 2006). Ces observations ont été décrites pour le tabac, le pétunia et le bananier (Mette et al., 2002 ; Richert-Pöggeler et al., 2003 ; Gayral et al., 2008). Ce sont dans ces zones chromosomiques que les mécanismes de régulations épigénétiques sont les plus importants afin de limiter, entre autres, les dommages liés à la ré-activation des transposons et des rétrotransposons.

Les raisons d'une insertion privilégiée dans ces zones ne sont pas encore clairement identifiées et plus particulièrement si ce choix est le fait des intégrations elles-mêmes ou s'il résulte d'une coévolution entre les EVE et la chromatine de l'hôte. Plusieurs hypothèses émanent des études faites sur les éléments transposables. La première hypothèse est que ces parties du génome sont très pauvres en gènes et peuvent absorber des EVE sans effets délétères sur la fitness de la plante. La deuxième hypothèse est que comme tous les EVE décrits sont issus de la même famille virale, les *Caulimoviridae*, des microhomologies existantes avec les autres intégrations ou les RE pourraient expliquer que les intégrations se retrouvent dans les mêmes zones.

4-3 Fixation des EPRV dans les génomes

Il est important de souligner que pour autant que l'intégration semble être un phénomène fréquent, la fixation dans le génome reste difficile. En effet, la première étape de fixation passe par une intégration dans les cellules germinales afin de pouvoir être transmise à la descendance. Pour cela, au moins une cellule ayant subi l'intégration, doit se retrouver dans

les gamètes et plus avant dans les cellules réalisant la gamétogénèse. Cette étape apparaît assez complexe car les cellules du méristème ont des mécanismes de défense très poussés empêchant l'infection virale. Il a donc été démontré que la RNA-dépendante RNA polymérase joue un rôle important dans la défense d'invasion du méristème (Blevin et al., 2006). Il est malgré tout possible que des virus puissent passer cette barrière et infecter le méristème. Un autre moyen que le virus a d'être fixé, est que la plante soit propagée végétativement. Une cellule possédant des intégrations peut se retrouver impliquée dans cette multiplication et permettre d'être à l'origine d'une nouvelle plante ayant l'intégration.

Une fois que l'EVE est présente dans le génome de toutes les cellules, la deuxième étape est de se maintenir dans la population hôte. Pour cela, l'intégration ne doit pas avoir d'effet négatif sur la fitness de la plante, elle doit être neutre ou apporter un avantage sélectif ce qui aurait pour conséquence d'accélérer cette fixation. Il est ainsi observé pour une très grande partie des EVE, des mutations délétères empêchant leur expression et la restitution de particules virales. Chez les animaux, il apparaît que des réarrangements génomiques post-insertionnels ont lieu au niveau des EVE se traduisant par une incapacité à être exprimés. Ces réarrangements mettent en jeu un mécanisme de recombinaison homologue non-allélique entre les EVE d'une même famille. Ce mécanisme produit des délétions, inversions et translocations au niveau des séquences EVE (Hughes et Coffin, 2001) et il est fort possible que ce même mécanisme ait conduit aux réarrangements observés chez les *NtoEPRV* du tabac (Gregor et al., 2004, Matzke et al., 2004).

4-4 Mécanismes de régulation

Les mécanismes mis en jeu pour réguler les EVE sont un facteur clé de leur évolution et de leur potentielle conservation dans les génomes. En effet, il est clairement observé que les EVE présents dans le génome des plantes sont très majoritairement non-fonctionnels. Seuls certains EVE ont conservé leur activité, et cela pour des plantes issues d'hybridation interspécifique et ayant subi des stress. Il semble que le silencing soit le mécanisme majeur sous-jacent à la régulation des EVE.

A la fin des années 90, la résistance virale de type « extinction génique » ou ARN interférent a été décrite (Vaucheret et al., 1998). La génomique moderne et l'analyse de transgènes prenant en compte les séquences non codantes ont permis la découverte de ce mécanisme. En effet comme tous les eucaryotes, les plantes ont la capacité de neutraliser des acides nucléiques aberrants ou invasifs (comme par exemple des éléments transposables, des virus

ou des transgènes) afin de lutter contre les effets délétères qu'ils induisent tels que l'augmentation de la taille des génomes ou l'extinction de gène. Ces mécanismes, tout d'abord découverts pour leur rôle de régulation de l'expression génique, peuvent agir à différents niveaux en diminuant ou empêchant la transcription (Transcriptional Gene Silencing, TGS) par l'intermédiaire de la méthylation des histones ou de l'ADN, ainsi que l'ubiquitination ou l'adénilation (Mette et al., 2000). Ces mécanismes agissent directement sur la structure de l'ADN la rendant peu ou pas accessible à la polymérase II. La régulation existe également au niveau post-transcriptionnel par des mécanismes dégradant les ARN messagers (Post-transcriptional gene silencing, PTGS) (Vaucheret et al., 2006). Ces différents phénomènes sont regroupés sous la terminologie de ARN interférant.

Ainsi, de nombreuses plantes ont développé des mécanismes de régulation spécifiques des séquences parasites, d'une part de type TGS, en favorisant l'intégration dans des zones d'hétérochromatine et/ou en méthylant les zones d'intégration (ceci a été montré pour la régulation des RE chez *Arabidopsis thaliana* (Huettel et al., 2007) et le maïs) ; d'autre part, il semble que les régulations de type PTGS qui agissent majoritairement sur les virus exogènes (Voinnet, 2005) comme défense virale, puissent avoir un rôle sur les séquences parasites de type RE chez *Arabidopsis thaliana* (Gazzani 2004).

En ce qui concerne les EPRV, il a été montré chez les *solanacées* genre tabac et pétunia, que la transcription des séquences virales intégrées proches du TVCV et du PVCV respectivement étaient régulée de manière épigénétique par des mécanismes de type TGS diminuant ainsi l'expression globale des EPRV (Mette et al., 2002 ; Noreen et al., 2007). Des petits ARNs ainsi qu'une méthylation des promoteurs viraux ont également été observés. Cette régulation aurait permis de mettre en place une résistance de la plante au virus qui semble avoir abouti à la disparition de ce dernier dans le cas du TVCV pour la plupart des solanacées. Un schéma de régulation potentielle a été proposé par Staginnus et Richert-Poggeler (2006) pour expliquer la régulation des EPRV des plantes. Ce schéma présente la régulation épigénétique des intégrations telle qu'elle a pu être observée mais aussi propose des régulations potentielles qui n'ont pas encore été validées. Il montre que certains EPRV seraient méthylés, et que d'autres copies resteraient accessibles à l'ARN polymérase, ce qui produirait un faible taux de transcrit utile pour de la régulation de type TGS et PTGS. Ce schéma montre également le rôle putatif de l'épigénétique dans la défense contre les particules virales libres, qui pourraient être régulées soit par du TGS ou du PTGS. Il est aussi très important de souligner que les intégrations infectieuses le sont dans des contextes très

particuliers et compatibles avec de la régulation épigénétique. En effet, les facteurs déclenchant le réveil des intégrations tels que sont les coupures (Richert-Poggeler et al., 2003), les chocs thermiques (Noreen et al., 2007), la culture cellulaire (Dallot et al., 2000) et l'hybridation génétique (Lheureux et al., 2003) correspondent à des stress génomiques. Dans la nature, les intégrations infectieuses ont été décrites chez des plantes provenant d'hybridation chez qui les parents n'étaient pas sujet à des réveils d'EPRV, comme pour les ePVCV (Richert-Poggeler et al., 2003), les eTVCV (Lockhart et al., 2000) et les différents eBSVs (Harper et al., 1999 ; Lheureux et al., 2003). Ces différents facteurs d'activation correspondraient à des stress génomiques qui peuvent conduire à la levée de silencing chez les EPRV et amener à la production de particules virales (Fisher et al., 2006). Cette partie ainsi que les schéma s'y rapportant seront développés dans l'introduction du chapitre 2 de cette thèse.

4-5 Coût/bénéfice pour le couple plante-virus, des intégrations virales

Le nombre d'EPRV découvert pour l'instant reste assez faible chez les plantes mais la comparaison avec les animaux et les découvertes récentes amènent à penser que leur quantité doit être bien supérieure mais que les moyens et méthodes utilisés pour les découvrir ne sont pas encore adaptés à ce type de recherche. De même, il est possible que la plasticité du génome des plantes puisse aussi jouer un rôle car il a été montré pour les RE que les plantes ne possédaient pas toutes la même capacité à accumuler les intégrations de séquences parasites (Kumar et Bennetzen, 1999).

Il semble que, tout comme pour les EVE animaux, des pressions de sélection soient appliquées sur les séquences EPRV en particulier pendant la période de colonisation des génomes de plante (Gifford et Tristem, 2003).Ces pressions proviennent des effets délétères des EVE sur la fitness de leur hôte. Que ce soit par leurs potentiels effets infectieux ou bien par le fait qu'ils puissent coloniser les génomes une fois l'intégration réalisée comme ce fut le cas chez le tabac (Jankowitsch et al., 1999). Comme nous l'avons vu, les plantes ont su s'adapter en développant des régulations de l'expression des EPRV basée sur des processus d'ARNi (Norreen et al., 2007). De plus, il apparaît que la majorité des séquences aient également accumulé des mutations ou subi des réarrangements conduisant à l'inactivation des EPRV et à leur pseudogénéisation. Cette régulation négative, comme pour les ERV, est partagée par tous les EPRV présents de manière naturelle dans le génome de l'hôte. En effet l'activation des EPRV n'a lieu que chez des hybrides interspécifiques qui sont pour la plupart non-naturels. Ce qui signifierait qu'à l'état sauvage les EPRV ne sont pas une contrainte pour

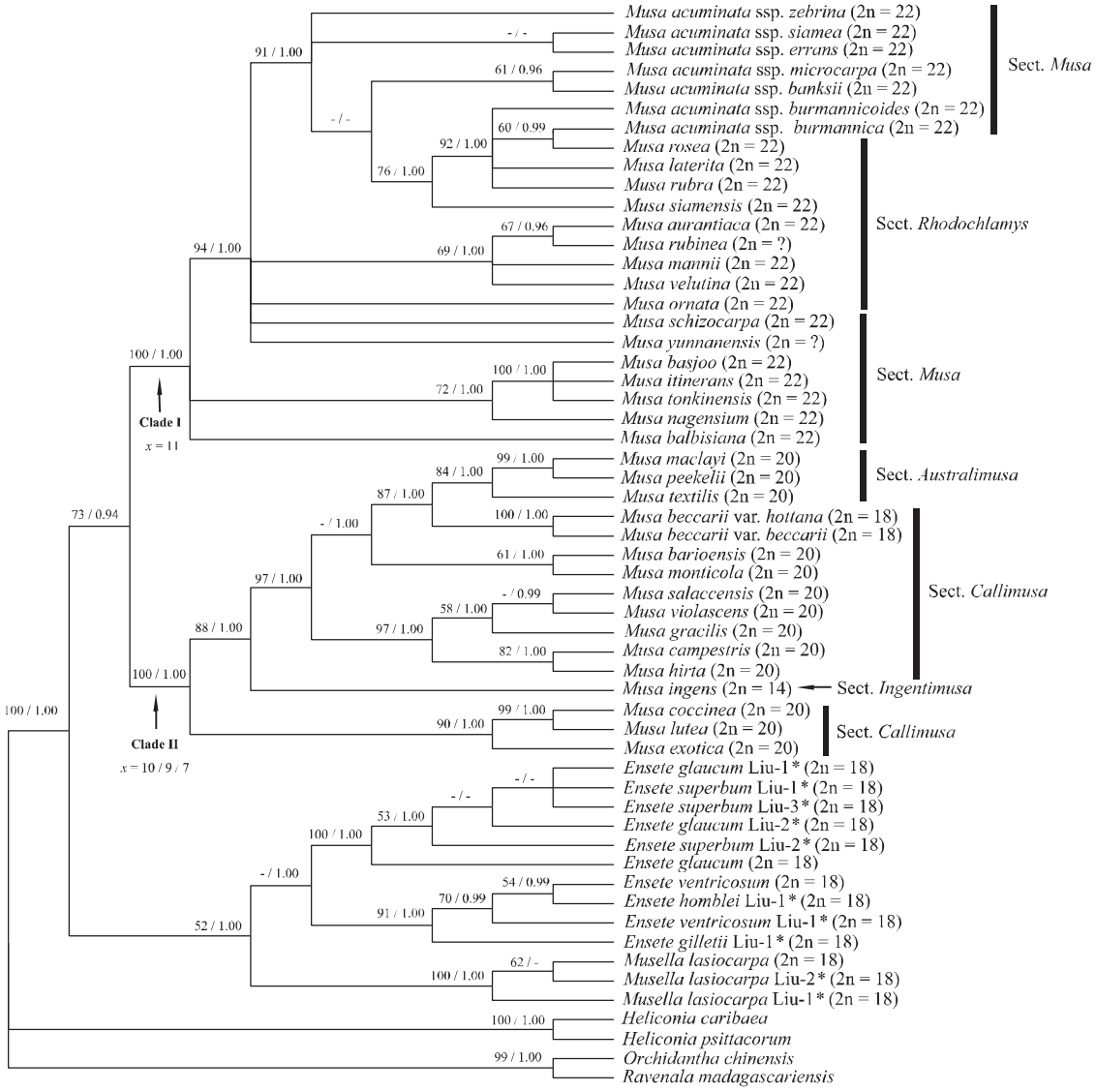


Figure 1-12 : Arbre phylogénétique des *Musaceae*

Arbre obtenu à partir de séquences nucléaires ITS et chloroplasitiques (atpB-rbcL spacer, trnL-F spacer, rps16 intron). Le premier chiffre sur les branches correspond aux valeurs de bootstrap obtenues lors du calcul par maximum de parcimonie (BS). Le deuxième chiffre correspond aux valeurs de probabilités postérieures obtenues lors de calculs de la phylogénie par inférence bayésienne (PP). (-) : BS<50% ou PP<0,5. Les deux clades du genre *Musa* sont reportés sur l'arbre ainsi que les nombres de chromosomes pour les différentes espèces. *D'après Li et al., (2010)*

les plantes. Sans contre-attaque, cette inactivation ne peut conduire qu'à l'extinction du virus. Cependant, il est possible que les virus, une fois dans le génome de l'hôte, puissent servir de réservoir de biodiversité aux virus exogènes (Stagginus et Richert-Pöggeler, 2006). Et pour certains, le fait que les mécanismes de régulation soient réversibles leur laisse la possibilité, en cas de stress, d'être libérés et de pouvoir à nouveau coloniser les plantes s'ils n'ont pas subi de mutation rédhibitoire.

Enfin, des études sur les ERV présents dans le génome des animaux ont montré que ces intégrations pourraient servir l'hôte et que les intégrations participeraient à la course à « l'armement » en fournissant des gènes viraux alimentant les mécanismes de défense (Aswad et Katzourakis, 2012). Enfin, nous avons développé dans le paragraphe 2-2-2 que certains gènes viraux ont été domestiqués par les animaux et participent à différents mécanismes (Feschotte et Gilbert, 2012) comme le gène de la syncitine par exemple. Ces effets bénéfiques contribueraient au maintien des ERV dans le génome des cellules. Chez les plantes, des hypothèses similaires ont été avancées afin d'expliquer la présence voire le maintien des EPRV dans les génomes. Mette et al, 2002 ont avancé un rôle dans l'acquisition de mécanismes de défense anti-viraux médiés par l'ARNi et ayant abouti à la disparition du virus ancestral (Stagginus et Richert-Pöggeler, 2006). Aucune de ces hypothèses n'a été à ce jour validée ni n'est même étudiée.

5-Le Bananier, l'espèce hôte du pathosystème d'étude

5-1 Taxonomie des bananiers

Les bananiers sont les hôtes uniques du virus de la mosaïque ou *Banana streak virus* (BSV), et la seule plante connue à posséder des intégrations virales correspondant au BSV. Le terme 'bananier' utilisé dans cette thèse concerne toutes les espèces du genre *Musa* qui appartiennent à l'ordre des *Zingiberales* et à la famille des *Musacea* (figure 1-12). Les bananiers sont des monocotylédones et sont des herbes géantes, généralement de plus de 3m de hauteur, qui ne réalisent pas de lignification (Tomlinson 1969). L'ordre des *zingibérales* appartient au clade des *Commelinidées* qui comprend également les *Poacés*, les *Commeliacés*, et les *Aracacés*. Des comparaisons génomiques sont donc possibles au sein de ce clade qui intègre les céréales largement étudiées et séquencées pour la plupart.

La famille des *Musacea* comprend deux genres, le genre *Ensete* et le genre *Musa*.

Le genre *Ensete* est morphologiquement proche du genre *Musa,* avec des plantes atteignant 5 à 7m de hauteur mais avec des capacités de propagation végétative par rejets de bananier

beaucoup plus limitées. Six espèces à 2n=18 chromosomes, ont été décrites (Simmonds, 1962). Elles se rencontrent à l'état sauvage depuis le sud de la Chine jusqu'à l'Afrique de l'est. Ensete n'a pas été domestiqué, sa culture traditionnelle dans une région d'Ethiopie pour la consommation des parties végétatives, après préparation par fermentation est rapportée.

Le genre *Musa* correspond aux bananiers et rassemble une cinquantaine d'espèces. Ce genre a été divisé en 5 sections : *Eumusa* (2n=22), *Rhodochlamys* (2n=22), *Australimusa* (2n=20), *Callimusa* (2n=20) (Cheesman, 1947), *Ingentimusa* (2n=14) (Argent, 1976). La subdivision au sein des 2n=22 et 2n=20 est aujourd'hui largement remise en cause, Li et al. (2010) montrent ainsi, à partir de séquences ITS et chloroplastiques, que les sections *Rhodoclamys* et *Eumusa* ont une origine monophylétique tout comme les sections *Australimusa et Callimusa (*Figure 1-12*).*

Mis à part, le groupe très particulier des Fehi cultivé dans les iles du Pacifique et appartenant à la section *Australimusa*, tous les bananiers cultivés appartiennent à la section *Eumusa*, et sont issus des espèces *Musa acuminata et Musa balbisiana* même si quelques introgressions d'autres espèces de la section sont parfois soupçonnées.

Les bananiers du genre *Musa* se rencontrent à l'état sauvage dans un vaste triangle, depuis l'est de l'Inde, au sud de la Chine et jusqu'au nord de l'Australie, la Papouasie et les iles Salomon dans le Pacifique (Simmonds, 1955). Ces bananiers sont présents dans les vallées ou les clairières humides des forêts de faible et moyenne altitude. Les plantes mesurent de quelques mètres jusqu'à plus de 15 mètres de haut, suivant les espèces. Ils se composent d'une tige souterraine appelée rhizome ou corme et d'un pseudo-tronc formé des gaines des feuilles serrées les unes contre les autres. Le méristème apical unique, est situé au centre du pseudo-tronc presque à hauteur du sol. La reproduction des bananiers peut se faire par voie sexuée ou par propagation végétative par rejetonnage. Les rejets se forment à partir de bourgeons apicaux secondaires de la tige souterraine, ils s'autonomisent ensuite après émission de racines axillaires. Il se forme alors des touffes de bananiers, regroupant des clones issus de la même plante. Les bananiers sauvages sont séminifères et se multiplient aussi de manière sexuée. Le méristème apical passe du statut végétatif au statut floral pour former une inflorescence, le dernier entre-nœud subit une élongation considérable pour former une hampe florale qui porte l'inflorescence en haut du pseudo-tronc. En général, les premières fleurs émises, en partie basale, sont femelles, le méristème passant ensuite en production continue de fleurs mâles au sein du bourgeon mâle (appelé popote aux Antilles) (Figure 1-13). Ce déphasage temporel limite alors les autofécondations, qui restent cependant

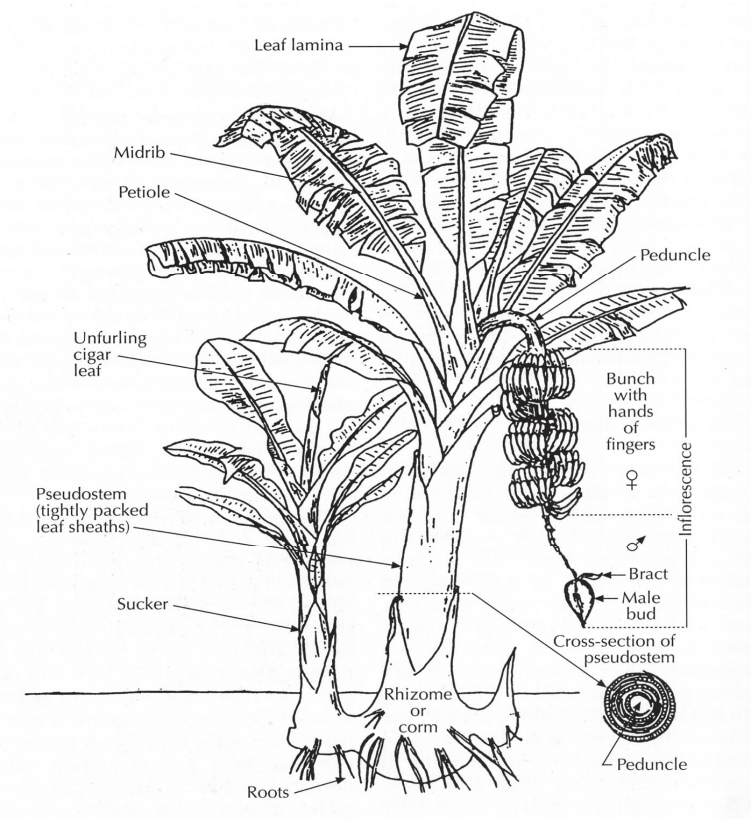


Figure 1-13 : Représentation schématique d'un bananier du genre *Musa*
D'après Jones, (2000)

Figure 1-14 : La diversité des bananes et des plantains en vente dans un magasin du sud de l'Inde (Varkala, Kerala)
Les cultivars sont indiquées par les lettres au-dessus des régimes. a- Cultivar 'Red' (génome AAA), un cultivar prisé de banane dessert. **b**- 'Palayam Codan' (AAB). **c**- 'Njalipoovan' AB (les couleurs vertes et jaunes correspondent au degré de murissement des fruit. **d**- 'Robusta' (groupe 'Cavendish', AAA). **e**- 'Nendran' (AAB), utilisé pour la cuisson et pour la fabrication de chips. **f**- 'Peyan' (ABB) utilisé comme légume pour les currys et pour les plats chauds. **g**- 'Poovan' (AAB).
D'après Heslop-Harrison et Schwarzacher, (2007)

possibles au sein d'une même touffe. Pour certaines espèces, les fleurs basales sont hermaphrodites, les autofécondations sont plus fréquentes, conduisant à des taux d'hétérozygotie faibles.

5-2 De la domestication à la culture des bananiers

La banane est la quatrième culture du monde pour les pays développés avec une production autour de 100Mt par an après le riz, le blé et le maïs. Les bananes sont cultivées dans toutes les zones tropicales et subtropicales autour du monde. La très grande majorité des bananes cultivées dérivent de croisements inter et intra spécifiques entre bananiers sauvages des espèces *M. balbisiana* (génome noté BB) et *M. acuminata* (génome noté AA). Le syndrome majeur de domestication des bananiers est une perturbation de la sexualité conduisant à un développement parthénocarpique du fruit en l'absence de graines, ces graines étant en effet très nombreuses et extrêmement dures, rendant le fruit très difficilement consommable. Ces variétés stériles sont depuis maintenues grâce à leur capacité de propagation végétative. Parmi des variétés stériles parthénocarpiques, la banane fruit dite « dollar », qui fait l'objet d'un vaste commerce international, est la plus connue. Cependant l'essentiel de la culture bananière (environ 87% des surfaces) est une production vivrière pour les pays du sud (Bioversity International, 2008) avec une diversité très importante de variétés en termes de couleur, forme, goût (Figure 1-14). Les variétés de bananes sucrées, celles du commerce international ou d'autres plus locales, sont cultivées pour l'autoconsommation ou l'alimentation des marchés locaux. Des superficies immenses, en particulier en Afrique et en Asie, sont consacrées à des bananes qui ne dégradent pas leur amidon en sucre et doivent être consommées cuites, avec une très grande diversité des préparations culinaires. Certaines variétés sont destinées à la fabrication de bière, en particulier dans la région des Grands Lacs en Afrique de l'est. Hormis la consommation des fruits, les bananiers sont traditionnellement utilisés pour de nombreux autres usages : tissage à partir de fibres, matériau de construction, emballage avec les feuilles, médecine, rituels religieux etc. Le bourgeon mâle est consommé dans certaines régions d'Asie, parfois même les fruits des types séminifères, avant durcissement des graines.

Pour mener à bien les programmes d'amélioration variétale, il s'est avéré indispensable de bien connaître la diversité génétique disponible au sein des formes parentales *M. acuminata* et *M. balbisiana* et des formes cultivées qui en découlent. Différentes études ont été

menées depuis les 20 dernières années afin de caractériser cette diversité. Des collections in vivo ont été constituées dans diverses régions du monde pour réunir la diversité locale, certaines ayant une vocation plus large comme celles du CIRAD en Guadeloupe ou du CARBAP au Cameroun. La communauté internationale s'est accordée pour créer une collection in vitro regroupant une part très importante des bananiers sauvages et cultivés disponibles (plus d'un millier aujourd'hui), ce centre (ITC International Transit Center) situé en Belgique, assure la conservation de ces accessions et assure leur libre diffusion.

Pour caractériser ces ressources, de nombreux marqueurs moléculaires des génomes nucléaires et cytoplasmiques ont été développés (Carreel et al., 1994 ; Lagoda et al., 1998 ; Crouch et al., 1998 ; Hippolyte et al., 2010 et 2012). Une synthèse récente de ces résultats a permis de dégager les principales structures au sein des formes sauvages ainsi que leur contribution aux formes cultivées diploides puis triploides (De Langhe et al., 2009, Perrier et al., 2009).

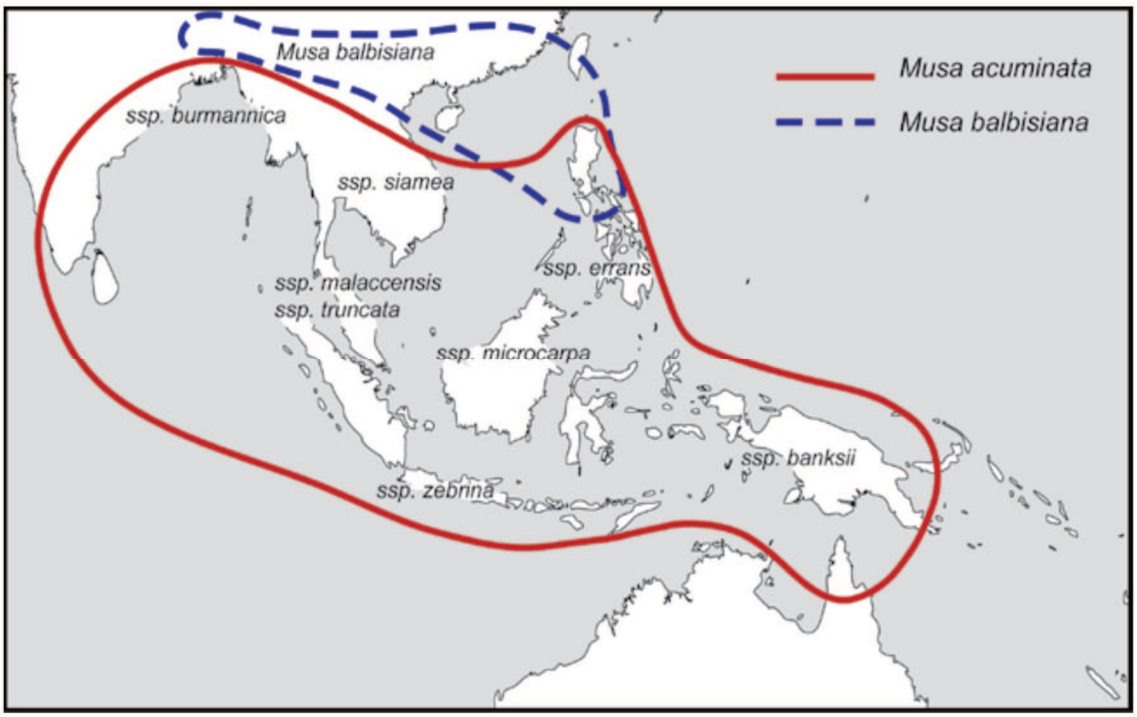


Figure 1-15 : Distribution géographique ancestrale des deux principales espèces de bananier et des sous-espèces associées en Asie du Sud Est

D'après Delanghe et al., (2009)

Ces études confirment ainsi la division de l'espèce *M. acuminata* en diverses sous-espèces géographiquement séparées (Shepherd, 1999) : Banksii en PNG, Zebrina à Java, Malaccencis dans la péninsule malaise et le SE asiatique continental, Errans aux Philippines, Burmanica et Siamea dans la zone birmane, jusqu'à l'Inde du nord (figure 1-15). Ces sous-espèces résulteraient d'isolements géographiques, suite aux variations du niveau marin, depuis une

forme ancestrale que l'on situe dans la zone de diversité maximale des *Musacées*, entre Chine du sud et Thaïlande. Les études de parenté entre formes sauvages et formes cultivées montrent qu'en général les formes diploides cultivées sont des hybrides entre diverses sous-espèces. Les hybridations restent effectivement possibles entre ces sous-espèces mais les divergences entre génome, en particulier par réaménagements chromosomiques (Shepherd, 1999), sont à l'origine de formes stériles AA parmi lesquelles l'homme a sélectionné un certain nombre de cultivars AAcv. Une sexualité résiduelle permet dans des cas rares la production de gamètes non réduits qui, combinés à des gamètes haploides de diploides, a permis l'apparition des formes triploides AAA. Certaines de ces formes triploides stériles ont eu un succès considérable et ont été préférées aux formes diploïdes et exportées, depuis le SE asiatique vers l'Inde, l'Afrique et l'Amérique centrale ou du sud.

L'espèce *M. balbisiana* est moins bien connue que l'espèce *M. acuminata* qui est, elle, à l'origine des triploides AAA de type Cavendish du marché international, et donc le plus anciennement et le plus largement étudiée. Les ressources génétiques disponibles en collection sont rares, moins de 25 accessions (dont l'origine géographique est dans certains cas assez mal connue), et il est souvent considéré que la diversité génétique y est faible. Il a été cependant montré que la diversité allélique de nature *M. balbisiana* des hybrides AAB et ABB était plus importante que celle rencontrée dans les seuls diploïdes BB, soit que les parents de ces triploides soient absents des collections, soient qu'ils se soient depuis éteints (Hippolyte et al., 2012). L'étude de Carreel et al. (1994), par marqueurs microsatellites, a permis de définir plusieurs groupes parmi les diploïdes *M. balbisiana* d'après les analyses microsatellites. Gayral et al. (2010) ont reconduit l'étude sur un échantillon sensiblement enrichi de *M. balbisiana*, ils confirment les groupes de Carreel et al., (1994) ainsi que la faible diversité génétique de ces individus. *M. balbisiana* a une distribution plus nordique que *M. acuminata* et se rencontre dans un croissant du nord de l'Inde jusqu'au sud de la Chine (Figure 1-15).Les limites restent cependant imprécises car des formes de *M. balbisiana* ont pu faire l'objet de déplacements par l'homme, cette espèce étant utilisée, en particulier pour la consommation du bourgeon mâle. On trouve ainsi des *M. balbisiana* dans des jardins et donc cultivés. Il est alors difficile de dire si les formes rencontrées aux Philippines par exemple, sont des formes endémiques ou des plants apportés par l'homme. Les zones de sympatrie avec *M. acuminata* se situent plutôt aux extrémités est et ouest du croissant de distribution, vers les Philippines et en Inde. On ne connaît pas de formes *M. balbisiana* domestiquées au sens précédent de la stérilité et de la parthénocarpie (Figure 1-16). En revanche les formes hybrides interspécifiques sont très répandues. Les hybrides AB sont

surtout cultivés en Inde, ils sont rares ailleurs. Les formes AAB et ABB sont beaucoup plus répandues, le génome B étant considéré comme apportant une vigueur, une meilleure adaptation aux stress abiotiques ainsi que des caractères de résistance à divers pathogènes (Bakry et al., 2009). Le mode de création de ces hybrides dans les zones de sympatrie, depuis les Philippines jusqu'à l'Inde, reste controversé. L'hybridation entre un gamète non réduit AA, comme évoqué précédemment, et un gamète B a été montré à l'origine de certains AAB mais conduit dans certains cas à des incompatibilités entre les génomes nucléaires et cytoplasmiques. Le passage par une forme diploïde AB permet d'expliquer la formation d'AAB comme d'ABB. Cependant ces formes AB sont rares et sont en général parfaitement stériles. Des schémas plus compliqués impliquant des backcross avec un des parents ont été proposés pour expliquer certains cas de génomes mitochondriaux et chloroplastiques (De Langhe et al., 2009). La divergence entre les génomes A et B permet encore des hybridations qui toutefois restent difficiles. Il a été montré qu'au moins 3 paires de chromosomes présentaient une très faible affinité (Jeridi et al., 2011). Il a également été observé par la méthode FISH que certains hybrides ne possédaient pas toujours les jeux complets de génomes A et B. Ce cas a été observé par D'Hont et al. (2000) pour l'hybride ABB 'pelipita' qui possède 25 B et 8 A au lieu de 22 B et 11 A.

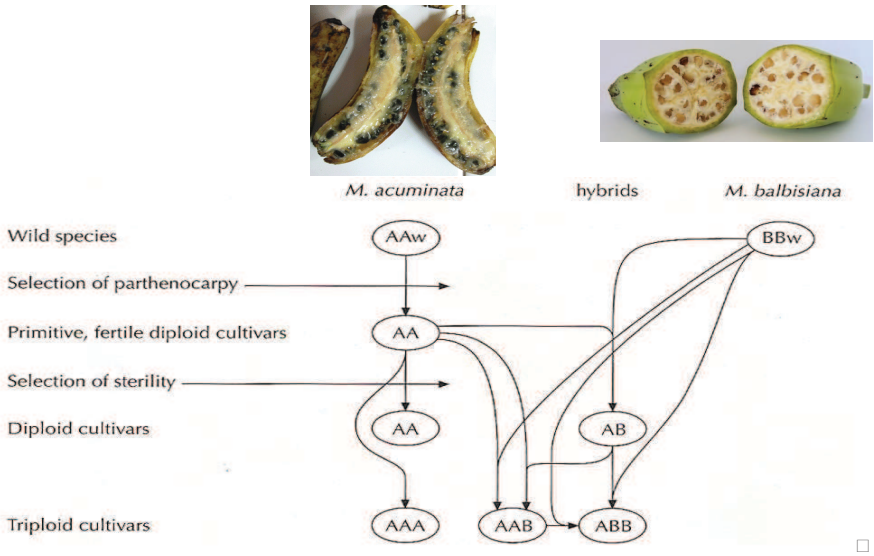


Figure 1-16 : Schéma de domestication des bananiers

Domestications des bananiers à partir des deux principales espèces *Eumusa* en fonction des critères de sélection pour la consommation. La photo au-dessus à gauche est celle d'une banane sauvage issue de l'espèce *M. acuminata* var. burmanica ; et la photo en haut à droite est celle d'une banane sauvage issue de l'espèce *M. balbisiana*. *Adapté de Carreel, (1994)*

De ces schémas de diversification, il apparait que l'homme est un facteur essentiel qui en déplaçant des formes normalement isolées a permis des hybridations (interspécifiques ou intersub-spécifiques) entre génomes suffisamment divergents pour induire des perturbations de sexualité. La confrontation de ces résultats avec des données de linguistiques a permis de montrer que ces voies de déplacements correspondaient effectivement à des migrations attestées. Des résultats d'archéologie attestant la présence de restes de bananiers (phytolithes) dans des niveaux datés par ailleurs ont permis d'ancrer les principaux évènements de la domestication (Perrier et al., 2011). On sait ainsi que les premiers AAcv apparaissent il y a au moins 7000 ans en PNG, que des AAcv générés vers Java ont eu un succès tel qu'on les a transportés jusqu'à la région Thaî où ils ont formé avec les AA locaux divers AAA (dont les AAA Cavendish du commerce international). Ils sont allés jusqu'en Inde où ils se sont hybridés avec des BB locaux pour donner diverses variétés de AAB. Ils sont encore aujourd'hui cultivés sous forme diploïde en Afrique de l'Est. On sait aussi que les AAB plantains d'Afrique viennent de la région des Philippines, où un BB descendu du sud de la Chine a rencontré une forme AAcv locale. Les marins austronésiens ont emporté un exemplaire de ces hybrides vers l'Afrique il y a plus de 2 500 ans, où il s'est ensuite répandu par multiplication végétative et s'est largement diversifié par mutations somaclonales (Figure 1-17).

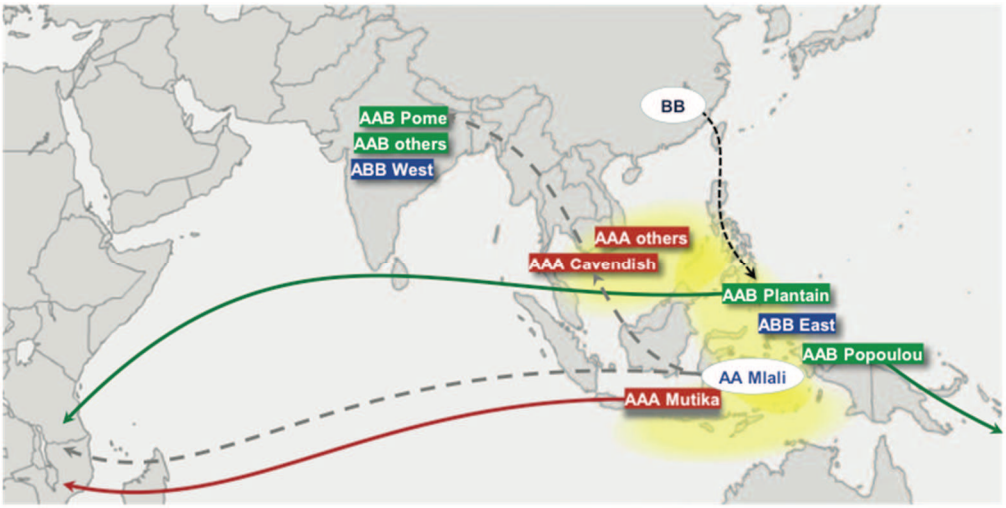


Figure 1-17 : Origines et migrations des principaux sous-groupes de bananiers triploïdes.

Les flèches pleines indiquent les migrations longues distances préhistoriques de cv triploïdes en Afrique et dans les îles du Pacifique. Les flèches grises en pointillé indiquent (i) les migrations du sous-groupe Mlali AAcv, qui ne se trouve plus en Asie du Sud-Est aujourd'hui, vers le continent Asiatique, où il a contribué à la formation du Cavendish AAA, puis en Inde, où il a rencontré *M. balbisiana* et donné les AAB Pome, et (ii) les migrations du sous-groupe Mlali vers la côte africaine. Les flèches noires en pointillé indiquent le parcours de *M. balbisiana* du sud de la Chine vers la Nouvelle-Guinée en passant pas Taiwan et les Philippines, dans l'hypothèse où les austronésiens auraient contribué à la dispersion de cette espèce. *D'après Perrier et al., (2011)*

Une avancée majeure pour l'étude du bananier a été réalisée très récemment. En effet la séquence du génome du bananier *Musa acuminata var. Pahang* haploïde doublé a été obtenue (D'Hont et al., 2012). Les données apportées par ce séquençage sont essentielles pour mieux comprendre le génome du bananier et son organisation. D'autre part, cela va permettre de rechercher les gènes impliqués dans les différents processus d'intérêt agronomique comme les résistances aux pathogènes, au stress et les qualités gustatives de la banane. De plus il s'agit là d'une porte d'entrée vers une meilleure compréhension des autres génomes du genre *Musa* car il existerait une synténie importante entre les génomes des espèces de ce groupe (Heslop-Harrison et Schwarzacher, 2007). Des approches de re-séquençage massif sont alors possibles pour la comparaison de génomes ou pour aider à la sélection.

Toutes ces données sur la diversité des bananiers, ainsi que sur la connaissance de leur génome servent de base à l'amélioration variétale pour faire face aux défis que pose la culture du bananier aujourd'hui. En effet le bananier est un enjeu très important d'un point de vue économique et social, comme nous l'avons vu précédemment. Un problème majeur de la culture bananière pour l'exportation est la monoculture. En effet les bananeraies du monde entier qui cultivent des bananes dessert pour l'exportation, cultivent la même variété à quelques variations somaclonales près : la banane *M. acuminata var cavendish* (AAA). Ceci pose le problème de la vulnérabilité de cette culture en cas d'attaque de pathogène, surtout dans un contexte où la réduction de la consommation d'intrant est un enjeu majeur. Ainsi, des maladies fongiques telles que les cercosporioses (raies noires et raies jaunes) qui détruisent les cultures, se déplacent à grande vitesse autour du monde. La solution de traitements aériens massifs pour leur contrôle ne sera plus acceptable à court terme (déjà interdit dans les Antilles françaises sauf dérogation). De même, les bananiers pour la consommation locale font l'objet d'attaques sévères qui réduisent considérablement les productions, dans un contexte où les traitements chimiques ne sont pas accessibles. Une solution pour les améliorateurs afin de lutter contre ces maladies tout en restreignant l'utilisation de pesticides est l'introgression de génome *M. balbisiana* dans des fonds génétiques *M. acuminata,* par des croisements inter spécifiques permettant d'apporter les caractères recherchés (tolérance aux stress hydrique, résistance aux maladies fongiques) tout en gardant les qualités agronomiques et gustatives demandées. Cependant, rapidement après la mise en place de stratégies de ce type, un problème majeur a été observé lié à

l'apparition et le développement du *Banana streak virus* (BSV) dans les hybrides nouvellement créés. Les hybrides interspécifiques présentent des intégrations du BSV dans le génome *M. balbisiana* ayant la capacité de restituer des particules virales et conduisant à l'apparition de la maladie de la mosaïque en tirets du bananier. L'utilisation de *M. balbisiana* dans les programmes d'amélioration a été suspendue en attendant la disponibilité de ressources génétiques exemptes d'insertions virales, ou du moins incapables de restiuer des particules virales.

6-*Le Banana streak virus* (BSV), le virus modèle

6-1 La maladie de la mosaïque en tiret du bananier

Les bananiers sont infectés par de nombreux pathogènes (bactéries et champignons) dont les virus. Parmi les cinq principaux virus connus comme responsables de maladies, le virus de la mosaïque en tirets ou *Banana streak virus* (BSV), responsable de la maladie de la mosaïque en tirets des bananiers (ou Banana Streak Disease ou BSD), est le plus atypique. En effet la maladie est le fait d'un complexe d'espèces de BSV. Ces virus dont l'hôte principal est le bananier présentent entre eux une diversité importante les structurant en espèces distinctes bien que responsable d'une même maladie. Ils sont transmis de plante à plante par cochenilles de manière non-circulante. Quatre espèces de cochenilles (*Hemiptera : Coccoidea ; Pseudicoccidae*) infectant les bananiers sont capables de transmettre les virus BSV : la cochenille des agrumes (*Planococcus citri*), la cochenille de la vigne (*P. ficus*), la cochenille rose de la canne à sucre (*Saccharicoccussacchari*) et la cochenille de l'ananas (*Dysmicoccusbevipes*) (Kubirica et al., 2001 ; Meyer et al., 2008). La cochenille est sédentaire et certains stades larvaires mobiles peuvent ainsi assurer une transmission de plante à plante (Lockhart, 1995). Mais les mouvements de cochenilles sont très lents, sur de courtes distances pouvant être augmentés par l'aide de fourmis qui prennent alors le relais. Habituellement, lorsque les symptômes sont détectés en champs un arrachage systématique de la zone impliquée permet de maîtriser la propagation de la maladie.

Bien qu'inféodés au genre *Musa*, le BSV peut être transmis expérimentalement par cochenille au genre *Ensete* (*E. ventricosum*) et à la canne à sucre (Lockhart et al., 1995). Des essais d'inoculation de BSV sur des bananiers du genre *Musa* comme *M. textilis* (Lockhart et Jones, 2000) ou *M. balbisiana* (Lheureux, 2002) n'ont pas abouti à une infection virale.

La maladie de la mosaïque en tirets du bananier a été décrite pour la première fois en 1966 en Côte d'Ivoire (Lassoudière, 1974 ; Yot-Dauthy et Bové, 1966). Cette maladie peut provoquer des pertes de rendements allant de 5% à 95% sur une parcelle, ces taux peuvent être inférieurs pour des cultivars résistants (Dahal et al., 1998 ; Daniells et al., 2001). Les symptômes de la maladie sont des tirets chlorotiques au niveau des feuilles qui évoluent en nécrose (Figure 1-18). Dans des cas sévères, l'infection peut aboutir à la mort des plants par nécrose du méristème apical conduisant à l'éclatement du pseudo-tronc (Lockhart et Jones, 2000). Plusieurs stress abiotiques, comme de fortes amplitudes de régimes hydriques ou de températures, semblent jouer un rôle sur l'expression des symptômes du BSD (Dahal et al., 1998 ; Daniells et al., 2001). Les techniques moléculaires de diagnostic ont permis de mettre en évidence que des plantes asymptomatiques peuvent être porteuses de particules virales. Peu d'épidémies ont été détectées à travers le monde, elles sont maintenant restreintes à l'Afrique de l'Ouest, principalement en Ouganda et au Nigéria bien que le virus ait une répartition mondiale (Kubiriba et al., 2000 ; Harper et al., 2004,2005) où il semblerait qu'elles soient maintenues par utilisation de rejets ou vitroplants contaminés plus qu'un contexte épidémique réel.

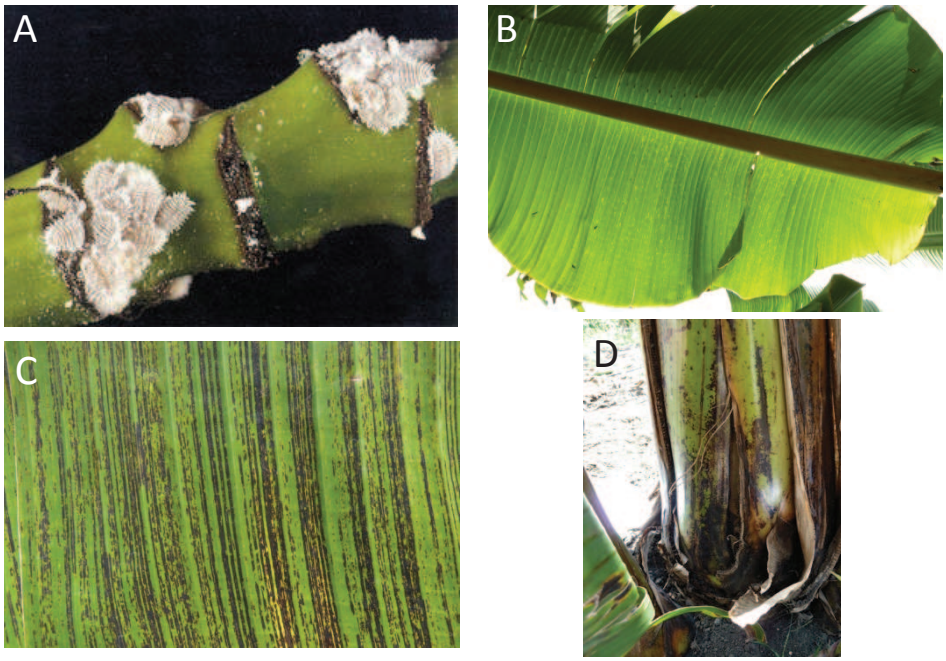


Figure 1-18 : Vecteur et symptômes de la maladie de la mosaïque en tirets des bananiers (BSD)

(A) Cochenilles *Planaccocus citri* présentes sur un pédoncule de bananier
(B) Chloroses en tirets jaunes sur le limbe des feuilles
(C) Nécrose sur le limbe des feuilles
(D) Eclatement du pseudo tronc à sa base entrainant la mort du plant

Le problème majeur lié au BSV est causé par sa biologie. Ce virus existe sous deux formes, l'une libre et l'autre intégrée au génome de *M. balbisiana*. C'est la forme intégrée qui aujourd'hui pose problème puisque, comme déjà évoqué, des séquences intégrées de virus (eBSV) sont capables d'être exprimées et d'induire des particules virales et en conséquence le développement de la maladie de la mosaïque en tirets. L'utilisation de bananiers ayant des eBSV pose le problème de leur expression potentielle non contrôlée. Ce phénomène prend son ampleur pour les programmes d'amélioration génétique utilisant des géniteurs à risque porteurs des intégrations. Ces géniteurs transmettent à la descendance l'eBSV qui peut par la suite, lors de la culture en champs, être à l'origine de l'apparition de la maladie (Hohn et al., 2008 ; Hull et al., 2000 ; Iskra-Caruana et al., 2003 ; Côte et al., 2010). Des épidémies ont été rapportées dans les années 1990 un peu partout autour du monde, en particulier en Ouganda où elles ont été les plus importantes, suite à la diffusion d'hybrides interspécifiques porteurs des intégrations infectieuses. Ce phénomène a été assimilé à une émergence voir une ré-émergence de la maladie de la mosaïque en tirets là où ce n'était qu'une explosion virale liée à la culture d'hybrides nouvellement créés qui se sont révélés être à risque. Ceci a amené le CIRAD à faire un moratoire sur l'utilisation des génomes B dans les programmes d'amélioration, malgré les avantages qu'ils peuvent apporter.

6-2 Biologie du BSV

Comme présenté dans la partie 3-4, le BSV est un pararetrovirus de plante non-enveloppé qui appartient au genre *Badnavirus,* et à la famille *Caulimoviridae* (Fauquet, 2005). Il possède des virions bacilliformes qui ont une taille de 30 par 150nm et un génome de 7,4 kpb à ADN circulaire double brin. La multiplication virale se fait par l'intermédiaire d'une reverse transcriptase. Comme tous les Badnavirus, il possède 3 ORF codantes sur le brin sens, à l'exception du *Cacao swollen shoot virus* (CSSV) avec 5 ORF et le *Citrus mosaic bacilliform virus (*CMBV*)* avec 6 ORF (Figure 1-10). Une des particularités de ces virus est de posséder pour l'ORF 1 un codon start particulier : CTG. Les ORF 1 et 2 codent pour deux petites protéines d'environ 22 et 14 kDa respectivement (Harper et Hull, 1998) dont la fonction n'est pas connue. Il a été montré pour le badnavirus *Commelina yellow mottle virus* (ComYMV) que la protéine synthétisée à partir de l'ORF 1 servait à la formation des virions immatures et que la protéine codée par l'ORF 2 aurait un rôle dans l'assemblage du virus (Cheng et al., 1996). Selon Jacquot et al. (1996), chez le CSSV la protéine codée par l'ORF 2 pourrait intervenir dans l'encapsidation de l'ADN, voire dans le transport et la protection des ARN viraux. L'ORF 3 code pour une polyprotéine de 208 kDa qui possède plusieurs motifs

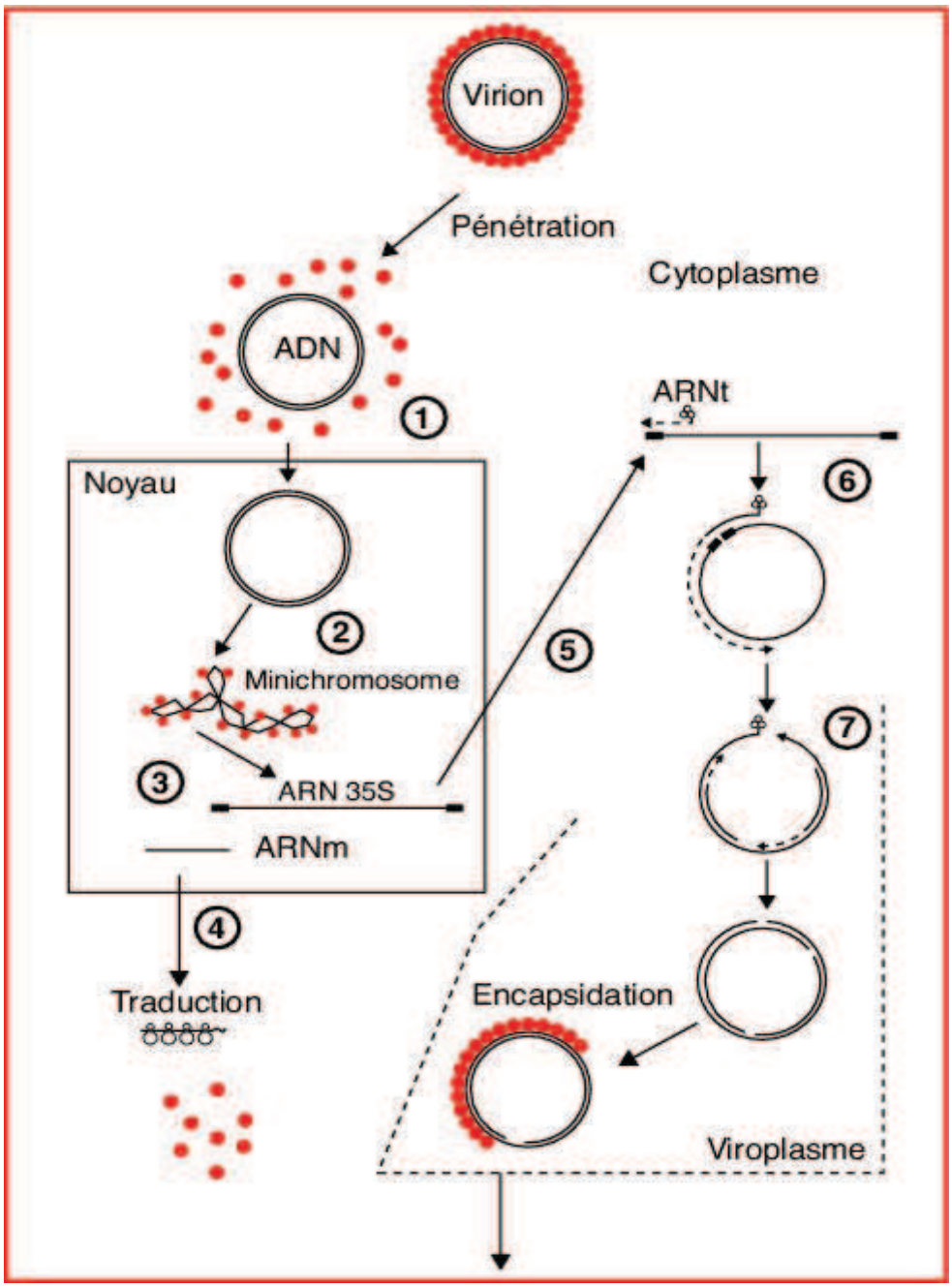


Figure 1-19 : Cycle de multiplication des *Caulimoviridae*

Basé sur les connaissances obtenues principalement sur le *Cauliflower mosaic virus* (CaMV)

1. Après la dissociation de la particule virale, l'ADN pénètre dans le noyau. **2.** L'ADN viral est alors transformé en une molécule super-enroulée par réparation des interruptions de séquence (Gap) à l'aide d'enzymes cellulaires. L'association de cette molécule avec des histones conduit à la formation d'un minichromosome. **3.** La transcription génère un ARN pré-génomique 35S de taille supérieure au génome et d'autres petits ARN dont l'ARN 19S. **4.** Exportation des ARN pré-génomiques dans le cytoplasme **5.** L'ARN pré-génomique sert à la fois d'ARNm pour la synthèse des protéines virales structurales et enzymatiques et de matrice à la transcriptase inverse pour la synthèse du génome viral. **6.** L'amorçage de l'étape de transcription inverse s'effectue à l'aide de l'ARNt^{met} permettant la synthèse dans un premier temps d'un brin d'ADN (-). L'ARN matriciel est dégradé en petits ARN par la RNAse H, sauf au niveau de certaines régions riches en purines. **7.** Ces petits ARN non dégradés sont utilisés comme amorces pour la synthèse des brins d'ADN complémentaires (+). L'intégration de séquences de certains *Caulimoviridae* dans le génome de la cellule résulterait d'une interaction entre le cDNA et l'ADN cellulaire par recombinaison illégitime. Remarque : Chez les *badnavirus*, seul l'ARN prégénomique est synthétisé.

Adaptée de Iskra-Caruana et al., (2003)

conservés et connus. Le clivage par protéolyse de cette polyprotéine restitue la protéine de capside virale (CP), l'aspartate protéase (AP), la reverse transcriptase (RT), la RNase H (RH) ainsi qu'une protéine putative de mouvement de cellule à cellule (Harper et Hull, 1998). Le cycle de réplication des BSV et des Badnavirus en général n'est pas directement connu et les hypothèses le concernant s'inspirent du cycle du *Cauliflower mosaic virus*(CaMV) espèce modèle des virus à ADN du genre *Caulimovirus*de la famille *Caulimoviridae*. Les principales étapes du cycle de réplication semblent conservées entre les différents *Caulimoviridae* donc nous pouvons penser que le cycle se passe de la même façon pour le BSV que pour le CaMV, même si certaines différences ont été mises en évidence, comme le nombre de transcrits (un seul pour le BSV) et la forme minichromosome qui n'a pas été identifiée pour le BSV (Lheureux, 2002). Ce cycle est décrit dans la figure 1-19.

6-3 La diversité du BSV

Dès les premières études il a été montré qu'il existe une forte hétérogénéité entre les différents isolats de BSV, que ce soit d'un point de vue sérologique, génétique et même biologique car ils peuvent induire des symptômes différents sur les mêmes plantes (Lockhart et Olszewski, 1993). L'International Commitee on Taxonomy of Viruses (ICTV) a déterminé que pour différencier deux espèces du genre *Badnavirus* il fallait une différence d'identité nucléique de plus de 20% au niveau de la région RT/RNase H de l'ORF 3. Des études de diversité des virus libres ont été réalisées en séquençant de manière aléatoire cette partie de la séquence lors d'épidémies de BSV en Ouganda, en Australie et à l'île Maurice (Harper et al., 2004-2005 ; Geering et al. 2005 ; Jaufeerally-Fakim et al., 2006). Ces études ont révélé qu'effectivement les isolats pouvaient avoir des différences génétiques supérieures à 30%. Une des hypothèses avancées pour expliquer cette diversité est que ces virus utilisent la RT pour la réplication du génome viral. Cette ADN polymérase ARN dépendante ne possède pas d'activité proof-reading et fait plus d'erreurs que les ADN polymérases ADN dépendantes (Duffy et al., 2008). Néanmoins les autres virus de ce genre et de cette famille utilisent la RT pour se multiplier et ne présentent pas une telle diversité d'espèces sur un même hôte à l'exception des Badnavirus de l'igname (Bousalem et al., 2008 ; Kenyon et al., 2008). Les différentes prospections lors d'épidémies de BSV ont permis de décrire 7 espèces de BSV libres distinctes et de séquencer leurs génomes complets : *Banana streak obino l'ewai virus* (BSOLV) (Harper et Hull, 1998), *Banana streak mysore virus* (BSMyV) (Geering et al., 2005b), *Banana streak vietnam virus* (BSVNV) (Lheureux et al., 2007), *Banana streak goldfinger virus* (BSGFV) (Gayral et al., 2008), *Banana streak imové virus*(BSImV) (Geering et al., 2011),

Banana streak Yunnan virus (Muller E., communication personnelle), *Banana streak perou virus* (Muller E., communication personnelle) et *Banana streak cavendish virus*(BSCaV) (James et al., 2011).

En plus des génomes viraux complets décrits et séquencés, il existe dans les bases de données de nombreuses séquences de RT/RNase H appartenant à des espèces BSV ou « BSV-like » pour lesquelles la correspondance avec des génomes épisomaux n'a pas été faite. Ces séquences ont été obtenues à partir d'échantillons de bananiers prospectés dans des zones d'épidémies de BSV en Ouganda ou en Australie. En effet, les techniques utilisées pour détecter ces séquences (à l'aide de primers dégénérés de séquences communes aux BSV) ne permettent pas de discriminer les séquences intégrées de celles qui ne le sont pas (Harper et al., 2005 ; Geering et al., 2000,2005).La diversité des BSV sera discutée dans la partie 7-2 avec celle des eBSV puisque en fait une partie de la diversité des BSV réside dans la diversité des séquences virales intégrées dans le génome des bananiers du genre *Musa.*

7- Les endogenous *Banana streak virus* (eBSV), le modèle d'étude des EPRV

7-1 Histoire des eBSV

Les BSV ont été longtemps considérés comme une menace mineure pour la culture bananière et ce n'est qu'avec l'explosion de foyers épidémiques à travers la zone mondiale de production de la banane, que cette menace a été prise au sérieux. En effet c'est lors de la culture après micro-propagation d'hybrides interspécifiques de bananiers sains et sans qu'il y ait de contaminations extérieures de BSV possible que ces foyers épidémiques sont apparus. L'utilisation de techniques de détection moléculaire tel que la Polymérase Chain Reaction (PCR)et le southern blot ont permis de mettre en évidence la présence de séquences virales intégrées chez les plantes du genre *Musa* (Lafleur et al., 1996), et de les associer avec l'émergence du virus. Cependant c' est seulement en 1999 que le lien entre intégrations BSV et développement de la maladie a été clairement établi et plus particulièrement la présence des intégrations dans le génome des plantes de l'espèce *M. balbisiana*. Ceci a amené à une prise de conscience de la contrainte majeure que représentait le BSV pour l'amélioration des bananiers, alors même que le génome B apporte de nombreux avantages, déjà évoqués.

Les séquences de BSV intégrées ont été nommées eBSV pour endogenous BSV, et c'est leur présence et leur capacité à induire la maladie qui a joué un rôle dans l'expansion

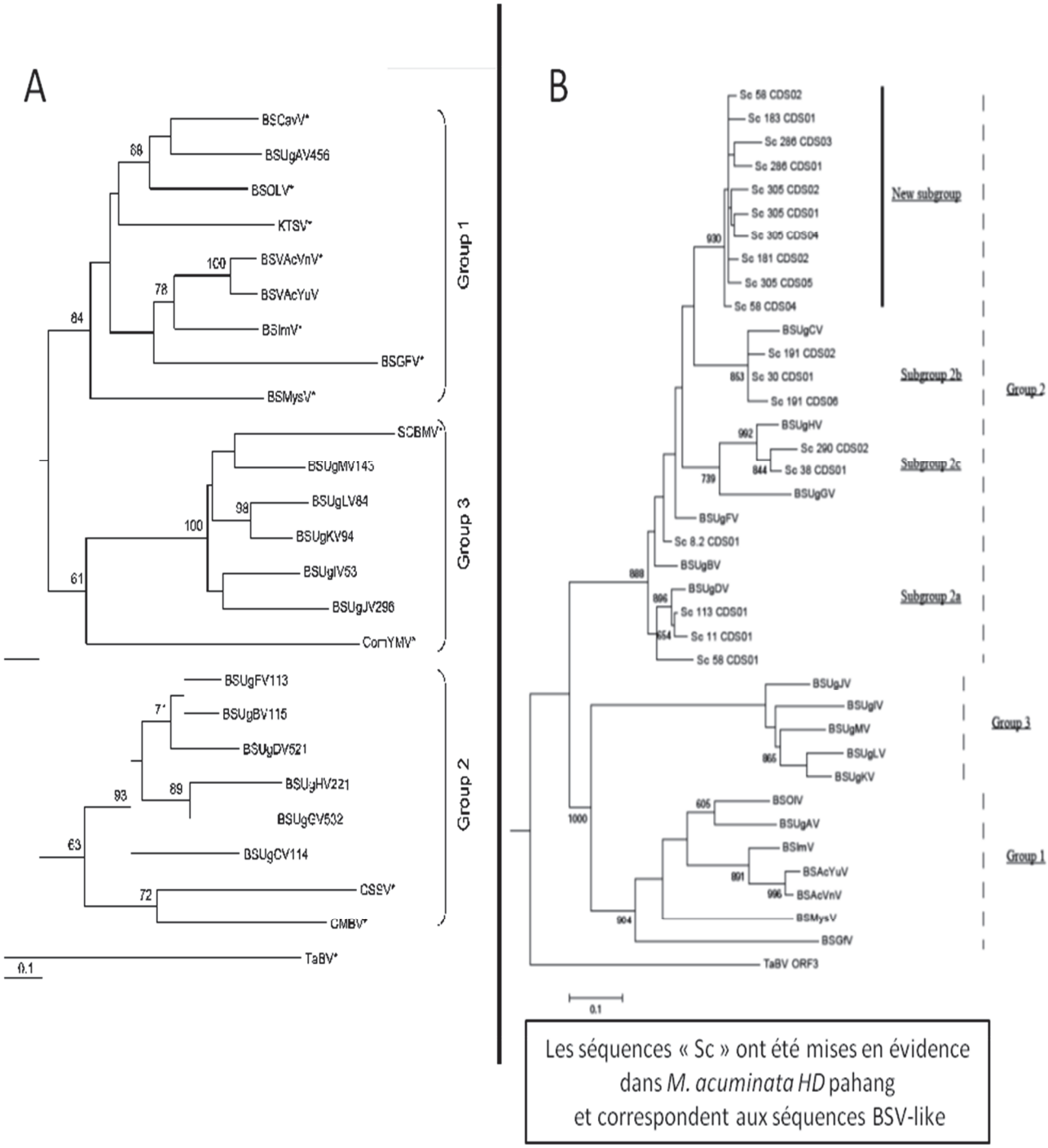


Figure 1-20 : Phylogénie des badnavirus basée sur les séquences RT/Rnase H

A : Phylogénie obtenus lors de l'étude de Gayral et Iskra-Caruana (2009), les valeurs bootstraps supérieures à 60% sont présentées sur les branches de l'arbre. Les virus libres totalement séquencés et décrits sont présentés avec une astérisque. Les séquences BSUg correspondent au séquences de BSV décrites par Harper et al (2002) sur des échantillons de bananiers lors d'épidémies en Ouganda. L'arbre est enraciné avec la séquences du *Taro bacilliform virus* (TaBV).

B : Phylogénie obtenus lors de l'étude de D'Hont et al., (2012) pour le séquençage du bananier *M. acuminata* HD pahang. Tous les séquences BSV-like retrouvés intégrés dans le génome de ce bananier ont été utilisés pour obtenir cette phylogénie, ils sont notés « Sc » et correspondent aux BSV-like cités dans cette thèse. Les séquences appartenant aux groupes mis en évidence en A ont été aussi utilisés, il s'agit des séquences notés « BSUg ». La séquence du Taro bacilliform virus a, là aussi, été utilisée pour enraciner l'arbre. Un nouveau sous-groupe a pu être mis en évidence il est noté New-subgroup.

D'après Gayral et Iskra-Caruana., (2009) et D'Hont et al., (2012)

géographique du BSV, suite à la production et la diffusion mondiale d'hybrides interspécifiques porteurs des eBSV (Côte et al., 2010 ; Dallot et al., 2000 ; Harper et al., 1999 ; Ndowora et al., 1999 et Lheureux et al., 2003). Par la suite, il a été montré que le génome de *M. balbisiana* était porteur d'eBSV de 4espècesBSV libres parmi lesquelles trois ont été décrites comme infectieuses ; BSOLV, BSImV et BSGFV avec la notation eBSOLV, eBSImV et eBSGFV pour les formes intégrées (l'eBSV non infectieux intégré correspond à BSMyV noté eBSMyV) (Gayral et al., 2008 ; Geering et al., 2005b ; Harper et al., 1999 ; Iskra-Caruana et al., 2003 ; Ndowora et al., 1999). Leur seule présence dans le génome ne permet pas la production de particules virales ; des facteurs extérieurs aux séquences seules sont de toutes évidence impliqués comme le facteur génétique BEL (BSV expressed Locus) mis en évidence par Lheureux et al., (2003).

7-2 Phylogénie des eBSV

Comme décrit dans la partie 6-3, de nombreuses séquences de BSV sont disponibles dans les bases de données, que ce soit des BSV libres ou intégrés. Bien que les séquences d'eBSV infectieux n'aient été décrites que chez les bananiers *M. balbisiana*, il a été mis en évidence des séquences de BSV intégrées sur d'autres espèces de ce genre comme *M. acuminata* et *M. schizocarpa* (Geering et al., 2005b ; Gayral et Iskra-Caruana, 2009).

La phylogénie basée sur les séquences de RT/RNase H proposée par Gayral et Iskra-Caruana en 2009 a permis d'établir la structuration en trois groupes du genre Badnavirus et confirmer la répartition polyphylétique de toutes les séquences BSV disponibles proposée par Harper et al., (2005) (Figure 1-20A). Concernant le BSV, le groupe 1 rassemble l'ensemble des séquences des espèces BSV décrites à l'état libre ainsi que les séquences virales intégrées (eBSV) correspondantes. On trouve également dans ce groupe les BSVUgA d'Ouganda décrits par Harper et al., (2005) pour lesquels les virions n'ont pas été séquencés Le groupe 2est constitué des séquence appelées « BSV-like badnavirus » qui sont intégrées dans les génomes *Musa* et pour la plupart dans le génome de *M. acuminata* et les BSUgMV, BSUgLV, BSUgKV, BSUgIV, BSUgJV d'Ouganda. Le groupe 3 rassemble exclusivement les BSUgFV, BSUgBV, BSUgDV, BSUgHV, BSUgVV, BSUgCV et correspond à des séquences de virus épisomaux pour lesquels aucune intégration virale dans les génomes bananiers n'existe à ce jour (E. Muller comm. personnelle)De façon intéressante, le groupe 2 intègre des Badnavirus infectant d'autres plantes que le bananier non décrits comme intégrés au génome de leur plante hôte comme le *Sugarcane baciliform mor virus* (infectant la canne à

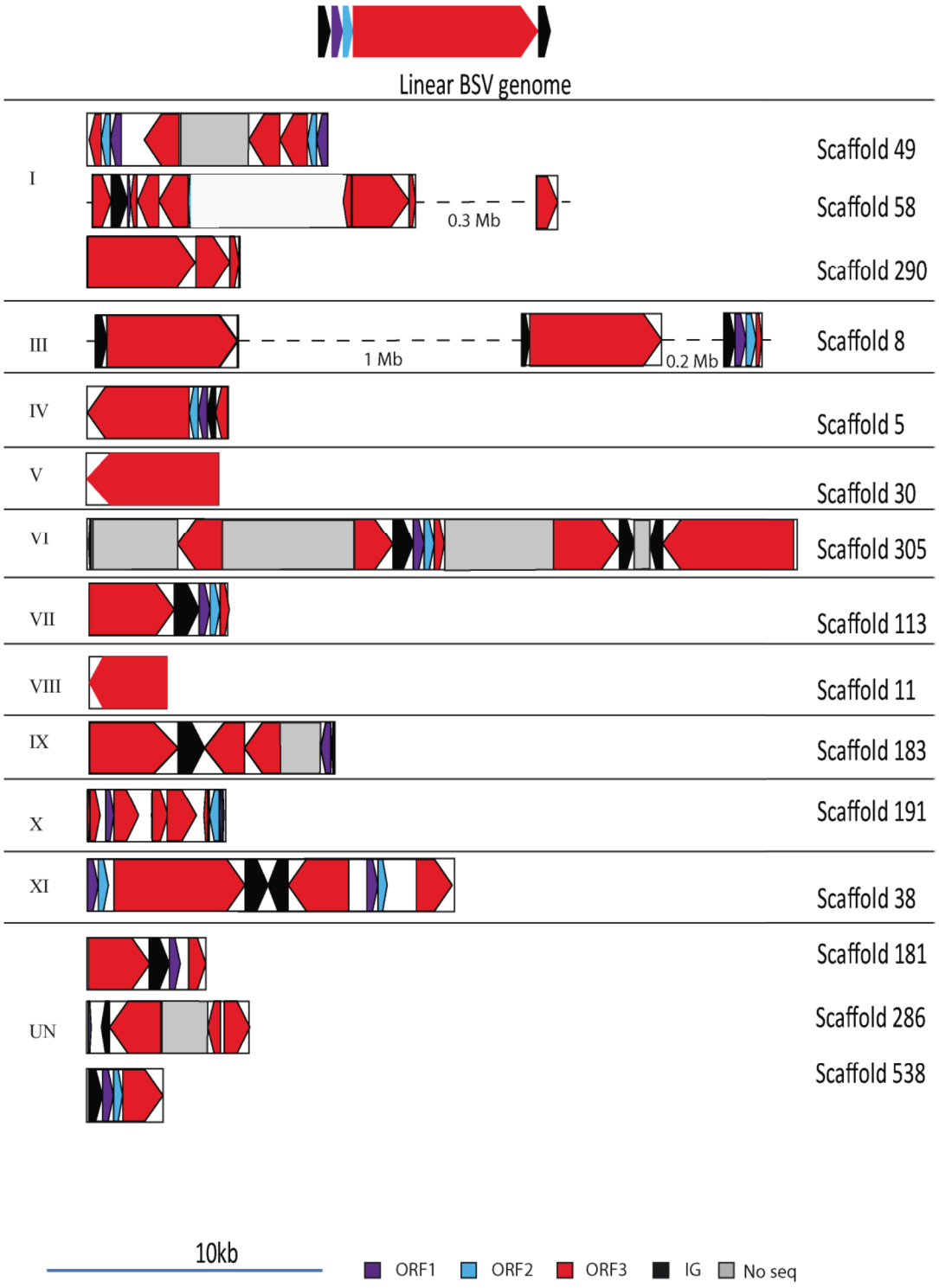


Figure 1-21 : Structure des séquences BSV-like présentes dans le génome *de M. acuminata*

Au-dessus est représentée la vue linéaire du génome de BSV. Les flèches indiquent l'orientation des séquences virales. Le nom des scaffolds sur lesquels ont été découvertes les séquences sont indiqués à droite et les chromosomes correspondants à gauche. *D'après D'Hont et al., (2012)*

sucre), *Commelina yellow mosaic virus*(infectant les *Commelinaceae*, le *Cacao Swollen Shoot Virus* (infectant le cacao) et le *Citrus mosaic bacilliform virus*(infectant les agrumes).

Cette phylogénie BSV a permis de montrer l'origine polyphylétique des BSV libres répartis dans les groupes 1 et 3 ainsi que celle des séquences intégrées réparties dans les groupes 1 et 2. Les BSV infectant les bananiers ont une origine commune au niveau du groupe 1. Les eBSV correspondant à ces BSV auraient quant à eux une origine différente des BSV-like présent dans le génome *Musa*. Cette étude a remis en cause la taxonomie des Badnavirus et mériterait donc d'être repensée (Gayral et Caruana, 2009).

Dernièrement, le séquençage du bananier *Musa acuminata* HD pahang a permis d'augmenter les connaissances vis à vis de la diversité des eBSV. Des BLAST ont été réalisés sur le génome de ce bananier afin d'évaluer la réalité de séquences BSV et BSV-like intégrées (figure 1-20B). Aucune intégration des espèces BSV du groupe 1 n'est décrite dans ce génome, ce qui confirme bien leur restriction aux génomes *M. balbisiana*. De façon très intéressante, il est décrit des intégrations eBSV à 24 loci, réparties sur 10 des 11 chromosomes (Figure 1-21). De plus, la taille des intégrations est importante allant de quelques centaines de paires de bases jusqu'à 18kpb. Les structures des intégrations sont fortement réarrangées et pourraient provenir pour beaucoup d'une intégration unique qui s'est ensuite dupliquée au niveau de ce même site (D'Hont et al., 2012). Une phylogénie basée sur la séquence RT/RNase H a été réalisée (figure 1-20B). Cette phylogénie montre très clairement que les eBSV découverts appartiennent tous au groupe 2 décrit précédemment. Ces eBSV correspondent à l'intégration d'au moins 4 Badnavirus différents et se répartissent pour certains dans les groupes BSVUgCV, BSUgDV et BSUgHV et pour d'autres forment un nouveau sous-groupe au sein du groupe 2. Des travaux en cours montreraient qu'une partie des séquences observées chez HD pahang et Ouganda sont ubiquitaires à *M. balbisiana* et *M. acuminata* (Muller et Chabannes communication personnelle). Ces résultats supporteraient l'hypothèse d'un évènement ancien d'intégration de BSV dans le génome de l'ancêtre commun au genre *Musa.*

7-3 Les eBSV infectieux

7-3-1 L'activation des eBSV infectieux

Différents stress abiotiques ont été identifiés comme étant des activateurs de la production de particules à partir des eBSV dans le génome de *M. balbisiana* : les stress hydriques ou thermiques, la culture in-vitro et les croisements interspécifiques (Côte et al., 2010 ; Dahal et

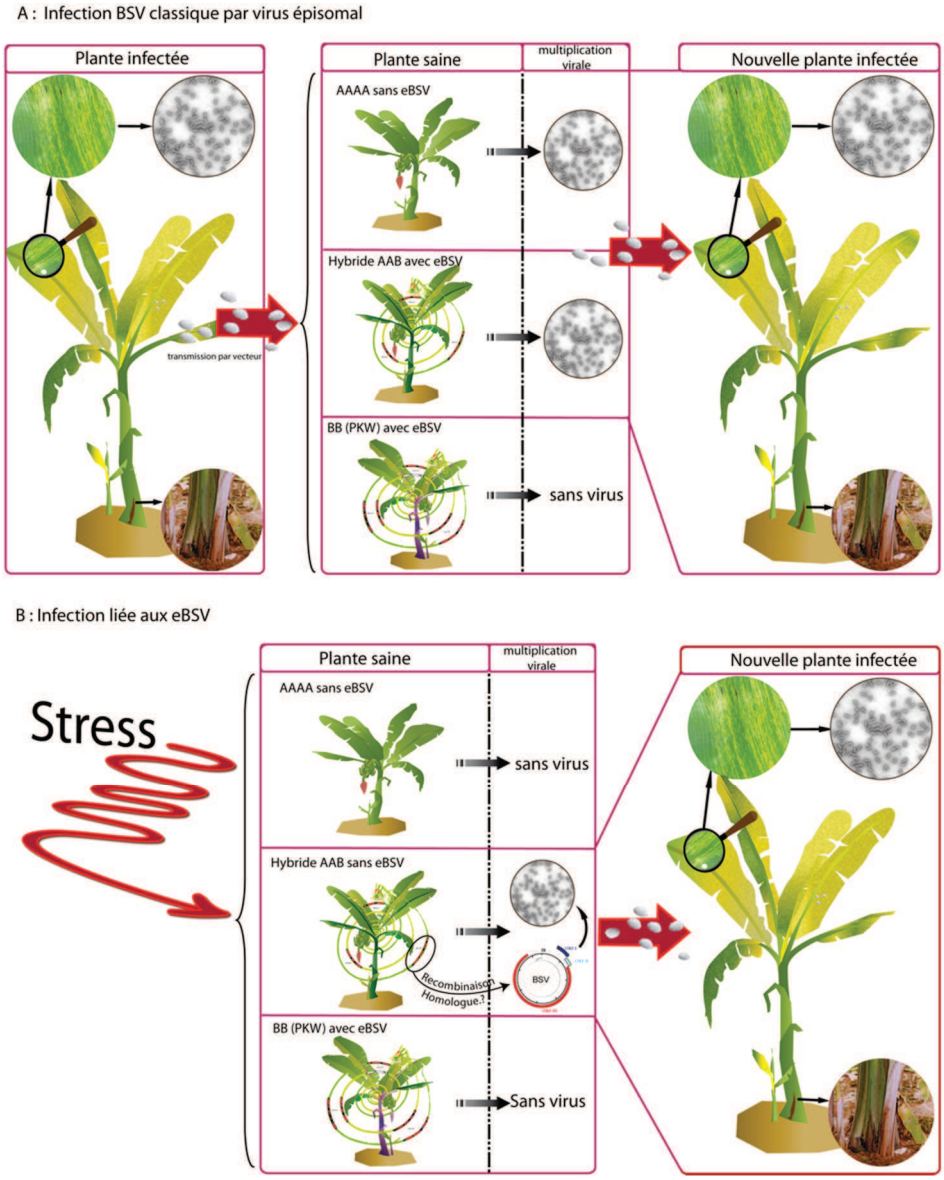


Figure 1-22 : Représentation schématique des deux types d'infections du BSV sur les bananiers

Soit A, une interaction triangulaire classique ou B, une interaction quadripartites inhabituelle

A : Dans l'interaction triangulaire classique, le BSV est transmis par un vecteur (les cochenilles) à partir de plantes infectées (source d'inoculum primaire) vers les plantes exemptes de virus, qui deviennent alors infectées et disponibles (source d'inoculum secondaire) pour propager la maladie. La propagation du BSV est illustré pour trois sortes de plantes indemnes de virus avec des génotypes différents: une plante sans eBSV (AAAA génome); une plante hybride interspécifique (AAB) qui est porteuse d'eBSV infectieux (une seule copie du génome B), et une plante sauvage diploïde Musa balbisiana (Pisang Klutuk Wulung [PKW]), qui est aussi un porteur des eBSV infectieux. Le développement de la maladie se produit dans deux plantes, celle sans les eEBSV et celles hybrides interspécifiques hébergeant les eBSV. Ces plantes deviennent alors des réservoirs de virus pour la transmission horizontale ou la transmission vectorielle. En revanche, l'infection ne se produit jamais dans la plante PKW par transmission vectorielle ; PKW semble résistant au BSV. Le mécanisme expliquant cette résistance est encore inconnue.

B : Le cas de l'interaction BSV quadripartites inhabituel, la primo-infection est due à la production de particules virales à partir des séquences eBSV. C'est le stress qui déclenche l'expression des eBSV et seules les plantes hybrides interspécifiques sont susceptibles d'être infectées. Le mécanisme expliquant la transformation de particules virales à partir des eBSV est encore non caractérisé, bien que la recombinaison homologue soit fortement suspectée. L'hybride infecté devient alors un réservoir pour la transmission horizontale, tel que décrit dans A. La plante diploïde BB PKW, qui abrite les mêmes intégrations infectieuses que les plantes hybrides mais à un niveau mono-ou di-allélique, présente une résistance naturelle à cette activation interne. Là encore, le mécanisme sous-jacent de cette résistance n'est pas connu.

Adapté de Iskra-Caruana et al., (2010)

al., 1998, 2000 ; Dallot et al., 2000 ; Lheureux et al., 2003). Le premier facteur influençant l'activation des eBSV a été déterminé par Dallot et al. (2000), qui ont démontré que la micro-propagation par culture de méristème influençait positivement l'expression de l'eBSOLV de l'hybride tétraploïde nouvellement créé interspécifique FHIA 21 (AAAB). Côte et al. (2010) ont confirmé ces résultats pour les eBSOLV et eBSGFV d'un autre hybride tétraploïde le CRBP39 (AAAB) ainsi que pour deux hybrides interspécifiques naturels triploïdes (AAB), Kelong Mekintu et Black Penkelon. Parmi les différents bananiers testés lors de cette étude, seul CRBP39 avait une activation de l'EBSGFV. Ceci peut s'expliquer par une activation différentielle de eBSOLV et eBSGFV faisant appel à des régulations différentes ou bien par des structures eBSGFV différente entre ces cultivars.

Le deuxième stress génomique induisant la production de particules virales par les eBSV est le croisement génétique interspécifique (Lheureux et al., 2003). Ce croisement entraine des bouleversements génomiques dus à la réunion de deux génomes différents et à la fréquente polyploïdisation qui y est associée. La production de particules virales à partir des séquences eBSV n'a été détectée que chez des hybrides interspécifiques résultant du croisement entre *M. balbisiana* et *M. acuminata.* De plus, cette production de particules virales n'a été observée que chez des hybrides ne possédant qu'un seul génome B (AAB et AAAB). Il semblerait donc qu'il s'agisse d'un pré-requis essentiel pour que l'activation puisse avoir lieu. On montre qu'à contrario les bananiers diploïdes *M. balbisiana* sont quant à eux résistants au BSV. Bien que ces bananiers soient porteurs des eBSV, il n'a jamais été observé de bananiers BB infectés par le BSV, que ce soit à l'état sauvage ou bien lors de tentative d'infection en laboratoire sur *M. balbisiana* Pisang Klutuk Wulung (PKW) et *M. balbisiana* Pisang Batu (Lheureux, 2002).

Les stress génomiques constituent donc le facteur déclenchant nécessaire à l'activation des eBSV infectieux. Ces facteurs semblent aussi identifiés pour les autres pathosystèmes, que ce soit pour l'eTVCV ou l'ePVCV chez le tabac et le pétunia respectivement. L'activation a lieu systématiquement chez un hybride interspécifique avec des parents porteurs sains des intégrations. On sait par ailleurs que les modifications épigénétiques provoquées par ces différents stress génomiques favorisent l'activation des rétroéléments (Capy et al., 2000 ; Stotkin et Martienssen, 2007).

L'activation a été étudiée d'un point de vue génétique avec la recherche d'un facteur qui serait aussi impliqué dans l'activation. Pour cela, Lheureux et al., (2003) ont utilisé une population triploïde AAB stérile issue d'un croisement interspécifique entre le diploïde

naturel (BB) *M. balbisiana* PKW et le cultivar tétraploïde *M. acuminata* IDN110 (AAAA), les deux parents étant sans BSV. Une ségrégation mendélienne monogénique de la répartition du BSV était observée pour la moitié de la population indiquant qu'un facteur génétique était impliqué dans l'activation des eBSV. L'analyse de la ségrégation de marqueurs AFLP (Amplified Fragment Length Polymorphism) dans la population a montré l'existence d'un locus BEL (BSV expression locus) associé à l'apparition du BSV. Ce locus était absent du parent *M. acuminata* et présent à l'état hétérozygote chez le parent *M. balbisiana*.

Il semble donc que différents facteurs soient indispensables à la production de particules virales à partir des eBSV. Ces différents facteurs peuvent être regroupés dans le terme de polyploïdie interspécifique, c'est à dire que l'expression des eBSV à lieu chez des hybrides interspécifiques avec un seul génome B. Le croisement génétique entraine des bouleversements génomiques dus à la réunion de deux génomes différents et à la fréquente polyploïdisation qui y est associée. Les stress abiotiques semblent être des prérequis également indispensables. Les différentes étapes permettant l'activation sont présentées dans la figure 1-22.

7-3-2 Structure génomique des eBSV infectieux

Afin de connaître la structure des eBSV intégrés dans le génome de *M. balbisiana*, une banque de chromosomes artificiels bactérien (BAC) intégrant le génome du bananier sauvage diploïde *M. balbisiana* PKW (BB) a été construite et caractérisée (Safar et al., 2004). La banque BAC a été hybridée avec des sondes eBSV représentant le génome complet de BSOLV, BSGFV, BSImV et BSMyV. Les résultats de ces hybridations montrent que ces différents eBSV sont intégrés en très faible nombre de copies. Les clones BAC portant ces eBSV ont été séquencés en totalité.

Chacune des séquences obtenues a été annotées. Toutes les intégrations correspondant à des virus et étant infectieuses ont été caractérisées et nommées : eBSVGF-7 et -9, eBSOLV-1 et -2 ainsi que eBSImV. Les séquences obtenues de l'eBSMyV dont les intégrations ne sont pas infectieuses, sont partielles et l'analyse est en cours. Il se trouve que toutes les intégrations sont plus longues que leur virus libre correspondant (7,26kb) : 13,28 et 15,58kb pour eBSVGFV, 15,8kb pour eBSImV et 22,9 et 23,2 kb pour eBSOLV. Toutes les intégrations présentent, à un degré plus ou moins grand, des réarrangements par rapport à l'organisation linéaire du virus libre, avec des régions dupliquées et réorientées. Seules les intégrations de BSGFV ont été clairement caractérisées et analysées (Gayral et al., 2008) (figure 1-23A). Les deux intégrations eBSGFV ne diffèrent entre elles que par la présence chez l'eBSGFV-9 d'une

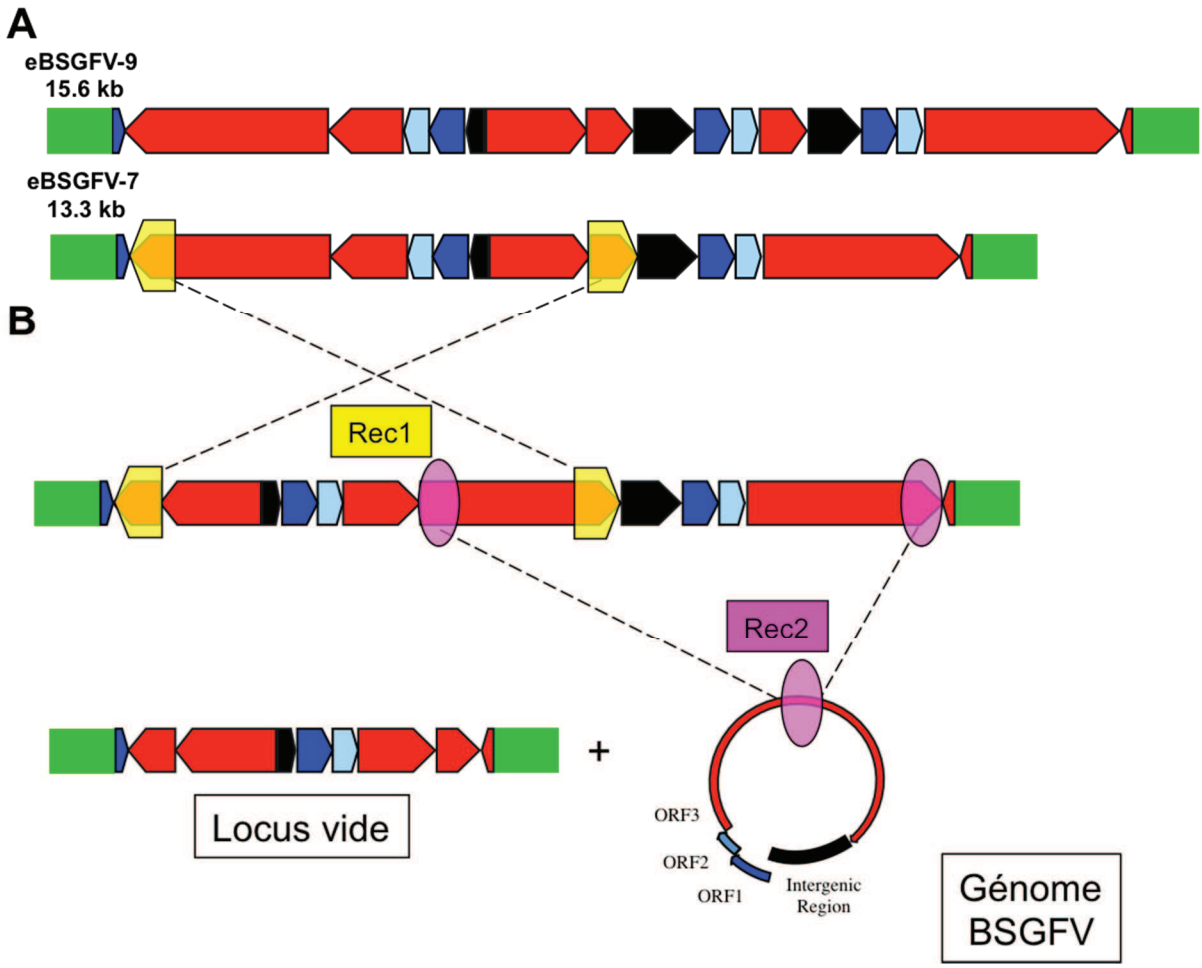


Figure 1-23 : Structure de l'eBSGFV présent dans le génome de PKW et schéma du scénario putatif de production de particules virales à partir de cet eBSV

A : Les rectangles colorés représentent les ORF 1, 2 et 3 du BSV en Bleu clair, Bleu foncé et rouge respectivement. Les rectangles noirs représentent la zone inter-génique (IG) du BSV. En vert est représenté le génome de la plante PKW. L'orientation des flèches indique le sens de lecture des séquences virales par rapport à BSGFV.

B : La recombinaison homologue (RH) est la base du modèle pour expliquer la libération d'un virus BSGFV à partir du génome de son homologue intégré. Les séquences répétées utilisées pour les RH sont grisées. 1- Structure théorique de la molécule d'eBSGFV-7 recombinée après la première RH (Rec1). 2- Structure théorique de la molécule recombinée après la deuxième RH (Rec2). 3- A gauche la molécule d'eBSGFV-7 résiduelle après la RH restant dans le génome de PKW. Et à droite la molécule virale excisée.

Adapté de Gayral et al., (2008) et Iskra-Caruana et al., (2010)

duplication. Toutefois il est important de noter que l'intégration eBSGFV-9 a accumulé des mutations ayant un impact sur la fonctionnalité *in silico* de l'intégration (Gayral et al., 2008). Des marqueurs de type PCR et dCAPs ont été développés pour eBSGFV et eBSImV afin de pouvoir suivre la ségrégation des intégrations dans la population AAB correspondant à la F1 du croisement PKW (BB) x IDN110 (AAAA). Les résultats ont permis de mettre en évidence que ces deux eBSV sont alléliques (Gayral et al., 2008, 2010) à un même locus. Cette étude a été jumelée avec la recherche de plantes infectées par BSGFV dans la population F1. La présence du virus BSGFV est corrélée seulement avec la présence de l'intégration de type eBSGFV-7 alors que les plantes porteuses de l'intégration eBSGFV-9 ne développent jamais d'infection. Les particules virales sont toutes porteuses d'une mutation présente uniquement sur l'intégration et mise en évidence par le marqueur dCAPS. Ces résultats ont permis de conclure que seule l'intégration eBSGFV-7 est responsable de l'infection (Gayral et al., 2008).

7-3-3La production de particules virales par les eBSV infectieux

Les différents résultats accumulés nous permettent de dire que les eBSV sont bien producteurs de particules virales, néanmoins les mécanismes expliquant la sortie des particules virales à partir de leur séquence restent inconnus. L'étude de la séquence eBSGFV nous permet de dire qu'une sortie du virus à partir de l'intégration par transcription directe est impossible comme c'est le cas chez le pétunia pour l'ePVCV inséré en tandem dans la plante (Norren et al., 2007). Dans notre cas, un modèle théorique d'activation a été proposé basé sur les recombinaisons homologues (RH) (figure 1-23B). Dans le modèle proposé par Iskra-Caruana et al. (2010), une première recombinaison a lieu (appelé REC1) entre deux séquences inversées-répétées de 642 pb, séparées de 6,942kpb. La deuxième recombinaison (REC2) se produit entre deux séquences de 633pb, flanquantes à cette séquence virale complète et conduit à l'excision d'un ADN viral de BSGFV-like qui peut ensuite se circulariser.

7-4 L'Evolution des eBSV

7-4-1 Les eBSV et l'évolution virale

Au cours de son évolution, le bananier a vraisemblablement eu à de multiples reprises des intégrations de Badnavirus, que nous avons nommés BSV-like de par leur éloignement phylogénétique avec les BSV exogènes actuels (Gayral et al., 2009 ; D'Hont et al., 2012). Ces données, ainsi que le fait d'avoir retrouvé dans le génome des plantes des EPRV très majoritairement de la même famille virale et dans le bananier du même genre, indiquent

que les événements d'intégrations ainsi que la fixation dans le génome de l'hôte sont des événements fréquents (Gayral et al, 2009 ; D'Hont et al., 2012).Ces intégrations pourraient faire partie d'un processus d'adaptation du virus. Stagginus et Richert-Pöggeler, (2006) ont émis l'hypothèse que ces intégrations participeraient à la diversité des BSV et serviraient de réservoirs génétiques en favorisant les échanges génomiques entre les BSV exogènes et endogènes au cours des infections.

Le cas des eBSV infectieux est quant à lui beaucoup plus rare car seuls à ce jour 4 eBSV possèdent encore cette capacité à être infectieux (démontré pour eBSGFV, eBSOLV, eBSImV et fortement suspecté pour eBSMyV) (Ndowora et al., 1999 ; Gayral et al., 2008, 2010 ; Geering et al., 2005b). La production de particules virales à partir de ces eBSV a été observée seulement chez des hybrides interspécifiques de type AAB appartenant à différentes variétés. Ces intégrations semblent avoir été à l'origine d'explosions de BSV principalement en Ouganda (Dahal et al., 1998, Harper et al, 2004,2005). Les eBSV pourraient donc participer au maintien de la diversité du BSV et participer à l'émergence de la maladie dans des zones encore épargnées, ce qui constitue un avantage évolutif certain pour le virus. L'intégration constitue néanmoins un risque d'élimination et /ou de neutralisation important pour le virus car comme nous l'avons vu, il semble que la majorité des intégrations non-infectieuses nommées BSV-like badnavirus sont distantes des espèces BSV actuelles et n'aient plus de correspondant épisomal.

7-4-2 Les eBSV et l'évolution du bananier

Les BSV-like badnavirus et les eBSV ont été retrouvés dans les génomes de différents bananiers de différentes espèces. Les BSV-like badnavirus ont été mis principalement en évidence dans le génome *M. acuminata* et les premières études indiquent que certains groupes phylogénétiques sont également présents dans le génome de *M. balbisiana* (Muller et Chabannes, communication personnelle). Ces informations sont en faveur d'une intégration et fixation virale avant la spéciation des deux espèces. Les 4 eBSV qui sont potentiellement infectieux semblent quant à eux spécifiques du génome B de l'espèce *M. balbisiana* (Ndowora et al., 1999 ; Geering et al, 2005b ; Gayral et al., 2010). Gayral et al (2010) ont montré que les eBSImV et eBSGFV étaient présents seulement dans les génomes B des bananiers *M. balbisiana* diploïdes, Ces informations indiquent que ces eBSV se sont intégrés après la spéciation de *M. balbisiana* (Gayral et al., 2010). Malgré tout, les eBSV ont gardé la capacité de produire des particules virales exprimées uniquement chez les hybrides interspécifiques et jamais chez les espèces *M. balbisiana* diploïdes (BB) ou *M. acuminata* (AA

ou AAA).Ces espèces ont des temps d'évolution beaucoup plus longs que les hybrides interspécifiques qui ont été principalement créés par l'homme durant les 7000 dernières années (Perrier et al, 2011). Les bananiers ont donc au cours de leur co-évolution avec les eBSV pu mettre en place des mécanismes régulant la production de particules virales à partir des eBSV et nécessitant de conserver une séquence infectieuse. Les effets négatifs de la sortie de virus sur la fitness de la plante peuvent expliquer que la plante ait pu développer ces mécanismes.

Le fait de ne jamais retrouver les bananiers de l'espèce *M. balbisiana* (BB) infectés par le BSV a posé question durant l'étude des eBSV. F. Lheureux durant sa thèse (2002) a mis en place des protocoles d'infection des bananiers par différentes espèces de BSV (BSGFV, BSOLV et BSMyV) en utilisant des cochenilles comme vecteur de la maladie. Les deux bananiers diploïdes *M. balbisiana* (PKW et Pisang Batu) utilisés n'ont jamais été infectés et donc développé la maladie alors que les témoins *M. acuminata* ont tous été infectés rapidement montrant des symptômes variables selon le génotype bananier. Ces observations ont amené à l'hypothèse que les eBSV pourraient apporter un avantage sélectif aux bananiers porteurs en apportant une résistance contre les virus exogènes. Cet avantage sélectif aurait donc participé à la conservation de ces séquences dans le génome du bananier.

7-Objectif de la thèse

Ce travail de thèse à pour objectif de préciser si les eBSV sont maintenus ou non dans le génome *Musa balbisiana* des bananiers et d'étudier les conséquences évolutives observables de l'intégration de ces séquences virales. Pour cela nous avons étudié les eBSV infectieux présents chez PKW et formulé deux hypothèses de travail. La première s'appuie sur le fait que les eBSV sont issus d'intégrations récentes et que chez les bananiers sauvages *M. balbisiana* ces intégrations ne produisent pas de particules virales. Ainsi ces informations tendent à montrer que les eBSV évoluent sous contraintes sélectives neutres et que les raisons de leur maintien dans le génome des bananiers s'expliquent par le fait que leur intégration est récente et située dans des zones du génome de la plante où la pression de sélection est faible. Ils tendent donc dans le futur vers une dégradation/pseudogénisation voire une élimination du génome. La deuxième hypothèse s'appuie sur les observations et études préliminaires menées pour le bananier et pour les autres EPRV qui montrent d'une manière générale que des stress génomiques sont à l'origine de leur activation et que le mécanisme d'interférence ARN est fonctionnel. Ces informations indiquent que les eBSV

seraient associés à un avantage évolutif pour les bananiers sous la forme d'une résistance constitutive induite par le maintien des eBSV. Cet avantage conduirait à une sélection positive des séquences eBSV et donc à un maintien dans le génome de la plante. Ces deux hypothèses caractérisent le paradoxe que constitue la présence de séquences virales dans le génome de leur hôte. En effet ces intégrations représentent un grand risque pour l'hôte au moment de leur intégration de par leurs potentiels infectieux. Néanmoins comme nous l'avons vu chez les animaux et comme il est suspecté pour les eBSV, ces intégrations peuvent apporter des avantages sélectifs essentiels pour la survie de l'hôte. Il a donc été indispensable pour l'hôte d'avoir su s'adapter et trouver l'équilibre entre les effets positifs et négatifs induits par les eBSV.

Ce travail de thèse a pour but de répondre aux principales questions qui se posent dans le contexte actuel de la recherche sur les EPRV concernant leur existence et leur maintien dans le génome des plantes hôtes, en prenant comme modèle biologique les séquences virales endogènes du *Banana streak virus* présentes dans le génome du bananier (*Musa* sp.)

Dans le premier chapitre nous avons tout d'abord caractérisé les trois intégrations infectieuses de BSV présentes dans le cultivar *M. balbisiana* cv. PKW, l'eBSOLV, l'eBSGFV et l'eBSImV. Cette étude a permis de décrire les structures, l'organisation génétique, le paysage génomique ainsi que les capacités infectieuses de ces intégrations. Sur la base de ces descriptions nous avons proposé un modèle d'intégration et d'évolution des eBSV dans le génome de PKW (Article 1).

Nous avons par la suite étudié la distribution des trois eBSV infectieux dans la diversité *M. balbisiana*. Sur la base des résultats de l'article 1, des outils moléculaires ont été développés afin de pouvoir suivre le polymorphisme d'intégration des structures eBSV de PKW dans le génome B des bananiers chez différents génotypes représentatifs de la diversité *M. balbisiana*. Nous avons ainsi pu retracer l'histoire évolutive de chacun des eBSV infectieux et proposer un modèle global d'évolution de ces eBSV infectieux dans le génome *M. balbisiana*.

Le deuxième chapitre est consacré à l'étude des mécanismes de régulation des eBSV. Ce travail a porté sur les mécanismes d'ARN interférent pouvant expliquer le maintien des eBSV dans le génome des bananiers. Nous nous sommes là encore basé sur les résultats de l'article 1 afin de développer un protocole d'analyse de la production de petit ARN depuis les eBSV présents dans le génome de PKW. Nous avons également recherché les mécanismes mis en place par les bananiers non-porteurs d'eBSV en cas d'infection afin de connaître les défenses constitutives des bananiers face à une attaque virale BSV. Cette étude nous a

permis de définir quelles sont les zones cibles des petits ARN sur les séquences BSV et eBSV dans ces différents contextes. Nous avons, sur la base de ces résultats, proposé un modèle de régulation des eBSV et des BSV.

En conclusion, nous discuterons de l'impact qu'a pu avoir la défense anti virale, basée sur une régulation par l'ARNi, sur l'évolution des eBSV avant de revenir plus globalement sur leur maintien dans le génome des bananiers fertiles *M. balbisiana*. Enfin nous examinerons les étapes évolutives qu'ont pu subir les eBSV dans le génome du bananier expliquant la diversité que nous observons aujourd'hui.

Les résultats obtenus au cours de cette thèse sont en cours de valorisation par les publications suivantes qui seront présentées dans cette thèse.

Article 1

Chabannes, M., Baurens, F.-C., Duroy, P.-O., Sidibe-Bocs, S., Vernerey, M.-S., Rodier-Goud, M., Barbe, V., Gayral, P., and Iskra-Caruana, M.-L. Three infectious viral species lying in wait in the banana. Submited to *Plos Pathogens*, 1-41.

Article 2

Duroy, P.-O., Perrier X., Laboureau N., Jacquemoud-Collet J.,P., and Iskra-Caruana, M.-L. How endogenous Banana streak virus (eBSV) could enlighten BSV and banana evolution. In prep.

Durant ma thèse deux autres publications concernant des travaux réalisés pendant mon cursus universitaire ont été publiées. Ne concernant pas les séquences intégrées de Bananas Streak Virus dans le génome du bananier ils ne seront pas présentés.

Hamon, P., Duroy, P.-O., Dubreuil-Tranchant, C., Mafra D'Almeida Costa, P., Duret, C., Razafinarivo, N.J., Couturon, E., Hamon, S., Kochko, A., Poncet, V., and Guyot, R. (2011). Two novel Ty1-copia retrotransposons isolated from coffee trees can effectively reveal evolutionary relationships in the Coffea genus (Rubiaceae). Molecular Genetics and Genomics **285,** 447-460.

Chaïr, H., Duroy, P.-O., Cubry, P., Sinsin, B., and Pham, J.L. (2011). Impact of past climatic and recent anthropogenic factors on wild yam genetic diversity. Molecular Ecology **20,** 1612-1623.

CHAPITRE 1

Caractérisation et distribution des eBSV fonctionnels de PKW

La caractérisation du contexte d'intégration des trois espèces BSV que sont BSGFV, BSOLV et BSImV dans le génome *Musa balbisiana* a été conduite à partir de la plante modèle, le bananier diploïde *M. balbisiana* fertile Pisang klutuk wulung ou PKW. Ce bananier présente la particularité d'être un porteur sain d'intégrations virales infectieuses et apparait résistant à toute multiplication virale BSV quelle qu'en soit l'origine endogène et exogène. Les premières études ont été réalisées suite au séquençage de clones BAC PKW (Safar et al, 2004) porteurs d'eBSV pour chacune des espèces BSV à étudier. Le contexte d'insertion de l'espèce BSGFV a été décrit précédemment par P. Gayral au cours de son travail de thèse, en utilisant la population hybride interspécifique BAA issue du croisement de PKW (BB) avec le bananier tétraploïde *M. acuminata* IDN110T (AAAA) (Gayral et al, 2008). Il a ensuite précisé sa distribution dans la diversité des bananiers diploïdes *M. balbisiana* (Gayral et al, 2010).

Le travail réalisé dans ce chapitre s'est inscrit dans la continuité des travaux menés pour BSGFV quant à la caractérisation fine du contexte d'intégration pour les espèces BSOLV, BSImV chez PKW et a complété ceux initiés sur BSGFV (Article 1). Ainsi, j'ai contribué, au sein de l'équipe, à la caractérisation des eBSOLV et eBSImV infectieux présents dans le génome de PKW. J'ai plus particulièrement validé les différents marqueurs moléculaires développés par M. Chabannes sur la population hybride BAA et proposé une signature eBSV pour suivre les 3 espèces virales. Ces travaux ont conduit à décrire les structures, l'organisation génétique allélique, le paysage génomique ainsi que les capacités infectieuses de ces intégrations. J'ai ensuite plus particulièrement développé en collaboration avec FC Baurens les analyses qui ont conduit aux propositions faites sur l'évolution de ces eBSV dans le génome de PKW. Les connaissances générées ont abouti à la proposition d'un schéma général d'intégration et d'évolution des eBSV chez les bananiers diploïdes *M. balbisiana*. Ce travail a fait l'objet d'une publication qui a été soumise à Ge*nome research*.

La globalité de ces connaissances eBSV chez PKW a constitué le socle méthodologique des recherches que j'ai menées par la suite sur la réalité du maintien des eBSV au sein de la diversité des bananiers *M. balbisiana* (Article 2). Nous avons pu, dans un premier temps, développer des outils performants pour étudier cette diversité en utilisant pour base les eBSV présents chez PKW : marqueurs PCR et dCAPs (espèces BSV et allèles spécifiques) et analyses southern blots après digestion par des enzymes de restriction. Gayral et al, (2010)

ayant montré l'existence d'un polymorphisme réduit au sein de l'échantillonnage disponible des bananiers diploïdes BB, nous avons étendu dans la mesure du possible la diversité du génome B en intégrant à notre échantillonnage des hybrides interspécifiques de ploïdie différente (AB, ABB, AAB). Cet élargissement avait deux objectifs : permettre l'accès à des génomes B non représentés dans les génomes des bananiers diploïdes et implémenter à notre étude des données phylogéographiques non-disponibles pour les bananiers *M. balbisiana* stricts. Ce travail que j'ai réalisé en totalité est présenté dans l'article 2 qui sera soumis à *Molecular Ecology* ou *Annals of Botany*. Nous avons ainsi pu retracer l'histoire évolutive de chacun des eBSV infectieux à travers l'histoire génétique des bananiers et inversement permettre un éclairage de l'histoire évolutive des bananiers en ancrant géographiquement les diploïdes *M. balbisiana*.

1- Article 1 : Three infectious viral species lying in wait in the banana genome

Three infectious viral species lying in wait in the banana genome

Chabannes M.*1, Baurens F. -C.*2, Duroy P. -O. 1, Sidibe-Bocs S. 2, Vernerey M. -S. 1, Rodier-Goud M. 2, Barbe V. 3, Gayral P. 4, and Iskra-Caruana M. -L. 1

1 CIRAD, UMR BGPI, F-34098 MONTPELLIER, France

2 CIRAD, UMR AGAP, F-34098 MONTPELLIER, France

3 GENOSCOPE, 2 rue Gaston Crémieux, BP5706, 91057 Evry, France

4 Institut de Recherche sur la Biologie de l'Insecte, UMR CNRS 7261

Université François Rabelais, Faculté des Sciences et Techniques

Parc Grandmont, Avenue Monge, 37200 Tours, France

* These authors equally contributed to this work

Corresponding author: Marie-line Iskra-Caruana CIRAD, UMR BGPI, Campus International de Baillarguet, TA A 54/K, F-34398 Montpellier Cedex 5, France. Phone: (33) 4 99 62 48 13 - Fax: (33) 4 99 62 48 08. E-mail: marie-line.caruana@cirad.fr

Running title: Banana genome harbors three infectious viral species

Keywords: Banana B genome, endogenous banana streak virus (eBSV), infectious, evolution

Abstract

Plant pararetroviruses integrate serendipitously into their host genome. The banana genome harbors integrated copies of Banana streak virus (BSV) named eBSV that are able to release infectious pararetrovirus. Here we characterize integrants of three BSV species—Goldfinger (eBSGFV), Imové (eBSImV) and Obino l'Ewai (eBSOLV)—in the seedy *Musa balbisiana* Pisang klutuk wulung (PKW) by studying their molecular structure, genomic organization, genomic landscape and infectious capacity. We first selected a BAC clone of each of the three eBSV species. eBSV segregation analysis on an F1 population of PKW combined with fluorescent *in situ* hybridization (FISH) analysis showed that eBSImV, eBSOLV and eBSGFV are each present at a single locus. eBSOLV and eBSGFV contain two distinct alleles, whereas eBSImV has two structurally identical alleles. Genotyping of both eBSV and viral particles expressed in the progeny demonstrated that only one allele for each species is infectious. The infectious allele of eBSImV could not be identified since both alleles are identical. Finally, we demonstrate that eBSGFV and eBSOLV are located on chromosome 1 and eBSImV on chromosome 2 of the reference *Musa* genome published recently. The structure and evolution of eBSVs suggest sequential integration into the plant genome, and haplotype divergence analysis confirms that the three loci display differential evolution. Based on our data, we propose a model for BSV integration and eBSV evolution in the *Musa balbisiana* genome. The mutual benefits of this unique host–pathogen association are also discussed.

INTRODUCTION

Endogenous retroviruses in animal genomes were discovered initially in the late 1960s and early 1970s (Weiss 2006). Integration into the host chromosome of a "provirus" mediated by a virus-encoded integrase is obligatory to the replication of retroviruses. Proviruses integrated into the DNA of germ line cells as endogenous provirus are indeed inherited by the host as Mendelian traits, and are named ERVs (endogenous retroviruses) to distinguish them from horizontally transmitted exogenous retroviruses (Vogt 1997). Until recently, retroviruses were the only endogenous viruses known. This feature has now been extended to many other viruses: from RNA viruses that do not require integration to replicate such as bornaviruses and filoviruses (Belyi et al. 2010a; (Taylor et al. 2010), to circoviruses and parvoviruses with single-stranded DNA genomes (Belyi et al. 2010b) and to hepadnaviruses with partially double-stranded DNA genomes (Gilbert and Feschotte 2010). Such integration events play no role in viral replication, cannot give rise to infectious virus even if their endogenous retroviruses is infectious, and constitute a fossil record useful to determine the age of viruses (Horie et al. 2010); (Katzourakis and Gifford 2010). Whether such sequences confer any biological advantage to the host remains an interesting question.

In contrast to animal viruses, no known plant virus encodes an integrase or requires integration into the host genome to replicate. Nowadays, access to plant genome sequencing data reveals a significant proportion of foreign sequences including viral sequences. All the viral sequences found so far belong to the *Caulimoviridae* family. These are double-stranded DNA viruses exploiting a reverse transcription process for replication. Interestingly, some of these integrated viral sequences are infectious, leading frequently to plant infection. Such sequences have undergone extensive viral genome rearrangements and contain more than one copy of viral genome (Hohn et al. 2008). Surprisingly, these sequences appear to be inherited by their host plant as Mendelian traits and to be transmitted vertically like ERVs. However, and unlike the animal kingdom, plants infected in this manner are always hybrids resulting from interspecific crosses. To date, three plant viruses have been described as endogenous and infectious viruses in hybrids: (1) *Tobacco vein clearing virus* (TVCV) in *Nicotiana edwardsonii*, which is an allohexaploid derived from a cross between *N. glutinosa* (n=24) and *N. clevelandii* (n=12) (Lockhart et al. 2000); (2) *Petunia vein clearing virus* (PVCV) in petunia hybrids *P. hybrida* resulting from a wild cross between *P. integrifolia* ssp. *inflata* and *P. axillaris* ssp. *axillaris* (Richert-Poggeler et al. 2003); and (3) *Banana streak virus* (BSV) in several interspecific banana hybrids obtained by crosses between *Musa balbisiana* (B genome) and *Musa acuminata* (A genome) (Lheureux et al. 2003; Gayral et al. 2008). Mechanisms that lead to integration, activation and subsequent episomal infection are complex and still largely unknown. Over the last decade, some data relating to integrated viral

sequences, the locus of integration and activation process occurring following different stresses such as wounding and tissue culture have been accumulating (Lockhart et al. 2000; Dallot et al. 2001; Richert-Poggeler et al. 2003; Cote et al. 2010), revealing several differences between the three different models described above. Indeed, the PVCV genome is integrated in tandem arrays in a complete, continuous form as a "retro provirus"; its activation resembles that of retroviruses, and involves the production of a greater than full-length transcript that could be reverse transcribed into DNA (Harper et al. 2002; Staginnus and Richert-Poggeler 2006). On the other hand, BSV *Goldfinger* species (BSGFV) is present in the diploid *M. balbisiana* (BB) Pisang klutuk wulung (PKW) plant as a di-allelic integration much longer than a single viral genome, exhibiting a succession of partial viral sequences that are sometimes inverted and partially duplicated, representing at least one total viral genome (Gayral et al. 2008). Based on the endogenous BSGFV sequence (eBSGFV), an activation process based on a direct transcription event seems unlikely to occur alone. Indeed, in the theoretical model proposed by Iskra-Caruana et al. (2010) two initial homologous recombination (HR) steps could be required to reconstitute a full-length circular BSGFV genome.

Of the three plant species affected by viral disease outbreaks from these endogenous pararetrovirus sequences, banana remains the most critical in terms of economic impact. Indeed, banana (*Musa* spp.) ranks as the world's fourth most important food crop in terms of gross value of production after rice, wheat, and maize. In the past 20 years, the emergence of BSV in all banana-producing countries resulted from the awakening of eBSVs, and correlates with the massive use of newly created inter-specific banana hybrids. Consequently, all reported inter-specific banana hybrids are considered at risk for the wide and rapid dissemination of BSV and, within a very short period of time, BSV has become the major constraint to banana breeding programs worldwide. Recently, we established that these problematic infectious eBSV are present in the B genome of banana only (Lheureux et al. 2003; Gayral and Iskra-Caruana 2009; Cote et al. 2010; Gayral et al. 2010). The B genome forms part of the genotype of many important banana cultivars, such as the famous plantain subgroup that is a staple food for millions of people in Africa and Latin America. Moreover, it is often associated with desirable traits of agronomic interest such as vegetative vigour, biotic and abiotic stress tolerance as well as resistance to pathogens such as the fungus responsible for the severe black sigatoka disease. Consequently, a better knowledge of eBSV structure, genomic organisation and chromosomal localisation in the B genome as well as infection capacity to produce functional viral genome will provide molecular tools that can be used to widely screen banana hybrids, genitors and germplasm. This step becomes a prerequisite not only for future crop-oriented breeding programmes aimed at producing safe interspecific

Table 1: Overview of BSV integrations in the PKW BAC library.

Probe		N° Hits	Fingerprint eBSV patterns	BAC selected
BSGFV	9	2	MBP_071C19, MBP_094I16 (Gayral et al. 2008)	
BSOLV	10	2	MBP_031O07, MBP_073B22 , MbP017D14	
BSImV	24	1	MBP_068C24	
BSMysV	15	2	none	
BSVNV	0	-	none	

banana hybrids but also to estimate and limit the risk of BSV outbreak on natural hybrids spread intensively in developing countries as a food source. Previous studies (Lheureux et al. 2003; Iskra-Caruana et al. 2010) have revealed the presence of three species of BSV: *Banana streak Obino l'Ewai virus* (BSOLV), *Banana streak Goldfinger virus* (BSGFV) and *Banana streak Imove virus* (BSImV) in a F1 triploid (AAB) population obtained from an inter-specific genetic cross between PKW, seedy diploid *Musa balbisiana* (BB), and IDN 110 tetraploid *Musa acuminata* (AAAA). Both are virus free, and PKW is the only one to harbour the eBSV counterpart of each BSV species. PKW is therefore solely responsible for virus transmission and expression among the progeny.

The aim of the present study was to fully characterise the three eBSVs present within the B genome that contribute to BSV epidemics worldwide. This in depth characterisation was performed on the diploid *Musa balbisiana* PKW genome by combining molecular, genomic, genetic and cytogenetic approaches. Based on our data, we propose a model representing the most probable scenario to have occurred from the initial eBSV integration to the picture observed nowadays in PKW, and discuss the mutual benefits of this particular plant–pathogen interaction.

RESULTS

1- Presence of eBSV in the genome of PKW

We screened high density filters of the 9x *Musa balbisiana* Pisang klutuk wulung (PKW) BAC library and two additional *Musa acuminata* BAC libraries (the seedy diploid Calcutta 4 and the triploid "Cavendish" cv. Grande Naine with a coverage of 9x and 4.5x, respectively) with probes covering the full length genome of four distinct BSV species - *Obino l'ewai* (BSOLV), *Imove* (BSImV), *Mysore* (BSMysV), *Vietnam* (BSVNV) - as described by Gayral et al. (2008) for BSGFV. The *M. acuminata* BAC libraries did not produce any signal with either BSV species, while the *M. balbisiana* BAC library had signals with BSOLV, BSImV and BSMysV (table 1). We analysed the fingerprints of all BAC clones positive for each BSV species after hybridization with the corresponding BSV probes. BSOLV, BSGFV and BSMysV had two distinct restriction patterns while BSImV exhibited a single pattern despite the use of four different enzymes (data not shown). All BSV species showed a number of hits that never exceeded three times the coverage of the BAC library. Furthermore, each eBSV showed only one or two distinct patterns. For this work, we focused on BSV species that are present, expressed and consequently infectious in the banana F1 population described by Lheureux et al. (2003), i.e. BSOLV, BSGFV and BSImV.

We set up a fluorescent in situ hybridization (FISH) experiment on metaphase chromosome preparations of PKW using probes corresponding to full-length viral genome. eBSV

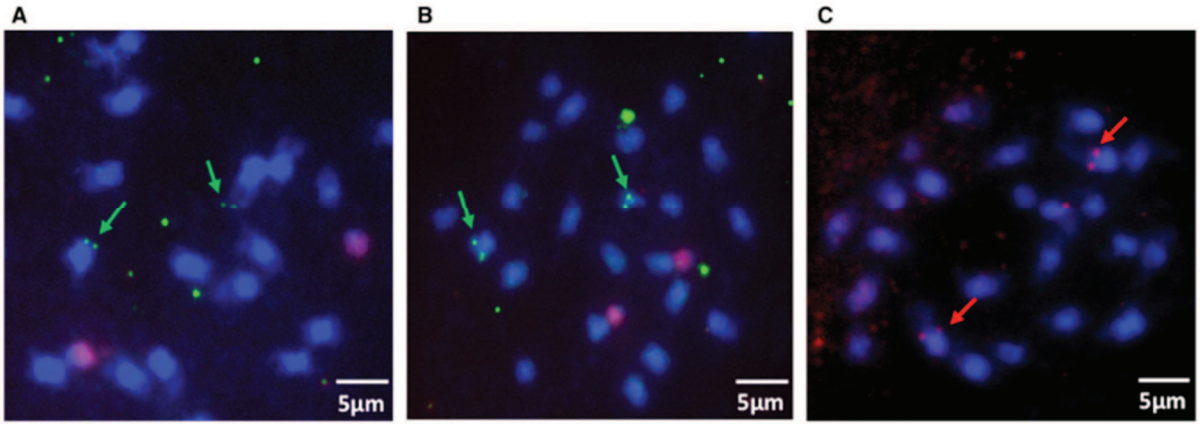


Figure 1: Localisation of the three BSV integrations by fluorescent *in situ* hybridization in root cells of PKW banana chromosomes.

Hybridizations were performed with full-length genome probes for each BSV species and detected with FITC (green A: BSGFV, B: BSImV) or rhodamine (red C: BSOLV). Pink diffuse signals on panels A and B correspond to the 45S rDNA sequence labeled with alexa 555. Chromosomes are counterstained with DAPI (blue).

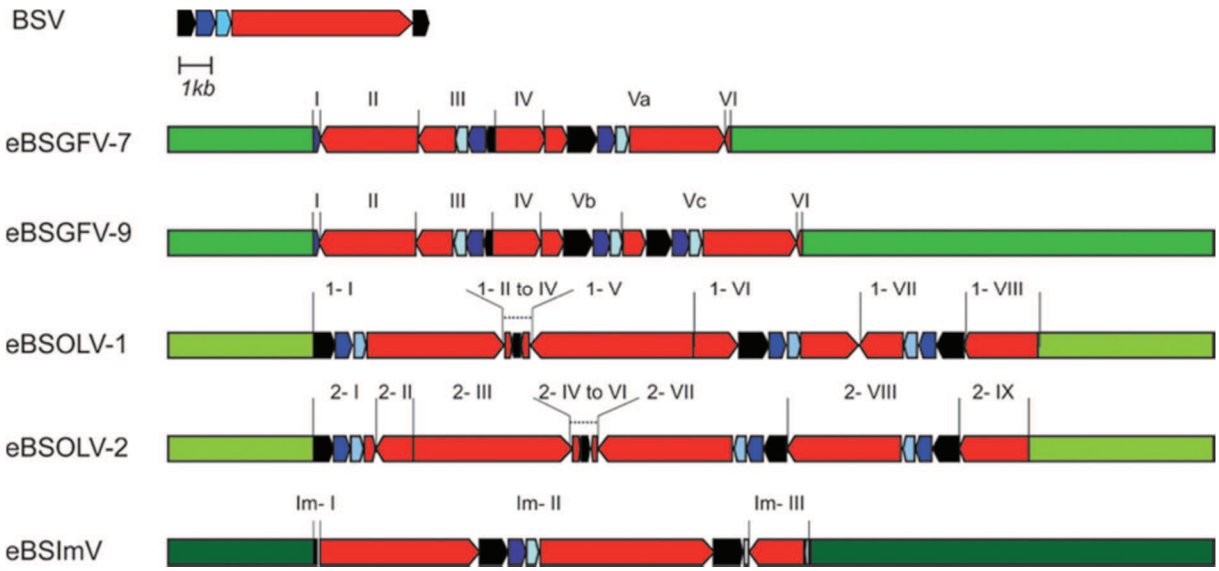


Figure 2: Overview of eBSV structures in PKW

Banana genomic sequences are in green. The BSV genome is represented in linear view with dark blue, light blue and red boxes indicating ORF 1, ORF2 and ORF 3 of the virus, respectively. The intergenic region (IG) is in black. eBSV fragments are indicated.

hybridization was carried out independently for each of the three BSV species (figure 1). We observed signals on two distinct chromosomes for each BSV species; double dots on chromosomes correspond to the hybridization of eBSV on both sister chromatids. eBSImV and eBSGFV locate in distinct chromosomes from those carrying rDNA. Taken together, our data clearly indicate low copy integration of all three BSV species.

2- Structure of eBSV in the genome of PKW

Of the three BSV species detected within the *Musa balbisiana* BAC library, we have already extensively described integrations of BSGFV termed eBSGFV (endogenous BSGFV) in Gayral et al. (2008). eBSGFV is located in a single locus of the genome and presents two different alleles with distinct structures and infectious properties.

Here, following sequencing of three BAC clones, we describe the structure of eBSOLV. The sequences of BAC clones MBP_73B22 and MBP_17D14, which exhibit identical restriction patterns, were grouped into a contig based on an identical sequence overlap of 69,938 bp to obtain a sequence of 185,044 bp named MBP_017D14c (HE983609). We observed eBSV in both BACs MBP_31O07 and MBP017D14c as a continuous stretch of sequences highly similar to BSOLV. The viral integrants in BACs MBP_31O07 and MBP017D14c are hereafter named eBSOLV-1 and eBSOLV-2, respectively. The integrants are much longer than a single BSV genome: 7,839 bp for BSOLV compared to 22,900 bp for eBSOLV-1 and 23,200 bp for eBSOLV-2. These integrants are composed only of viral sequences, with no *Musa* embedded genome sequences (figure 2). Both exhibit a complex rearrangement of viral sequences, with most viral regions being duplicated and therefore present in several copies within each eBSV, either in the same or opposite orientation with respect to the organization of the episomal BSV genome.

eBSOLV-1 is composed of 8 fragments numbered 1-I to 1-VIII that are structurally identical to BSOLV (figure 3A). Fragments 1-I, 1-VI and 1-VII contain intergenic sequence (IG) followed by a complete ORF 1 and ORF 2 and part of ORF 3. In fragment 1-VI, this is preceded by part of ORF 3. IG is complete in fragment 1-VI, and truncated by 315 bp in 1-I, 668 bp in 1-III and 91 bp in 1-VII. All IGs are similar and differ from the BSOLV intergenic region by a 9 bp insertion (52-ATAGCTGTA -32) at position 90 (except in fragment 1-VI) and a 12 bp insertion (52-GACTGGCTAGGT-32) at position 434 of the virus. Fragments 1-II and 1-IV have only a small part of ORF 3 (200-300bp) while fragments 1-V and 1-VIII harbour larger parts (>1kb). No full-length ORF3 is present in any of the 7 fragments that contain it. However, considering all fragments, the entire ORF3 can be reconstituted (figure 3A).

eBSOLV-2 is composed of 9 fragments (figure 3A). Fragments 2-I, 2-VII and 2-VIII contain IG followed by complete ORFs 1 and 2 and part of ORF 3. All IG are truncated, by 315 bp in

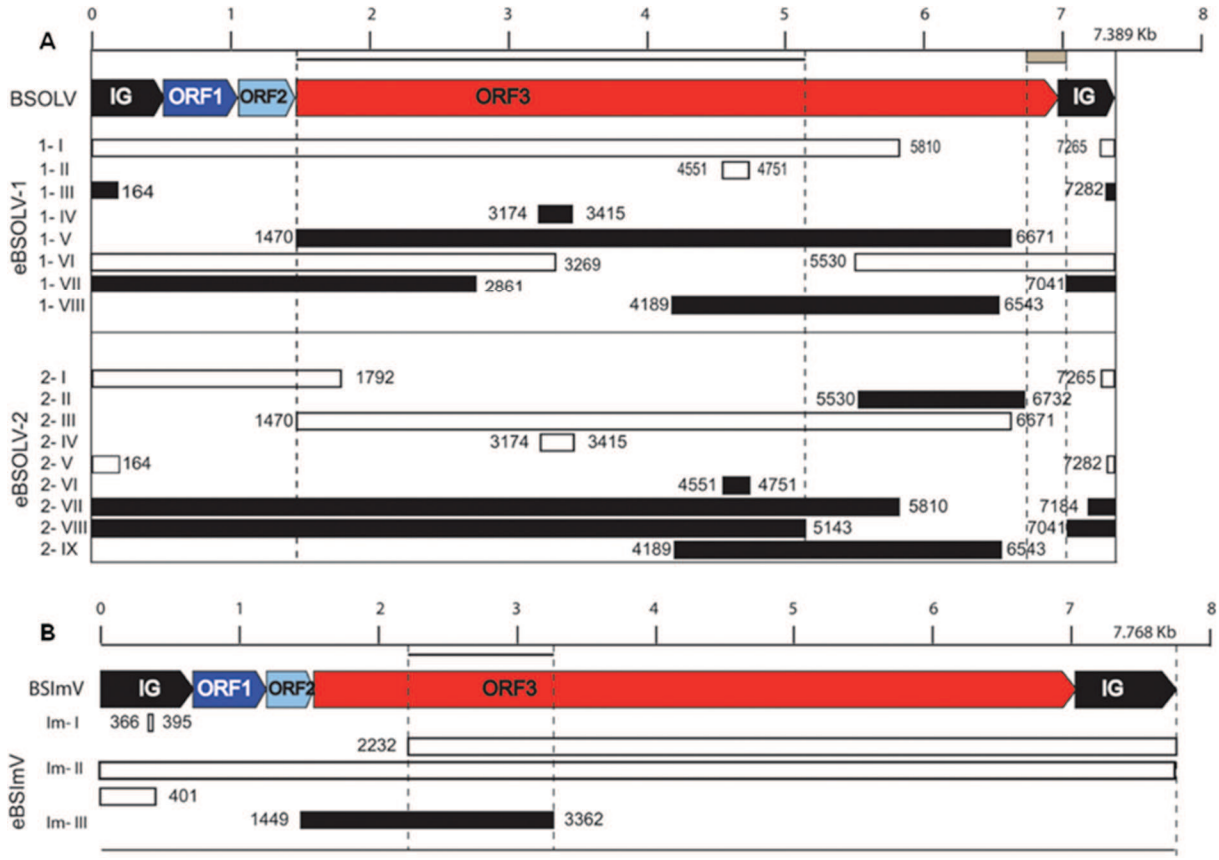


Figure 3: Positions/distribution of eBSV fragments in the counterpart BSV genome.

BSOLV (A) and BSImV (B) genomes are each represented in linear view with a kbp scale bar above. IG, ORF1, ORF 2 and ORF3 are depicted by black, dark blue, light blue and red boxes, respectively. Boxes below the BSV genome represent fragments of eBSV in the same orientation (white boxes) or in reverse orientation (black boxes). Coordinates indicate boundaries of eBSV fragments in reference to the viral genome. The name of fragments is indicated on the left part of the figure: eBSVOL-1 (1-I to 1-VIII), eBSVOL-2 (2-I to 2-IX), eBSImV (Im-I to Im-III). The grey box below the scale bar indicates the missing fragment of ORF 3 in eBSOLV-2.

Black lines below the scale bar represent the zone of ORF3 used for dating analysis of both BSOLV and BSImV.

2-I, 668 bp in 2-V, 234 bp in 2-VII and 91 bp in 2-VIII, and are similar to those found in eBSOLV-1 including both 9 and 12 bp insertions. Unlike eBSOLV-1, none of them is complete. Indeed, 91 bp at the 5' end of IG are always missing. Fragments 2-IV and 2-VI correspond only to a small part of ORF 3 while fragments 2-II, 2-III and 2-IX correspond to larger parts. No full length ORF3 is present in any of the 8 fragments that contain it. Moreover, and in contrast to eBSOLV-1, a fragment of 217 bp (corresponding to nucleotides 6732-6949 on the BSOLV genome) at the 3' end of ORF3 is always missing.

The 5' and 3' flanking regions of eBSOLV-1 and eBSOLV-2 are similar over 6063 bp and 2426 bp, respectively, despite the structural reorganisation of eBSOLVs. This indicates a common locus of insertion in the *Musa* genome.

We described eBSImV based on sequencing of BAC MBP_068C24 (HE983625). This 121,693 bp sequence contains one eBSV as a unique stretch of sequence similar to BSImV (figure 2). The integrant is again much longer than a single BSV genome: 7827 bp for BSImV and 15,800 bp for eBSImV. eBSImV is composed of 3 fragments: Im-I, Im-II and Im-III. Im-I is composed of 35bp of IG. Im-II is a long stretch of sequences containing more than 1 copy of the complete circular viral genome (1.76 viral genomes). Starting in the middle of ORF3, this fragment contains the 32 part of ORF3, followed by the complete sequences of IG, ORF1, ORF2 and ORF3 and finishing with a truncated IG sequence missing about 240 bp at the 32 end. Im-III contains only the 32 end of ORF2 and 1.9 kbp of ORF3 in reverse orientation (figure 3B).

3- Genomic organisation of eBSV in the genome of PKW

BSGFV is integrated at a single locus as two alleles (Gayral et al. 2008). We characterised the genomic organisation of eBSOLV and eBSImV to see whether or not they follow a similar organisation.

The two eBSOLVs containing BAC clones exhibit very high sequence identity (99.998%) on the 108.6 kbp of overlapping *Musa* regions. This high gene and transposable element (TE) structure conservation is consistent with the hypothesis of allelic insertion for BSOLV as described for BSGFV, where eBSGFV-7 and eBSGFV-9 are located on homologous chromosomes.

To further analyse allelic insertion in PKW, we first monitored the segregation of eBSOLV-1, eBSOLV-2 and eBSImV among the triploid (AAB) F1 progeny of a genetic cross between PKW (BB) (female parent) and *M. acuminata* cv. IDN 110 4x (AAAA) (male parent). The parents were confirmed as virus free by multiplex-Immuno-Capture PCR (Le Provost et al. 2006). As a high degree of sequence conservation between the integrated and episomal forms of BSV exists, we developed molecular markers [PCR and Derived Cleaved Amplified

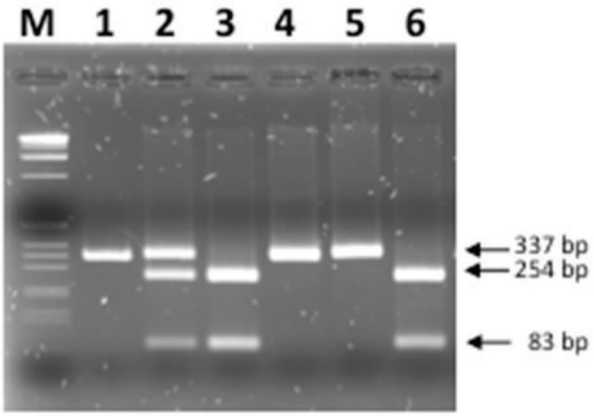


Figure 4: Genotyping of eBSOLV-1 and eBSOLV-2 in the F1 triploid (AAB) population using a dCAPs marker.

The PCR product obtained from eBSOLV-1 is cut into two bands of 83 and 254 bp by the endonuclease *Hae*III (New England Biolabs) whereas the one from eBSOLV-2 is not digested. Lane M: 1 kb DNA ladder (invitrogen), Lane 1: undigested PCR product of the diploid *M. balbisiana* PKW, Lanes 2-6: digested PCR product from PKW carrying both eBSOLV alleles, BAC MBP_31007 carrying eBSOLV-1, BAC MBP_73B22 carrying eBSOLV-2 and two triploid (AAB) F1 progeny of a genetic cross between PKW and *M. acuminata* cv. IDN 110 4x (AAAA) carrying eBSOLV-2 and eBSOLV-1, respectively.
Digested DNA was loaded onto a 2.5% agarose gel stained with ethidium bromide, and bands were visualized under UV light.

Table 2: Number of mutations accumulated within each eBSV fragment compared to the reference genome of BSOLV (A) and BSImV (B)

eBSV	Total Mutations *	Syn	nSyn	Mutation rate
eBSOLV-1	48	15	26	0.0021
eBSOLV-2	65	19	37	0.0028
eBSImV	23	10	9	0.00145

* Total mutations include mutations in intergenic region; Syn: Synonymous mutation; nSyn: Non synonymous mutation

Polymorphic Sequences (dCAPs)] to genotype each eBSV, discriminating them from their viral counterparts (see Materials and methods, and supplementary data). These PCR markers are specific to either *Musa* junctions or internal structure and, where possible, specific to each allele. The genotyping of parents confirms the absence of eBSOLV-1, eBSOLV-2 and eBSImV within the *M. acuminata* parent (data not shown).

We developed a dCAPs (Dif-OL-HaeIII) marker to distinguish between eBSOLV-1 and eBSOLV-2 (figure 4). A PCR amplification test using a set of primers specific to eBSOLV-2 and located on another part of eBSV allowed confirmation of the allelic profile with presence/absence scoring. eBSOLV-1 and eBSOLV-2 segregate in 53.5% and 46.5% among the F1 progeny, respectively, which is compatible with the 50/50 ratio expected for single locus segregation ($df = 2$; $X2 = 0.40$; $p = 0.70$). As for eBSGFV, this indicates a monogenic segregation of eBSOLV.

A unique eBSV is described for eBSImV and its genotyping by PCR revealed its presence in the entire progeny. Next, we developed various molecular markers and strategies to detect a potential second eBSImV allele: sequencing of the eBSImV ends for the 24 BAC clones positive with BSImV probe, development of two dCAPs markers based on point mutations detected on eBSImV (see § 4) and 20 SSR derived from *Musa* sequences flanking eBSImV. Unfortunately, all SSR were monomorphic in PKW (data not shown) and point mutations were detected in all BAC clones (data not shown). Finally, no single nucleotide polymorphism (SNP) was observed between BAC end sequences and the MBP_068C24 sequence (data not shown).

Based on these genetic results, we demonstrate that the two chromosomes labelled with each BSV species using the FISH technique (figure 1) are homologous chromosomes. Integration of the three BSV species is confirmed as allelic, all resulting from a monolocus integration event within the genome of PKW. The integration of BSGFV and BSOLV is di-allelic, whereas that of BSImV is mono-allelic.

4- Which allele is infectious?

In the case of BSGFV, Gayral et al. (2008) demonstrated that only eBSGFV-7 is infectious. For BSOLV and BSImV, we first searched *in silico* for the presence of a full-length BSV genome in all eBSV sequences and further analyzed the type of mutations accumulated in each eBSV allele (table 2). We noticed a deletion of 309 bp at the junction of ORF 3 and IG in eBSOLV-2 (figure 3A). This deletion corresponds to nucleotides 6732 to 7041 of the BSOLV genome. Conversely, eBSOLV-1 contains the full-length BSV genome at least once (figure 3A).

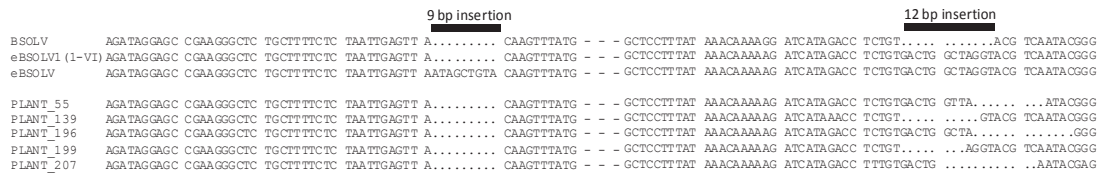


Figure 5: Comparison of IG of BSOLV infecting triploids offspring and eBSOLV.

Primers (BSV2, 5′GTA TCA GAG CAA GGT TCG TTT TT 3′ and BSV525, 5′ATC CCA AGT TTT CTC GAC CAT AA 3′) surrounding the 9 and 12 bp deletions in the IG of BSOLV were designed and used in IC-PCR on 5 independent infected interspecific AAB hybrids (Plants 55, 139, 196, 199 and 207) using the following PCR program: denaturation stage at 94°C during 5min followed by 30 cycles (30 s at 94°C, 30 s at 60°C, 1 min at 72°C) and a final extension at 72°C for 10 min. PCR products were purified and sequenced. BSOLV: Intergenic sequence (IG) of episomal virus (accession number: AJ002234); eBSOLV1 (1-VI): IG present in the fragment VI of eBSOLV allele 1; eBSOLV: IG present in both alleles of the wild diploid *M. balbisiana* PKW (except in fragment VI of eBSOLV1); Plant 55 to 207: IG of episomal virus produced from eBSOLV in the different hybrids.

We compared the homologous region of each eBSOLV-1 fragment to the corresponding BSV genome. Despite the very close similarity (>99% identity on average) between sequences, some differences observed in eBSV sequences may affect viral gene functions. There are 15 synonymous and 26 non-synonymous mutations (Table 2) but none of these led to either premature stop codons or frameshifts. Based on these *in silico* analyses, eBSOLV-1 is the only possible infectious allele.

We then characterized the entire AAB progeny to identify plants infected with BSOLV using immunocapture PCR (IC-PCR). In parallel, we genotyped the same population with the molecular markers developed to discriminate plants harboring eBSOLV-1 or eBSOLV-2 alleles and found that, for all 69 infected plants, BSOLV is always associated with the eBSOLV-1 allele. In addition, we sequenced the IG of viral particles infecting five independent triploid offspring to search for the 12 bp and 9 bp insertions present in the IGs of eBSOLV. No insertion was observed at position 90 whereas partial insertions at position 434 were seen in all released virions. These insertions were all shorter than 12 bp, and differed from one virion to another, but all had 100% sequence identity to the reference BSOLV genome (figure 5). Altogether, our data demonstrate unambiguously that BSOLV infecting the progeny originate from the eBSOLV-1 allele and that fragment 1-VI is involved in release of the functional episomal genome since it is the only allele to display an IG without the 9 bp insertion at position 90. One interesting possibility is that the region of the 12 bp insertion could be a hot-spot of recombination involved in release of the BSOLV genome.

Regarding eBSImV, no comparison of the two alleles was possible as they are so far identical. We identified at least a full-length viral genome within eBSImV as described in §2. As with eBSOLV and eBSGFV, our *in silico* analysis (Gayral et al. 2008) showed a very high degree of sequence conservation between integrated and episomal forms of the virus present in the database (Geering et al. 2011). Similarly, identity was higher than 99%, with only 23 mutations found among the 15.8 kbp of eBSImV (table 2). Synonymous and non-synonymous mutations are represented equally, with 10 and 9 mutations, respectively. In addition, we found two deleterious mutations in the two ORF 3s in fragment II: a deletion of an adenosine at position 95,192 on BAC MBP_068C24 leading to a frameshift and premature stop codon (61 bp downstream of the deletion), and a substitution of guanine to adenosine at position 102,946 leading to a premature stop codon. The presence of these two mutations was confirmed in the 24 BAC clones isolated during screening of the PKW library. The presence of these two point mutations precludes reconstitution of an episomal virus through simple transcription as the point mutations are 7754 bp apart whereas a full length BSImV genome is 7768 bp long. In addition, mutations at both positions would seem to interrupt production of a complete protein.

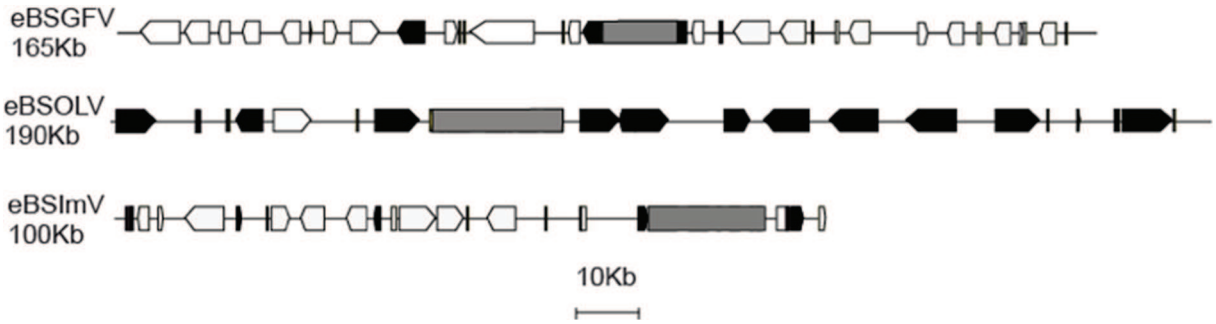


Figure 6: Gene and transposable element (TE) content surrounding eBSVs in the *Musa* genome present in BAC clones.

The black boxes correspond to TE, the grey boxes to eBSV and the white boxes to genes in BAC clones containing eBSGFV, eBSImV and eBSOLV.

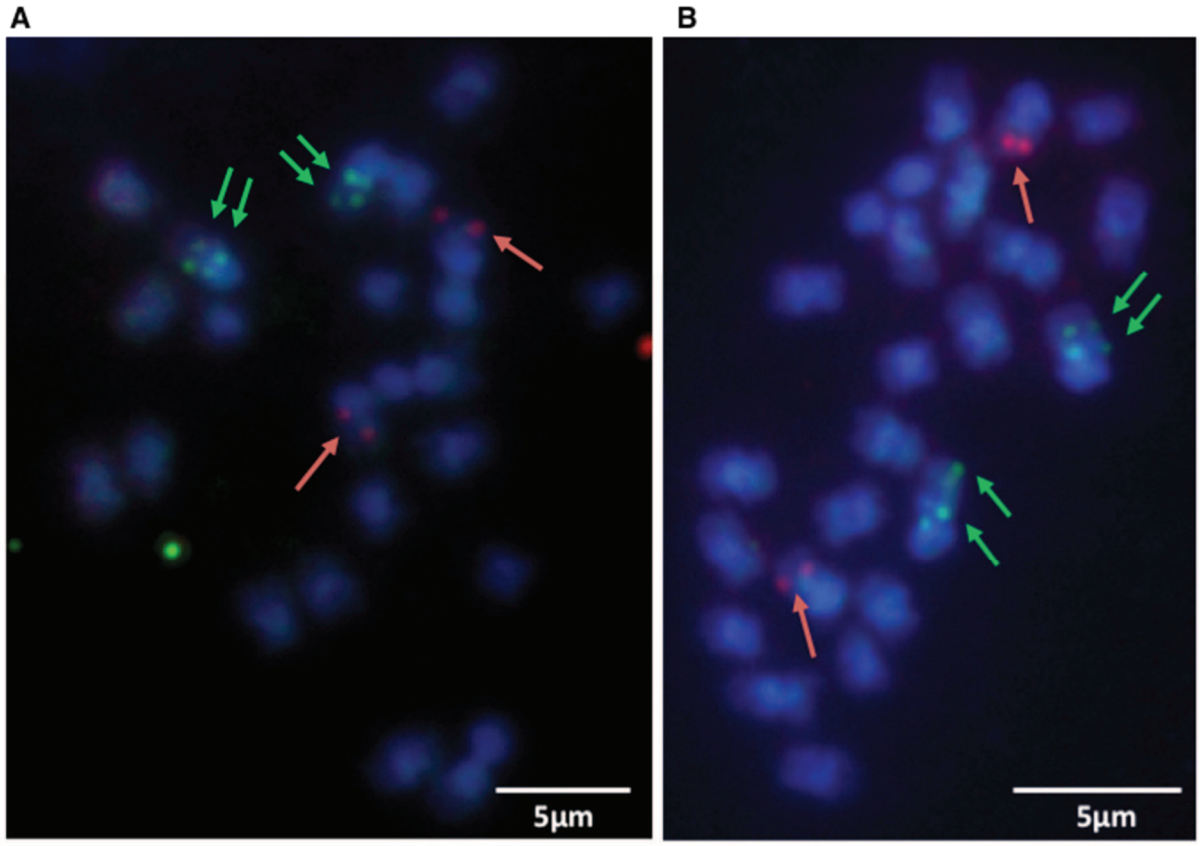


Figure 7: Localization of the three eBSV on the chromosomes of PKW.

Two independent metaphases are shown (Panels A and B). Hybridizations were performed with full-length genome probes for each BSV species and detected with FITC (green) for BSGFV and BSOLV and with alexa 594 for BSImV. Chromosomes are counterstained with DAPI (blue).

We then performed PCR and sequencing of the part of ORF 3 originally containing the two point mutations in BSImV particles obtained from five triploid offspring. None of the BSImV sequences obtained harboured either of the two point mutations. The detection by IC-PCR of BSImV particles expressed in the AAB progeny shows that 88% of the diseased hybrids are infected with BSImV (Lheureux 2002) (data not shown). Although the BSImV particles present in the hybrids come from eBSV we cannot say which allele is infectious as they are structurally identical. Nevertheless, the absence of both point mutations in the episomal virus released lead us to think that either the two alleles recombined during gamete formation or that reverse transcriptase of the virus may change one of the point mutations very early after the first transcription event.

5- Genomic landscape and eBSV chromosome localization

At first glance, gene and TE content appear highly variable between the different eBSVs (figure 6). Indeed, eBSGFV and eBSImV are inserted within two gene-rich regions with an average of 0.17 and 0.15 gene per kbp, while eBSOLV is inserted within a TE-rich region with only one predicted gene among the 190 kbp of total BAC sequence (0.005 gene per kb).

We investigated the 2 kb upstream and downstream of each eBSV more precisely by using multiple alignment tools (MAFFT, CLUSTAL) in order to search for common structure at the insertion locus. We previously established (Gayral et al. 2008) that eBSGFV is inserted into a Ty3/*gypsy-like* retrotransposon, which is itself inserted into the fifth intron of the *mom* gene. No similarities or notable structures exist that suggest a common target insertion site for the three BSV species. We performed additional analyses of the 20 first bases upstream and downstream of each eBSV to search for target site duplication (TSD). No similarity was recorded.

eBSOLV is inserted in a TE-rich region between two recently inserted retrotransposons based on LTR similarities (data not shown). Unlike eBSGFV, neither gene fragments nor retrotransposon-interrupted structures are detected flanking eBSVs, indicating that insertion occurred in intergenic sequences. We identified *gypsy-like* retrotransposons and a *copia-like* retrotransposon (RE06) close to the 52 and 32 flanking regions of eBSOLVs, respectively; however, the *gypsy-like* retrotransposon was not similar to the one containing eBSGFV.

For eBSImV, the 52 flanking region is composed of a truncated *RE02* LARD element while the 32 region is composed of intergenic sequence. We observed a similar environment to eBSOLV although the BAC is rich in genes.

We then performed FISH analysis using the three BSV species as probes to hybridize to the same chromosome preparation. eBSGFV and eBSOLV co-localised on the same chromosome

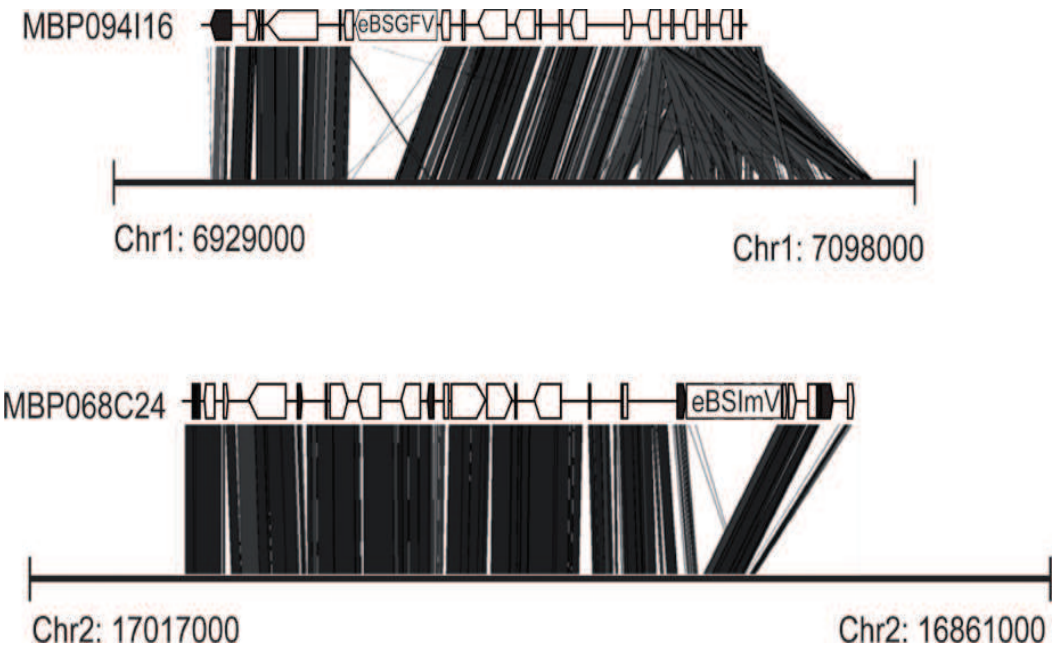


Figure 8: eBSV locus synteny with the *Musa acuminata* reference genome (D'Hont et al. 2012).

MBP094I16 and MBP068C24 refer to BACs containing eBSGFV and eBSImV, respectively. The black boxes correspond to TE and the white boxes to genes. Thick black lines represent the part of chromosomes 1 and 2 of the reference *M. acuminata* genome matching with BAC containing eBSGFV and eBSImV, respectively. Positions on chromosome are indicated with coordinates.

(as demonstrated by the presence of a pair of double green dots on the same chromosomes) whereas eBSImV is on a different chromosome (figure 7).

Blast alignment of all BAC clones to the recent *Musa acuminata* reference genome sequence (D'Hont et al. 2012) allowed synteny analysis to determine potential chromosome localisation (figure 8). We noted strong similarity for BACs containing eBSGFV with chromosome 1 and for BACs containing eBSImV with chromosome 2. Regarding the two regions surveyed, breaks in synteny are located precisely in the neighbourhood of eBSVs and in the vicinity of some TE. This confirms the distinct localisation of both integrations. No match was found for BAC containing eBSOLV. This is probably due to the high TE content of the BAC clone. Unfortunately, the only gene present on the BAC is not yet anchored in the *Musa acuminata* reference genome sequence.

6- Divergence and evolution of eBSV since their integration.

Faced with the genomic and genetic organization of eBSV within PKW described in this paper, we conclude that the genetic organization is complex and can result from multiple and/or sequential insertions of BSV at the same locus. In all cases, the initial hemizygous integration has been duplicated onto the homologous chromosome but, depending of scenario, the timing of this duplication is more or less recent. By comparing the nucleotide diversity of similar sequences, we then calculated the p-distance and Ks coefficient between BAC haplotypes (tables 3 and 4) and the p-distance within eBSV alleles (table 5).

The haplotype divergence of gene sequences surrounding eBSV sites is presented in table 3. No divergence was recorded for the only gene present in eBSOLV BACs, indicating recent duplication or a high selection constraint. Similarly, we observed weak divergence among the six genes present in eBSGFV BACs, with only two genes containing three mutations (2 non synonymous and 1 synonymous). The divergence of nucleic acid sequences flanking eBSVs is 0.0026 for the eBSGFV locus and 0.0013 for the eBSOLV locus. These values are similar to those of eBSVs themselves, in which divergence between eBSGFV-7 and eBSGFV-9 is 0.0037 and divergence between eBSOLV-1 and eBSOLV-2 is 0.0013 (Supplementary File 1). Consequently, the data presented here do not allow calculation of the time of divergence based on synonymous mutations accumulated into gene sequences. However, the nucleotide divergence data obtained on almost 100kb of sequence may indicate that divergence of the eBSGFV locus may be twice as old as that of the eBSOLV locus.

Within eBSV, the divergence recorded varies from gene to gene. The divergence recorded within eBSGFV-7 and eBSGFV-9 in ORF3 is higher than that of other genes (ORF1 and ORF2) or IG. The divergence observed within eBSGFV-9 is systematically higher than that observed within eBSGFV-7. This certainly indicates a differential evolution of each allele.

Table 3: *Musa* haplotype gene divergence of eBSV BAC clones dating

BSGFV Haplotypes	Size	nSyn	Syn	Ks	My
PAL	2139	0	0	0	0.00
Zonadhesin	3846	1	1	0.0011	0.12
HP1	1239	0	0	0	0.00
HP2	1419	0	0	0	0.00
Auxin respnsiveprotein	420	0	0	0	0.00
Actin depolymerizing factor	432	1	0	0	0.00
BSOLV Haplotypes	**Size**	**nSyn**	**Syn**	**Ks**	
Histidine kinase	3001	0	0	0	0

Syn: Synonymous mutation; nSyn: Non synonymous mutation; Ks: Synonymous mutations/synonymous sites; MY: Million years

Table 4: *Musa* haplotype nucleic acid divergence of eBSV BAC clones

Element	size (bp)	Nucleic acid differences between haplotypes	p-distance
Colinear region except eBSGFV	82,579	210	0.002554
Colinear region except eBSOLV	101,400	130	0.00129

Table 5: Divergence (p-distance) within eBSV

	ORF1	ORF2	ORF3	IG
eBSGFV-7	0.0000	0.0000	0.0042	0.0000
eBSGFV-9	0.0025	0.0040	0.0074	0.0056
eBSOLV-1	0.0012	0.0000	0.0011	0.0010
eBSOLV-2	0.0000	0.0016	0.0014	0.0031
eBSImV	NA	NA	0.0018	0.0010

NA: not applicable

The values obtained for both genes and IG within eBSOLV-1 and eBSOLV-2 are similar, with the exception of IG in eBSOLV-2, which accumulates more mutations. The divergence observed within eBSOLV-2 is systematically higher than that observed within eBSOLV-1 except for ORF 1, indicating a differential evolution of each allele here also, as observed for eBSGFV. Values are globally lower than those recorded for eBSGFV.

The eBSImV divergence of duplicated ORF3 and IG within the available allele is similar to that recorded for ORF3 of eBSOLV.

DISCUSSION

A similar integration pattern for each BSV species in PKW

Here, we fully characterized for the first time the integration of three widespread BSV species present within the genome of PKW and restricted to *Musa balbisiana* species. We found that these eBSV are absent not only from the two other common *Musa acuminata* accessions screened during this study but also from the *M. acuminata* spp. malaccensis DH Pahang reference sequence (D'Hont et al. 2012) and more widely from the other *Musa* species (Gayral et al. 2010; P.O. Duroy, X. Perrier and M.L. Iskra-Caruana, unpublished data). Both genetic studies and FISH analysis confirmed that integration of BSOLV, BSGFV and BSImV each resulted from independent events at a single locus. The PKW genome harbors at least three different BSV species integrated once each, which is rare compared to integration events reported for other plant pararetroviruses. Lockhart et al. (2000) described more than a hundred clustered copies of viral integrants in tobacco (*N. edwardsonii*), Richert-Poggeler et al. (2003) identified ca. 100–200 copies of PVCV integrated at five loci in petunia. No other viral integrants or molecular fossils of BSOLV, BSImV and BSGFV, such as those reported for endogenous retroviruses (Lovisolo et al. 2003) were found in the PKW genome other than those described here. These examples illustrate the diversity of infectious integrant patterns known nowadays in plants, ranging from a few copies to several hundred copies.

Unlike the patterns described in PKW, non-infectious BSV-like sequences ranging from 700 bp to 18,000 bp have been discovered recently within the *Musa acuminata* genome of the newly sequenced DH-Pahang (D'Hont et al. 2012). These correspond to 24 loci spread over the entire banana genome, on 10 out of 11 chromosomes. Similar structures are reported in rice where Kunii et al. (2004) found 29 endogenous non-infectious RTBV-like sequences, and in tobacco (*N. tabacum*) where thousands of non infectious TVCV-like insertions have been uncovered (Jakowitsch et al. 1999). In contrast to infectious integrants, non-infectious integrants seem to exhibit a similar pattern, with numerous integrations all over the plant genome.

We noted that the 52 flanking region of eBSOLV in PKW is identical (100% identity) to the musa6 clone previously sequenced (gb: AF106946) from the polyploid cultivated AAB banana clone Obino l ′ Ewai (Ndowora et al. 1999). In addition, based on the use of specific PCR markers for each integrant, we have previously reported that eBSGFV and eBSImV conserve the same locus of integration among all the *Musa balbisiana* diploids available worldwide (Gayral et al. 2010). Using the same approach, our group (P.O. Duroy, X. Perrier and M.L. Iskra-Caruana, unpublished data) confirmed that the eBSV/*Musa balbisiana* genome junctions are extremely well conserved in all three BSV species in all B genomes available among 77 accessions of diploids and triploid banana plants representing the entire *Musa balbisiana* diversity available. All together, our data indicate that BSV integration occurred in natural *Musa balbisiana* diploids before domestication, and highlight the fact that current diploids and A/B interspecific cultivars originated from the same *M. balbisiana* ancestor carrying eBSVs.

The vicinity of endogenous pararetroviruses (EPRVs) with retrotransposons, mainly *Ty3-gypsy* elements of the family *Metaviridae*, has been reported frequently for several plants (Gregor et al. 2004; Richert-Poggeler et al. 2003; Staginnus et al. 2007). In PKW, this is supported by the systematic presence of TE closely surrounding eBSV. All eBSVs are inserted into, or very close to, repeated elements. This suggests that the process of BSV integration may be opportunistic. Quasi-LTR (QTR) regions similar to those flanking ePVCV in the petunia genome do not exist in banana. No specific sequence signatures required for retroelement integration (e.g. target site duplication or inverted repeats) is observed flanking eBSVs. We found that the three BSV integrations occur in different types of *Musa* regions (gene-rich for eBSGFV and eBSImV, and a TE-rich region for eBSOLV) and on different chromosomes (chromosome 1 for eBSOLV and eBSGFV, and 2 for eBSImV). Clearly, little is known about the molecular mechanisms of integration of pararetroviruses into plant genomes and additional data are required to elucidate this process. However, as viral DNA gets into the plant nucleus during infection, it has the chance of becoming integrated during DNA break repairs as suggested for eBSGFV inserted into Ty3/gypsy retrotransposon (Gayral et al. 2008) or by illegitimate recombination following the mechanisms of non homologous end-joining (NHEJ) as reported for several endogenous viral elements (EVE) (Feschotte and Gilbert 2012; Holmes 2011). In this latter case, integration is thought to occur during the minichromosome viral phase due to either the two gaps existing within the open circular viral DNA allowing access to single-stranded DNA or to single-stranded overhanging sequences (flaps) constituting the end of the open circular viral DNA being readily available to initiate the recombination process (Jakowitsch et al. 1999; Richert-Poggeler et al. 2003; Kunii et al. 2004; Hohn et al. 2008). This could explain the structure of eBSImV, i.e. 1.76 linear

genomes. Taken together, these data support the hypothesis of stochastic insertion of each BSV into a single locus of the *Musa* genome.

Sequential integration of BSV species into the genome of PKW.

It is difficult to estimate the time of BSV integration because BSV probably evolves with a faster substitution rate than that of its host and, consequently, of the eBSV counterpart, which is subject to the host rate. Consequently, as described for other EVE (Feschotte and Gilbert 2012), eBSV should reflect an ancestral state of the virus genome. We recorded strong identity at the nucleic acid level between eBSV and BSV for the three species (>99%).

The nucleotide divergence recorded between BSV and eBSV is between 1.5x10-3 and 2.8x10-3 mutations per site among the three BSV species (table 2). This divergence is due to the differential rate of mutation between viruses and the plant genome. With a rate for retroviruses assumed to be between 10-6 and 10-4 mutations per site per cell infection (Menendez-Arias 2009), we can estimate the divergence time between BSV and eBSV at 15 to 2,800 cell infection cycles (Sanjuan et al. 2010). This could indicate a fairly recent integration, whereas our data suggest an integration event after the speciation of A/B but before *M. balbisiana* diversification. This is in agreement with Gayral et al. (2008), who estimated BSGFV integration at or after 0.640 MYA. Therefore the low divergence observed between eBSV and BSV more probably reflects the massive and almost exclusive contribution of eBSVs to the current viral population since no epidemic of BSOLV, BSImV and BSGFV is reported worldwide.

The nucleotide divergence recorded within eBSVs (table 5), between eBSVs (table 5) and between BSV and eBSV (table 2), as well as the structural organization of integration, suggest sequential integration of BSV species into the PKW genome. It is difficult to distinguish which arrived first between BSOLV and BSGFV since the p distance seems to indicate an older insertion for BSGFV but the structure of eBSOLV, which is more re-arranged, may suggest the opposite. However, eBSImV contains fewer mutations that other eBSVs, shows a relative linear viral genome and is mono-allelic, all of which clearly suggests that eBSImV is the most recent.

Integration and rapid evolution of eBSV loci.

The different patterns of integration in PKW for the three BSV species range from structures similar to the concatenated linear viral genome for eBSImV, to the more complex organisation of eBSGFV and eBSOLV. Integrant size is always over a full-length viral genome, with no embedded *Musa* sequences, and each eBSV locus contain one functional genome. All three eBSV have two allelic copies.

Integration of tandem copies of a virus genome has been documented in plants (Richert-Poggeler et al. 2003). In eBSImV we found a continuous structure of a 1.76-length viral genome, suggesting that a tandem integration process occurred also in banana. This duplicated structure is inserted in the neighbourhood of repetitive elements that may promote rapid evolution of eBSV as has been documented for resistance gene cluster evolution (Mazourek et al. 2009, David et al. 2009). Indeed, the current structure of the eBSV locus may be compared to that described extensively in the same plant for the RGA08 locus (Baurens et al. 2010). Unlike the RGA08 locus, where intergenic sequences and TE are clearly implicated in the evolution of the locus, we found that only viral sequences are involved in evolution of eBSV loci structure. We observed that the variation in eBSV structure concerns only viral sequences whereas flanking sequences are remarkably conserved in all contexts, especially at the eBSOLV integration locus, which is flanked by dozens of TE. Indeed, we estimated nucleotide diversity at 0.25 % (0.03% for coding sequences) for eBSGFV haplotypes, and 0.13% (no variation for coding sequences) for eBSOLV haplotypes (tables 3 and 4).

To arrive at the current picture of integration in PKW, we assume that the most probable scenario is based on the unequal recombinations that follow duplication of the initial integration on the homologous chromosome. However, we cannot totally rule out other scenarios. Indeed, sequential targeted insertion of virus at the same locus might occur. We observed that point mutations accumulate in eBSVs at the same rate as in flanking *Musa* sequences, indicating that novel viral sequences have not integrated since haplotype divergence. Moreover, we searched for accumulation of mutations within duplicated viral fragments of eBSV. The absence of any difference allows us to discard the scenario of sequential waves of virus integration at the same locus, contrary to what is observed in the *Musa acuminata* HD genome (D'Hont et al. 2012; M. Chabannes and F.C. Baurens, unpublished data). A further possibility concerns recombination between eBSV and virions; however, this is very unlikely because it requires at least two crossing overs within a very short distance (virus genome length) to produce viable gametes.

Current integration structures are driven by both virus and host

Gayral et al. (2010) established that eBSGFV, eBSOLV and eBSImV are present in the genomes of all *M. balbisiana* surveyed. We established here by genomic, genetic and cytogenetic analysis that each eBSV is present on a homologous chromosome in the PKW genome. In a wild population with N random-mating diploids, if the new DNA does not confer any selective advantage to the host plant, the probability that the mutation becomes fixed (i.e. that the entire population contains homozygous new DNA) is 1/2N. The fixation

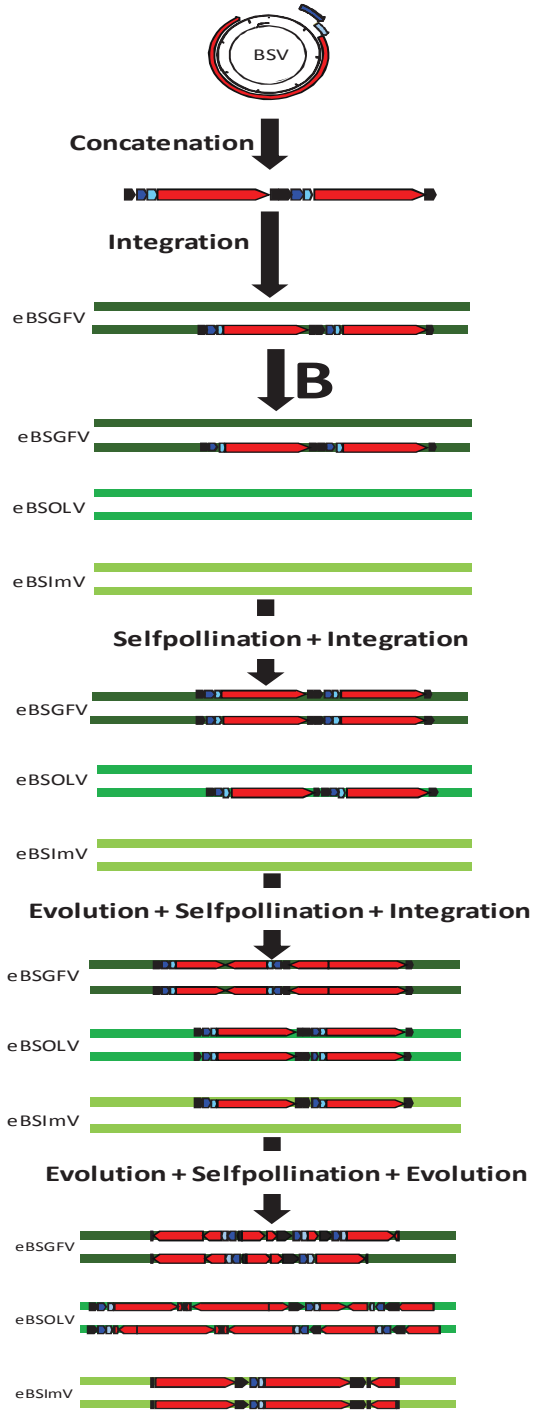


Figure 9: Proposed scenario for BSGFV, BSOLV and BSImV integrations into the *M. balbisiana* nuclear genome.

The life cycle of the virus produces concatenated viral genomes in the nucleus of the cell. This concatenated genome is integrated into the nuclear genome of PKW by illegitimate recombination. The three BSV species integrate sequentially into the PKW genome. After each BSV integration, the homozygous state is obtained by self-pollination and selected. Between each new integration and self-pollination, the previously integrated BSV evolves continously by unequal recombination and point mutation insertions, leading to the present-day structure. The BSV genome is represented in linear view with dark blue, light blue and red boxes indicating the three ORFs of the virus. The intergenic region (IG) is in black.

process takes on average 4N generations (Innan and Kondrashov 2010). Fixation is unlikely to occur without selective pressure in the *M. balbisiana* context due to pollen dispersion over a large geographical area, which implies an extensive banana population (Ge et al. 2005). In addition, the flower morphology of *Musa* does not usually allow self-pollination in a single bunch because male and female flowers bloom not at the same time but sequentially. However, vegetative multiplication of bananas produces clump of plants composed of the same genotype that can produce multiple flowers at the same time, thus allowing self-fertilization. Our data support the early statement that wild bananas are outbred but tolerate occasional generations of inbreeding (Simmonds 1962).

PKW eBSV loci are widely conserved in almost all banana cultivars with B genome (P.O. Duroy, X. Perrier and M.L. Iskra-Caruana, unpublished data). This is explained easily if a small number of plants form the origin of these cultivars. Indeed, new multi-disciplinary findings concerning domestication of banana (Perrier et al. 2011) suggest that the species *M. balbisiana* has been selected by mankind and transferred out of its geographical area of origin before contributing to some important groups of interspecific cultivars. However, the finding of similar eBSV structures also in seedy *M. balbisiana* populations (Gayral et al. 2010) suggests that a strong bottleneck might have occurred as previously suggested by (Ge et al. 2005) to explain the relatively low diversity observed in *M. balbisiana* species nowadays.

Based on the data collected and elements discussed in this article, we proposed a model explaining the integration and evolution process of BSV from virions to the currently observed eBSV (Figure 9).

Our model uses the property of BSV to concatenate its DNA during replication as a template for the initial integration in germ cells. This is followed by a duplication process (self-pollination) of eBSV on the homologous chromosome. Alleles then evolved by both unequal recombination and point mutation accumulation. This process took place independently and sequentially for the three eBSV described here.

Why do eBSVs persist?

It is difficult to imagine maintenance of a functional viral sequence in the host genome without any selective advantage. The initial context of BSV integration favours a harmonious co-existence between the two partners. This proximity allowed passive and stochastic movement of DNA. It is therefore likely that the initial integration conferred a better fitness to the infected plant, leading to fixation of eBSVs in the banana population. Once fixed, the endogenous viral genome duplications and inversions observed may indicate a plant strategy to disarm any viral activity with potentially lethal effect. The presence of the viral genome in

the plant allows the establishment of a resistance mechanism based on sequence homology between eBSV and BSV to prevent further external infection. Gene-silencing-based resistance has been reported in tobacco and petunia for TVCV and PVCV, respectively (Mette et al. 2002; Richert-Poggeler et al. 2003; Noreen et al. 2007).

However, we observed that a differential selection between eBSV alleles exists (table 5). This differential selection seems to favour the infectious allele as more mutations accumulate in the non infectious allele in both eBSOLV and eBSGFV. Moreover, the infectious allele (eBSGFV-7) is more prevalent in the seedy BB plant diversity than the non-infectious allele (Gayral et al. 2010). This suggests that selection tends to retain functional alleles. We assume that, after the initial integration of viral DNA into the plant genome, two antagonist forces act on the eBSV locus: one aiming to keep functional integrations in the population and the other aiming to disrupt the viral genome to prevent "self infection".

In PKW, eBSV structures imply a minimum of two steps of recombination to release a functional viral genome. These recombination steps have been described extensively by Iskra-Caruana et al. (2010) for eBSGFV, and our data suggests that recombination also occurs in the expression process for both eBSOLV and eBSImV. The structure of eBSOLV does not permit direct transcription and virions obtained in *Musa* population always contain fragment 1-VI and a hotspot of recombination existing in the IG that is probably involved in viral genome recircularisation (figure 5). Functional BSImVmost likely arise from recombination because of the presence of deleterious mutations in eBSV (frameshift or a stop codon) that prevent direct transcription of functional viral genomes. Thus, we assume that a first step of structural evolution, leading to viral genome shuffling in eBSV, is necessary to prevent the large scale production of viral genomes and viruses from eBSVs, which might threaten plant populations.

In the case of PKW, this process must have occurred with the three BSV species at different times, and probably in response to different epidemic or abiotic constraints.

MATERIALS AND METHODS

Screening of BAC library.

Bacterial artificial chromosome (BAC) libraries were obtained from diploid *M. balbisiana* PKW (Safar et al. 2004) and two *M. acuminata* banana plants: the diploid Calcutta 4 (AA) (Vilarinhos et al. 2003) and the triploid "Cavendish" subgroup cv. Grande Naine (AAA) (Piffanelli et al. 2008). Clones of BAC libraries were spotted onto high-density Hybond N+ filters (AP Biotech, Little Chalfont, United Kingdom) using a Flexys robot. The filters were hybridized with BSV probes covering the entire viral genome of four BSV species [*Obino l'ewai* - BSOLV (NC_003381.1/Harper and Hull 1998), *Imove* - BSImV (HQ659760

/Geering et al. 2011), Mysore - BSMysV (AY805074/Geering et al. 2005), *Vietnam* - BSVNV(AY750155/Lheureux et al. 2007)] as described in Gayral et al. (2008) for Goldfinger - BSGFV. BSV-positive BAC DNA was digested with four different enzymes (*Hin*dIII, *Bam*HI, *Pst*I, and *Xho*I), and separated on a 0.8% agarose gel in 1X Tris-acetate-EDTA at 60 V, run for 20 h. The separated fragments were denatured and transferred to nylon membrane (Hybond-N+; Amersham Pharmacia Biotech). Filters were hybridized with probes corresponding to the full length genome of each BSV species. Restriction profiles were scored manually. One BAC clone was selected by fingerprint profiles for sequencing.

BAC sequencing.

BACs containing eBSGFV were obtained from GenBank (accession numbers AP009325 and AP009326 corresponding to MBP_71C19 and MBP_94I16, respectively; Gayral et al. 2008). BAC MBP_31007 containing eBSOLV was obtained from GenBank (accession number AP009334).

Selected BAC clones containing eBSOLV (BAC_73B22 and BAC_17D14) as well as eBSImV (BAC_68C24) were sequenced at Genoscope (http://www.genoscope.cns.fr/spip/). Libraries from the three BAC clones were obtained after mechanical shearing of BAC DNA and cloning of 5 kbp and 10 kbp fragments into pcdna 2.1 (Invitrogen) and pCNS (pSU18-derived) plasmids, respectively. Vector DNAs were purified and end-sequenced using dye terminator chemistry on ABI 3730 sequencers (Applied Biosystems, France) until 12X coverage. BACs 73B22 and 17D14 were assembled using the Phred/Phrap/Consed software package (http://www.phrap.com), whereas Arachne assembler was used for BAC 68C24. Primer walks and PCR were needed to complete the final phases.

For BACs harboring eBSImV (MBP_68C24) or eBSOLV (MBP_17D14c) sequences, the Genbank accession numbers are HE983625 and HE983609, respectively.

Gene model prediction, sequence alignment and synteny analysis

Plant and virus genes structures were predicted using the EuGène combiner release 3.2 (Foissac et al. 2008) with rice-specific parameters that integrate several lines of evidence. Gene models were predicted with the *ab initio* gene finders, EuGèneIMM and Fgenesh (Salamov and Solovyev 2000). Translation start sites and splice sites were predicted by SpliceMachine (Degroeve et al. 2005). Available monocotyledon ESTs from EMBL were aligned on the genome using Sim4 (Florea et al. 1998). Similarities to protein sequences were identified using BLASTX (NCBI-BLASTALL) (Altschul et al. 1997) on (Consortium 2009).

Polypeptide functions were also predicted by integrating several lines of evidence. Protein similarities were searched for using tBLASTn on translated monocotyledon ESTs, and

BLASTp on UniProt. Protein domains were predicted with InterproScan (Quevillon et al. 2005). Clusters of orthologous genes between the predicted polypeptides and the proteomes of *Oryza sativa* (TIGR release 5.0) and *Sorghum bicolor* (JGI release 1.0) were also identified using the pipeline GreenPhyl (Conte et al. 2008). Predicted genes were annotated manually using Artemis (Carver et al. 2008). A gene is considered complete when its coding sequence (CDS) is canonical and significantly matches a known sequence in the public databanks with coverage parameters Qcov and Scov greater or equal to 0.8. Under these parameters, CDS is also predicted to be functional. When a gene contains mutations that could prevent correct expression (i.e. a missing start or stop codon, non canonical splice site, frameshift or in-frame stop codon), it is considered a pseudogene. A polypeptide was annotated as a fragment when its coverage (Qcov) was less than 0.8 when comparing its length to the length of the match with the best significant hit. We annotated a gene as a remnant when it was composed of a small fragment (Qcov <0.3), more than three fragments (Qcov <0.5) and/or when it had more than two mutations preventing correct CDS expression. BAC sequence annotation is available on a genome browser at http://gnpannot.musagenomics.org/cgi-bin/gbrowse/musa/. Sequence comparison between haplotypes was performed using dotplot analysis (Krumsiek et al. 2007). Local alignments were performed using BLASTN (Altschul et al. 1997) and visualised with the Artemis Comparison Tool (ACT) (Carver et al. 2008). Synteny analysis with *Musa acuminata* reference genome (D'Hont et al. 2012) was performed using blast analysis. BAC sequences were compared to pseudomolecules using a blastn search with an e value of 10-10 and visualised with ACT. The synteny between BAC proteomes and the reference *Musa* genome was performed using blastp analysis against translated CDS at http://www.banana-genome.cirad.fr/blast.html.

Fluorescent in situ hybridization - FISH

Plant material and chromosome preparations:

Roots tips of *M. balbisiana* diploid PKW grown in a tropical greenhouse were collected at different times in the morning from 8.45 a.m to 11.30 a.m. The young growing roots were treated with 8-hydroxyquinoleine 0.04% for 4 hours, fixed in a solution of ethanol: glacial acetic acid (3:1 v/v) for 24 hours (12 hours at room temperature and 12 hours at 4°C) then kept in 70% ethanol. Chromosome preparations were prepared as described in D'Hont et al. (1996) with slight modifications: the enzymatic mixture was composed of 1% cellulase (SIGMA), 1% cytohelicase (SIGMA) and 1% pectolyase (SIGMA). Before proceeding to the squash, the roots were kept in water overnight at 4°C.

The hybridizations were performed with full-length genome probes for each BSV species and the 45S rDNA sequence as a control. The virus probes were labelled by random priming with

biotin-14-dUTP (Invitrogen life technology) and digoxigenin-11-dUTP (High Prime DNA Labeling Kit, Roche), and with biotin for 45S. The hybridisation protocol was described by D'Hont et al. (2000) with slight modifications: selected slides were pre-treated with 1 µg/mL of RNase in 2x SSC for 45min at 37°C, then rinsed twice in 2XSSC at room temperature. Slides were then denatured in a solution of 70% formamide (v/v) in 2XSSC for 2 min at 70°C in a bath, immersed for 3 min in 2XSSC in a bath previously put on ice and finally dehydrated through an alcohol series (5 min in each in ethanol baths of 70%, 90% and 100%, respectively, at –20°C). The hybridization mixture contained: 50% (v/v) deionised formamide, 10% (w/v) sodium dextran sulfate, 2XSSC, 0.5 µg/µL of sonicated salmon sperm DNA, 0.66% (w/v) of SDS and 5 µL of virus probes. This mixture was denatured in a boiling bath for 10 min and kept on ice for at least 15min; 50 µL of hybridization mixture was applied to each slide and covered with a plastic coverslip. Hybridization was carried out overnight at 37°C in a moist chamber. The next day, the slides were washed in 2XSCC, 0.5XSSC, 0.1XSSC for 10 min each at 42°C, and 2XSSC at room temperature. Slides were pre-treated with 5% (w/v) BSA (bovine serum albumin) in 4XSSC with Tween20 for 10 min at 37°C. The detection solution contained 6ng/µL of avidin-rhodamine and 10ng/µL of sheep anti-Dig FITC diluted in the pre-treatment BSA solution. Fifty microliters of this mixture were loaded on each slide and detection was carried out for 45 min at 37°C in a moist chamber. Slides were rinsed 3 times in 4x SSC with Tween20 for 5 min at 42°C. Then 50 µl of a solution of 7.5 ng/µl of rabbit anti-sheep-FITC antibodies and, if amplification was required for probes labeled with biotin (in the case of virus biotin probes), 5ng/µL of biotinylated anti-avidin antibodies diluted in goat serum 4XSSC with Tween20 were loaded on each slide and left for 45 min in a moist chamber at 37°C. Slides were rinsed 3 times in 4XSCC with Tween20 for 5 min at 42°C. Then, chromosomes were counter-stained with Vectashield mounting medium with DAPI (Vector Laboratories). In the case of virus biotin probes, before this step, a final detection step with a solution of 6ng/µl of avidine-rhodamine diluted in 5% (w/v) BSA, 4XSSC with Tween20 was applied for 45 min in a moist chamber at 37°C followed by 3 rinses in 4XSCC with Tween20 for 5 min at 42°C.

Observations were carried out on an epi-fluorescence microscope (LEICA DMRXA 2), images are captured using a cooled Hamamatsu Orca AG camera, and processed using Velocity software (Perkin Elmer).

Genetic analysis

Interspecific genetic cross of PKW with cv. IDN 110 4*x*.

The plant population used in the present study consisted of 165 F1 allotriploid hybrids (AAB). This population derives from an interspecific genetic cross between the virus-free

diploid (BB) *Musa balbisiana* female parent PKW and the virus-free autotetraploid (AAAA) *Musa acuminata* male parent cv. IDN 110 4*x*. Absence of viruses in both parents was confirmed by immunosorbent electron microscopy and by immunocapture PCR (IC-PCR) (Lheureux et al. 2003). This genetic cross is fully described and characterized in Lheureux et al. (2003). Leaf samples were stored at -80°C.

DNA extraction.

Total DNA was extracted by the method described in Gawel and Jarret (1991) from leaf tissue of AAB progeny stored at -80°C. The quality and amount of DNA was estimated visually after separation of 5 µl of DNA extraction in a 0.8% agarose gel, staining with ethidium bromide, and visualizing on a UV transilluminator.

PCR.

All PCRs were performed on 5 to 20 ng of template DNA using a common mix composed of 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 0.1 mM each deoxynucleoside triphosphate, 1.5 mM MgCl*2*, 400 nM of the forward and reverse primers, and 1 U of *Taq* DNA polymerase (Eurogentec, Seraing, Belgium) in a final volume of 25 µl. DNA was amplified using the following program: one cycle at 94°C for 4 min, 35 cycles at 94°C for 30 s, primers annealing temperature for 30 s, 72°C for 1 min per kb, and a final extension at 72°C for 10 min. Amplicons were visualized after migration of 8 µl of PCR products on an agarose gel (1–3% according to the expected size of the amplicon) in 0.5X TBE (45 mM Tris-borate, 1 mM EDTA [pH 8]). The gel was stained with ethidium bromide and amplified bands visualized under UV light.

eBSV genotyping

To perform genetic analysis of eBSV loci, segregation analysis was performed on the triploid progeny (AAB) described above. Allelic segregation was estimated in the eBSV region using two types of molecular markers: (i) markers derived from the eBSV structure itself when the specific structure of integration allows for specific amplification of the integration locus (see below), (ii) SSR markers defined from the BAC sequence with the dedicated pipeline SAT (http://sat.cirad.fr/sat) (Dereeper et al. 2007) of the Southgreen bioinformatic platform (http://southgreen.cirad.fr/), using default parameters.

The targeted eBSV, the type of marker developed, the name and the sequence of the primer as well as the size of the PCR product are given in the table provided in supplementary data 1. For allelic markers, discrimination between alleles is also described in detail in supplementary data 1.

BSV genotyping

BSV was genotyped by immunocapture PCR. The immunocapture step consisted of coating sterile polypropylene thin-walled 0.2 ml microfuge tubes (Axygen, Union City, CA) for 4 h at 37°C with 25 µl of immunoglobulin G purified from polyclonal antiserum raised against BSV species and *Sugarcane bacilliform virus* species (a kind gift from Dr B.E.L. Lockhart), diluted at 2 µg/ml in carbonate coating buffer (15 mM sodium carbonate, 34 mM sodium bicarbonate [pH 9.6]). The tubes are then washed three times with 100 µl of PBT washing buffer (136 mM NaCl, 1.4 mM KH2PO4, 2.6 mM KCl, 8 mM NA2HPO4, 0.05% Tween 20 [pH 7.4]). Plant extracts were prepared by grinding 0.5 g leaf samples in 5 ml of grinding buffer (2% polyvinylpyrrolidone 40, 0.2% sodium sulfite, and 0.2% bovine serum albumin prepared in PBT) using a manual bead grinder and plastic grinding bags (Bio-Rad Phytodiagnostics, Marnes-la-Coquette, France). Portions (1 ml) of plant extracts were transferred to microfuge tubes and clarified by centrifugation at room temperature for 5 min at 7,000 rpm. Then, 25 µl of the supernatant was loaded into coated tubes, followed by incubation for 1 h 30 min at room temperature. The tubes were washed five times with 100 µl of PBT, three times with 100 µl of sterile water, and then dried briefly.

To avoid any contamination by plant genomic DNA, a DNAse I treatment (RNase-Free DNAse from PROMEGA) was performed. 30µl of DNAse mix (3µl 10Xbuffer (400mM Tris-HCl [pH 8.0 at 25°C], 100mM MgSO4, 10mM CaCl2); 3µl of DNAse I and 24 µl of water) was added to the coated tubes and incubated for 1h at 37°C. The supernatant was removed and the tubes washed once with water. DNAse I was inactivated by incubation at 95°C for 10 min.

BSV was genotypedby PCR directly in tubes using specific BSV species primers. The following primers were used for BSOLV (OL-R [5'-GCT CAC TCC GCA TCT TAT CAG TC-3'] and OL-F [5'-ATC TGA AGG TGT GTT GAT CAA TGC-3']) ; BSImV (Im-R [5'-CAC CCA GAC TTT TCT TTC TAG C-3'] and Im-F [5'-TGC CAA CGA ATA CTA CAT CAA C-3']) and BSGFV (GF-R [5'-TCG GTG GAA TAG TCC TGA GTC TTC-3'] and GF-F [5'-ACG AAC TAT CAC GAC TTG TTC AAG C-3']). PCR reactions were performed as described above at a *Ta* of 58°C for 25 cycles (BSImV) or 30 cycles (BSOLV and BSGFV). Genomic DNA contamination was controlled using *Musa* sequence tagged microsatellite site primers AGMI025 [5'-TTA AAG GTG GGT TAG CAT TAG G-3'and AGMI026 [5'-TTT GAT GTC ACA ATG GTG TTC C-3'] (Lagoda et al. 1998). These primers were used in multiplex PCR with the specific BSV primers described above.

Haplotype divergence of eBSV in PKW.

The divergence time between allelic *Musa* regions corresponding to overlapping BAC sequences was calculated based on genic, intergenic and repetitive sequence divergence according to the formula T = K/(2r), where T is the time of divergence, K is the number of base substitutions per site, and r is the substitution rate (SanMiguel et al. 1998). Nucleotide substitutions were calculated using MEGA4 (Tamura et al. 2007) with the substitution model of Kimura 2-parameter. A rate 2-fold higher than that determined for coding sequences in banana (Lescot et al. 2008) was used based on the assumption that non-coding sequences evolve more rapidly (Ma and Bennetzen 2004).

In order to estimate if sequential insertions of BSV occurred in eBSV, we calculated the p-distance (proportion of nucleotide differences between two sequences) within each eBSV allele for their duplicated ORFs and IG. The same sequence was used to calculate the p-distance between eBSV alleles. Due to the high reorganization of each eBSV, we compared only parts of the viral genome, calculated pairwise distances and finally used the average p-distances to compare their evolution. For eBSGFV, we compared ORF1 and ORF2 included in fragments III, Va, Vb and Vc; 0.6kb of ORF3 included in fragment II, IV, Va and Vc and 0.24kb IG in fragments III, Va, Vb and Vc. For eBSOLV we compared ORF1 and ORF2 in fragments 1-I, 1-VI, 1-VII, 2-I, 2-VII and 2-VIII; 3kb of ORF3 included in fragments 1-I, 1-V, 2-III, 2-VII and 2-VIII; 1kb of ORF3 included in fragments 1-I, 1-V, 1-VIII, 2-III, 2-VII, 2-VIII and 2-IX and 0.66kb IG included in fragments 1-I, 1-VI, 1-VII, 2-I, 2-VII and 2-VIII. For BSImV, we compared 1kb of ORF3 included twice in fragments Im-II and once in Im-III, and 1kb of IG included twice in fragment Im-II.

The complete distance matrix is available in supplementary data 2.

Data access

GenBank accession numbers for BAC harboring eBSImV sequence (MBP_68C24) and eBSOLV sequences (MBP_17D14c) are HE983625 and HE983609, respectively.

Acknowledgments

This work was supported by a CIRAD and Genoscope grant under the project “Sequencing and molecular mapping of EPRV-BSV in banana”. The authors wish to thank Gaëtan DROC for the set up of the Gnpannot annotation platform, Nathalie Laboureau for technical help in population segregation analysis and Sophie Mangenot for technical help with BAC sequencing data. We are very grateful to P. Capy and A. D’Hont for critical reading of the manuscript. P.O. Duroy is supported by a CIRAD Ph.D grant and P. Gayral by a CIRAD/Région Languedoc Roussillon Ph.D grant.

References

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**(17): 3389-3402.

Baurens FC, Bocs S, Rouard M, Matsumoto T, Miller RN, Rodier-Goud M, D MB-A-M, Yahiaoui N. 2010. Mechanisms of haplotype divergence at the RGA08 nucleotide-binding leucine-rich repeat gene locus in wild banana (Musa balbisiana). *BMC Plant Biol* **10**: 149.

Belyi VA, Levine AJ, Skalka AM. 2010a. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. *PLoS Pathog* **6**(7): e1001030.

Belyi VA, Levine AJ, Skalka AM. 2010b. Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old. *J Virol* **84**(23): 12458-12462.

Carver T, Berriman M, Tivey A, Patel C, Bohme U, Barrell BG, Parkhill J, Rajandream MA. 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**(23): 2672-2676.

Consortium U. 2009. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* **37**(Database issue): D169-174.

Conte MG, Gaillard S, Lanau N, Rouard M, Perin C. 2008. GreenPhylDB: a database for plant comparative genomics. *Nucleic Acids Research* **36**: D991-D998.

Cote FX, Galzi S, Folliot M, Lamagnere Y, Teycheney PY, Iskra-Caruana ML. 2010. Micropropagation by tissue culture triggers differential expression of infectious endogenous Banana streak virus sequences (eBSV) present in the B genome of natural and synthetic interspecific banana plantains. *Mol Plant Pathol* **11**(1): 137-144.

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M et al. 2012. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*.

D'Hont A, Grivet L, Feldmann P, Rao S, Berding N, Glaszmann JC. 1996. Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. *Mol Gen Genet* **250**(4): 405-413.

D'Hont A, Paget-Goy A, Escoute J, Carreel F. 2000. The interspecific genome structure of cultivated banana, Musa spp. revealed by genomic DNA in situ hybridization. *Theoretical and Applied Genetics* **100**(2): 177-183.

Dallot S, Acuna P, Rivera C, Ramirez P, Cote F, Lockhart BE, Caruana ML. 2001. Evidence that the proliferation stage of micropropagation procedure is determinant in the expression of banana streak virus integrated into the genome of the FHIA 21 hybrid (Musa AAAB). *Arch Virol* **146**(11): 2179-2190.

David P, Chen NW, Pedrosa-Harand A, Thareau V, Sevignac M, Cannon SB, Debouck D, Langin T, Geffroy V. 2009. A nomadic subtelomeric disease resistance gene cluster in common bean. *Plant Physiol* **151**(3): 1048-1065.

Degroeve S, Saeys Y, De Baets B, Rouze P, Van de Peer Y. 2005. SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics* **21**(8): 1332-1338.

Dereeper A, Argout X, Billot C, Rami JF, Ruiz M. 2007. SAT, a flexible and optimized Web application for SSR marker development. *BMC Bioinformatics* **8**: 465.

Feschotte C, Gilbert C. 2012. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet* **13**(4): 283-296.

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Research* **8**(9): 967-974.

Foissac S, Gouzy J, Rombauts S, Mathe C, Amselem J, Sterck L, Van de Peer Y, Rouze P, Schiex T. 2008. Genome annotation in plants and fungi: EuGene as a model platform. *Current Bioinformatics* **3**(2): 87-97.

Gawel NJ, Jarret RL. 1991. A modified CTAB DNA extraction procedure for *Musa* and *Ipomoea*. *Plant Mol Biol* **Rep 9**: 262-266.

Gayral P, Blondin L, Guidolin O, Carreel F, Hippolyte I, Perrier X, Iskra-Caruana ML. 2010. Evolution of endogenous sequences of banana streak virus: what can we learn from banana (Musa sp.) evolution? *J Virol* **84**(14): 7346-7359.

Gayral P, Iskra-Caruana ML. 2009. Phylogeny of Banana streak virus reveals recent and repetitive endogenization in the genome of its banana host (Musa sp.). *J Mol Evol* **69**(1): 65-80.

Gayral P, Noa-Carrazana JC, Lescot M, Lheureux F, Lockhart BE, Matsumoto T, Piffanelli P, Iskra-Caruana ML. 2008. A single Banana streak virus integration event in the banana genome as the origin of infectious endogenous pararetrovirus. *J Virol* **82**(13): 6697-6710.

Ge XJ, Liu MH, Wang WK, Schaal BA, Chiang TY. 2005. Population structure of wild bananas, Musa balbisiana, in China determined by SSR fingerprinting and cpDNA PCR-RFLP. *Mol Ecol* **14**(4): 933-944.

Geering AD, Parry JN, Thomas JE. 2011. Complete genome sequence of a novel badnavirus, banana streak IM virus. *Arch Virol* **156**(4): 733-737.

Geering AD, Pooggin MM, Olszewski NE, Lockhart BE, Thomas JE. 2005. Characterisation of Banana streak Mysore virus and evidence that its DNA is integrated in the B genome of cultivated Musa. *Arch Virol* **150**(4): 787-796.

Gilbert C, Feschotte C. 2010. Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS Biol* **8**(9).

Gregor W, Mette MF, Staginnus C, Matzke MA, Matzke AJ. 2004. A distinct endogenous pararetrovirus family in Nicotiana tomentosiformis, a diploid progenitor of polyploid tobacco. *Plant Physiol* **134**(3): 1191-1199.

Harper G, Hull R. 1998. Cloning and sequence analysis of banana streak virus DNA. *Virus Genes* **17**(3): 271-278.

Harper G, Hull R, Lockhart B, Olszewski N. 2002. Viral sequences integrated into plant genomes. *Annu Rev Phytopathol* **40**: 119-136.

Hohn T, Richert-Poeggeler KR, Staginnus C, Harper G, Schwarzacher T, Chee How T, Teycheney PY, Iskra Caruana ML, Hull R, ed. 2008. *Evolution of integrated plant viruses*. Springer Berlin.

Holmes EC. 2011. The evolution of endogenous viral elements. *Cell Host Microbe* **10**(4): 368-377.

Horie M, Honda T, Suzuki Y, Kobayashi Y, Daito T, Oshida T, Ikuta K, Jern P, Gojobori T, Coffin JM et al. 2010. Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* **463**(7277): 84-87.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* **11**(2): 97-108.

Iskra-Caruana ML, Baurens FC, Gayral P, Chabannes M. 2010. A four-partner plant-virus interaction: enemies can also come from within. *Mol Plant Microbe Interact* **23**(11): 1394-1402.

Jakowitsch J, Mette MF, van Der Winden J, Matzke MA, Matzke AJ. 1999. Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. *Proc Natl Acad Sci U S A* **96**(23): 13241-13246.

Katzourakis A, Gifford RJ. 2010. Endogenous viral elements in animal genomes. *PLoS Genet* **6**(11): e1001191.

Krumsiek J, Arnold R, Rattei T. 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**(8): 1026-1028.

Kunii M, Kanda M, Nagano H, Uyeda I, Kishima Y, Sano Y. 2004. Reconstruction of putative DNA virus from endogenous rice tungro bacilliform virus-like sequences in the rice genome: implications for integration and evolution. *BMC Genomics* **5**: 80.

Lagoda PJ, Noyer JL, Dambier D, Baurens FC, Grapin A, Lanaud C. 1998. Sequence tagged microsatellite site (STMS) markers in the Musaceae. *Mol Ecol* **7**(5): 659-663.

Le Provost G, Iskra-Caruana ML, Acina I, Teycheney PY. 2006. Improved detection of episomal Banana streak viruses by multiplex immunocapture PCR. *J Virol Methods* **137**(1): 7-13.

Lescot M, Piffanelli P, Ciampi AY, Ruiz M, Blanc G, Leebens-Mack J, da Silva FR, Santos CM, D'Hont A, Garsmeur O et al. 2008. Insights into the Musa genome: syntenic relationships to rice and between Musa species. *BMC Genomics* **9**: 58.

Lheureux F. 2002. Etude des mécanismes génétiques impliqués dans l'expression des séquences EPRVs pathogènes des Bananiers au cours de croisements génétiques interspécifiques. In *Ecole Nationale Supérieure Agronomique de Montpellier*, p. 102. Université Sciences et Techniques du Languedoc USTL, Montpellier.

Lheureux F, Carreel F, Jenny C, Lockhart BE, Iskra-Caruana ML. 2003. Identification of genetic markers linked to banana streak disease expression in inter-specific Musa hybrids. *Theor Appl Genet* **106**(4): 594-598.

Lheureux F, Laboureau N, Muller E, Lockhart BE, Iskra-Caruana ML. 2007. Molecular characterization of banana streak acuminata Vietnam virus isolated from Musa acuminata siamea (banana cultivar). *Arch Virol* **152**(7): 1409-1416.

Lockhart BE, Menke J, Dahal G, Olszewski NE. 2000. Characterization and genomic analysis of tobacco vein clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. *J Gen Virol* **81**(Pt 6): 1579-1585.

Lovisolo O, Hull R, Rosler O. 2003. Coevolution of viruses with hosts and vectors and possible paleontology. *Adv Virus Res* **62**: 325-379.

Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A* **101**(34): 12404-12410.

Mazourek M, Cirulli ET, Collier SM, Landry LG, Kang BC, Quirin EA, Bradeen JM, Moffett P, Jahn MM. 2009. The fractionated orthology of Bs2 and Rx/Gpa2 supports shared synteny of disease resistance in the Solanaceae. *Genetics* **182**(4): 1351-1364.

Menendez-Arias L. 2009. Mutation rates and intrinsic fidelity of retroviral reverse transcriptases. *Viruses* **1**(3): 1137-1165.

Mette MF, Kanno T, Aufsatz W, Jakowitsch J, van der Winden J, Matzke MA, Matzke AJ. 2002. Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. *EMBO J* **21**(3): 461-469.

Ndowora T, Dahal G, LaFleur D, Harper G, Hull R, Olszewski NE, Lockhart B. 1999. Evidence that badnavirus infection in Musa can originate from integrated pararetroviral sequences. *Virology* **255**(2): 214-220.

Noreen F, Akbergenov R, Hohn T, Richert-Poggeler KR. 2007. Distinct expression of endogenous Petunia vein clearing virus and the DNA transposon dTph1 in two Petunia hybrida lines is correlated with differences in histone modification and siRNA production. *Plant J* **50**(2): 219-229.

Perrier X, De Langhe E, Donohue M, Lentfer C, Vrydaghs L, Bakry F, Carreel F, Hippolyte I, Horry JP, Jenny C et al. 2011. Multidisciplinary perspectives on banana (Musa spp.) domestication. *Proc Natl Acad Sci U S A* **108**(28): 11311-11318.

Piffanelli P, Vilarinhos A, Safar J, Sabau X, Dolezel J. 2008. Construction of bacterial artificial chromosome (BAC) libaries of banana (Musa acuminata and Musa balbisiana). *Fruits* **63**: 375-379.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: protein domains identifier. *Nucleic Acids Research* **33**: W116-W120.

Richert-Poggeler KR, Noreen F, Schwarzacher T, Harper G, Hohn T. 2003. Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. *EMBO J* **22**(18): 4836-4845.

Safar J, Noa-Carrazana JC, Vrana J, Bartos J, Alkhimova O, Sabau X, Simkova H, Lheureux F, Caruana ML, Dolezel J et al. 2004. Creation of a BAC resource to study the structure and evolution of the banana (Musa balbisiana) genome. *Genome* **47**(6): 1182-1191.

Salamov AA, Solovyev VV. 2000. Ab initio gene finding in Drosophila genomic DNA. *Genome Research* **10**(4): 516-522.

Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. *J Virol* **84**(19): 9733-9748.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. *Nat Genet* **20**(1): 43-45.

Simmonds NW, ed. 1962. *The evolution of the bananas.*, London.

Staginnus C, Gregor W, Mette MF, Teo CH, Borroto-Fernandez EG, Machado ML, Matzke M, Schwarzacher T. 2007. Endogenous pararetroviral sequences in tomato (Solanum lycopersicum) and related species. *BMC Plant Biol* **7**: 24.

Staginnus C, Richert-Poggeler KR. 2006. Endogenous pararetroviruses: two-faced travelers in the plant genome. *Trends Plant Sci* **11**(10): 485-491.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**(8): 1596-1599.

Taylor DJ, Leach RW, Bruenn J. 2010. Filoviruses are ancient and integrated into mammalian genomes. *BMC Evol Biol* **10**: 193.

Vilarinhos AD, Piffanelli P, Lagoda P, Thibivilliers S, Sabau X, Carreel F, D'Hont A. 2003. Construction and characterization of a bacterial artificial chromosome library of banana (Musa acuminata Colla). *Theor Appl Genet* **106**(6): 1102-1106.

Vogt PK. 1997. Historical Introduction to the General Properties of Retroviruses.

Weiss RA. 2006. The discovery of endogenous retroviruses. *Retrovirology* **3**: 67.

SUPPLEMENTARY DATA

Supplemental data 1

Target	Type of markers	Primer name	Primer sequence (5′—3′)	Size of PCR product (bp)
eBSGFV	Integration locus	VM1F	TTGTCCAAAATCTGCTCGTG	481
		VM1R	TGTAATTCCTGCTCCTGCAA	
		VM2F	TTCTCCCTTTTCGATCCGTA	374
		VM2R	TTTTGATGCATCTCCAGCAG	
	Structure	VV1F	ACAGCTCCAGGAGATTGGAA	268
		VV1R	CTGAAGTGTGCCTGTGGAGA	
		VV2F	TCTGAGATCTCCAGCCAGGT	639
		VV2R	GACAGTTCCAGCACAGCAGA	
		VV3F	TTGCCAAGAATTCCTCCAAG	376
		VV3R	AAGTTCTTGTCGGCAAGGTG	
		VV4F	GAGCAACACGAGTCAACGAA	784
		VV4R	TCTCCACAGGCACACTTCAG	
		VV6F	GCATGAAGCATGACTGGAGA	264
		VV6R	AATGCATAAGGGCCTCGAAT	
	Allelic	VV5F	CCATGGAGGTTGACCTGTCT	628
		VV5R	ACCCCTCTGTCTTCCCAACT	
		DifGfF	TTGCAGGAGCAGGAATTACA	670
		DifGfR	GGATGGAAGATGAGCTCTTTG	
eBSOLV	Integration locus	Musa-Ol jonction1 F	TGCATTAGATGGTCTGGGAAA	563
		Musa-Ol jonction1 R	ACTTCACGATGCCCATGTTT	
		Musa-Ol jonction2 F	GAGCTGTTTCCTCCGTGTCT	590
		Musa-Ol jonction2 R	CCTGGAAGAAAGCAGACGAG	

	Structure	sig1 eBSOLV F	TTCGAGGAGTCAACGGAGTC	606
		sig1 eBSOLV R	CCTGGTCTGCACAGAGATGA	
		sig2 eBSOLV F	CTTGCTCTGTGGGCAAGACT	426
		sig2 eBSOLV R	CCATTTTTCTCGCAGATTGTC	
	Allelic	Marker1-BSOLV(2) F	ATACGAAGCCCAACGAATTG	601
		Marker1-BSOLV(2) R	ATGGCTTGCCTTCACAGATT	
		Marker2-BSOLV(2) F	ACTCGCACAAAGTGAACTCG	399
		Marker2-BSOLV(2) R	ACAGTACAAGCCCCACCAAT	
		Marker2-BSOLV(1) F	GTGGTGGTTCTTGATCCGGT	1469
		Marker2-BSOLV(1) R	CACGTGGTAGGGGTCCGCCA	
		Dif-OL(F) (HaeIII)	GAATCATTATTCGAGGAGTCAACGG	337
		Dif-OL(R) (HaeIII)	CGAGTAGAGCGCAAGATCCTAGTTC	
		Dif-OL (F) (AhdI)	TTGGAACAAGACAGATTGACTTCCT	500
		Dif-OL (R) (AhdI)	GGTTCGTTTTTATGGCTTTCATGG	
eBSImV	Integration locus	Musa/F2-F	ACTCAGCAAAGGCAAGCAGT	561
		Musa/F2-R	TCTGGTGTGAGTTTTAATAATACCG	
		F5/Musa-F	GTATGGTTCTTGCCCGATGA	594
		F5/Musa-R	TCGTGCAGACCCCTTACTCT	
	Structure	F1/F3-F	TTCGGTATTATTAAAACTCACACCA	490
		F1/F3-R	GCTGCTAACTGAGGATAATCGAA	
		F3/F4-F	TCCCACGCAAGCTTACTTCT	600
		F3/F4-R	GAAGCTGTCCAAGCCTATATCA	
		F4/F5-F	TGGACAGCTTCTGGTGTGAG	540
		F4/F5-R	AGCAGCTACAACCCTGGAGA	

For eBSGFV:

Primers DifGfF and DifGfR (annealing temperature *Ta*=60°C) amplify a PCR product of 670 bp. This PCR product was digested with the restriction enzyme *Taa*I (Fermentas) in a final volume of 20 µl according to the manufacturer's instructions. Digested DNA was loaded onto a 2.5% agarose gel stained with ethidium bromide, and the bands visualized under UV light. Digestion of the PCR product obtained from eBSGFV9 and eBSGFV7 yields two (442 bp + 227 bp) and three (366 bp + 227 bp + 76 bp) bands, respectively.

The second set of primers (VV5F/VV5R, *Ta*=60°C) hybridizes with eBSGFV9 only and yields a 628-bp amplification product.

For eBSOLV:

Primers Marker1-BSOLV(2) F/Marker1-BSOLV(2) R and Marker2-BSOLV(2) F/Marker2-BSOLV(2) R (Ta=65°C) hybridize with eBSOLV2 only and yield amplification products of 601-bp and a 399-bp, respectively.

The third set (Marker2-BSOLV(1) F/ Marker2-BSOLV(1) R (Ta=65°C) hybridizes with eBSOLV1 only and yields a 1469-bp amplification product.

Primers Dif-OL(F) (HaeIII) and Dif-OL(R) (HaeIII) (*Ta*=65°C) amplify a PCR product of 337 bp. This PCR product was digested with the restriction enzyme *Hae*III (New England Biolabs) in a final volume of 20 µl according to the manufacturer's instructions. Digested DNA was loaded onto a 2.5% agarose gel stained with ethidium bromide, and the bands visualized under UV light. The PCR product obtained from eBSOLV2 was not digested and that from eBSOLV1 was cut into two bands of 83 and 254 bp.

Primers Dif-OL (F) (AhdI) and Dif-OL (R) (AhdI) (*Ta*=65°C) amplify a PCR product of 500 bp. This PCR product was digested with the restriction enzyme *Ahd*I (New England Biolabs) in a final volume of 20 µl according to the manufacturer's instructions. Digested DNA was loaded onto a 1.5% agarose gel stained with ethidium bromide, and the bands visualized under UV light. The PCR product obtained from eBSOLV1 was not digested and that from eBSOLV2 was cut into two bands of 202 and 298 bp.

For eBSImV:

12 SSR markers defined from the BAC sequence (MBP_68C24) according to the procedure described in the paragraph "eBSV genotyping" were tested on parent *M. balbisiana* PKW in order to check polymorphism for subsequent segregation analysis.

Moreover, two dCaps markers were developed from the BAC sequence (MBP_68C24) and tested on the 23 other BAC clones that hybridized with BSImV probes during the BAC library screening in order to identify a second eBSImV allele. The markers were developed based on point mutations located at position 4,091 and 11,845 of eBSImV, which lead to stop codons. Both point mutations are present in the ORF3 fragment of eBSImV. Primers EPRV-Im 4091 For [5′-AGA AGA ATG AAT AGT CAA GAT TGG AAG ATT GTA CCA T-3′] and EPRV-Im 4091 Rev [5′-GCT TTG CCT TCC ATT TGC AAA-3′] amplify a 158 bp fragment whereas primers EPRV-Im11845 For 2 [5′-AGC CCC ACA TCA TCA AGA AG-3′] and EPRV-Im11845 Rev 2 [5′-ACC TGA GTT TTG ATG TTT TGT ACA ATC CA-3′] amplify a fragment of 217 bp. Both PCRs are performed as described in paragraph "PCR" in the Materials and Methods sections at a *Ta* of 60°C for 35 cycles and PCR product are digested with the restriction enzyme *Bcc*I (New England Biolabs) in a final volume of 20 µl according to the manufacturer's instructions. Digested DNA was loaded onto a 3% agarose gel stained with ethidium bromide, and the bands were visualized under UV light. Digestion of the PCR products yielded two bands (116 bp + 42 bp) and (183 bp + 34 bp) for EPRV-Im 4091 and EPRV-Im 11845, respectively.

Supplemental data 2

Estimates of evolutionary divergence between sequences of eBSVs found in PKW.

For each eBSV species, pairwise distances were computed separately for each virus component (ORF3, IG, ORF1 and ORF2). In the data matrix, the number of base differences per site between sequences is shown. For each virus component, we calculated the average divergence within the sequence of each allele, between the two alleles and finally between BSV and eBSV sequences (Excel spreadsheet). Overall divergences deduced from these data are listed in Table 5.

2- Article 2 : How endogenous Banana streak virus (eBSV) could enlighten BSV and banana evolution

How endogenous Banana streak virus (eBSV) polymorphism could enlighten BSV and Banana evolution

Pierre-Olivier Duroy[1], Xavier Perrier[2], Nathalie Laboureau[1], Jean-Pierre Jacquemoud-Collet[2] and Marie-line Iskra-Caruana[1]

1 CIRAD, UMR BGPI, F-34398 Montpellier Cedex 5.
2 CIRAD, UMR AGAP, F-34398 Montpellier Cedex 5.
Corresponding author: Marie-Line Iskra-Caruana, email: marie-line.caruana@cirad.fr

Abstract

The nuclear genome of banana plants harbours numerous copies of viral sequences derived from banana streak virus (BSV)—a DNA virus belonging to the family *Caulimoviridae.* These integrated viral sequences are mostly defective as a result of "pseudogenization" driven by evolution of the host genome. However, some integrations can release a functional infectious viral genome following activating stresses. We have previously characterized infectious endogenous BSV (eBSV) for three BSV species (BSOLV, BSGFV and BSImV) present within the *Musa balbisiana* B genome of the seedy diploid Pisang Klutuk Wulung (PKW). Our aim is to study PKW-related BSV integrations among the diversity of banana B genomes in order to retrace the evolutionary history of BSV and banana.
Here, we extend the diversity of sampling of BB seedy diploids of *M. balbisiana* by including interspecific hybrids with *M. acuminata* exhibiting different levels of ploidy for the B genome (ABB, AAB, AB) in order to include unsampled or extinct *M. balbisiana* resources. We also focused our analysis on two areas of sympatry between *M. acuminata* and *M. balbisiana* representing the centers of origin of the most widely cultivated AAB cultivars: one in India and the other in East Asia, ranging from the Philippines to New Guinea (Perrier et al., 2009).
We characterized PKW-related eBSV allelic polymorphism in 77 accessions using PCR markers (Chabannes et al., 2012) and Southern hybridisation. We coded the results of Southern and PCR analysis to create a common dissimilarity matrix with which to interpret eBSV distribution. As a result, three dendrograms of PKW-related eBSV on the 77 banana accessions were constructed for each BSV species using the neighbor joining (NJ) method, as well as one dendogram resulting from NJ analysis of all three BSV species together. We found that the known phylogeny of banana accessions based on *M. acuminata* genomes can help elucidate eBSV structural diversity, and that eBSV polymorphisms can shed light on the particularly unresolved question of *M. balbisiana* diversity. We propose for the first time a banana phylogeny driven by the *M. balbisiana* genome. A scheme of BSV/eBSV banana evolution is also presented.

Key words: *Musa sp.*, *Banana streak virus* (BSV), Phylogeny, endogenous pararetrovirus (EPRV).

Introduction

Modern sequencing technologies have led to the discovery that, in addition to retroviruses, which have an integration step during their life cycle and are a common component of eukaryotic genomes, many other virus sequences have integrated into the genome of their host. In the animal kingdom this is now well documented, with endogenous viral elements (EVE) having been reported in a multitude of animal genomes and possibly representing a large proportion of some of them (Feschotte and Gilbert 2012; Horie et al., 2011). Although no retroviruses have been found in the plant kingdom, plant genomes also contain integrated viruses. Most viruses found within plant genomes are members of the *Caulimoviridae* family. From the six genera of this family, so far five have been found to integrate into the plant genome (Hohn et al., 2008).

The *Caulimoviridae*—DNA viruses that replicate via reverse transcription and have no obligatory integration step into the host genome in their life cycle—are classed as pararetroviruses (PRV), and their integrants are known as endogenous pararetroviruses (EPRV) (Harper et al. 2002). The large-scale sequencing of plant genomes since the 1990s has revealed extensive invasion by EPRVs, which have been found in more than ten species of plants so far, e.g. pineapple (Gambley et al. 2008), rice (Kunii et al. 2004), tomato (Bejarano et al. 1996), tobacco (Jakowitch et al. 1999), potato (Staginnus et al. 2007), petunia (Richert-Pöggeler et al. 2003) and banana (Harper et al. 1999; Ndowora et al. 1999) and there will certainly be more. Some EPRVs contribute a significant proportion to their host genome, e.g. *Tobacco vein clearing virus* (TVCV) is present in the tobacco genome at more than 10^3 copies (Jakowitsch et al. 1999). Others, such as *Petunia vein clearing virus* (PVCV) (Richert-Pöggeler and Shepherd, 1997) and *Banana streak virus* (BSV) (Gayral et al. 2008) are present in fewer copy number. Although integration events are probably frequent during viral infections, few become fixed and maintained in the host genome, and this will occur only when their impact on plant fitness is not deleterious or negative (Hohn et al. 2008). In order to be fixed in the plant population, EPRV have to integrate into the germ cells and become positively selected in the plant population following hybridization events. EPRV are integrated preferentially into heterochromatin, mainly in pericentromeric regions (Richert-Pöggeler et al. 2003; Hansen et al. 2005; Staginnus et al. 2007), and they frequently colocalize with transposable elements (Chabannes et al. 2012; Richert-Pöggeler et al. 2003). Silencing mechanisms preferentially regulate expression of these seqences from genomic sites as demonstrated by Noreen et al. (2007), who showed that gene silencing down regulated endogenous PVCV (ePVCV).

In specific cases, EPRVs can be still infectious after integration, i.e. EPRV sequences can produce functional viral genome and viral particles that can infect the host plant. Such EPRVs are rare and only three cases have been reported to date in the plant kingdom:

eTVCV in tobacco (Lockhart et al. 2000), ePVCV in petunia (Richert-Pöggeler et al. 2003) and eBSV in banana (Ndowora et al. 1999). The latter pathosystem is the model used in this study.

BSV—a member of the family *Caulimoviridae*—is a circular double-stranded DNA virus that replicates by reverse transcription (Fauquet et al. 2005). Like almost all others members of the badnavirus genus, the BSV genome comprises 3 ORFs. BSV is a complex of different virus species that all induce the same disease in banana plants: banana streak mosaic disease (BSD) (Lockhart and Olszewski, 1993). To date, seven full-length BSV species have been described and sequenced completely (Harper and Hull, 1998; Geering et al. 2005, 2011; Lheureux et al. 2007; Gayral et al. 2008).

BSD is propagated slowly by mealybugs, and infections are circumscribed easily by digging out infected plants. Today, the major problem attributable to BSV is caused by endogenous BSV sequences (eBSV), which can release functional viral genomes producing episomal viruses and infection. These eBSV are named infectious eBSV and are so far reported to be restricted to *Musa balbisiana* genomes (denoted B) only (Gayral et al. 2008; Lheureux et al. 2003). Breeders usually used diploid *M. balbisiana* as progenitors in their breeding programs, and the genotypes of cooking bananas such as plantains—the main staple food of several African, Latin America and Asian countries—harbour a single B genome.

At the beginning of the 2000s, outbreaks of BSD were reported from different countries worldwide due to the massive use of newly created interspecific hybrids with single B genomes, revealing the capacity of eBSV to wake up. Genomic and abiotic stresses such as those experienced during interspecific crosses and micropropagation by in vitro culture play a major role in the production of viral particles (Harper et al. 1999; Ndowora et al. 1999 Lheureux et al. 2003; Côte et al. 2010; Dallot et al. 2000). Recently, eBSVs of three BSV species [Obino l'Ewai (BSOLV), Goldfinger (BSGFV) and Imove (BSImV)] present within the B genomes of the seedy *M. balbisiana* diploid Pisang klutu wulung (PKW) have been fully characterized (molecular structure, genomic organization, genomic landscape and infectious capacity) (Gayral et al. 2008, 2010; Chabannes et al. 2012). The results revealed that eBSV of each BSV species are present in an complex allelic insertion at a single locus resulting from a single integration event. There are two different alleles for eBSOLV (1 and 2) and eBSGFV (7 and 9), and the same allele twice in the case of eBSImV. The alleles eBSOLV-1 and eBSGFV-7 are able to produce viral particles and are thus termed infectious.

Examining the level of sequence divergence between eBSV and its counterpart BSV, combined with the differential rearrangements of eBSV alleles, Chabannes et al. (2012) assumed sequential BSV integrations, placing eBSOLV as older than both eBSGFV and eBSImV. The integration time of eBSGFV can be estimated at circa 640,000 years ago due to the integration locus in an LTR retrotransposon (Gayral et al. 2008). Molecular markers specific to the eBSGFV and eBSImV alleles were developed initally and used to screen

PKW-related eBSV distribution in *M. balbisiana* species (Gayral et al. 2010). Preliminary results have shown that integration is limited strictly to the B genome. eBGFV is present in all BB tested while eBSImV is absent or mutated in some varieties. These results favor an integration event shortly after the speciation of *M. balbisiana* in the *Musa* genus and before species diversification.

Gayral et al. (2010) established a microsatellite-based phylogeny of *M. balbisiana* diploids (BB), revealing few polymorphisms. between the six identified clusters. However, the absence of geographical data and the limited number of *M. balbisiana* accessions (only 20) currently preclude a complete description of eBSV evolution on *Musa* sp.

In order to delineate the early history and evolutionary fate of eBSV in relation to the banana (*Musa* sp.) host plant, in this manuscript we describe the distribution not only of eBSGFV and eBSImV but also of eBSOLV, on a larger sample better representing *Musa* species diversity, in particular many wild or cultivated *M. acuminata* accessions and interspecific hybrids (Hippolyte et al., 2012). In particular, we extend *M. balbisiana* sampling diversity with the addition of interspecific hybrids with *M. acuminata* showing different levels of ploidy for the B genome (ABB, AAB, AB) in order to include unsampled or extinct *M. balbisiana* resources. We also based our analysis on two areas of sympatry between *M. acuminata* and *M. balbisiana* that are the centers of origin for the most widely cultivated AAB cultivars. One is in India and the other in East Asia, stretching from the Philippines to New Guinea (Perrier et al. 2009). However, the process of *Musa* hybrid creation is not still well understood, and different hypothesis exist. Carrel et al. (2004) used cytoplasmic markers to show that several AAB and ABB derive from a preliminary AB hybrid combining later to A or B gametes, whereas Perrier et al. (2009) showed that AA parthenocarpic cultivars are still able to produce gametes, although these are often non-reduced 2N gametes. These AA gametes are confirmed not only to be at the origin of several AAA cultivars but also to be associated with the B gametes of AAB hybrids such as the Indian cv Pome accession. The mechanism is certainly more complex since, based on FISH analysis, D'Hont et al. (2000) showed that the genotype of the Pelipita (ABB) accession contained 8 A chromosomes and 25 B chromosomes instead of the expected 11 and 22 chromosomes, respectively. Unbalanced chromosome numbers are also suspected for AAB hybrids such as the Pisang Nangka or Kunaimp accessions (unpublished results), suggesting this is probably not a rare event.

To complement the preliminary studies of Gayral et al. (2010), we are developing additional markers and techniques to further characterize eBSV allelic polymorphism. We show here that the known phylogeny of banana accessions can help elucidate eBSV structural diversity and that, by reversal of roles, eBSV polymorphisms can help us understand the particularly unresolved question of *balbisiana* diversity. Finally, we test whether eBSV markers can be adapted to describe *M. balbisiana* phylogeny.

Section	Species/Hybrids	Subspecies[a]/ phylogenetic group[b]	Genome[c]	Ploidy[c]	Accession name	Abbreviation	Origine[c]	Accession number[d]	Heterozygosity[e]
Australimusa	*M. textilis* Née		TT	2	Textilis 1072	OG Tex	ND	NEU0001	
Rhodochlamys	*M. laterita*			2	Laterita 0627	OG Lat	ND	ITC0627	
	M. ornata			2	Ornata	OG Orn	ND	ITC0370	
	M. mannii			2	Mannii H. Wendl	OG Man	ND	NEU0011	
	M. velutina			2	Velutina	OG Vel	ND	NEU0006	
	Rhod. ind.			2	Not Named Mainz	OG NNM	ND	ITC0241	
Eumusa	*M. basjoo* linuma			2	Basjoo	OG Baj	ND	NEU0060	
	M. acuminata	AAw banksii	AAw	2	Banksii 0620	AA Bank	ND	ITC0620	
		AAw errans	AAw	2	Agutay	AA Agu	Philippines	NEU0033	
		AAw burmannica	AAw	2	Long Tavoy	AA LoTA	Thailand	NEU0016	0.588
		AAw zebrina	AAw	2	Maia Oa	AA Maia	Martinique	NEU0029	0.294
		AAw malaccensis	AAw	2	Pahang	AA Pah	Southeast Asia	NEU0013	0.471
		Desert banana	AAcv	2	IDN 110	AA IDN	Indonesia	NEU0137	0.706
	Hybrids	Pisang Awak	ABB	3	Fougamou 1	ABB Foug	Gabon	ITC0101	
		Pisang Awak	ABB	3	Kluai Namwa Khom	ABB Khom	ND	ITC0526	
		Pelipita	ABB	3	Pelipita	ABB Peli	Philippines	NEU0360	
		Bluggoe	ABB	3	Burro Cemsa	ABB Burro	South America	NEU0339	
		Bluggoe	ABB	3	Dole	ABB Dole	ND	NEU0334	
		Ney Mannan	ABB	3	Blue Java	ABB Blue	Fiji	ITC0361	
		ABB ind.	ABB	3	Pisang Kepok Bung	ABB Bung	Java	NEU0359	
		ABB ind.	ABB	3	Daru	ABB Daru	PNG	ITC0795	
		ABB ind.	ABB	3	Bengani	ABB Beng	PNG	ITC0855	
		ABB ind.	ABB	3	Auko	ABB Auko	PNG	NEU0365	
		Saba	ABB	3	Saba	ABB Saba	Philippines	NEU0361	
		Silk	AAB	3	Figue Pomme Géante	AAB Figue	ND	NEU0285	
		Silk	AAB	3	Tay Tia	AAB Tay	Vietnam	ITC1365	
		Pome	AAB	3	Foconah	AAB Foco	Cameroun	NEU0298	
		Pome	AAB	3	Prata ana	AAB Prata	ND	NEU0310	
		Nadan	AAB	3	Lady Finger (Nelson)	AAB Lady	ND	NEU0297	
		AAB ind.	AAB	3	Yangambi n°2	AAB Yang2	Zaire	ITC1275	
		Maia Maoli/Popoulu	AAB	3	Mai'a Popo'ulu Moa	AAB Moa	ND	ITC1169	
		Maia Maoli/Popoulu	AAB	3	Popoulou	AAB Pop	ND	NEU0279	
		Iholena II	AAB	3	Tigua	AAB Tig	PNG	ITC0875	
		AAB ind.	AAB	3	Boung Fu	AAB Bong	PNG	ITC0940	
		AAB ind.	AAB	3	Gamaha	AAB Gama	PNG	ITC1006	
		Plantain	AAB	3	Corne 1	AAB Corn	ND	ITC0754	
		Plantain	AAB	3	Kelong Mekintu	AAB KM	Cameroon	NEU0251	
		Plantain	AAB	3	Orishele	AAB Ori	ND	NEU0256	
		Iholena I	AAB	3	Luba	AAB Luba	PNG	NEU0448	
		Laknao	AAB	3	Pisang Kapas	AAB Kap	Java	NEU0316	
		Pisang Raja	AAB	3	Pisang Raja Bulu	AAB PRB	Java	NEU0276	
		Nendra Padaththi	AAB	3	Pisang Radjah	AAB PR	ND	NEU0282	
		AAB ind.	AAB	3	Kunaimp	AAB Kuna	PNG	ITC0836	
		AAB ind.	AAB	3	Chuoi Mat	AAB ChuM	Vietnam	ITC1381	
		AAB ind.	AAB	3	Pisang Nangka	AAB Nang	Java	ITC1062	
		Mysore	AAB	3	Pisang Ceylan	AAB Ceyl	ND	NEU0284	
		AAB ind.	AAB/AABB	3	Pisang Slendang	AAB Slen	Indonésie	NEU0382	
		AAB ind.	AAB	3	Muracho	AAB Mur	Philippines	NEU0327	
		AAB ind.	AAB / AAAB	3	Porp	AAB Porp	PNG	NEU0376	
		AAB ind.	AAB	3	Dimaemamosi	AAB Dima	PNG	ITC0920	
		Ney Poovan	AB	2	Safet Velchi	AB Safet	India	NEU0152	0.647
		Ney Poovan	AB	2	Kunnan	AB Kunn	India	NEU0155	0.706
		ABcv	AB	2	Figue Pomme(Ekona)	AB Eko	Cameroon	NEU0153	0.647
	M. balbisiana	Msat-1	BBw	2	Honduras	BB Hond	ND	NEU0049	0.176
		Msat-1	BBw	2	Balbi HDN 211	BB 211	ND	ITC0211	0.176
		Msat-1	BBw	2	Balbi 1016	BB 1016	PNG	ITC1016	0.000
		Msat-1	BBw	2	Balbi 0626	BB 626	PNG	ITC0626	0.176
		Msat-1	BBw	2	Balbi I63-080	BB 63-80	ND	ITC0080	0.118
		Msat-1	BBW	2	Balbi I 63	BB I63	ND	ONN0154	0.118
		Msat-2	BBW	2	Montpellier	BB Mont	ND	ONN0152	0.588
		Msat-2	BBw	2	Cameroun	BB Cam	ND	NEU0050	0.294
		Msat-3	BBw	2	Balbi 0545	BB 545	ND	ITC0545	0.412
		Msat-3	BBw	2	Balbi 10852	BB 852	ND	ITC0094	0.529
		Msat-3	BBw	2	Lal Velchi	BB Lal	India	NEU0051	0.412
		Msat-4	BBw	2	Klue Tani	BB KT	Thailand	ITC1120	0.294
		Msat-4	BBW	2	Pisang Batu	BB Batu	ND	NEU0055	0.412
		Msat-4	BBW	2	Pisang Klutuk	BB PK	ND	NEU0056	0.353
		Msat-4	BBW	2	Pisang Klutuk Wulung	BB PKW	ND	NEU0054	0.294
		Msat-5	BBw	2	Butuhan	BB But	ND	NEU0057	0.294
		Msat-5	BBw	2	Eti Kehel	BB EK	Sri lanka	ITC0271	0.353
		Msat-6	BBw	2	Balbi LBA-342	BB 342	ND	ITC0342	0.353
		Msat-6	BBW	2	Los Banos	BB LBA	ND	ONN0151	0.412
		BBw	BBw	2	Butuhan inter apex	BB ButIA	ND	ITC0565	0.294
		Msat-7	BBw	2	Chinois 1	BB Chi1	China	ND	0.176
		Msat-7	BBw	2	Chinois 2	BB Chi2	China	ND	0.118
		Msat-7	BBw	2	Chinois 3	BB Chi3	China	ND	0.353
		Msat-7	BBBcv	3	Lep Chang Kut	BB LCK	ND	ITC0647	

Table 1: *Musa* accessions used in this study

[a] subspecies for *M. acuminata* are identified according to Perrier et al (2009)

[b] phylogenetic groups for hybrids from Hippolyte et al., (2012) and for *M. balbisiana* from Carrel et al., (2002)

[C] Information from MGIS database, http://www.crop-diversity.org/banana,, according to the current agro-morphological classification (IPGRI-INIBAP(Bioversity), 2003) and ploidy level determined by flow cytometry (Dolezel et al, 1997)

[d] Collections : ITC, International Transit center ; NEU, CIRAD-Neufchateau

[e] Heterozygosities were calculated for diploids samples from microsatellite data

Materials and methods

Plant material and DNA extraction

Sampling of *Musa* diversity for eBSV characterization was based initially on genotyping with 22 SSR nuclear markers in a population of more than 500 accessions (Hippolyte et al. 2012). To increase the diversity of B genomes, this sampling was complemented with 23 new accessions including several recently collected *M. balbisiana* and interspecific hybrids AAB and ABB of interest, which were absent in the previous analysis. A sample of 77 accessions was thus defined (24 BB, 11 ABB, 26 AAB, 3 AB, 6 AA, 7 Out Groups-OG) (table 1) representative of (i) the two species at the origin of all cultivated bananas: *M. acuminata* and *M. balbisiana*, (ii) diploid and triploid hybrids of these two species, (iii) some other *Musa* species as outgroups. This sample was characterized for both its nuclear genome and its eBSV insertions. Fresh leaf samples were kindly supplied by the in vivo germplasm collections of CIRAD in Guadeloupe and the International Institute of Tropical Agriculture (IITA) in Nigeria; the INIBAP Transit Center (ITC) in Leuven (Belgium) supplied plantlets from in vitro culture. Each genotype was documented with its genome constitution and subgroup classification according to the current agro-morphological classification (IPGRI-INIBAP (Bioversity), 2003), and ploidy levels were estimated by flow cytometry (Dolezel et al. 1997).

Total genomic DNA was extracted from banana leaf tissue by the method of Gawel and Jarret (1991). The quality of DNA was assessed visually under UV light after migration of 5µl of DNA sample in a 1% agarose gel in 0.5X TBE (45mM Tris-borate, 1mM EDTA [pH 8]), stained with ethidium bromide, and by PCR amplification of the housekeeping *Musa actin* gene (see below).

Microsatellite genotyping

Among the 22 SSRs of the previous analysis (developed from *M. acuminata* cv. 'Gobusik' and *M. balbisiana* cv. PKW), 17 were selected for the present study (table 2). Microsatellite analysis followed the protocol developed by Hippolyte et al. (2012). The 17 SSRs were shown to be independent and to be distributed among 10 of the 11 linkage groups (Hippolyte et al., 2010). For all SSR loci, the forward primer was designed with a 5′ -end M13extension (5′-CACGACGTTGTAAAACGAC-3′). This extension enabled the generation of fluorescent amplicons following fluorescent dye hybridization. PCR was performed in 10 µl of a mixture containing 25ng of DNA, 0.40µM reverse primer, 0.40µM M13-tailed forward primer, 1U of GoTaq DNA Polymerase (Promega™, Madison, WI), 0.1mM concentrations of each deoxynucleoside triphosphate (dNTP), 1.5mM $MgCl_2$, 20mM Tris-HCl (pH 8.4), 50mM KCl,

Marker Name	Synonym	Motif	Primer Sequence	Annealing Tm	Min allele (+M13 tail)	Max allele (+M13 tail)	Accession GenBank	References
mMaCIR01F	AGMI24	GA (20)	CACGACGTTGTAAAACGACTTAAAGGTGGGTTAGCATTAGG	55	238	314	X87262	Lagoda et al., 1998
mMaCIR01R	AGMI25		TTTGATGTCACAATGGTGTTCC					
mMaCIR03F	AGMI35		CACGACGTTGTAAAACGACTGACCCACGAGAAAAGAAGC	55	110	138	X87263	Lagoda et al., 1998
mMaCIR03R	AGMI36	(GA)10	CTCCTCCATAGCCTGACTGC					
mMaCIR07F	AGMI93		CACGACGTTGTAAAACGACAACAACTAGGATGGTAATGTGTGGAA	53	146	184	X87258	Lagoda et al., 1998
mMaCIR07R	AGMI94	(GA)13	GATCTGAGGATGGTTCTGTTGGAGTG					
mMaCIR08F	AGMI95		CACGACGTTGTAAAACGACACTTATTCCCCCGCACTCAA	55	252	298	X87264	Lagoda et al., 1998
mMaCIR08FR	AGMI96	(TC)6N24(TC)7	ACTCTCGCCCATCTTCATCC					
mMaCIR13F	AGMI105		CACGACGTTGTAAAACGACTCCCAACCCCTGCAACCACT	53	270	298	X90745	Lagoda et al., 1998
mMaCIR13R	AGMI108	(GA)16N76(GA)8	ATGACCTGTCGAACATCCTTT					
mMaCIR24F	AGMI131		CACGACGTTGTAAAACGACACTCTTTCTTATCCTTCTAACG	48	237	297	Z85972	Lagoda et al., 1998
mMaCIR24R	AGMI132	(TC)7	ATTAGATCACCGAAGAACTC					
mMaCIR39F	AGMI189		CACGACGTTGTAAAACGACAACACCGTACAGGGAGTCAC	52	329	369	Z85970	Lagoda et al., 1998
mMaCIR39R	AGMI190	(CA)5GATA(GA)5	GATACATAAGGCAGTCACATTG					
mMaCIR40F	AGMI191		CACGACGTTGTAAAACGACGGCAGCAACACATACTACGAC	54	168	206	Z85977	Lagoda et al., 1998
mMaCIR40R	AGMI192	(GA)13	CATCTTCACCCCCATTCTTTTA					
mMaCIR45 F	AGMI203		CACGACGTTGTAAAACGACTGCTGCCTTCATCGCTACTA	57	272	294	Z85968	Lagoda et al., 1998
mMaCIR45R	AGMI204	(TA)4CA(CTCGA)4	ACCGCACCTCCACCTCCTG					
mMaCIR150F			CACGACGTTGTAAAACGACATGCTGTCATTGCCTTGT	54	257	270	AM950440	Hippolyte et al., 2010
mMaCIR150R		(CA) X10	GAATGCTGATACCTCTTTGG					
mMaCIR152F			CACGACGTTGTAAAACGACCCACTTTGATTCTCTCC	54	158	194	AM950442	Hippolyte et al., 2010
mMaCIR152R		(CTT) X18	TTTCCCTCTTCGATTCTGT					
mMaCIR164F			CACGACGTTGTAAAACGACAAGACAAGTTCCATTGCTTG	55	255	409	AM950454	Hippolyte et al., 2010
mMaCIR164R		(AC) X14	GTTCGGGCTTTCGGT					
mMaCIR231F			CACGACGTTGTAAAACGACGCAAATATGTCAAGGAATCA	55	238	286	AM950497	Hippolyte et al., 2010
mMaCIR231R		(TC) X10	ACCCAGGTCTATCAGGTCA					
mMaCIR260F			CACGACGTTGTAAAACGACGGATGTTTGGGCTGTTTCTT	55	194	230	AM950515	Hippolyte et al., 2010
mMaCIR260R		(TG) X8	AAGCAGGTCATGATTGTTCC					
mMaCIR264F			CACGACGTTGTAAAACGACAGGAGTGGGAGCCTATTT	53	234	274	AM950519	Hippolyte et al., 2010
mMaCIR264R		(CT) X17	CTCCTCGGTCAGTCCTC					
mMaCIR307F			CACGACGTTGTAAAACGACAGACTTGTATCGCTTGGTAAA	54	160	172	AM950533	Hippolyte et al., 2010
mMaCIR307R		(CA) X6	ACGCTGCACCAGTCAA					
Ma1-32F		(GA)17AA(GA)8AA (GA)2	CACGACGTTGTAAAACGACGGAACAGGTGATCAAAGTGTGA	60	227	270	NA	Kaemmerer et al., 1997
Ma1-32R			TTGATCATGTGCCGCTACTG					

Table 2: Microsatellite loci used in this study

From Lagoda et al., 1998 and Hippolyte et al., 2010

5μl and 0.40μM M13 primer fluorescently labeled with FAM™ (Blue), NED™ (Green), PET™ (Red) or VIC™ (Yellow) (Applied Biosystems™). An initial denaturing step of 2 min at 94°C was followed by 10 touchdown cycles with a rate of –1°C for each cycle, with a first step at 94°C for 30 s, 55°C for 30 s, and 72°C for 1 min, and then 25 cycles of 94°C for 60 s, 55°C for 30 s, and 72°C for 1 min 10 min at 72°C. PCR products were diluted 15 times before measurement. The sizes of the amplified fragments were measured using a capillary sequencer 3500xL Genetic Analyzer (Applied biosystems™). Alleles were scored by using GeneMapper™ v4.1 software (Applied Biosystems™). Sizes were standardised by the incorporation of 0.1μl of GeneScan 600 LIZ Size Standard v2.0 (Applied Biosystems) in the diluted PCR product before analyzing. Alleles included in final consensus genotypes were observed at least twice. Two samples with known genotypes served as positive controls and were included in each run of 77 PCRs to standardize genotyping across experiments.

Diversity tree of banana accessions from microsatellite markers

Whatever the method of construction, the accuracy of a diversity tree relies on the representativeness of the sample analyzed. In order to increase this representativeness, the results for the 17 SSR markers on the 77 sampled accessions were concatenated with the results of the analysis developed by Hippolyte et al. (2012). The band level notations were adjusted on the subset of 54 accessions common to both analyses. The resulting data matrix on 567 accessions was used to calculate dissimilarities between pairs of accessions. The dissimilarity estimated from codominant SSR markers can be estimated from the proportion of shared alleles, which has proven effective in reconstructing correct genealogical relationships. However, this measure cannot be applied directly to our data, which mixes diploids and triploids. So, an extended index as defined in Hippolyte et al. (2012) was used to apply to two diploids and two triploids as well as a diploid and a triploid.

A diversity tree was built from the dissimilarity matrix on 567 accessions, using the neighbour-joining (NJ) algorithm (Saitou and Nei, 1987) implemented in DARwin v5.0.155 software (X. Perrier and J. P. Jacquemoud-Collet [http://darwin.cirad .fr/darwin], 2006). The tree was too large to be shown here, thus a subtree of our 77 accessions of interest was extracted from the whole tree.

We used PowerMarker software to calculate the hetrozygosity of diploids (Liu and Muse, 2005).

eBSV diversity, screening by PCR amplification

PCR screening conducted made according to the protocol developed by Gayral et al. (2010) and Chabannes et al. (2012).

eBSV amplified		Name	Sequence	Annealing temperature	Amplification size (bp)
eBSGFV	Jonction Markers	VM1F	TTGTCCAAAATCTGCTCGTG	60	481
		VM1R	TGTAATTCCTGCTCCTGCAA	60	
		VM2F	TTCTCCCTTTTCGATCCGTA	60	374
		VM2R	TTTTGATGCATCTCCAGCAG	60	
		VM2bis-F	GAGGCCCTTATGCATTGTTG	60	159
		VM2bis-R	TCGACCGTACCGATATCCTC	60	
	Internal Markers	VV1F	ACAGCTCCAGGAGATTGGAA	60	268
		VV1R	CTGAAGTGTGCCTGTGGAGA	60	
		VV2F	TCTGAGATCTCCAGCCAGGT	60	639
		VV2R	GACAGTTCCAGCACAGCAGA	60	
		VV2bis-F	GCTGGCAGTGGAATTCAGTT	60	395
		VV2bis-R	CATGGTGGGAGAAGAGGAAG	60	
		VV3F	TTGCCAAGAATTCCTCCAAG	60	376
		VV3R	AAGTTCTTGTCGGCAAGGTG	60	
		VV4F	GAGCAACACGAGTCAACGAA	60	784
		VV4R	TCTCCACAGGCACACTTCAG	60	
		VV4bis-F	GGAAAACTCTGGGTTGGTGA	60	766
		VV4bis-R	GGAGTACGGCATTTTCTCCA	60	
		VV6F	GCATGAAGCATGACTGGAGA	60	264
		VV6R	AATGCATAAGGGCCTCGAAT	60	
		VV6bis-F	AGGCCACTACGCATCAGAAT	60	712
		VV6bis-R	GGCCTCGAATTTATCATTGG	60	
	Allelic Markers	VV5F	CCATGGAGGTTGACCTGTCT	60	628
		VV5R	ACCCCTCTGTCTTCCCAACT	60	
		VV5bis-F	CGCACCTTCATCACAGAAGA	60	588
		VV5bis-R	TACCAGATGGGGAGAAATCG	60	
		DIF GF F	TTGCAGGAGCAGGAATTACA	60	
		DIF GF R	GGATGGAAGATGAGCTCTTTG	60	
eBSImV	Jonction Markers	Musa F2 F	ACTCAGCAAAGGCAAGCAGT	60	561
		Musa F2 R	TCTGGTGTGAGTTTTAATAAACCG	60	
		Musa/F2bis-F	AGCTGAAGTGATGCGAACCT	60	937
		Musa/F2-R	-	60	
		F5 Musa F	GTATGGTTCTTGCCCGATGA	60	594
		F5 Musa R	TCGTGCAGACCCCTTACTCT	60	
		F5/Musabis-F	CCACCTGGTATCCCTGAAGA	60	905
		F5/Musabis-R	TGTCAAGCTGTTGGTTGCTC	60	
	Internal Markers	F1 F3 F	TTCGGTATTATTAAAACTCACACCA	60	490
		F1F3 R	GCTGCTAACTGAGGATAATCGAA	60	
		F1/F3-F	-	60	630
		F1/F3bis-R	TTCTTGGGGTACTGGTTTCG	60	
		F3 F4 F	TCCCACGCAAGCTTACTTCT	60	600
		F3 F4 R	GAAGCTGTCCAAGCCTATATCA	60	
		F3/F4bis-F	GGTGCAAATCAGAGTCATGC	60	987
		F3/F4-R	-	60	
		F4 F5 F	TGGACAGCTTCTGGTGTGAG	60	540
		F4 F5 R	AGCAGCTACAACCCTGGAGA	60	
		F4/F5-F	-	60	927
		F4/F5bis-R	AGCATCCGCTTTGGAGACTA	60	
eBSOLV	Jonction Markers	Musa-OI jonction1 F	TGCATTAGATGGTCTGGGAAA	45	564
		Musa-OI jonction1 R	ACTTCACGATGCCCATGTTT	58	
		Musa-OI jonction2 F	GAGCTGTTTCCTCCGTGTCT	62	590
		Musa-OI jonction2 R	CCTGGAAGAAAGCAGACGAG	62	
	Internal Markers	Marker2-BSOLV(2) F	ACTCGCACAAAGTGAACTCG	60	399
		Marker2-BSOLV(2) R	ACAGTACAAGCCCCACCAAT	60	
	Allelic Markers	sig1 eBSOLV F	TTCGAGGAGTCAACGGAGTC	62	606
		sig1 eBSOLV R	CCTGGTCTGCACAGAGATGA	62	
		sig2 eBSOLV F	CTTGCTCTGTGGGCAAGACT	62	426
		sig2 eBSOLV R	CCATTTTTCTCGCAGATTGTC	45	
		Dif-OL (F) (AhdI)	TTGGAACAAGACAGATTGACTTCCT	60	499
		Dif-OL (R) (AhdI)	GGTTCGTTTTTATGGCTTTCATGG	60	

Table 3: PCR markers used to genotype PKW eBSOLV-eBSGFV-eBSImV

From Chabannes et al, 2012 and Gayral et al, 2010

All PCRs were performed with 25–100ng of DNA, 5µl of 5X Green GoTaq® Reaction Buffer (Promega™, Madison, WI), 100mM of each dNTP, 10pmol of each primer, and 0.1U of GoTaq® DNA polymerase (Promega™, Madison, WI) in a total reaction volume of 25µl. PCR conditions were as follows: 1 cycle at 94°C for 5 min, followed by 35 cycles at 94°C for 30 s, 60 or 65°C for 30 s, and 72°C for 15s to 1min30s, and then one elongation cycle at 72°C for 10 min. Time of elongation and temperature of hybridization are given in table 3 for each primer pair. PCR products were visualized under UV light after migration of 10µl of PCR products on a 1% agarose gel in 0.5 X TBE (45 mM Tris-borate, 1mM EDTA, [pH 8]) stained with ethidium bromide.

For the derived cleaved amplified polymorphism sequences (dCAPs) markers, the DifGf-TaaI and DifOl-AhdI methods were used. The PCRs were performed with primer pairs DifGf F/DifGf R and DifOl F/DifOl R (figure 2), using the same conditions as for the PCR screen of eBSV previously described. PCR products (7µl; 0.2 to 1.5µg of DNA) were digested with 2U of *Taa*I (Fermentas) in 1X Tango buffer (Fermentas) for the DifGf product, and with 2U of *Ahd*I (NEB) in 1X NEBuffer 4 (NEB) for DifOl in a final volume of 10µl. Incubations were performed at 65°C for 2 h for DifGf and at 37°C for 2h for DifOl. Digested DNA was loaded onto a 2.5% agarose gel; after staining with ethidium bromide; the bands were visualized under UV light.

For amplification of the housekeeping actin gene, the following primers and conditions were used: Actine1F (5'-TCCTTTCGCTCTATGCCAGT-3'), Actine1R (5'-GCCCATCGGGAAGTTC ATAG- 3'), and a Tm of 58°C for 25 cycles with 1.5 mM $MgCl_2$.

eBSV diversity, screening with Southern blot analysis

The Southern blot method used here was developed specifically for banana DNA analysis of eBSV. The genomes of BSOLV, BSGFV and BSImV cloned in plasmids were used as probes for hybridization on restricted genomic DNA of the different banana accessions. Restriction enzymes were chosen based on the DNA sequence of eBSGFV, eBSOLV and eBSImV alleles present in PKW. From the restriction map of these sequences, enzymes were chosen that would cleave the eBSV into at least five fragments of different sizes (between 1000bp and 8000bp to facilitate observation on gels) to allow them to be discriminated from each other and to permit differentiation of each allele. Several polymorphic probe/enzyme combinations were identified.

Total genomic DNA of the samples (40 µg) and BAC DNA from the BAC bank of PKW carrying eBSV (1.5µg) were digested separately with 1 U/µg of DNA for each enzyme, *Hin*dIII and *Bam*HI for BSOLV-specific Southern blot and *Spe*I and *Dra*I for BSGFV- and BSImV-specific Southern blot. Proteins and salts were removed by precipitation with isopropanol and washing with 70% ethanol. Total genomic DNA (around 20µg) and digested BAC DNA

(around 800ng) were separated by electrophoresis on 1% agarose gels run overnight at 40 V in 0.5 X TBE (45 mM Tris-borate, 1 mM EDTA, [pH 8]), together with a 1 kb marker DNA ladder (Invitrogen®, Carlsbad, CA, USA). According to the Hybond N+ membrane Southern blot protocol (Amersham Biosciences®, Piscataway, NJ, USA), capillary transfer in 20X SSC transfer solution was realized overnight after rinsing the gel in three baths (DNA depurination, DNA denaturation and DNA neutralization). The nucleic acids were then fixed onto the membrane using a UV crosslinker (70,000µJ/cm²). The membranes were then placed in 20ml pre-hybridization buffer (50mM Tris HCl pH 8, 25mM ETDA, 5 SSC, 1% SDS, Denhardt's solution 2.5X and 2mg of denatured salmon sperm DNA) and incubated for 3h at 65°C in a rotisserie oven. Plasmids containing episomal forms of BSV served as probes for this experiment. These probes were prepared from plasmids harbouring the DNA of BSOLV (7389bp), BSGFV (7263bp) and BSImV (7768bp); 50 ng of BSV probes were labeled with 2 µl of α-^{32}P dCTP with a random priming protocol (Prime-a-Gene kit, Promega®). The labeled probes (50ng) were added to 20 ml of hybridization solution (50mM Tris HCl pH 8, 25mM ETDA, 5 SSC, 1% SDS, Denhardt's 2.5X, 2mg of denaturated salmon sperm DNA and 5% Dextran sulfate) and incubated overnight at 65°C. In order to remove non-specific background following hybridization, membranes were washed at 65°C for 10 min, twice in wash solution 1 (1 SSC and 0.1%SDS), once in wash solution 2 (0.5X SSC and 0.1X SDS). Membranes were air dried and observed after both overnight exposures on a filmless autoradiography Storm 820 imaging system (Amersham Biosciences®, Piscataway, NJ, USA). Scorable fragment length polymorphisms were examined for each BSV probe/restriction enzyme combination.

PCR and Southern blot-based data analysis

We conducted a PCR and dCAPs analysis of all three BSVs studied. The data were scored according to presence or absence of amplification fragments for each accession. We chose to score as present only amplifications showing the PKW expected size. We detected amplification with different sizes especially for eBSImV and eBGFV, consequently we developed “bis” primers, which amplified the same fragments with a shift of 5 to 10 nt in order to validate or refute the presence of these fragments. This analysis was evaluated according to the presence or absence of the amplification fragment for each PCR primer pair and for each accession.

Hybridized membrane pictures from Southern blot analysis (figure 3B, 4B, 5B) were analyzed using the software ImageQuant TL (GE Healthcare ®). This software permits automatic fragment detection from the membrane image and caculates their size by referring to the ladder and positive controls present on each membrane. A visual control of hybridization was made in order to validate data obtained through the software outputs. We developed two

types of DNA digestion for all samples as described above. Each sample was recorded according to the presence or absence of fragments obtained following hybridization with the three BSV probes. We chose to score any clearly detectable additional bands adding when necessary new fragment sizes to those in PKW reference.. Such cases were rare and a new fragment was added only for eBSOLV. In the case of fragments corresponding to non-complete DNA digestion were discared, such fragments were not included together with the real one in the final analysis. Incomplete digestion can be determinating if the two fragments corresponding to complete digestion are also present in the sample. When band intensities were to low or too difficult for an accurate size determination, the data were considered as missing. This strategy increases the number of missing data, but reinforced the robustness of the results. For each banana accession, we obtained presence/absence data on the fragment composition for each eBSV present in their genome (figure 3B, 4B, 5B).

Southern blot and PCR results were obtained separately but, since they concern the same subject material, they can be analyzed jointly, and a specific method was developed for this purpose.

We divided each eBSV into several fragments based on those already defined during Southern blot analysis (figures 3, 4, 5) and, where possible, we coded analysis from PCR and Southern blot together for each fragment. In others cases, PCR or Southern blot data were coded separately. The dissimilarity between two accessions was calculated as the proportion of cases where the two accessions were not in agreement (presence/absence). However, it was considered that Southern blot data were more informative than PCR data regarding our questions on conservation of eBSV structure. So, when estimating the dissimilarity between two accessions, disagreement in Southern blot results were considered as full differences, contributing a value of 1 to the dissimilarity, while disagreements in PCR data contributed a weight lower than 1. Several values were tested and a weight of 0.2 was retained because it best resolved some typically spurious associations between accessions. For dissimilarity calculations, Southern blot fragments were coded 1 for presence and 0 for absence for convenience, and PCR or dCAPs amplifications were also coded 1 for presence but 2 for absence. This scoring method for 5 accessions is presented in figures 3, 4 and 5 and data are shown in supplementary tables 1, 2 and 3.

A specific procedure was developed to calculate the dissimilarity matrices according to our double weight system. A matrix was calculated separately for each eBSV and used to build a diversity tree using the neighbour-joining (NJ) algorithm (Saitou and Nei, 1987) with 1000 bootstrap replicates (DARwin v5.0.155 software, Perrier and Jacquemoud-Collet, 2006). For a joint analysis of all three eBSVs, a synthetic dissimilarity matrix was calculated as the sum of the three eBSV dissimilarities, each with a specific weight to compensate for the unequal number of observed fragments. A NJ tree was built from this overall dissimilarity.

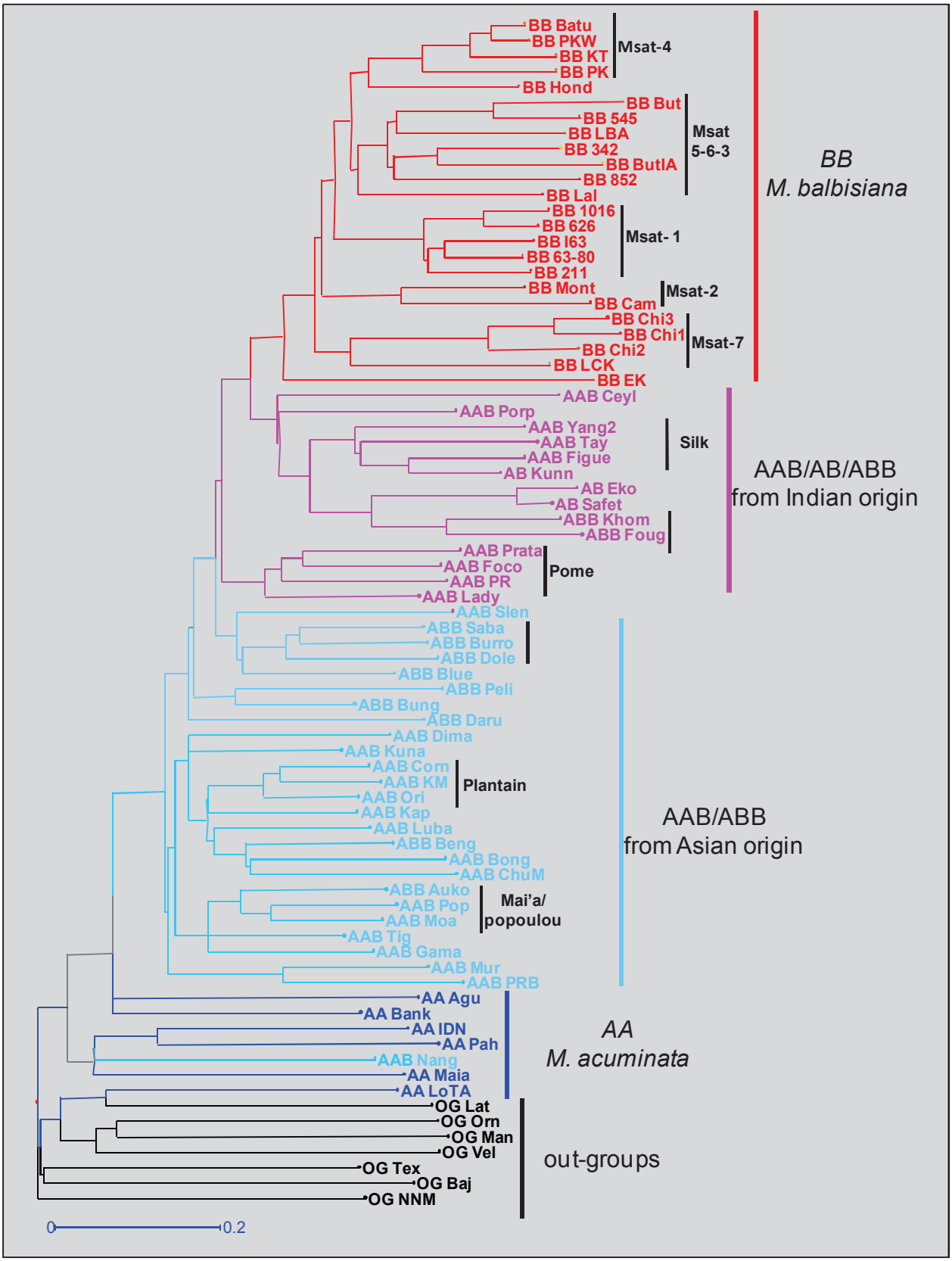


Figure 1: Neighbor joining diversity tree for the sampling of 77 accessions, extracted from-a larger NJ tree based on 17 microsatellite markers for a representative sampling of 569 banana accessions.

Tree rooted on other *Musa* species as outgroups

Colors of accessions show their affiliation to either genotypic groups or geographic origin when possible. Thin black lines indicate sub-groups according to Gayral et al (2010) for BB accessions and Perrier et al (2009) for hybrids (AAB, ABB and AB). Accession identifiers concatenate genotypic group and abbreviated names according to table 1.

Results

Development of a microsatellite phylogeny

A *Musa* phylogeny representative of the available diversity of wild and cultivated *M. acuminata*- and *M. balbisiana*-derived types and based on SSR markers was proposed by Perrier et al. (2009, 2011) and Hippolyte et al. (2012). Since this phylogeny was less representative of *M. balbisiana* diversity, we genotyped 23 additional accessions: seedy *M. balbisiana* diploids as well as interspecific accessions, diploid (ABB) or haploid (AAB) for the *M. balbisiana* genome. Combining the two data sets, we performed a novel microsatellite-based analysis on a total of 567 accessions. From this overall phylogeny (data not shown) we extracted a sub-tree corresponding to our 77 accessions and rooted on the outgroup accessions (figure 1).

The main structure of the phylogeny revealed a contrast between *M. acuminata* and *M. babisiana* genomes. The *M. acuminata* genome pole included all *M. acuminata* accessions (AA); *M. laterita* associated to *M. acuminata* Long Tavoy and the hybrid AAB Pisang Nangka are also present in this group. The former is known to be very close to *M. acuminata* (and should be reclassified), the latter is suspected to have only a partial B genome.

Seedy BB diploids formed a monophyletic group structured into six groups showing little diversity; their geographic origins are often not known and a possible geographic structure could not be verified. This result is in agreement with those of with Gayral et al. (2010). We observed similar compositions for microsatellite groups 1, 2 and 4 whereas microsatellite groups 5, 6 and 3 were not differentiated. However, the structure was not as robust as that observed through low bootstrap values in Gayral et al. (2010). We described an additional Msat group, named 7, grouping accessions absent in the previous analysis. It included three *M. balbisiana* accessions, recently collected in South China (Chi 1 to 3) and the Lep Chang Kut accession also originating from China. These accessions are very close and probably belong to the same genetic population (figure 1).

Between the pure *M. acuminata* and *M. balbisiana* poles, all interspecific hybrids were organized in successive clusters. Two main forces—genotype and geography—explained the observed structure. Accessions grouping in the same cluster had the same genotype (AAB/AB, ABB) (figure 1). Several of these clusters corresponded to triploid subgroups defined on morphological characters such as Pome, Plantains, Awak, Silk and so on. The second level of structure was based on the geographical origin of the accessions. This clustered the interspecific hybrids schematically into accessions from India, and accessions from a large SE Asia region ranging from Papua New Guinea to the Philippines and Indonesia. The geographic indication refers to the origin of the accession and not its current distribution; for example, AAB plantains are found exclusively in Africa but originated from the Philippines region.

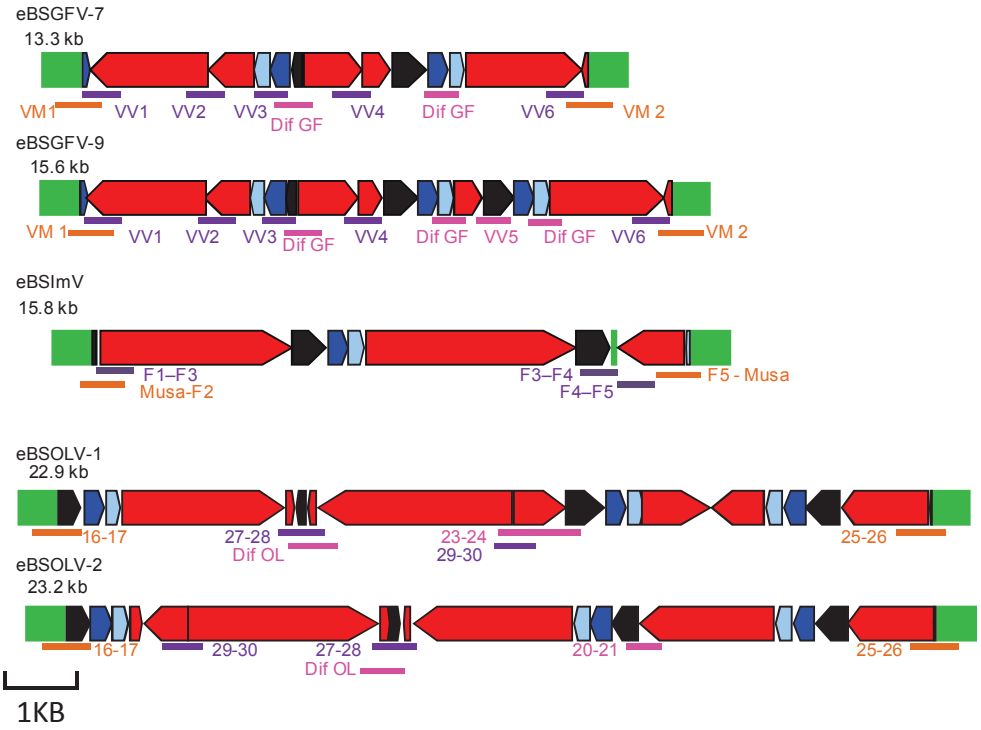


Figure 2 : Overview of PCR and dCAPs loci within PKW eBSGF -eBSOLV-eBSImV

Banana genomic sequences are in green. BSV is represented in linear view. Dark blue, light blue and red boxes indicating ORF1, ORF2 and ORF3 of the virus respectively for BSV and eBSV. The intergenic region (IG) is in black. Arrow boxes indicate the orientation of the fragment in the genome PCR amplification fragments are represented by orange, purple and pink lines corresponding to *Musa* junction, internal and allelic amplifications fragments respectively.

These PCR markers were developed by Gayral et al (2010) and Chabannes et al (2012)

The heterozygosis of diploid samples was estimated from SSR alleles (table 2). AA and AB diploids showed a heterozygosis twice that recorded for seedy BB diploids (table 2), in agreement with the reduced diversity observed. The Balbisiana 1016 accession was shown to be fully homozygous, and accessions of the Msat group 1 showed a very low level of genetic diversity.

Genotyping PKW-related diversity of eBSGFV, eBSOLV and eBSImV within *M. balbisiana* diversity

PCR-based eBSV hallmarks

The full-length structures of the three BSV species present within the seedy diploid *M. balibisiana* PKW were established definitively by Gayral et al. (2008) for BSGFV and by Chabannes et al. (2012) for BSOLV and BSImV. Knowledge of the eBSV structures in PKW allowed the development of specific PCR and dCAPs markers. The PCR markers are specific to either *Musa* junctions or internal structures and, where possible, specific to each allele (figure 2); they do not react with any episomal viral counterparts. The dCAPs markers (denoted Dif OL and Dif GF) discriminate eBSV alleles of BSOLV and BSGFV, respectively, in PKW; eBSImV is monoallelic. Consequently, the PCR results can testify not only to the presence of eBSVs but also their preliminary structure. The results recorded with the various markers provide an eBSV hallmark for each BSV species.

We set up the genotyping as described in the Materials and methods and established the PKW-related eBSV genotype of the 77 accessions presented above.

No eBSOLV, eBSGFV and eBSImV hallmarks existed in either the out-group or the AA accessions, confirming the real dual specificity of our markers. PKW-related eBSV hallmarks were recorded for the majority of the other accessions concerning BSGFV as well as BSOLV and BSImV, indicating the wide BSV colonization of *M. balbisiana*. PCR genotyping revealed a polymorphism in structure for PKW-related eBSV rather than a polymorphism of integration. eBSGFV hallmarks were strongly conserved whereas eBSOLV and eBSImV hallmarks appeared more diverse and rearranged. The modifications concerned mostly internal rearrangements, going from a small amount to a lot of missing PCR amplification. Nevertheless, the majority of accessions yielded systematic PCR amplifications with the two *Musa*-junction PCR markers. This may indicate a common locus for integration of each BSV species into the B genome. Interestingly, some accessions lacked one or more of the eBSV hallmarks examined. This was the case for the AAB Pisang Nangka accession, which had no eBSV hallmarks; AAB Kunaimp, which had neither eBSOLV nor eBSImV hallmarks; four AAB accessions, which had neither eBSImV and eBSGFV hallmarks; three AAB and one AB accessions without eBSGFV hallmarks; and one BB, one ABB, two AB and eight AAB accessions that were also apparently eBSImV hallmark-free (supplementary data).

	1-Im	**2-Im**	**3-Im**	**4-Im**	5-Im	6-Im	7-Im
eBSImV	1	**3**	**2**	**2**	1	1	1

	1-GF	2-GF	3-GF	4-GF	5-GF
eBSGFV-7	**2**	1	1	1	
eBSGFV-9	**2**	1	1	1	1

	1-OL	**2-OL**	3-OL	4-OL	5-OL	6-OL	7-OL	8-OL
eBSOLV-1	**2**	**2**	1	1	1	1		
eBSOLV-2		1	1			1	1	1

Table 4: Numbers of the rectriction enzyme fragments in PKW eBSGFV-eBSOLV-eBSImV. Grey box : no fragment for the eBSV allele
Bold characters: repeated fragment

Southern blot-based eBSV hallmarks

eBSV hallmarks gave us a preliminary picture of the eBSV structure when PCR results were positive. However, a negative PCR result could be due to either no fragment or the presence of another fragment of BSV not recognized by the PCR primers. In order to obtain a more complete picture of PKW related-eBSV structure, we implemented the additional approach of Southern blot analysis. We finally selected two restrictions enzymes per eBSV due to the fact that eBSV structure often comprises the same repeated viral fragments. We used *Bam*H1/*Hin*dIII for eBSOLV and *Dra*1/*Spe*1 for eBSGFV and eBSImV. The restriction enzyme patterns of each PKW eBSV are presented figures 3A, 4A and 5A.

The eBSV structure produced several common fragments regardless of the restriction enzymes used due to the same viral sequences being repeated within each eBSV allele and between alleles (table 4, figures 3A, 4A, 5A). Thus, the eBSGFV alleles were very similar. Each allele had fragment 1-GF repeated once, and all alleles shared most of the 6 digested fragments except fragment 5-GF, which was harbored only by eBSGFV-9 (table 4). Conversely, eBSOLV alleles showing more structural differences shared only 3 of the 8 potentially reconstituted fragments. eBSOLV-1 showed only fragments 1-OL and 2-OL, repeated twice. eBSImV had 7 fragments in total. Fragment 2-Im was present three times whereas fragments 3-Im and 4-Im were present twice. To overcome these difficulties, we first repeated identical Southern blots several times for the same BSV species in order to confirm the PKW-related eBSV patterns observed. Next, the *Dra*1/*Spe*1 membranes were cross-hybridized with the BSOLV probe, and those of *Bam*H1/*Hin*dIII with either BSImV or BSGFV (supplementary data/data not shown) in order to cross-check the analysis.

Controls were used to calibrate fragments sizes: the diploid BB PKW harboring the reference eBSV alleles and, when possible, BAC clones (each containing one PKW eBSV allele) were included to visualise the expected patterns, and a size ladder. We treated all fragment accessions using Image QuanTL software (GE Healthcare©) to determine the exact size of fragments after calibration. We analysed Southern blots by referring to the known eBSV patterns given by PKW for the three BSV species (figures 3A, 4A and 5A).

No hybridization pattern occurred in either the out-group or the AA accessions according to the eBSV hallmark genotyping results.

Among the other accessions, patterns recorded for BSGFV appeared well conserved (figure 3B). This reflects the similarity in structure of both eBSV alleles (table 4), which differ only after digestion to yield fragment 5GF. In addition, we recorded several over-size fragments for all accessions (noted * figure 3B). We determined their sizes using QuantTL software and hypothesized a partial digestion of eBSGFV. All the oversized fragments corresponded to non-digested fragments as determined by *in-silico* analysis of the PKW eBSGFV sequence. We concluded that we had a general digestion problem during the experiment and we decided to not count these fragments for analysis.

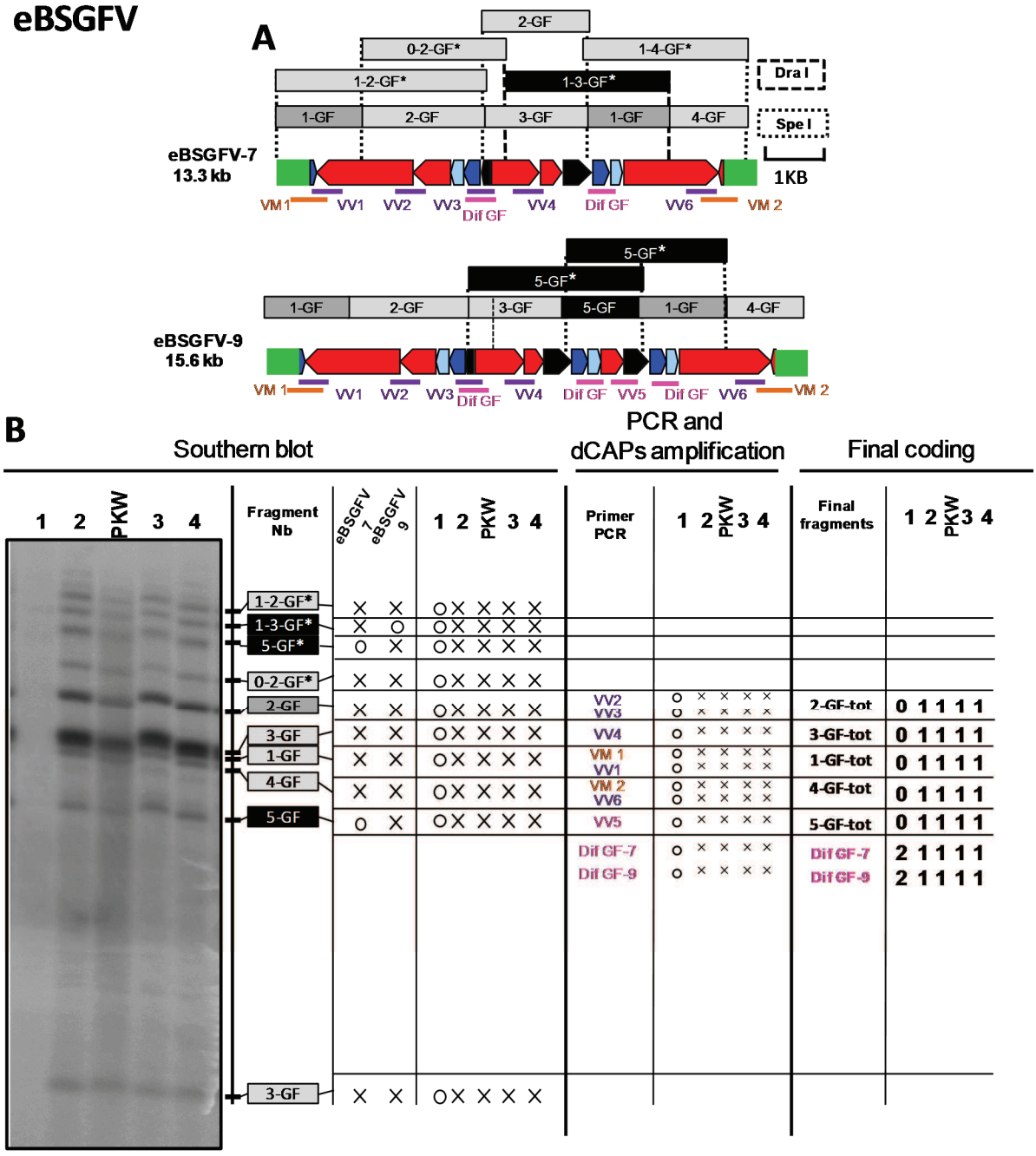


Figure 3 : Southern blot and PCR-based hallmarks to establish the final eBSGFV genotype

A. **Overview of the restriction enzyme digestion on PKW eBSImV**.

Dot and streak lines correspond to Bam H1 and Spe1 digestion respectively. Grey, dark grey and black boxes represent fragments obtained following enzyme digestion and corresponding to unique fragment existing in both alleles, fragment repeated within the allele and common with the other allele, and unique fragment for the eBSV respectively.

* correspond to initial fragment non digested.

B. Southern blot pattern and analysis method

Southern blot patterns of digested DNA of *M. balbisiana* diploids accessions (noted 1, 2, 3 and 4 for Porp, Chinois 1, Klue tani and Pisang batu accessions respectively) revealed by hybridization with the full length BSGFV probe were on the left part . PKW is the positive control.

Tables presented the coding method employed of southern blot and PCR/dCAPs- based results were on the right part. Cross show presence of fragment and of correct PCR amplification (0) absence. Final coding table presented the final results of the line, (1) presence (0) absence and (2) presence of only southern blot fragment without PCR amplification. For dCAPs markers, like for dif GF-7and -9, (2) represent the absence of PCR amplification.

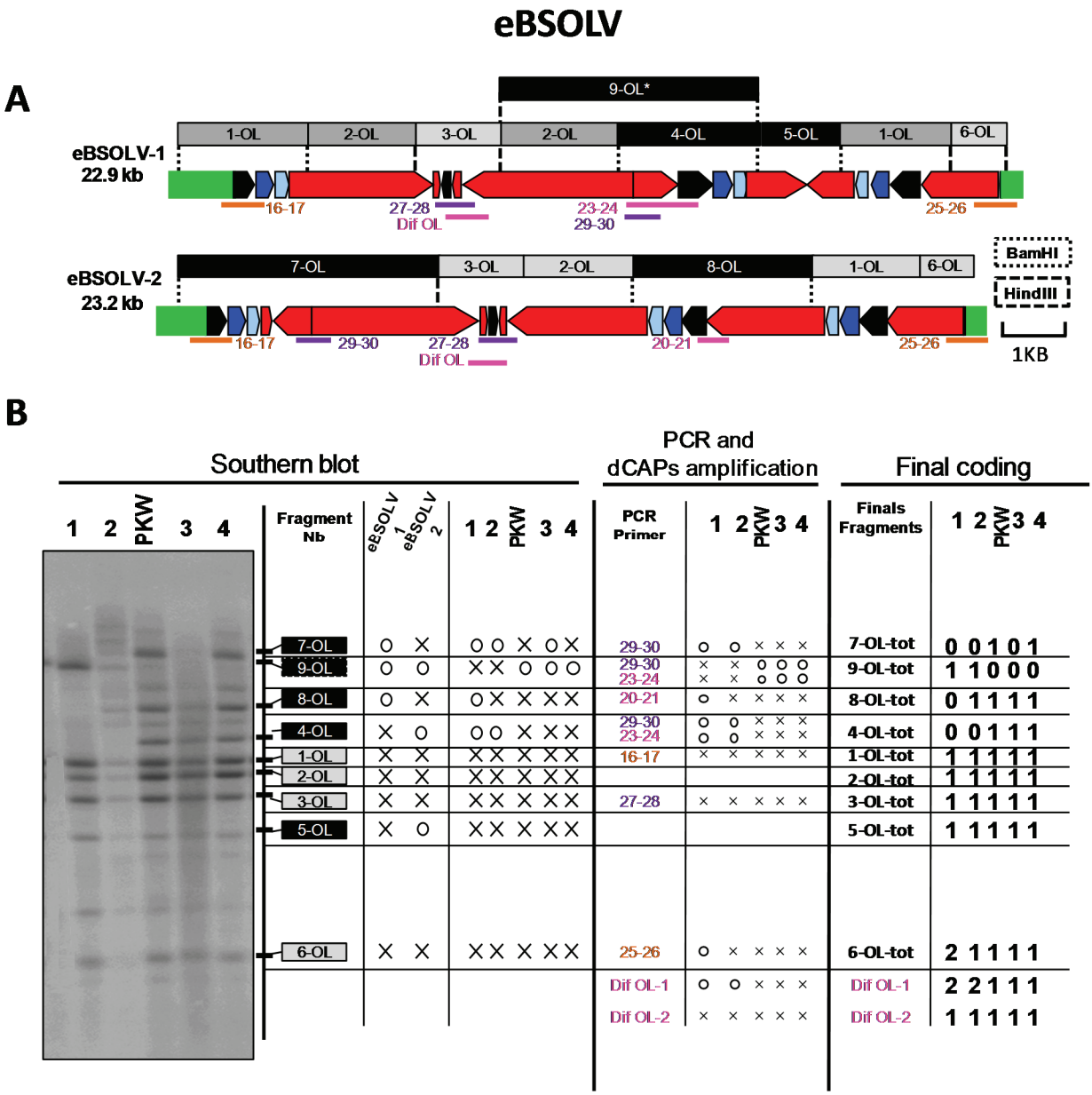


Figure 4 : Southern blot and PCR-based hallmarks to establish the final eBSOLV genotype

A- Overview of the restriction enzyme digestion on PKW eBSImV.

Dot and streak lines correspond to Bam H1 and Spe1 digestion respectively. Grey, dark grey and black boxes represent fragments obtained following enzyme digestion and corresponding to unique fragment existing in both alleles, fragment repeated within the allele and common with the other allele, and unique fragment for the eBSV respectively. * correspond to initial fragment non digested.

B. Southern blot pattern and analysis method

Southern blot patterns of digested DNA of *M. balbisiana* diploids accessions (noted 1, 2, 3 and 4 for Porp, Chinois 1, Klue tani and Pisang batu accessions respectively) revealed by hybridization with the full length BSGFV probe were on the left part . PKW is the positive control.

Tables presented the coding method employed of southern blot and PCR/dCAPs- based results were on the right part. Cross show presence of fragment and of correct PCR amplification (0) absence. Final coding table presented the final results of the line, (1) presence (0) absence and (2) presence of only southern blot fragment without PCR amplification. For dCAPs markers, like for dif OL-1 and -2, (2) represent the absence of PCR amplification.

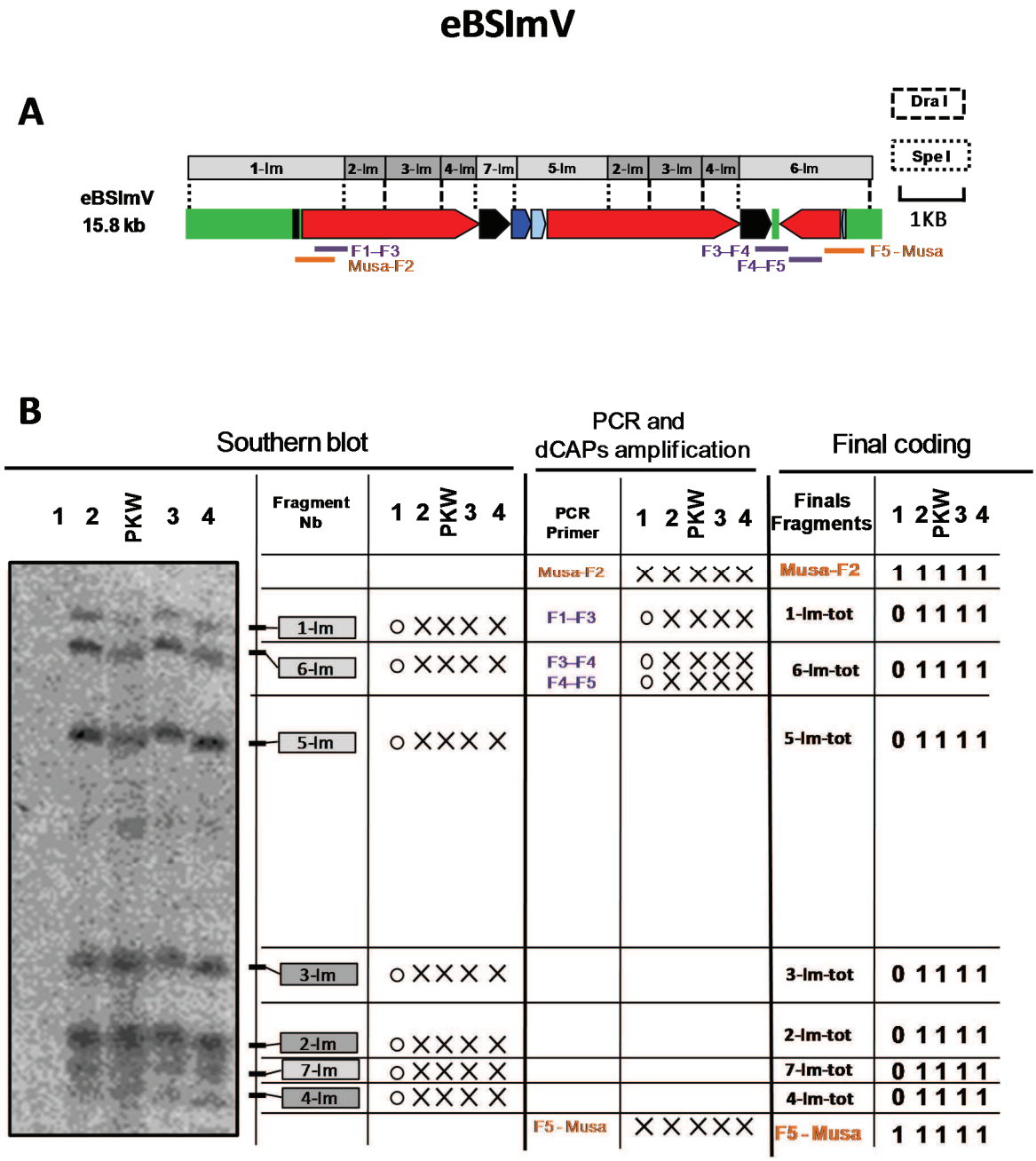


Figure 5 : Southern blot and PCR-based hallmarks to establish the final eBSImV genotype

A. **Overview of the restriction enzyme digestion on PKW eBSImV and PCR markers**

Dot and streak lines correspond to to Dra1 and Spe1 digestion respectively. Grey and dark grey boxes represent fragments obtained following enzyme digestion and corresponding to unique fragment existing in both alleles and fragment repeated within the allele.

B. Southern blot pattern and analysis method

Southern blot patterns of digested DNA of *M. balbisiana* diploids accessions (noted 1, 2, 3 and 4 for Porp, Chinois 1, Klue tani and Pisang batu accessions respectively) revealed by hybridization with the full length BSGFV probe were on the left part . PKW is the positive control.

Tables presented the coding method employed of southern blot and PCR/dCAPs- based results were on the right part. Cross show presence of fragment and of correct PCR amplification (0) absence. Final coding table presented the final results of the line, (1) presence (0) absence and (2) presence of only southern blot fragment without PCR amplification. When PCR amplification are code alone, like for F5-Musa and Musa-F2, (2) represent the absence of PCR amplification.

BSOLV Southern blot analysis showed a large diversity of patterns that certainly reflected allele differences. As with BSGFV, we observed one over-sized fragment (noted 9-OL) for several accessions (figures 4B). However, the presence of the 9-OL fragment was always correlated to the absence of the 4-OL fragments and sometimes to the absence of the 2-OL fragment when the genotype was haploid for the B genome. This argument, together with the size of the fragment, allowed us to determine that this fragment corresponded to a non-digestion between 2-OL and 4-OL due to an SNP at this restriction site. So we decided to count this fragment in our analysis because it really does correspond to a difference between accessions.

BSImV patterns ranged from similar to those of PKW eBSImV, to totally different, to absent (figure 5B). The differences likely represent different integrated alleles as in the case of BSOLV.

A few additional bands remained non-determined regarding PKW-related eBSV patterns and were not taken into consideration as they could not be reproduced for the same accession or between accessions.

Structure of the PKW-related eBSV within M. balbisiana diversity

Justification of Southern blot method used

PCR and Southern blot-based hallmarks give complementary data that is both useful and relevant to proposing a possible representation of PKW-related eBSV for each accession. Indeed, all restriction enzyme fragments are associated with PCR markers for both eBSGFV and eBSOLV, except fragments 2-OL and 5-OL for eBSOLV (figures 3A, 4A and 5A). However, the linear eBSImV structure, which resembles a tandem linear episomal viral genome, makes analysis difficult. Flanking region fragments of eBSImV are associated with PCR markers whereas the internal structure, with the same organization as the episomal genome, precludes the design of specific PCR markers for specific fragments. Nevertheless, our main objective was not to propose a definitive eBSV structural organization but rather to identify components related to PKW eBSVs in order to propose an accurate picture of PKW-related eBSV allele diversity in the *M. balbisiana* species. On this basis, we coded our Southern blot data together with PCR-based results to interpret eBSV distribution. The final analysis method is presented for 5 accessions in figures 3B, 4B and 5B and explained in Materials and methods.

Allelic information reported on the dendrogram resulted from both the specific fragments from Southern blot analysis and specific PCR amplification of the different alleles reported in the supplementary data. Three dendrograms of PKW-related eBSV distribution were inferred by the Neighbor Joining method for each BSV species from the total results of the 77 accessions.

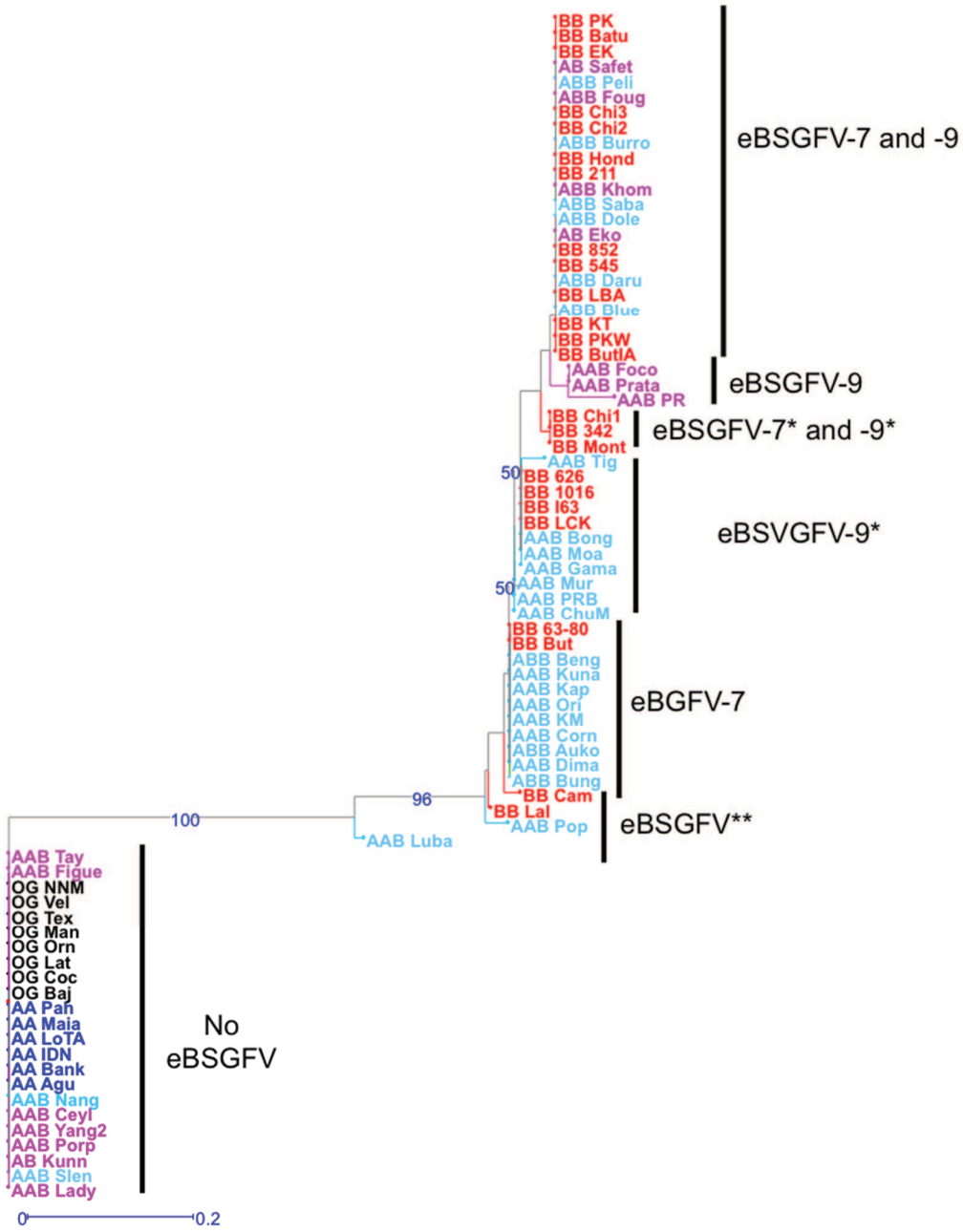


Figure 6 : Neighbor joining tree built from eBSGFV genotypes of the 77 sampled *Musa* accessions

The dissimilarity matrix in input is estimated from PCR and southern blot based hallmarks (cf figure 5). Black lines indicated clusters of accessions with identical eBSGFV alleles: 7 and 9 are standard alleles present in PKW, 7* and 9* are lightly modified alleles, ** indicates alleles widely modified in sequence and structure Colors of accessions are according to fig 2. Only bootstrap values greater than 50 are displayed on the tree.

PKW-related eBSGFV distribution

The PKW-related eBSGFV dendrogram is separated into two parts: with and without eBSGFV (figure 6). All AA and out-group accessions grouped with eight AAB and one AB accession that did not present any eBSGFV. The AAB Nangka accession was already grouped in the *Musa* phylogeny (figure 1) with AA diploid accessions. The others (AAB and AB) originate from India.

The group with eBSGFV showed a remarkable conservation of PKW eBSGFV since, with the exception of the AAB Luba accession, all *M. balbisiana* diploid as well as *M. balbisiana* haploid accessions are distributed into one well supported group (strong bootstrap-96). This accession lacked the 2-GF fragment, and had a different mutation that led to non-amplification in the case of 3 PCRs. All other accessions presented a similar allelic structure as eBSGFV-7 and eBSGFV-9. These were structured into 6 very close allelic-based clusters (short branches). The differences are due mainly to point mutations, which not only prevented PCR amplification but also precluded assessment of ploidy level. This is illustrated by a group of 10 accessions that share a mutation in the PCR marker VV5 while the specific fragment of allele 9 is present (labelled 2 in the table of supplementary data for the 5-GF fragment). The clusters are named according to their allelic information.

Two AB accessions—Ney Poovan and Safet Velchi—harboured 7 and 9 alleles in their genomes, respectively, despite having only one B genome. They grouped with the main BB diploid accessions harboring both PKW eBSGFV.

PKW-related eBSGFV is very well conserved among the diversity of *M. balbisiana* accessions. We observed that both eBSGFV-7 and -9 alleles are present in the majority of BB diploid accessions with the exception of 4 that lackt eBSGFV-7, and another 3 lacking eBSGFV-9. Among Indian accessions, eBSGFV-7 and eBSGFV-9 both occur in AAB accessions, whereas eBSGFV-7 only is reported in ABB accessions. Among Asian accessions, eBSGFV-9 is the only one reported in AAB accessions whereas both are reported in AB and ABB accessions.

PKW-related eBSOLV distribution

The PKW-related eBSOLV dendrogram split the accessions into three parts, i.e. those with PKW-related eBSOLV, those with modified PKW-related eBSOLV and those without PKW-related eBSOLV (figure 7). All the AA and out-group accessions, including the two AAB accessions (Nang and Kuna), did not present any eBSOLV. The flanking fragments were always conserved despite strong changes in internal organization.

Seven accessions were highly diverged. A large number of fragments were missing for AAB Slen, AAB Luba, AAB PRB, AAB Mur, BB LCK, BB Cam, AAB Tig accessions.

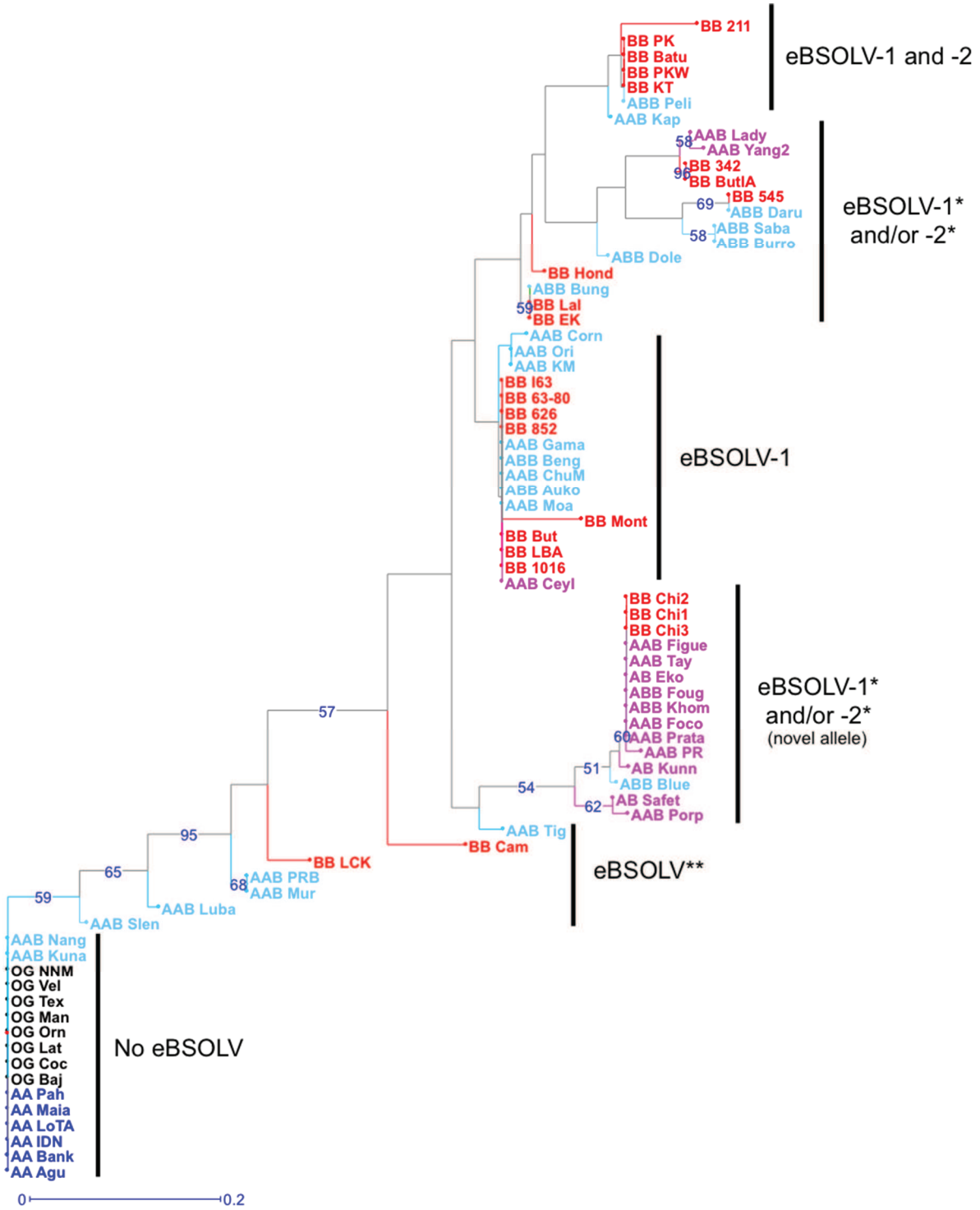


Figure 7 : Neighbor joining tree built from eBSOLV genotypes of the the 77 sampled *Musa* accessions

The dissimilarity matrix in input is estimated from PCR screen and southern blot based hallmarks (cf figure 5). Black lines indicate clusters of accessions with identical eBSOLV alleles: 1 and 2 are standard alleles present in PKW, 1* and 2* are lightly modified, ** indicates wide divergences, in sequence and structure, with the PKW alleles which are no longer. Colors of accessions are according to fig 2. Only bootstrap values greater than 50 are displayed on the tree.

The other accessions were distributed into two allelic-based divergent clusters. The first grouped into three sub clusters according to PKW eBSOLV alleles (5 accessions), and two slightly modified PKW related alleles (13 and 17 accessions). For the first time, we observed a non-expected fragment of 4300bp for the BB Montpellier accession corresponding to a novel fragment. The second modified allele corresponded to structural changes in fragments 4-OL, 5-OL, 8-OL, where we had already detected a difference between the two PKW alleles. The 7-OL fragment was observed to be lacking when the modifications were large, including the extreme case of the absence of fragments 1-OL and 2-OL. We first assumed that all these modifications corresponded to novel PKW-related eBSOLV alleles; however, these novel structures were in fact a mix between eBSOLV-1 and eBSOLV-2. The observed eBSOLV diversity is large, as at least 22 alleles are reported.

PKW eBSV alleles were found together only in the PKW clade (3 accessions), and in BB 211 and two ABB accessions from the India Group. eBSOLV-1 is the only allele to be massively represented in BB, AAB and ABB accessions from the Asian Group and the AAB Ceylan accession from the Asian Group. Newly inserted BB diploid accessions (Ch1, 2 and 3) forming a new microsatellite group presented specific eBSOLV alleles. These alleles are shared with 10 other AAB, ABB, and AB accessions, all from the Indian group.

PKW-related eBSImV distribution

The dendrogram constructed using data obtained for PKW-related eBSImV proposes three clusters (figure 8). A large cluster aggregated the entire AA and out-group accessions including 15 AAB, 1 ABB, and 2 AB as well as one BB accession that did not have any PKW-related eBSImV. The Honduras accession was the only BB without eBSImV.

The second cluster grouped 9 accessions showing flanking fragments and lacked the others. As for BSOLV, these insertions may be considered as lost. Four BB accessions have lost a large part of eBSImV.

The third cluster is split into two sub-clusters: a big one of 27 accessions harboring PKW eBSImV alleles including all ABB accessions, and a smaller one containing 8 accessions. The integration appeared identical to PKW. The structural differences observed in these accessions may be correlated with different stages of eBSImV loss. In the hybrids group, we observed either the complete absence of eBSImV in the plantains group or an almost complete absence in the Indian Group. The presence of just small parts of eBSImV in particular at the junction level for certain accessions from the Indian Group may correspond to degeneration of the integration, which leads to a complete loss of eBSImV in these genomes. We observed degeneration in the Mai'a/ Popoulou group in the other accessions.

For the whole sample, the results show that PKW-related eBSV present in hybrid accessions are more diverse than those derived from integration in diploid BB accessions. We also

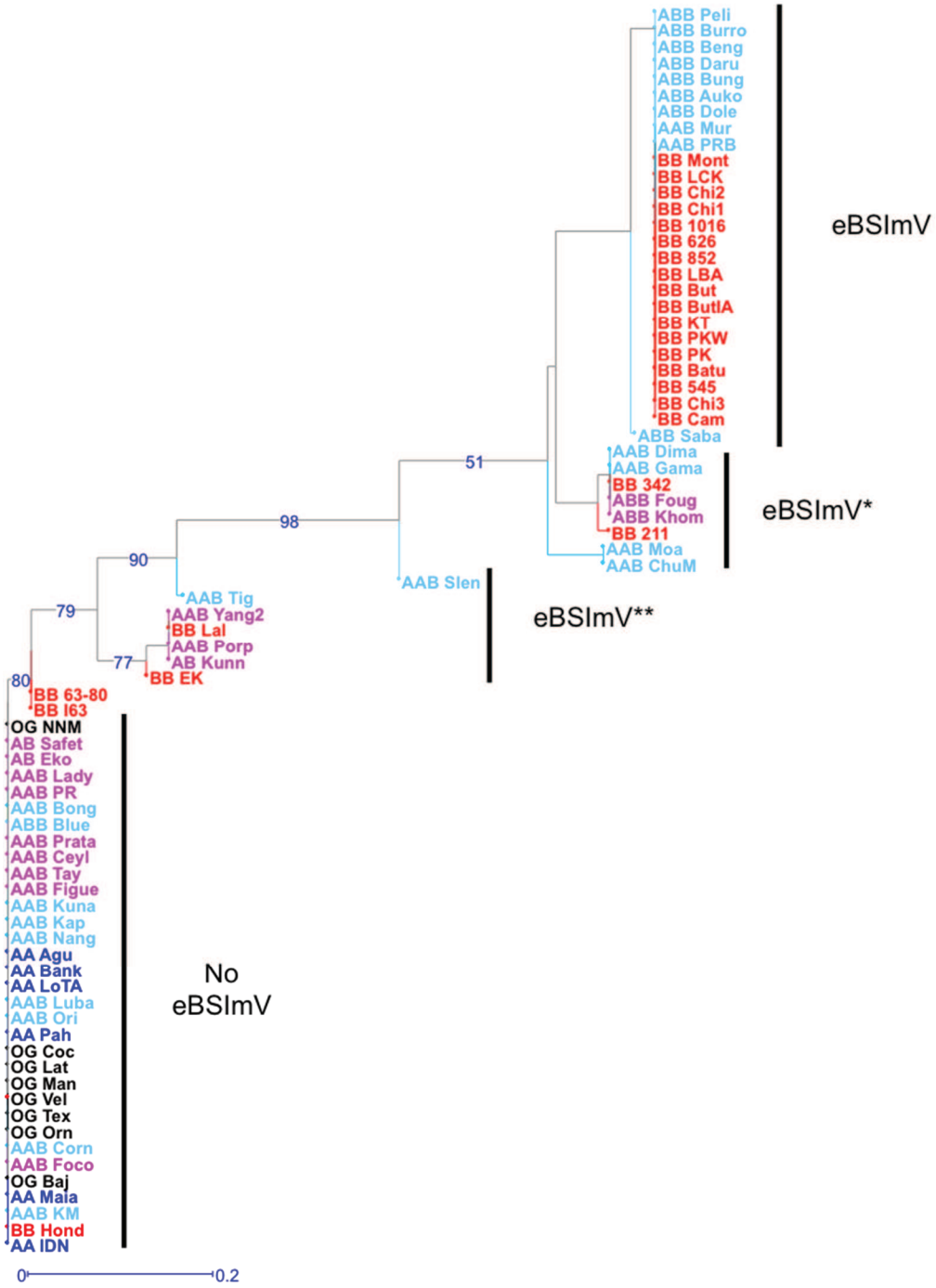


Figure 8 : Neighbor joining tree built from eBSImV genotype of the 77 sampled *Musa* accessions

The dissimilarity matrix in input is estimated from PCR and southern blot based hallmarks. Black lines indicated clusters of accessions with identical eBSImV compare to that present in PKW, * are lightly modified, ** indicates wide divergences, in sequence and structure, with the PKW eBSImV which are no longer. Colors of accessions are according to fig 2. Only bootstrap values greater than 50 are shown on the tree.

observe a good correlation between eBSV structures and phylogeny. Indeed, the majority of phylogenetic groups correlate well with either the same type of eBSV allele or the absence of integration.

eBSV based analysis of *M. balbisiana* diversity

A dendrogram was inferred from all eBSV data (figure 9). The synthetic tree, which was built on the weighted sum of dissimilarities of EBSV, was structured logically by the presence and absence of the different eBSV sequences (figure 9). The upper part of the tree was rooted on divergence between *M. acuminata* and *M. balbisiana* grouped accessions with only eBSOLV, next came accessions with eBSOLV and eBSGFV and, finally, accessions with all three eBSVs. Downstream this structure, the groups of accessions, grouped for their SSR markers (figure 1), were also present in this tree, accessions from these groups sharing the same eBSV structures. Genotypes of these groups also seemed structuring in this tree. AAB accessions were all present in the upper part, with few BB and all the other BB and ABB accessions present in the lower part of the tree. This means that a large fraction of AAB accessions bring just some of the eBSV types, while ABB and BB bring all three eBSV. Geographic hybridization areas pointed out by Perrier et al. (2009) were also represented but less clearly because of the large diversity of eBSV structure present in each group. Indeed, AAB and AB accessions from India were found in the same group. However, plantains from Africa and Mai'a/Popoulou from Oceania, which have a similar genetic origin, formed two distinct groups that were linked only with BB plants. The others branches were composed essentially of BB and ABB accessions. As in the phylogeny built from *Musa* genome markers, BB accessions from groups Msat-1 and 4 formed a specific branch. Pelipita ABB accession, (ABB-Peli) was also in the Msat-4 group and might have the same B genomes as the BB accessions in this group. ABB accessions formed a specific group linked both with the Msat-4 group and with another BB. We can observe that ABB plants from India were grouped with BB-211, although in the phylogenetic tree (figure 1) they were with the other AAB from India. Apart for the AAB Silk subgroups, all AAB and ABB grouped with one or more BB accessions. Accessions with very specific eBSV constituted a specific branch encompassing Tigua (AAB-Tig), Kunaimp (AAB-Kuna) and Luba (AAB-Luba). As already described for each eBSV, AAB Pisang Nangka is in the root of the tree with the out-group and AA accessions.

Discussion

The aim of this work was first to characterize the eBSV diversity of three BSV species among *Musa balbisiana* accessions, and to set eBSV in an adapted host phylogenetic context in

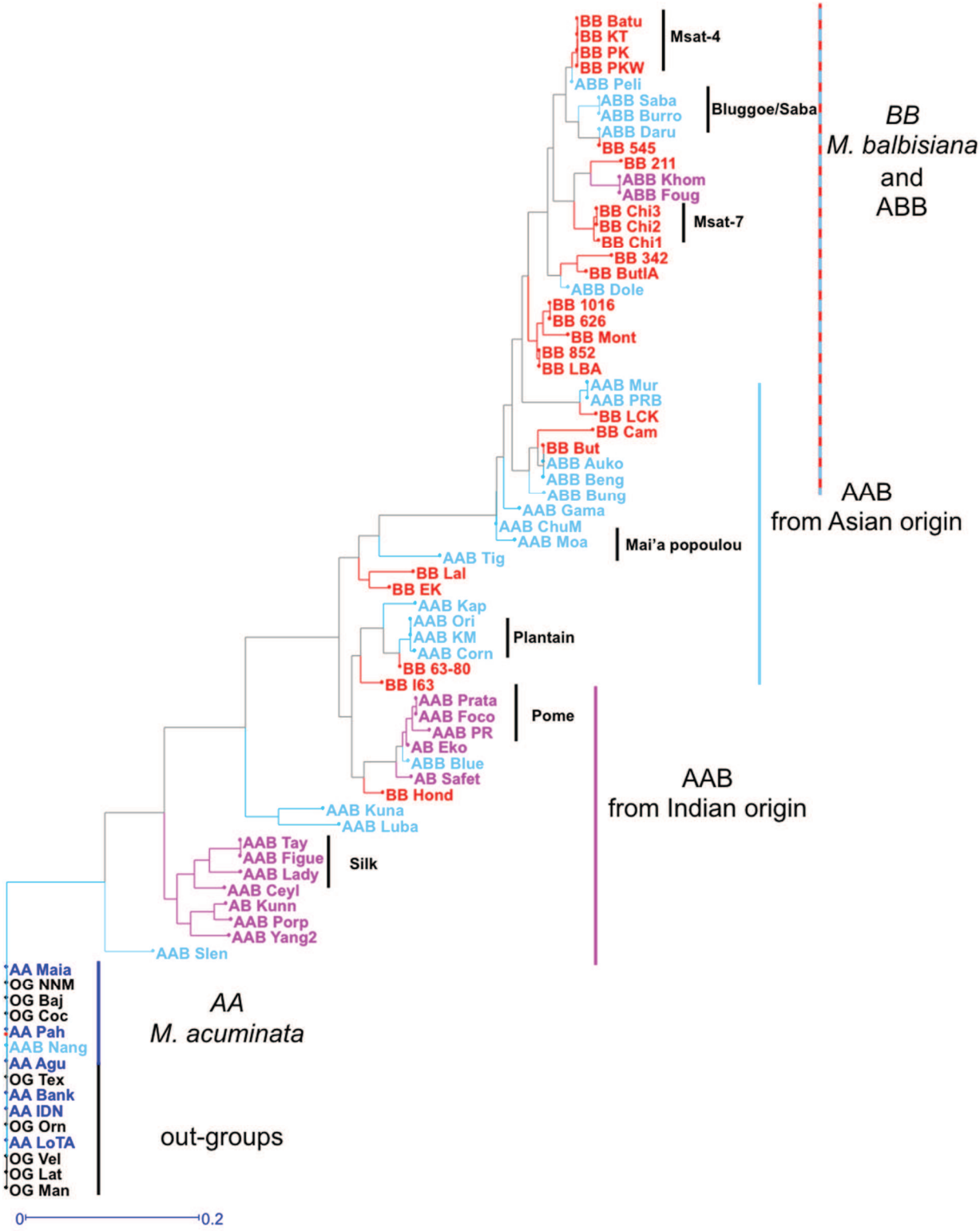


Figure 9 : Neighbor joining tree from a joint analysis of the the 3 eBSV for 73 sampled *Musa* accessions

The dissimilarity in input is a weighted sum of dissimilarities calculated separatly for eBSGFV, eBSOLV and eBSImV. Colors of accessions show their affiliation to either genotypic groups or geographic origin as on fig. 5. Thin black lines indicate sub-groups according to Gayral et al (2010) for BB accessions and Perrier et al (2009) for hybrids (AAB, ABB and AB). Accession identifiers concatenate genotypic group and abbreviated names according to table 1.-Bootstraps were not estimable for this analysis.

order to investigate the evolutionary history of eBSV. Next, we wanted to test whether eBSV can be used as a phylogenetic marker to contribute to resolving B genome diversity.

Development of an original method for the analysis of eBSV distribution

A preliminary analysis of eBSV diversity in the B genome among the 20 seedy BB accessions then available was perfomed by Gayral et al. (2010). The latter authors showed that eBSImV and eBSGFV polymorphism is weak, although PCR markers detected some variations in eBSImV. Nevertheless, the majority of plants showed the same integrations as in PKW. Further study of eBSGFV sequences showed that the nucleotide diversity was also very limited and did not really discriminate the different eBSVs present in BB accessions, most likely due to the narrow genetic basis of *M. balbisiana* diversity. To extend these studies, we first increased the *M. Balbisiana* diversity of the banana sample by including AB, ABB and AAB hybrid accessions as suggested by Hippolyte et al. (2012) who showed that a lot of SSR balbisiana alleles found in AAB or ABB interspecific hybrids were not present in the analyzed *M. balbisiana* accessions. We also took account of the BSV species story by including the common BSOLV species. Finally, we also modified the technical approaches used: we used Southern blot analysis to reveal eBSV polymorphism and to allow interpretation of negative PCR results. Indeed, a more accurate picture of integration is given by the joint analysis of the both data on the same parts of eBSV. Although the Southern blots proved difficult to perform, a lot of information has been recorded and analysed so far. In the majority of cases, we observed a close correlation between the results of PCR and Southern blot analysis. In cases where the PCR result was negative, such as for eBSGFV studies, Southern blots were able to confirm the presence of expected fragments and permitted conclusions regarding point mutations to be made. On the other hand, Southern blots allowed us to pinpoint some parts of the eBSV structure not covered by PCR markers, such as the hot spot zone of recombination for eBSOLV (figure 4A) and the internal structure of eBSImV (figures 5A). Thus, our analysis revealed that information gleaned from Southern blots augmented and enhanced the information obtained from PCR analysis.

eBSGFV distribution among diverse B genome

BSGFV analysis shows that eBSGFV alleles are highly conserved among diverse B genomes, particularly for the BB and ABB accessions. Most of these accessions harbor both eBSGFV_7 and -9 alleles in their genome. This is in agreement with the early appearance of both alleles because all phylogenetic groups described for these plants possess both alleles.

BB accessions with only one allele can be explained by weak heterozygoty. For example, the BB Balbi 1016 accession, which is homozygous for microsatellite markers, has eBSGFV-9 only.

Surprisingly, two AB accessions appeared with both eBSGFV-7 and -9 alleles. A first explanation would be that these accessions might not be true AB genotypes, and that they might have a supernumerary chromosome bringing the second allele. Such cases of unbalanced chromosome representation of the two genomes have been observed in triploid interspecific hybrids but never in diploid context. Another explanation could be spurious interpretation. Indeed, allelic genotyping is tricky as the eBSGFV are similar, and dCAPs markers are based on point mutations present only in the additional fragment of eBSGFV-9. The eBSGFV can thus be different in these two AB plants due to recombination, and the AB plants could harbor the two types of mutation in their sequence.

Several AAB Pome and Silk accessions as well as AB accessions, all from the Indian origin, did not have any eBSGFV in their genome. Moreover, the others AAB Pome accessions showed only the non-infectious eBSGFV-9 allele, sometimes with mutations whereas ABB and AB accessions from this Indian area have both eBSGFV-7 and -9 alleles as PKW. This suggests that the two alleles are, or were in the past, effectively present in the parental Indian BB accessions. Nevertheless, BB Lal velchi—the parent of AAB subspecies Pome or AB Kunnan accessions of the Indian Group—is the BB accession with most mutations in comparison to the sequences of eBSV in PKW.

A majority of accessions from Asian region, including AAB hybrids like African Plantains, brings the eBSGFV-7 allele which has been shown still infectious (Côte et al., 2010). Nevertheless, in the Mai'a Maoli / Popoulou AAB subspecies of Oceania, genetically close to African plantains, the eBSGF-9 allele is dominant

eBSOLV distribution among diverse B genomes

eBSOLV analysis revealed this integration to be the most diverse, with many novel eBSOLV alleles being discovered here for the first time compared to those of PKW. Indeed, only the BB plants from group Msat-4 and ABB Pelipita accession share PKW eBSOLV-1 and 2 alleles exactly. Almost all the plants with the B genome bring at least traces of eBSOLV. The newly discovered alleles are different in terms of structure compared to those described in PKW. Some correspond to a mix between alleles eBSOLV-1 and -2, making it difficult to discern whether these novel alleles are infectious or not. However, we observed that diversity converged on the same part of the eBSV sequence. Indeed, the *Musa* integration sites are conserved among all the B genomes tested. Fragments 1-OL, 2-OL, 3-OL and 6-OL are present in almost all plants (figure 4A). The other fragments (4-OL, 5-OL and 8-OL) usually used to discriminate eBSOLV-1 and -2 alleles are also involved in the observed allele diversity. These results could be correlated with the presence of a hot spot of recombination in this part of the eBSOLV structure. We suspect that this area, which is composed of a complete BSV sequence in eBSOLV-1, contributes to the release of functional viral genomes by homologous recombination. eBSOLV is located within a transposable-element-rich

genomic area (Chabannes et al., 2012), i.e. an area known as an area with a tendency to recombine (D'Hont et al., 2012 ; Mézard, 2006).

eBSImV distribution among diverse B genome

Analysis of eBSImV is simplified by the fact that we cannot discriminate between the two alleles present in PKW. Thus, it is not possible to differentiate between the hemizygous and homozygous states of the accessions.

This integration is completely absent from 20 accessions with the B genome (AB, AAB, ABB and BB). For the first time, we observed five BB accessions with no or only a few traces of PKW-related eBSImV. Only two AAB plants have exactly the same integration pattern as PKW, and the majority showed a large structural loss. AAB plants from the Maia popoulou subspecies and their relatives harbored almost complete integration, with some mutations in the eBSImV sequences. Unlike eBSGFV, we observed some accessions where PCR markers, and particularly PCR markers at junctions, were the only viral sequences recorded in genotype eBSImV, and no viral fragments were detected in Southern blots. This indicated an elimination or loss of a significant fraction of eBSV following integration, with the exception of the region flanking the integration. This is a strong argument that BSImV integration happened as for other BSV species in all the BB accessions, albeit with a large degree of polymorphism of integration recorded.

The loss could be explained by a different locus of integration of eBSGFV and eBSOLV, but on the same chromosome. Integration could result in a negative impact on the fitness of the plant that may be lethal for BB accessions, resulting either in partial or total elimination from the B genome at an early time point following integration, and before banana hybridization steps.

This might be the case for the hybridization area of the Indian accessions and also for plantains. However, this suppression cannot be really correlated with the production of viral particles because wild BB plants also lack eBSImV and the large modifications, particularly with plants of Eti kehel and Lal velchi, which are known to be phylogenetically linked with the Indian group (Hippolyte et al., 2012).

We should also mention the possibility that BSImV never occured in certain BB accessions; the only survivor of this BB group could be BB-Honduras. This seedy BB ancestor may be the B parent of the AAB accessions without eBSImV. This hypothesis might explain, for example, the absence of eBSImV in the plantain group. However, only two AAB plants have complete integration, and the majority of plants with two B genomes as ABB and BB have complete integration. Thus, we assume that this integration has no deleterious effect on these plants and could be conserved.

Synthetic eBSV evolutionary history in M. balbisiana genomes

The results of these studies confirm massive BSV integration of at least one of the three BSV species into all B genomes among diverse *M. balbisiana* accessions. We always observed the same locus of integration for each BSV species, which speaks in favor of an integration event of BSOLV, BSGFV and BSImV before *M. balbisiana* diversification. All plants with *M. acuminata* or out-group genotypes had no eBSV of these BSV species. This included the having particular genotype of AAB Pisang Nangka accession, which has no trace of any integration of BSOLV, BSGFV and BSImV. These observations confirm an integration event after the speciation of *M. acuminata*/*M. balbisiana*.

For the three BSV, we observed that interspecific hybrids allowed us to access eBSV diversity than BB accessions, as suspected most notably for eBSOLV. This was observed especially for AAB hybrids with a very specific eBSV structure or complete lack of eBSV. This means that several modified eBSV structures found in hybrids are not found in our *M. balbisiana* sample. However, it can be inferred that this diversification of eBSV structures happened in the frame of wild diploid *M. balbisiana* long before the creation of interspecific hybrids. Historically, the divergence between the two species is dated around 4,6 Mya (Lescot et al., 2008), meaning a very long period of sexual diversification, natural selection, drift and diffusion from the original area, probably in continental South Asia. This diversification is attested by the presence in our sample of BB accessions with important structural rearrangements. Creation of interspecific hybrids, which are steril and necessarily maintained by human activities, cannot be dated earlier than 7000 years (Perrier et al., 2009). However, even if the creation of diversity is negligible in this recent phase, the selection in this diversity has certainly been active, particularly for AAB hybrids known to produce viral particles when bringing infectious allele of eBSV (Lheureux et al., 2003, Cote et al., 2010).

Therefore, the rearrangement or complete absence of eBSV as observed for AAB accessions, especially from the Indian group, can be linked to a long selection process, post-hybridization. This selection has maintained only those hybrids derived from particular accessions with defective or missing eBSV in a BB population where all alleles were present, as attested by Indian ABB accessions bringing the same eBSV as PKW and even infectious alleles of eBSGFV.

These results reveal a marked evolution of eBSGFV in the Indian region, and indeed India was, and remains, a very active secondary center of banana diversification, with a long-term process of selection leading to a wide range of diploid and triploid hybrid cultivars. For AAB, it could be suggested that this strong selection has eliminated infectious eBSV alleles able to produce viral particles in favour of truncated non-infectious eBGFV structures.

As previously illustrated for eBSGFV, this observation cannot be extended directly to hybridization areas in South East Asia, where AAB hybrids harbouring infectious alleles can

be found, even if most eBSV structures in AAB accessions are more rearranged than in ABB or BB. AAB African plantains as well as Pacific plantains are the issue of ancient hybridizations, at least 2500 years old (Perrier et al., 2011), and are found far from their areas of origin. They have been fixed in the state of their initial genotype when it moved from SE Asia to Africa or Oceania, without subsequent sexual recombination. We cannot exclude that eBSV variants were later selected in SE Asia local populations, as in India, but that have now become extinct or, more probably, absent in the main *Musa* collections. It is also possible that, in these original areas, the large diversity of cultivated bananas, and particularly the many AA cultivars, have not led to mass effects of specific cultivars generating a selection pressure sufficient for BSV to become a problem.

Among Asian AAB, Kunaimp and Pisang Nangka are the only accessions without eBSOLV and eBSGFV integrations. Chabannes et al. (2012) showed that these two eBSVs are localized on the same chromosome. From molecular markers on their nuclear genome, it has already been suggested that they might have an incomplete *M. balbisiana* genome. Therefore, it can be suspected that these accessions lack the chromosome containing the two eBSVs integrations. It might be an extreme case of selection in order to lose the infectious eBSV.

eBSV hallmarks as markers of B diversity

We observed that phylogenetic relationships and geographical information on banana accessions has helped us to better understand the evolutionary history of eBSV. Conversely, we demonstrated that eBSV distribution can also help *Musa* phylogeny by proposing for the first time a phylogenetic relationship based on the B genome.

Previous microsatellite-based analysis (Gayral et al., 2010; Hippolyte et al., 2012) has been extended in this study with new B genomes through BB plants and AAB or ABB hybrids, to produce the most resolved picture of phylogeny of *M. balbisiana* and *M. acuminata* species and their hybrids to date.

This novel phylogeny was extracted for our 77 analysed accessions and results mainly from the genetic differences between *M. babisiana* and *M. acuminata*—two very divergent species in the *Musa* genus (Li et al., 2010). Between the pure *M. acuminata* and the pure *M. balbisiana* poles, interspecific hybrids form a series of successive clusters. A first level of clustering depends on the genotype: AB, AAB or ABB. The geographical origin of the accessions defines a second level of clustering, with the two hybridization areas in SE Asia and in India (Perrier et al., 2011).

This phylogeny does not provide evidence of relationships between diploid AA or BB and their interspecific hybrids. Previous studies have shown some phylogenetic links between particular *M. acuminata* sub-species and some triploid AAA or AAB cultivars. In some cases,

it was even possible to identify the parents of these triploids (Perrier et al., 2009). However, it was not possible to show links between BB accessions and interspecific diploid or triploid hybrids, except in the particular case of the Indian BB Lal Velchi, which was shown to be a good candidate as parent of Indian AAB Pome cv and Kunnan AB cv.

This study provides evidence that eBSVs, which are distributed specifically in B genomes, can help resolve B genome phylogeny. The tree mixing data obtained for the three eBSVs (figure 9) gives a picture of *Musa* diversity based on viral integration and coevolution. As already mentioned, the structure of this tree is approximately coherent with the tree from *Musa* SSR markers, but with two main differences. The first divergence concerns the structure of the geographic origin of the accessions, which is locally disturbed. The general structure opposing *M. acuminata* to *M. balbisiana* is represented in this tree firstly by an opposing number of eBSV, with none in *M. acuminata* and, on the other hand, the *M. balbisiana* Msat-4 group, which has exactly the same integrations as PKW. Clustering of hybrid groups at higher levels is also a function of the number of integrations. So, when natural selection led to the absence of eBSV in some AAB accessions, these losses occurred independently in different areas, inducing groupings based on convergence rather than on true phylogenetic relationships. For example, this could be the case for AAB African plantains of SE Asian origin, and AAB Pome, from India, which are close in the tree because of the absence of eBSImV in their genomes.

The second difference is of particular interest. The organization of BB accessions into distinct groups is limited to Msat-4 and Msat-7, with all other BB being dispatched into ABB or AAB groups. A large number of AAB and ABB groups is closely linked to grouped or single BB accessions, and we assumed that these BB shared the same ancestor BB plants at the origin of these triploid hybrids. For example, ABB Pelipita has exactly the same integrations as PKW; like BB, accessions of the Msat-4 group, Saba, Burro Cemsa and Daru accessions of the ABB group Bluggoe are also linked to this subset. BB 63-80 is linked closely to AAB plantains, BB Honduras to Indian AAB Pome, BB Butuhan to ABB Auko, Bengani, and Pisang Kepok Bung, BB 211 to Indian ABB Awak. The hybrid component is sometimes restricted to just one accession, like ABB Dole with BB 342 and BB LBA. Moreover, the grouping around particular BB accessions leads to the merging of several hybrid groups sharing the same BB ancestor. For example, AAB Kapas, which belongs to the group Laknao from the Philippines, is very close to AAB Plantains, which are thought to originate from the same region. The AAB group Pome includes AB Ekona and AB Safet Velchi, but also ABB Blue Java, which was collected in Fiji but characterized as belonging to the Ney Mannan ABB group. All these accessions shared the same origin in India, which supports the hypothesis of an identical *M. balbisiana* ancestor. To the best of our knowledge, this is the first time that *M. balbisiana* ancestors of the main triploid cultivars can be assumed. The theoretical implications for further elucidation of the historical and geographical process of

Musa domestication, as well as practical implications for genetic improvement programmes, are obviously numerous.

Some BB groups remain isolated; they were not recruited in hybridizations with *M. acuminata* genomes. It is also possible that the hybrids generated are absent from our sample. Conversely, a single hybrid group, the AAB Silk from India, is not linked to any particular BB accession. A first hypothesis is that the ancestral BB is extinct or absent from the sample. An other possibility might be a specific modification within the triploid genome. The intense agricultural selection in this area has already been mentioned and could explain this specificity. Several accessions classified as indeterminate for SSR markers (Pisang Slendang, Luba, Kunaimp, Tigua) were also confirmed; indeed, these plants, which exhibit very specific eBSV structures, are positioned on specific intermediate branches. The integrity of their *M. balbisiana* genome has already been questioned.

In conclusion, eBSV markers appear to be efficient tools with which to elucidate the poorly resolved phylogeny of *M. balbisiana.* This efficiency could be explained by the analogy of eBSV structure to that of transposable elements, which allows their high polymorphism to be exploited in REMAP technologies to study populations with low diversity (Hamon et al., 2011).

References

Bejarano ER, Khashoggi A, Witty M, Lichtenstein C (1996) Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *PNAS* **93**, 759-764.

Bioversity Musa Germplasm Information System (MGIS) http://www.crop-diversity.org/banana/.

Carreel F, de Leon DG, Lagoda P*, et al.* (2002) Ascertaining maternal and paternal lineage within Musaby chloroplast and mitochondrial DNA RFLP analyses. *Genome* **45**, 679-692.

Carreel F, Fauré S, González de León D*, et al.* (1994) Evaluation de la diversité génétique chez les bananiers diploïdes (Musa sp). *Genetics Selection Evolution* **26**, 125-136.

Chabannes M, Baurens F-C, Duroy P-O*, et al.* Three infectious viral species lying in wait in the banana genome. *Genome Research*, 1-41.

Côte FX, Galzi S, Folliot M*, et al.* (2010) Micropropagation by tissue culture triggers differential expression of infectious endogenous Banana streak virus sequences (eBSV) present in the B genome of natural and synthetic interspecific banana plantains. *Molecular plant pathology* **11**, 137-144.

D'Hont A (2005) Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. *Cytogenetic and Genome Research* **109**, 27-33.

D'Hont A, Denoeud F, Aury J-M*, et al.* (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, 1-7.

D'Hont A, Paget-Goy A, Escoute J, Carreel F (2000) The interspecific genome structure of cultivated banana, Musa spp. revealed by genomic DNA in situ hybridization. *TAG Theoretical and Applied Genetics* **100**, 177-183.

Dallot SS, Acuña P, Rivera CC*, et al.* (2000) Evidence that the proliferation stage of micropropagation procedure is determinant in the expression of banana streak virus integrated into the genome of the FHIA 21 hybrid (Musa AAAB). *Archives of Virology* **146**, 2179-2190.

Dolezel J, Lysák MA, Van den Houwe I, DolezelovÃ M, Roux N (1997) Use of flow cytometry for rapid ploidy determination in Musa species. *Infomusa* **6**, 6-9.

Fauquet CM (2005) *Virus Taxonomy: VIIIth Report of the International Committee on Taxonomy of Viruses* Academic Press.

Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. *Nature Publishing Group* **13**, 283-296.

Gambley CF, Geering ADW, Steele V, Thomas JE (2008) Identification of viral and non-viral reverse transcribing elements in pineapple (Ananas comosus), including members of two new badnavirus species. *Archives of Virology* **153**, 1599-1604.

Gawel NJ, Jarret RL (1991) A modified CTAB DNA extraction procedure for Musa and Ipomoea. *Plant Molecular Biology Reporter* **9**, 262-266.

Gayral P, Blondin L, Guidolin O*, et al.* (2010) Evolution of Endogenous Sequences of Banana streak virus: What Can We Learn from Banana (Musa sp.) Evolution? *Journal of Virology* **84**, 7346-7359.

Gayral P, Noa-Carrazana JC, Lescot M*, et al.* (2008) A Single Banana Streak Virus Integration Event in the Banana Genome as the Origin of Infectious Endogenous Pararetrovirus. *Journal of Virology* **82**, 6697-6710.

Geering ADW, Pooggin MM, Olszewski NE, Lockhart BEL, Thomas JE (2005) Characterisation of Banana streak Mysore virus and evidence that its DNA is integrated in the B genome of cultivated Musa. *Archives of Virology* **150**, 787-796.

Hamon P, Duroy P-O, Dubreuil-Tranchant C*, et al.* (2011) Two novel Ty1-copia retrotransposons isolated from coffee trees can effectively reveal evolutionary relationships in the Coffea genus (Rubiaceae). *Molecular Genetics and Genomics* **285**, 447-460.

Hansen CN, Harper G, Heslop-Harrison JS (2005) Characterisation of pararetrovirus-like sequences in the genome of potato (Solanum tuberosum). *Cytogenetic and Genome Research* **110**, 559-565.

Harper G, Hull R, Lockhart B, Olszewski N (2002) Viral sequences integrated into plant genomes. *Annual Review of Phytopathology* **40**, 119-136.

Harper G, Osuji JO, Heslop-Harrison JS, Hull R (1999) Integration of banana streak badnavirus into the Musa genome: molecular and cytogenetic evidence. *Virology* **255**, 207-213.

Harper G, Hull R (1998) Cloning and sequence analysis of banana streak virus DNA. *Virus Genes* **17**, 271-278.

Hippolyte I, Bakry F, Seguin M*, et al.* (2010) A saturated SSR/DArT linkage map of Musa acuminata addressing genome rearrangements among bananas. *BMC Plant Biology* **10**, 65.

Hippolyte I, Jenny C, Gardes L*, et al.* (2012) Foundation characteristics of edible Musa triploids revealed from allelic distribution of SSR markers. *Annals of Botany* **109**, 937-951.

Hohn T, Richert-Pöggeler KR, Staginnus C*, et al.* (2008) Evolution of integrated plant viruses. *Plant Virus Evolution. Berlin: Springer*, 53-81.

Horie M, Tomonaga K (2011) Non-Retroviral Fossils in Vertebrate Genomes. *Viruses* **3**, 1836-1848.

Jakowitsch J, Mette MF, Van der Winden J, Matzke MA, Matzke AJ (1999) Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. *PNAS* **96**, 13241-13246.

Kunii M, Kanda M, Nagano H*, et al.* (2004) Reconstruction of putative DNA virus from endogenous rice tungro bacilliform virus-like sequences in the rice genome: implications for integration and evolution. *BMC Genomics* **5**, 80.

Lagoda PJ, Noyer JL, Dambier D*, et al.* (1998) Sequence tagged microsatellite site (STMS) markers in the Musaceae. *Molecular Ecology* **7**, 659-663.

Lescot M, Piffanelli P, Ciampi AY*, et al.* (2008) Insights into the Musa genome: syntenic relationships to rice and between Musa species. *BMC Genomics* **9**, 58.

Lheureux F, Carreel F, Jenny C, Lockhart BEL, Iskra-Caruana ML (2003) Identification of genetic markers linked to banana streak disease expression in inter-specific Musa hybrids. *TAG Theoretical and Applied Genetics* **106**, 594-598.

Lheureux F, Laboureau N, Muller E, Lockhart BEL, Iskra-Caruana ML (2007) Molecular characterization of banana streak acuminata Vietnam virus isolated from Musa acuminata siamea (banana cultivar). *Archives of Virology* **152**, 1409-1416.

Li L-F, Häkkinen M, Yuan Y-M, Hao G, Ge X-J (2010) Molecular phylogeny and systematics of the banana family (Musaceae) inferred from multiple nuclear and chloroplast DNA fragments, with a special reference to the genus Musa. *molecular phylogenetics and evolution* **57**, 1-10.

Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* **21**, 2128-2129.

Lockhart BEL, Menke J, Dahal G, Olszewski NE (2000) Characterization and genomic analysis of tobacco vein clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. *Journal of General Virology* **81**, 1579-1585.

Lockhart BEL, Olszewski NE (1993) Serological and genomic heterogeneity of banana streak badnavirus: implications for virus detection in Musa germplasm. *… on Genetic Improvement of Bananas for ….*

Mézard CC (2006) Meiotic recombination hotspots in plants. *Biochemical Society Transactions* **34**, 531-534.

Ndowora T, Dahal G, LaFleur D*, et al.* (1999) Evidence that badnavirus infection in Musa can originate from integrated pararetroviral sequences. *Virology* **255**, 214-220.

Noreen F, Akbergenov R, Hohn T, Richert-Pöggeler KR (2007) Distinct expression of endogenous Petunia vein clearing virus and the DNA transposon dTph1 in two Petunia hybrida lines is correlated with differences in histone modification and siRNA production. *Epigenomes of endogenous PVCV* **50**, 219-229.

Perrier X, Bakry F, Carreel F, Jenny C, Horry JP, Lebot V, and Hippolyte I, (2009) Combining biological approaches to shed light on the evolution of edible bananas. *Ethnobotany Research & Applications* **7**, 199-216.

Perrier X, De Langhe E, Donohue M*, et al.* (2011) Multidisciplinary perspectives on banana (Musa spp.) domestication. *Proceedings of the National Academy of Sciences* **108**, 11311-11318.

Perrier X, Jacquemoud-Collet JP (2006) DARwin software. *http://darwin.cirad.fr/*.

Richert-Pöggeler KR, Noreen F, Schwarzacher T, Harper G, Hohn T (2003) Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. *EMBO Journal* **22**, 4836-4845.

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406-425.

Staginnus C, Gregor W, Mette MF*, et al.* (2007) Endogenous pararetroviral sequences in tomato (Solanum lycopersicum) and related species. *BMC Plant Biology* **7**, 24.

Supplementary data

	Finals fragments					dCAPs marker	
	1-GF-tot	2-GF-tot	5-GF-tot	3-GF-tot	4-GF-tot	Dif-GF (7)	Dif-GF (9)
	Southern blot fragments						
	1-GF	2-GF	5-GF	3-GF	4-GF	NA	NA
Accession	PCR markers						
abreviation	VM1/VV1	VV2/VV3	VV5	VV4	VM2/VV	NA	NA
OG Tex	0	0	0	0	0	2	2
OG Lat	0	0	0	0	0	2	2
OG Orn	0	0	0	0	0	2	2
OG Man	0	0	0	0	0	2	2
OG Vel	0	0	0	0	0	2	2
OG NNM	0	0	0	0	0	2	2
OG Baj	0	0	0	0	0	2	2
AA Bank	0	0	0	0	0	2	2
AA Agu	0	0	0	0	0	2	2
AA LoTA	0	0	0	0	0	2	2
AA Maia	0	0	0	0	0	2	2
AA Pah	0	0	0	0	0	2	2
AA IDN	0	0	0	0	0	2	2
ABB Foug	1	1	1	1	1	1	1
ABB Khom	1	1	1	1	1	1	1
ABB Peli	1	1	1	1	1	1	1
ABB Burro	1	1	1	1	1	1	1
ABB Dole	1	1	1	1	1	1	1
ABB Blue	1	1	1	1	1	1	1
ABB Bung	1	1	0	1	1	1	2
ABB Daru	1	1	1	1	1	1	1
ABB Beng	1	1	0	1	1	1	2
ABB Auko	1	1	0	1	1	1	2
ABB Saba	1	1	1	1	1	1	1
AAB Figue	0	0	0	0	0	2	2
AAB Tay	0	0	0	0	0	2	2
AAB Foco	1	1	1	1	1	2	1
AAB Prata	1	1	1	1	1	2	1
AAB Lady	0	0	0	0	0	2	2
AAB Yang2	0	0	0	0	0	2	2
AAB Moa	1	1	2	1	1	1	2
AAB Pop	2	1	0	1	2	1	2
AAB Tig	2	1	2	1	1	1	2
AAB Bong	1	1	2	1	1	1	2
AAB Gama	1	1	ND	1	1	1	2
AAB Corn	1	1	0	1	1	1	2
AAB KM	1	1	0	1	1	1	2
AAB Ori	1	1	0	1	1	1	2
AAB Luba	2	0	0	1	1	2	2
AAB Kap	1	1	0	1	1	1	2
AAB PRB	1	1	ND	1	1	1	2
AAB PR	1	1	1	1	2	2	2
AAB Kun	1	1	0	1	1	1	2
AAB ChuM	1	1	ND	1	1	1	2
AAB Nang	0	0	0	0	0	2	2
AAB Ceyl	0	0	0	0	0	2	2
AAB Slen	0	0	0	0	0	2	2
AAB Mur	1	1	ND	1	1	1	2
AAB Porp	0	0	0	0	0	2	2
AAB Dima	1	1	0	1	1	1	2
AB Safet	1	1	1	1	1	1	1
AB Kunn	0	0	0	0	0	2	2
AB Eko	1	1	1	1	1	1	1
BB Hond	1	1	1	1	1	1	1
BB 211	1	1	1	1	1	1	1
BB 1016	1	1	2	1	1	1	2
BB 626	1	1	2	1	1	1	2
BB 63-80	1	1	0	1	1	1	2
BB I63	1	1	2	1	1	1	2
BB Mont	1	1	2	1	1	1	1
BB Cam	1	1	0	1	1	2	2
BB 545	1	1	1	1	1	1	1
BB 852	1	1	1	1	1	1	1
BB Lal	2	1	0	1	1	1	2
BB KT	1	1	1	1	1	1	1
BB Batu	1	1	1	1	1	1	1
BB PK	1	1	1	1	1	1	1
BB PKW	1	1	1	1	1	1	1
BB But	1	1	0	1	1	1	2
BB EK	1	1	1	1	1	1	1
BB 342	1	1	2	1	1	1	1
BB LBA	1	1	1	1	1	1	1
BB ButIA	1	1	1	1	1	1	1
BB Chi1	1	1	2	1	1	1	1
BB Chi2	1	1	1	1	1	1	1
BB Chi3	1	1	1	1	1	1	1
BB LCK	1	1	2	1	1	1	2
	1=presence / 0=Absence 2 = presence of only southern blot fragment					1=presence / 2=Absence	

Table 1sup: Data for PKW-related eBSGFV genotypes

Accession abbreviated names according to table 1. ND: no determinated data.

	Finals Fragments										dCAPs marker	
	1-OL-tot	2-OL-tot	3-OL-tot	7-OL-tot	8-OL-tot	4-OL-tot	5-OL-tot	9-OL-tot	6-OL-tot	10-OL-tot	Dif-OL (2)	Dif-OL (1)
	Southern blot fragments										NA	NA
	1-OL	2-OL	3-OL	7-OL	8-OL	4-OL	5-OL	9-OL	6-OL	10-OL		
Accession abreviation	PCR markers										NA	NA
	16-17	/	27-28	16-17/29-30	20-21	23-24/29-30	/	/	25-26	/		
OG Tex	0	0	0	0	0	0	0	0	0	0	2	2
OG Lat	0	0	0	0	0	0	0	0	0	0	2	2
OG Orn	0	0	0	0	0	0	0	0	0	0	2	2
OG Man	0	0	0	0	0	0	0	0	0	0	2	2
OG Vel	0	0	0	0	0	0	0	0	0	0	2	2
OG NNM	0	0	0	0	0	0	0	0	0	0	2	2
OG Baj	0	0	0	0	0	0	0	0	0	0	2	2
AA Bank	0	0	0	0	0	0	0	0	0	0	2	2
AA Agu	0	0	0	0	0	0	0	0	0	0	2	2
AA LoTA	0	0	0	0	0	0	0	0	0	0	2	2
AA Maia	0	0	0	0	0	0	0	0	0	0	2	2
AA Pah	0	0	0	0	0	0	0	0	0	0	2	2
AA IDN	0	0	0	0	0	0	0	0	0	0	2	2
ABB Foug	1	1	1	0	1	0	1	1	1	0	1	2
ABB Khom	1	1	1	0	1	0	1	1	1	0	1	2
ABB Peli	1	1	1	1	1	1	1	0	1	0	1	1
ABB Burro	1	1	1	1	1	1	0	0	1	0	1	2
ABB Dole	1	1	1	0	1	1	0	0	1	0	2	1
ABB Blue	1	1	1	0	2	0	1	1	1	0	1	1
ABB Bung	1	1	1	0	2	1	1	0	1	0	2	1
ABB Daru	1	1	1	1	1	0	0	0	1	0	1	2
ABB Beng	1	1	1	0	0	1	1	0	1	0	2	1
ABB Auko	1	1	1	0	0	1	1	0	1	0	2	1
ABB Saba	1	1	1	1	1	1	0	0	1	0	1	2
AAB Figue	1	1	1	0	1	0	1	1	1	0	1	2
AAB Tay	1	1	1	0	1	0	1	1	1	0	1	2
AAB Foco	1	1	1	0	1	0	1	1	1	0	1	2
AAB Prata	1	1	1	0	1	0	1	1	1	0	1	2
AAB Lady	1	1	1	0	2	0	0	0	1	0	2	1
AAB Yang2	1	1	1	0	2	0	0	0	2	0	2	1
AAB Moa	1	1	1	0	0	1	1	0	1	0	2	1
AAB Pop	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND
AAB Tig	1	1	1	0	0	0	1	0	2	0	2	1
AAB Bong	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND
AAB Gama	1	1	1	0	0	1	1	0	1	0	2	1
AAB Corn	1	1	2	0	0	1	1	0	2	0	2	1
AAB KM	1	1	1	0	0	1	1	0	2	0	2	1
AAB Ori	1	1	1	0	0	1	1	0	2	0	2	1
AAB Luba	2	0	0	0	0	0	0	0	2	0	2	2
AAB Kap	1	1	1	1	1	1	1	0	1	0	2	1
AAB PRB	1	0	0	0	0	0	1	0	1	0	2	2
AAB PR	2	1	1	0	1	0	1	1	1	0	1	2
AAB Kun	0	0	0	0	0	0	0	0	0	0	2	2
AAB ChuM	1	1	1	0	0	1	1	0	1	0	2	1
AAB Nang	0	0	0	0	0	0	0	0	0	0	2	2
AAB Ceyl	1	1	1	0	0	1	1	0	1	0	2	1
AAB Slen	2	0	0	0	0	0	0	0	0	0	2	2
AAB Mur	1	0	0	0	0	0	1	0	1	0	2	2
AAB Porp	2	1	1	0	0	0	1	1	1	0	1	2
AAB Dima	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND
AB Safet	1	1	1	0	0	0	1	1	1	0	1	2
AB Kunn	1	1	1	0	2	0	1	1	1	0	1	2
AB Eko	1	1	1	0	1	0	1	1	1	0	1	2
BB Hond	1	1	1	0	1	1	1	0	1	0	1	1
BB 211	1	1	1	1	1	1	1	1	1	0	1	1
BB 1016	1	1	1	0	0	1	1	0	1	0	2	1
BB 626	1	1	1	0	0	1	1	0	1	0	2	1
BB 63-80	1	1	1	0	0	1	1	0	1	0	2	1
BB I63	1	1	1	0	0	1	1	0	1	0	2	1
BB Mont	1	1	1	0	0	1	1	0	1	1	2	1
BB Cam	1	1	0	0	0	1	0	0	1	0	2	2
BB 545	1	1	1	1	1	0	0	0	1	0	1	2
BB 852	1	1	1	0	0	1	1	0	1	0	2	1
BB Lal	1	1	1	0	2	1	1	0	1	0	2	1
BB KT	1	1	1	1	1	1	1	0	1	0	1	1
BB Batu	1	1	1	0	1	1	1	0	1	0	1	1
BB PK	1	1	1	1	1	1	1	0	1	0	1	1
BB PKW	1	1	1	1	1	1	1	0	1	0	1	1
BB But	1	1	1	0	0	1	1	0	1	0	2	1
BB EK	1	1	1	0	2	1	1	0	1	0	2	1
BB 342	1	1	1	0	1	0	0	0	1	0	2	1
BB LBA	1	1	1	0	0	1	1	0	1	0	2	1
BB ButIA	1	1	1	0	1	0	0	0	1	0	2	1
BB Chi1	1	1	1	0	1	0	1	1	1	0	1	2
BB Chi2	1	1	1	0	1	0	1	1	1	0	1	2
BB Chi3	1	1	1	0	1	0	1	1	1	0	1	2
BB LCK	1	0	0	0	2	0	1	0	1	0	2	2
	1=presence / 0=Absence / 2 = presence of only southern blot fragment										1=presence / 2=Absence	

Table 2sup: Data for PKW-related eBSOLV genotypes

Accessions abbreviated names according to table 1. ND: no determinated data.

Accession abreviation	Finals fragments 1-Im-tot	2-Im-tot	3-Im-tot	7-Im -tot	4-Im-tot	5-Im-tot	6-Im-tot	PCR markers only M/F2	F5-M
	Southern blot fragments 1-Im	2-Im	3-Im	7-Im	4-Im	5-Im	6-Im	NA	NA
	PCR markers F1-F3	/	/	/	/	/	F3-F4/F4-F5	M/F2	F5-M
OG Tex	0	0	0	0	0	0	0	2	2
OG Lat	0	0	0	0	0	0	0	2	2
OG Orn	0	0	0	0	0	0	0	2	2
OG Man	0	0	0	0	0	0	0	2	2
OG Vel	0	0	0	0	0	0	0	2	2
OG NNM	0	0	0	0	0	0	0	2	2
OG Baj	0	0	0	0	0	0	0	2	2
AA Bank	0	0	0	0	0	0	0	2	2
AA Agu	0	0	0	0	0	0	0	2	2
AA LoTA	0	0	0	0	0	0	0	2	2
AA Maia	0	0	0	0	0	0	0	2	2
AA Pah	0	0	0	0	0	0	0	2	2
AA IDN	0	0	0	0	0	0	0	2	2
ABB Foug	1	1	1	1	0	1	1	1	1
ABB Khom	1	1	1	1	0	1	1	1	1
ABB Peli	1	1	1	1	ND	ND	1	1	1
ABB Burro	1	1	1	1	ND	ND	1	1	1
ABB Dole	1	1	1	1	1	1	1	1	1
ABB Blue	0	0	0	0	0	0	0	2	2
ABB Bung	1	1	1	1	1	1	1	1	1
ABB Daru	1	1	1	1	1	1	1	1	1
ABB Beng	1	1	1	1	1	1	1	1	1
ABB Auko	1	1	1	1	1	1	1	1	1
ABB Saba	1	1	1	1	1	0	1	1	1
AAB Figue	0	0	0	0	0	0	0	2	2
AAB Tay	0	0	0	0	0	0	0	2	2
AAB Foco	0	0	0	0	0	0	0	2	2
AAB Prata	0	0	0	0	0	0	0	2	2
AAB Lady	0	0	0	0	0	0	0	2	2
AAB Yang2	0	0	0	0	0	0	0	1	1
AAB Moa	1	0	1	0	1	1	1	1	1
AAB Pop	ND	ND	ND	ND	ND	ND	ND	ND	ND
AAB Tig	0	0	0	0	0	0	1	1	1
AAB Bong	0	0	0	0	0	0	0	2	2
AAB Gama	1	1	1	1	0	1	1	1	1
AAB Corn	0	0	0	0	0	0	0	2	2
AAB KM	0	0	0	0	0	0	0	2	2
AAB Ori	0	0	0	0	0	0	0	2	2
AAB Luba	0	0	0	0	0	0	0	2	2
AAB Kap	0	0	0	0	0	0	0	2	2
AAB PRB	1	1	1	1	1	1	1	1	1
AAB PR	0	0	0	0	0	0	0	2	2
AAB Kun	0	0	0	0	0	0	0	2	2
AAB ChuM	1	ND	ND	ND	ND	1	1	1	1
AAB Nang	0	0	0	0	0	0	0	2	2
AAB Ceyl	0	0	0	0	0	0	0	2	2
AAB Slen	1	0	0	0	0	1	1	1	1
AAB Mur	1	1	1	1	1	1	1	1	1
AAB Porp	0	0	0	0	0	0	0	1	1
AAB Dima	1	1	1	1	0	1	1	1	1
AB Safet	0	0	0	0	0	0	0	2	2
AB Kunn	0	0	0	0	0	0	0	1	1
AB Eko	0	0	0	0	0	0	0	2	2
BB Hond	0	0	0	0	0	0	0	2	2
BB 211	0	1	1	1	0	1	1	1	1
BB 1016	1	1	1	1	1	1	1	1	1
BB 626	1	1	1	1	1	1	1	1	1
BB 63-80	0	0	0	0	0	0	0	1	2
BB I63	0	0	0	0	0	0	0	1	2
BB Mont	1	1	1	1	1	1	1	1	1
BB Cam	1	1	1	1	1	1	1	1	1
BB 545	1	1	1	1	1	1	1	1	1
BB 852	1	1	1	1	1	1	1	1	1
BB Lal	0	0	0	0	0	0	0	1	1
BB KT	1	1	1	1	1	1	1	1	1
BB Batu	1	1	1	1	1	1	1	1	1
BB PK	1	1	1	1	1	1	1	1	1
BB PKW	1	1	1	1	1	1	1	1	1
BB But	1	1	1	1	1	1	1	1	1
BB EK	0	0	0	0	0	0	0	1	2
BB 342	1	1	1	1	0	1	1	1	1
BB LBA	1	1	1	1	1	1	1	1	1
BB ButIA	1	1	1	1	1	1	1	1	1
BB Chi1	1	1	1	1	1	1	1	1	1
BB Chi2	1	1	1	1	1	1	1	1	1
BB Chi3	1	1	1	1	1	1	1	1	1
BB LCK	1	1	1	1	1	1	1	1	1
	1=presence / 0=Absence / 2 = presence of only southern blot fragment							1=presence / 2=Absence	

Table 3sup: Data for PKW-related eBSImV genotypes

Accessions abbreviated names according to table 1. ND: no determinated data.

Points clés du chapitre 1

L'article 1 a permis une caractérisation fine des eBSV présent dans le génome de PKW :

- **PKW** présente seulement trois eBSV infectieux dans le génome: **eBSGFV, eBSOLV et eBSImV**.

- Chacune des intégrations est à **un seul locus** indépendant les uns des autres.

- Leur structure est **réarrangée** par rapport aux séquences virales libres de BSV.

- L'intégration est **allélique** : eBSOLV et eBSGFV possèdent un allèle capable de produire des particules virales fonctionnelles (eBSGFV-7 et l'eBSOLV-1) et un allèle pour lequel ce n'est pas possible (l'eBSGFV-9 et l'eBSOLV-2). Bien qu'allélique nous n'avons pas pu différencier les 2 allèles de l'eBSImV.

- L'eBSGFV et l'eBSOLV co-localisent sur **le même chromosome** (1), l'EBSImV est localisé sur le chromosome 2.

- Ils sont intégrés dans des régions génomiques **variées** qui peuvent être riches en gène (eBSGFV et eBSImV) ou riches en éléments transposables (eBSOLV).

- Leurs séquences divergent très peu des séquences **BSV exogènes**, ainsi qu'entre les allèles.

- La signature globale des eBSV de PKW montre des **intégrations séquentielles** : BSOLV, BSGFV et BSImV.

- Le bananier PKW a dû subir une **autofécondation**, seule explication à la présence des eBSV à l'état homo ou hétérozygote.

L'article 2 a permis de mettre en évidence les diversités de structures des eBSV infectieux dans la diversité *M. balbisiana.*

- **Tous** les bananiers ***M. balbisiana*** **diploïdes ont les 3 eBSV** à l'exception de l'accession Honduras pour l'eBSImV. Insertion des trois eBSV après la spéciation entre *M. acuminata* et *M. balbisiana* mais avant la **diversification** de l'espèce *M. balbisiana.*

- **Un seul événement** d'intégrations est à l'origine de la fixation de chacun des eBSV et la divergence est le fait **de réarrangements internes**.

- Les trois eBSV possèdent des histoires évolutives **distinctes** dans la diversité.

- L'eBSImV est absent d'un grand nombre de bananiers hybrides présents dans plusieurs zones géographiques. **La convergence évolutive** conduisant à la perte de cet eBSV peut expliquer ces similitudes.

- **L'eBSOLV** est présent chez tous les bananiers porteurs de génome B et possède un hot-spot de recombinaison sur une partie de sa séquence. C'est l'eBSV qui présente le plus de diversité.

- **L'eBSGFV** est très conservé dans la diversité mais toute une partie des hybrides provenant de la même zone d'hybridations ne possèdent pas l'intégration.

- Les réarrangements des eBSV ont eu lieu lors de l'évolution des bananiers fertiles *M. balbisiana* et non au sein des hybrides. Les pressions de sélection appliquées sur ces séquences semblent avoir été importantes chez certains bananiers diploïdes.

- **Les hybrides interspécifiques** susceptibles de produire des particules virales possèdent majoritairement des eBSV dégradés par rapport à ceux de PKW voire pas d'eBSV du tout. Il semble qu'une pression de sélection ait existé lors de la création de ces hybrides pour modifier les eBSV fonctionnels, alors que les hybrides de type ABB, qui ne produisent pas de particules virales, possèdent la même diversité d'eBSV que les bananiers *M. balbisiana* fertiles.

- L'utilisation des eBSV comme **marqueurs moléculaires** du génome de *M. balbisiana* a permis de mettre en évidence des relations phylogénétiques entre les bananiers hybrides et des diploïdes *M. balbisiana*. Nous avons ainsi pu assigner certains génomes B à des zones géographiques

CHAPITRE 2

Mode de régulation des eBSV

La découverte de séquences de virus dont le cycle de multiplication n'oblige pas à une intégration et qui colonisent le génome des êtres vivants dont les plantes, interroge sur l'origine et les conséquences d'une telle interaction avec le génome hôte. De nombreuses hypothèses ont été formulées afin de tenter d'expliquer cette présence tout d'abord en s'interrogeant sur les évènements et contextes à l'origine de l'intégration. Ces hypothèses, ainsi que les travaux s'y rapportant, sont présentés dans la partie 7-5 de l'introduction et le chapitre I. La complexité pour de telles séquences, une fois intégrées, de se retrouver fixées dans les lignées germinales est discutée dans l'article 1 et 2 de cette thèse. Ces séquences, à l'état hétérozygote et/ou homozygote chez PKW et pour les autres diploïdes *M. balbisiana* disponibles, seraient ainsi le témoignage d'autofécondation et de sélection de bananiers ancestraux porteur d'eBSV améliorant (ou sans effet sur) la fitness de la plante.

L'existence de génomes viraux complets au sein des génomes *M. balbisiana* questionne néanmoins sur le bénéfice/coût lié, en regard des contraintes induites par leurs présences dans ces génomes. En effet il existe un paradoxe à maintenir des eBSV dans le génome de ces bananiers car ces intégrations peuvent restituer des génomes viraux fonctionnels produisant des virions. Il apparaît que PKW et certains génotypes diploïdes pour le génome B sont des porteurs sains d'eBSV, alors que leurs intégrations se révèlent infectieuses pour des génotypes bananiers haploïdes pour le génome *M. balbisiana*. Chez les bananiers diploïdes tels que PKW, il a été également montré que ni les stress activateurs connus tels que la mise en culture in vitro (Umber, com personnelle, Cote et al, 2010), des stress hydriques ou des écarts de températures pas plus que des tests de transmissions utilisant les cochenilles vectrices de la maladie (Lheureux, 2002) ne permettaient d'induire une infection.

Le but de ce 2ème chapitre est d'identifier les raisons pour lesquelles les eBSV seraient conservés dans le génome malgré les effets négatifs qu'ils peuvent avoir sur la fitness de certains bananiers ayant un génome haploïde B. Nous nous sommes donc intéressés à identifier et caractériser les mécanismes susceptibles de contribuer à la conservation de ces eBSV chez les bananiers porteurs sains. La seule hypothèse formulée jusqu'à présent est celle d'une régulation épigénétique de type ARN interférant (ARNi) impliqué dans les mécanismes de défense virale des plantes (Staginnus et Richert-Pöggeler, 2006, Hohn et al, 2008, Geering et Teycheney, 2011). Cette régulation empêcherait non seulement l'activation des eBSV permettant la production de virions, mais elle confèrerait une résistance à la multiplication de virus issus d'une contamination extérieure. La régulation des eBSV induirait ainsi une résistance constitutive de la plante aux virus. Il apparaît que

de tels mécanismes de régulation, réversibles et transmissibles aux diverses générations, soient maintenus pour les séquences endogènes aux génomes de plante (Lisch et al, 2009).

Introduction

Il existe dans le génome des plantes différents types de séquences parasites ayant un lien avec les virus. Ces séquences sont capables de se multiplier et d'envahir le génome hôte. Les éléments transposables (ET) par exemple peuvent représenter une très grande partie du génome de certaines plantes comme c'est le cas chez le maïs où 85% du génome est composé d'ET (Schnable et al., 2009). Nous avons d'ailleurs vu dans l'introduction générale (§2-1) les différentes raisons de la présence des séquences virales ou assimilées dans le génome des êtres vivants notamment comme vecteurs majeurs de diversité au sein des génomes hôtes (Holmes, 2011). L'étude de leurs implications potentielles dans la défense immunitaire chez les animaux en est à ses débuts mais ils semblent que ces séquences soient utilisées dans certains cas pour la lutte antivirale (Perron et Lang, 2009 ; Aswad et Katzourakis, 2012). Chez les plantes, l'implication de séquences pararétrovirales endogènes (EPRV) dans les mécanismes de défense virale est l'hypothèse principale ces dernières années pour expliquer leur conservation dans les génomes (Staginnus et Richert-Pöggeler, 2006 ; Hohn et al., 2008).

Malgré tout, ces séquences proches des virus quelles qu'elles soient peuvent avoir des effets délétères sur la fitness de leur hôte parce qu'elles ont la capacité de se multiplier et d'envahir les génomes de différentes manières. Les rétroéléments (RE) à Longue région Terminale Répétée (LTR), qui font partie des ET, par exemple produisent tous les éléments nécessaires à la synthèse de particules virales excepté la protéine d'enveloppe, ils vont donc rester dans la cellule qu'ils colonisent sans pouvoir passer aux cellules environnantes. Leurs effets sont donc limités sur l'hôte car ils se restreignent à la cellule dans laquelle ils se multiplient. L'impact sur l'hôte est différent entre RE à LTR et Endogenous Viral Elements (EVE), avec des conséquences beaucoup plus néfastes sur l'hôte pour les EVE mais aussi beaucoup plus rares. En effet, potentiellement les EVE au moment de l'intégration au moins ont la capacité de produire des particules virales et donc d'infecter l'hôte de manière systémique. Malgré tout chez les animaux par exemple aucun EVE n'a été découvert comme pouvant directement produire des virions (Stoye, 2012). Du coté des plantes par contre il a été découvert que des EVE de type endogenous pararetrovirus (EPRV) peuvent, pour certains d'entre eux, produire des particules virales qui vont coloniser d'autres cellules et petit à petit envahir l'hôte, jusqu'à l'infection systémique (Hohn et al., 2008).

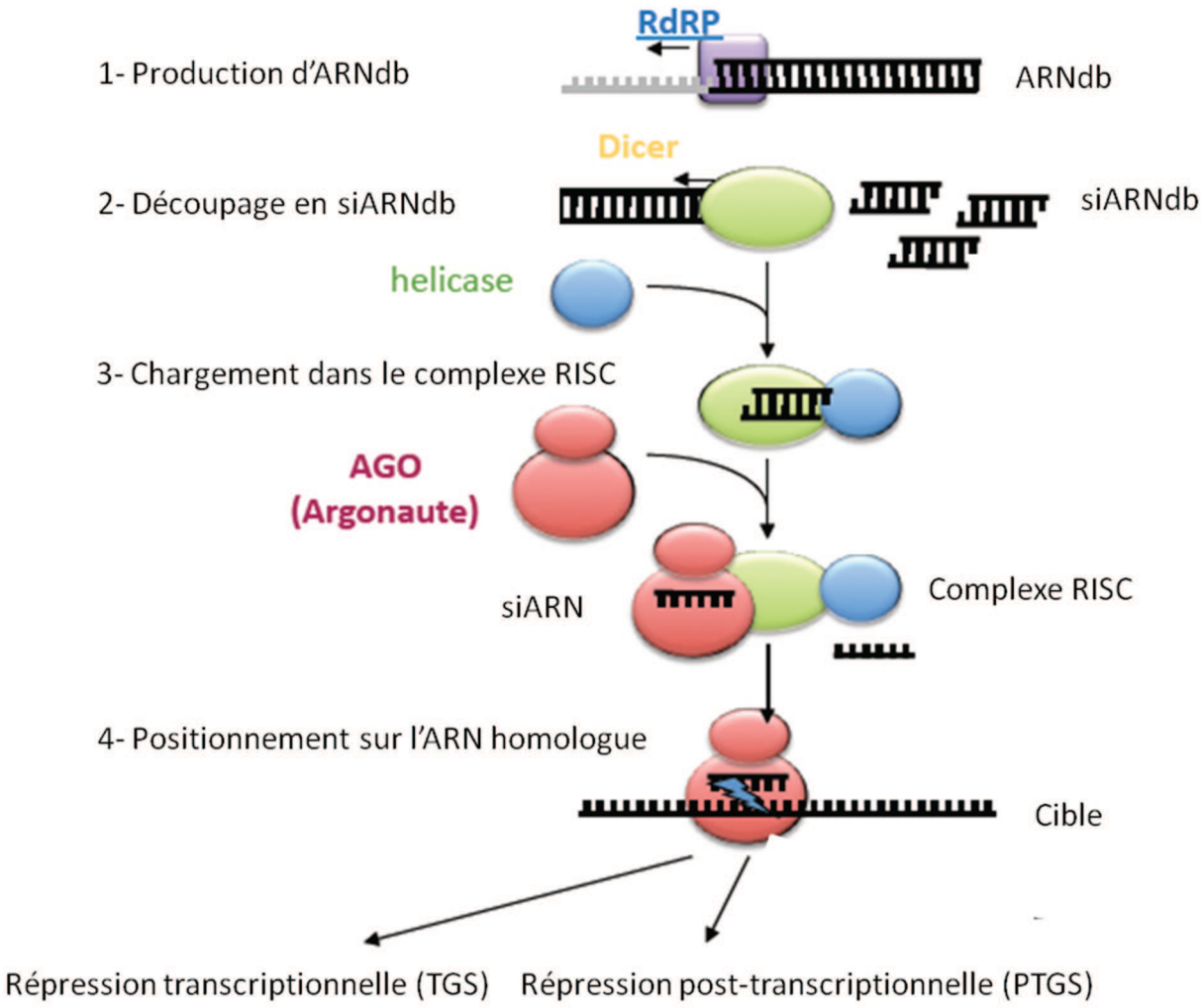


Figure 2-1 : Représentation des principales voies de l'ARN interférent (ARNi)

1- La production d'ARN double brin (ARNdb) par l'ARN polymérase ARN dependante (RdRP).

2-Découpage de l'ARNdb aberrant en siARNdb par des endonucléases de type DICER.

3-Formation du complexe RISC (RNA-Induced Silencing-Complex) qui se compose d'une protéine d'accroche avec les enzymes DICER, d'une hélicase permettant l'ouverture des siRNAdb, et d'une protéine Argonaute (AGO) pour l'accrochage et la reconnaissance du siARN. Le complexe recrute les siARNdb et élimine le brin sens.

4-Présentation des siARN à l'ARN homologue et reconnaissance. Selon la taille des siRNA et la nature de la protéine AGO, les régulations mises en place seront soit de nature transcriptionnelle (TGS) soit post-transcriptionnelle (PTGS).

Adaptée N. Baumberger, Sainsbury Laboratory, (2005)

1- L'implication de l'épigénétique dans la régulation des séquences virales

Afin de contrecarrer les effets délétères des séquences parasites, les plantes et les êtres vivants plus généralement ont mis en place des mécanismes empêchant l'expression des EVE et des ET dans leur génome. Un premier moyen de contrôler ces séquences est d'accumuler des mutations ou d'effectuer des réarrangements par recombinaison homologue afin d'éviter toute multiplication/amplification et in fine d'aboutir à leur disparition du génome hôte par sélection naturelle purifiante. Ce mécanisme a été étudié dans le chapitre 1 de cette thèse pour les eBSV. Nous avons ainsi pu montrer que malgré des réarrangements de séquences importants les eBSV étaient particulièrement bien conservés au sein du génome des bananiers *M. balbisiana* et que l'expression virale était possible pour certains allèles.

Le deuxième mécanisme fait appel à la régulation épigénétique et en particulier l'ARN interférent ou ARNi. Cette régulation est tout d'abord un moyen de défense contre l'invasion par des ADN étrangers (ET, virus, transgènes) mais elle est aussi exploitée largement pour la régulation des gènes. Nous avons étudié dans ce chapitre les mécanismes épigénétiques qui peuvent être impliqués dans la régulation de l'expression des eBSV. Nous avons recherché en quoi l'ARNi peut influer sur la structure et le maintien des eBSV dans les génomes B des bananiers.

1-1 Les mécanismes clés de l'ARNi

Le phénomène épigénétique fait appel à un ensemble de mécanismes qui ont pour conséquence la diminution ou l'extinction de l'expression de gènes ou « silencing », au niveau transcriptionnel (TGS) ou post-transcriptionnel (PTGS), obligeant toujours à des interactions séquences spécifiques médiées par l'ARN (figure 2-1) (Dunoyer et al., 2009). Ces différents processus sont regroupés sous la terminologie d'interférence ARN ou ARNi. L'ARNi est hautement conservé chez la quasi-totalité des organismes eucaryotes de l'algue unicellulaire *Chlamydomonas reinhartii* jusqu'à l'Homme (Dunoyer, 2009) où il existe sous différentes formes au sein des espèces et entre espèces. L'ARNi, principalement actif dans l'hétérochromatine, a ainsi un rôle clé dans la stabilité des génomes en réprimant le mouvement des éléments mobiles comme les rétrotransposons situés en grande partie dans ces zones du génome (Lippman et al., 2004). Les mécanismes clés de l'ARNi sont présentés dans la figure 2-1. La molécule clé initiatrice de l'ARNi est un ARN double brin (ARNdb), présent dans le cytoplasme des cellules (Fire et al., 1998). Cette molécule aberrante peut

être produite *in vivo* soit par la transcription de séquences génomiques organisées en répétitions inversées, à partir de promoteurs convergents, ou via l'action d'ARN polymérase ARN-dépendantes (RdRp) d'origine endogène ou virale qui convertiront un ARN simple brin en ARNdb (Bernstein et al., 2001). Cet ARNdb est segmenté via une endonucléase de type ARNase III appelée Dicer générant la seconde catégorie de molécules ubiquitaires de l'ARNi que sont les petits ARN (small RNA ou sARN) (Hamilton, 1999 ; Elbashir et al., 2001). Ces sARN de 21 à 24nt de longueur vont être recrutés par le complexe enzymatique du « silencing » nommé RISC (RNA induced silencing complex). RISC est doté d'une activité hélicase ATP dépendante et ARNase. Il élimine dans un premier temps les sARN sens et dirige les sARN antisens vers la cible et les y aligne de manière séquence-spécifique (Hammond et al., 2000) afin d'induire le clivage de l'ARNm cible. Selon la nature du complexe RISC impliqué ainsi que la taille des sARN produits, cette reconnaissance entrainera soit du PTGS par clivage endonucléotidique de l'ARN messager (ARNm) ciblé ou une inhibition de sa traduction (Mallory et al., 2008), soit du TGS, par méthylation de l'ADN ou modifications des histones (acétylation, alkylation) (Ekwall, 2004) (Figure 2-1).

1-2 L'ARNi et la régulation des éléments transposables de plantes

De nombreux résultats montrent l'importance de l'ARNi pour réguler les séquences parasites de toutes natures.

La majorité des travaux ont porté sur l'étude de la régulation des ET. Ils ont notamment permis de démontrer que le RNAi induisait la répression de l'expression des ET (pour revue voir Rigal et Mathieu, 2011) via différents processus tels que la méthylation des cytosines de l'ADN, la modification de la région N terminale des histones H3 lysine 9 par diméthylation et/ou acétylation et/ou phosphorylation (Johnson et al., 2002). Ces processus sont médiés principalement par la machinerie de méthylation de l'ADN guidée par l'ARN (RdDM). Cette machinerie est guidée par des sARN de 24nt. Ces sARN sont issus de la transcription des ET par l'ARN polymérase IV (Huettel et al., 2006). Cette polymérase qui ne possède pas de fonction exonucléase 3' vers 5' a la particularité de produire des ARN à partir de n'importe quelle séquence ADN avec une préférence pour les séquences ADN non-codantes (Wierzbicki, 2012). L'existence de cette combinaison de marques épigénétiques permet de définir l'état chromatique d'une zone de génome. Méthylation de l'ADN et modification des histones sont étroitement interconnectées. L'étude des mécanismes impliqués dans la mise en place ou le maintien de ces processus a montré les liens étroits entre présence de sARN de 24 nt, méthylation de l'hétérochromatine et silencing des ET (Zilberman et al., 2006),

indiquant que la méthylation était le mode principal de contrôle des ET. Cette régulation de type TGS est également suspectée pour la régulation des EPRV.

La transcription de certains ET peut cependant être accompagnée de la production de sARN de 21nt (McCue et al., 2012, Mirouze et al., 2009). Ces sARN de 21nt sont considérés comme des marques de PTGS qui peuvent être activés lorsque le TGS est « relâché » (Bourchis et Voinnet, 2010). De manière intéressante Llave (2010) a montré que le PTGS est le plus fréquemment impliqué dans les mécanismes de défense antiviraux ce qui suggère que ces deux mécanismes (TGS et PTGS) coopéreraient pour contrôler les ET. Cependant le rôle de ces sARN de 21nt est encore mal connu et nécessite des travaux complémentaires pour comprendre leur réel effet sur le contrôle des ET.

1-3 La régulation des EPRV : similitudes avec les ET

L'étude de la régulation des EPRV n'en est encore qu'à ses débuts, mais les similarités qui existent entre les EPRV et les ET invitent à un rapprochement de ces deux types de séquences. Tout d'abord, ET et EPRV co-localisent fréquemment dans les zones hétérochromatiques et péricentromériques (Richert-Pöggeler et al., 2003; Hansen et al., 2005; Staginnus et al., 2007). Ensuite le contexte de stress responsable de l'activation des EPRV montre une grande similarité avec ceux responsables de l'activation des ET. En effet, tous deux (EPRV et ET) sont activés suite à des croisements interspécifiques. Cela est bien documenté pour les EPRV (Richert-Pöggeler et al., 2003 ; Lockhart et al., 2000 ; Lheureux et al., 2003) et a été mis en évidence pour les ET lors de croisements interspécifiques entre *Arabidopsis thaliana* et *A. lyrata* (Lockton et Gaut, 2010). Lors de ces croisements, chaque individu de la descendance possède un contexte génomique et/ou une ploïdie au niveau des EPRV et des ET différents de ceux du parent initialement porteur. Cela constitue un premier stress génomique pouvant provoquer la levée du contrôle épigénétique des EPRV/ET existant chez le parent. Outre ce stress génomique, on observe aussi un réveil des EPRV et des ET lors de divers stress abiotiques. Pour les EPRV, ces stress peuvent être multiples comme les blessures et le rabattage des plants (Richert-Pöggeler et al., 2003), la mise en culture in-vitro (Dallot et al., 2000) ou les stress hydriques et thermiques (Noreen et al., 2007). Ces stress connus pour interagir sur les mécanismes de transposition pourraient affaiblir le contrôle épigénétique des EPRV et provoquer la production de virus. Côté ET, il a été montré chez *Arabidopsis thaliana* par exemple que le rétroélément à LTR ONSEN devenait actif et se multipliait dans le génome en cas de stress thermique. Ce stress enlève l'état de méthylation initial qui empêche l'expression du rétroélément (Ito et al., 2011). De

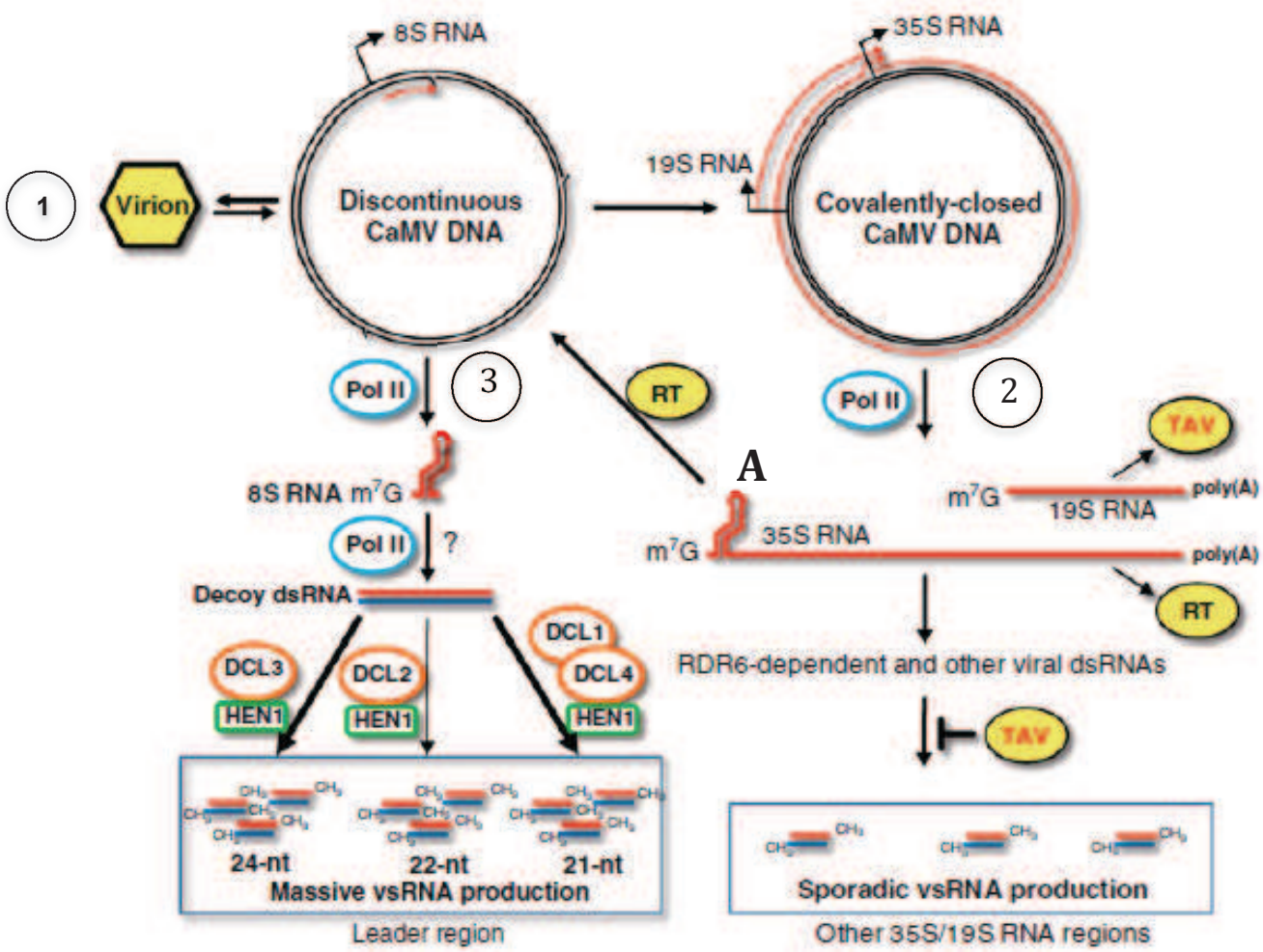


Figure 2-2 : Modèle d'interactions du *Cauliflower mosaic virus* (CaMV) avec la machinerie silencing génératrice des petits ARN

1- L'ADN viral est libéré du virion dans le noyau de la cellule hôte. Les zones non double brins ou « gap » de cet ADN viral laissé par la reverse transcription (RT) durant le cycle de réplication sont réparés par les enzymes hôtes pour créer de façon covalente des molécules fermées.

2- Cet ADN viral est transcrit par la Pol II en ARN 35S pré-génomique qui est un ARNm polycistronique codant plusieurs protéines virales, dont la RT, et qui sert de matrice à la transcription inverse en ADN. L'ADN viral est également transcrit par la Pol II en ARN 19S subgénomique, qui est l'ARNm codant la protéine virale transactivatrice/viroplasm (TAV).

3- L'arrêt brutal de la transcription par la Pol II (qui est potentiellement due à un gap d'ADN non réparé (Met-ARNt gap)) se traduit par la production d'ARN aberrants 8S ayant perdu leur queue poly A. Cet ARN 8S correspond à la région leader de l'ARN 35S et peut former une structure secondaire ressemblant à celle des viroïdes qui peut être convertie en ARN double brin (db) par Pol II. Cet ARNdb issu de l'ARN 8S sert de leurre pour mobiliser les quatre Dicer (DCL) dans la production massive de petits ARN viraux de 21, 22, 23 et 24-nt (vsARNs). Ces vsARN entrent ainsi en concurrence au niveau des protéines AGO avec les vsARNs provenant d'autres régions du génome viral. *D'après Rajeswaran et Pooggin, (2012)*

façon intéressante des ET sont retrouvés quasi-systématiquement à proximité des EPRV ce qui laisse dire à Hohn et al., (2008) que ces ET pourraient jouer un rôle dans l'activation des EPRV.

Ces différents travaux ont laissé suggérer que l'ARNi induirait une répression de l'expression des EPRV chez les plantes sauvages qui serait levée en cas de stress et induirait la production de particules virales. Ainsi, il a été montré tout d'abord par Noreen et al., (2007) que l'expression des séquences endogènes de *Petunia vein clearing virus* (ePVCV) était corrélée à des modifications des histones et la production de sARN. Les auteurs ont montré que les séquences d'ePVCV ainsi que le transposon dTph1 sont méthylés au niveau des histones de manière différentielle chez les plantes où les ePVCV sont activées et chez les plantes où elles ne le sont pas. Ces résultats impliquant une régulation basée sur la méthylation et la production de sARN tendent à montrer une régulation principale de type TGS telle que celle décrite chez les ET. La présence de sARN de 21nt chez les plantes infectées par le PVCV semble indiquer que d'autres voies de régulation épigénétique pourraient être impliquées dans la défense anti-virale comme celle du PTGS.

De la même manière, Mette et al., (2000, 2002) ont montré sur le pathosystème Tabac/*Tabacco vein clearing virus* que le gène rapporteur GUS, mis sous contrôle d'un promoteur dérivé des endogenous TVCV (eTVCV), était très rapidement méthylé et induisait la production de sARN lorsqu'il était inséré chez le tabac porteur d'eTVCV alors qu'il était exprimé normalement chez une plante non hôte sans eTVCV telle *qu'Arabidopsis thaliana*.

Ces différents résultats confirment l'existence d'une régulation ARNi des EPRV. Ces données ne permettent cependant pas une analyse fine de la compréhension des mécanismes régissant la régulation des EPRV. En 2006, Staginnus et Richert-Pöggeler ont néanmoins proposé un modèle reprenant les différentes régulations auxquelles les EPRV peuvent être confrontés. Ce modèle qui est toujours d'actualité est présenté figure 2-2. Il montre que les EPRV infectieux ou non-infectieux seraient régulés par des mécanismes de type TGS entrainant une méthylation de l'ADN ou des histones suite à la production de sARN de 21 à 25nt. Dans le cas d'apparition de particules virales suite à l'affaiblissement des régulations ARNi en condition de stress, les mécanismes de type TGS ou PTGS pourraient être impliqués dans la défense antivirale. Les auteurs suspectent l'existence de mécanismes de suppression du silencing à partir des EPRV mais ceci reste très spéculatif car aucune information n'est disponible à ce jour.

1-4 L'ARNi dans la défense antivirale

Il a été mis en évidence que les plantes au delà des systèmes de défense classiques possédaient des mécanismes de défense spécifiques aux virus basés sur la dégradation des ARNm viraux par le mécanisme de type PTGS (Voinnet, 2005 ; Li F et al., 2006). Cette observation a été faite de manière non-ambiguë suite à la détection systématique de quantité importante de sARN dérivant de génomes viraux dans les tissus de plantes infectées par des virus très variés. Ces sARN ont été nommés « virus derived sARN » ou vsARN. Ce mécanisme, découvert avec l'utilisation de transgènes activateurs du système, permet de protéger la plante contre les virus exogènes et contre la production d'ARN viraux (Voinnet, 2005). L'ARNi peut donc être considéré comme LE système immunitaire des plantes dont un aspect fondamental est qu'il s'agit d'un système totalement inné puisque la réponse est uniquement programmée par les caractéristiques et la séquence du génome infectieux et non par le génome de l'hôte (Voinnet, 2005). L'hypothèse du maintien par la plante des EPRV dans son génome pour permettre une résistance au virus prend toute sa dimension à la lumière de ces observations. Cette résistance serait une adaptation constitutive de la défense antivirale où la plante utiliserait les EPRV pour produire des sARN afin de lutter de manière précoce contre les virus libres pénétrant dans les cellules et empêcher ainsi leur multiplication. Il a en effet été montré chez *Arabidopsis thaliana* que l'intégration de séquences partielles des virus *Turnip yellow mosaic virus* et *Turnip mosaic virus* avec des précurseurs de micro ARN (miARN) (type de sARN particuliers constitutifs des plantes et synthétisés à partir d'ARNdb formant des motifs de type hairpin) permet d'obtenir une résistance en cas d'infection contre ces deux virus par la production de vsARN (Niu et al., 2006). Des expériences similaires ont été menées chez *Arabidopsis* afin d'obtenir une résistance au *Cucumber mosaic virus* (Qu et al., 2007) et chez la tomate pour lutter contre le *Potato spindle tuber viroid* (Schwind et al., 2009). Ces différentes expériences démontrent tout l'intérêt que peut avoir une plante à posséder des séquences virales dans son génome afin de lutter contre les virus.

Dans le cas de virus à ADN double brin la majorité des études a été réalisée sur le pathosystème *Arabidopsis thaliana*/*Cauliflower mosaic virus* (CaMV). Le CaMV appartient à la famille *Caulimoviridae* comme le BSV. Les travaux réalisés sur le CaMV montrent que les mécanismes de défense liés au ARNi, impliquent les 4 enzymes Dicer présentes dans le génome d'*Arabidopsis* qui agissent de concert pour lutter contre les infections virales (Blevin et al., 2006), en impliquant majoritairement l'enzyme Dicer 3 (DCL3). Contrairement aux virus à ARN pour lesquels il semble que l'enzyme Dicer 4 (DCL4) soit la seule impliquée dans la production de vsARN de 21nt. Les plantes infectées par le CaMV accumulent donc majoritairement des sARN de 21 à 24 nt

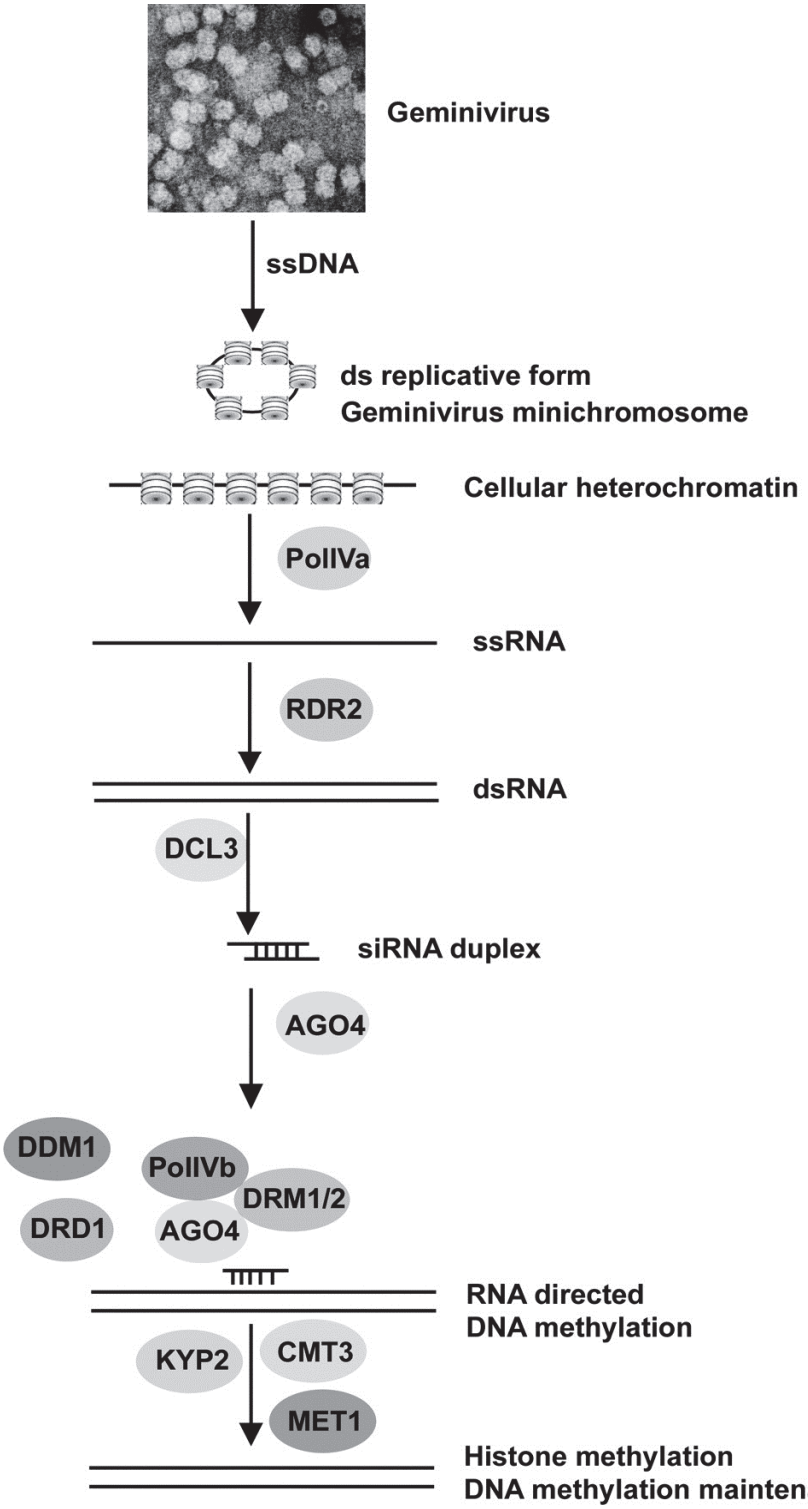


Figure 2-3 : Voie supposée de méthylation des geminivirus chez *Arabidopsis thaliana*

Des séquences cibles du génome viral et de la plante peuvent être transcrites par le complexe ARN polymérase IVa (Pol IVa). L'ARN simple brin produit (ssRNA) est converti en ARN double brin (dsRNA) par l'ARN polymérase ARN dependante RDR2. Les siARN de 24-nt issus de l'ARNdb suite à l'action de DCL3 sont chargés par les complexes contenant AGO4 où ils sont associés à Pol IVb. L'AGO4 associée au siARN cible va se positionner sur les séquences ADN homologues où les cytosines méthyltransférases sont recrutées. Cette étape de méthylation implique également un complexe de remodelage de la chromatine faisant intervenir DRD1 et DDM1. Les cytosines méthyltransférases CMT3 et MET1 sont principalement impliquées dans la maintenance de la méthylation des sites CNG et CG, respectivement. La méthylation CNG par CMT3 est également liée à la méthylation H3K9 des histones. *D'après Raja et al., (2008)*

produits par les gènes DCL1 et DCL4 pour les sARN de 21nt, DCL3 pour ceux de 24nt et DCL2 pour ceux de 22nt (figure 2-2). De plus, Moissard et Voinnet, (2006) ont montré que la production de vsARN au cours de l'infection du CaMV était initiée principalement à partir de la région appelée leader du transcrit polycistronique 35S. Cette région a été caractérisée par Pooggin et al., (1999) comme étant une zone produisant un transcrit nommé 8S sans rôle connu mais qui a la particularité de posséder une structure secondaire formant un hairpin qui servirait de substrat pour les enzymes Dicer. Ceci n'est pas le cas des autres types de virus pour lesquels la production de vsARN est répartie sur toute la séquence virale (Garcia-Ruiz et al., 2010). Une des hypothèses avancée par Blevin et al., (2011) pour expliquer cette production de vsARN très ciblée au niveau du transcrit 8S est que l'enzyme DCL1 qui est normalement impliquée dans la production de siRNA verrait son rôle détourné par le virus. En effet, celui-ci mobiliserait sa région intergénique (la leader région) non utilisée pour la réplication virale pour produire une quantité importante de transcrits 8S. Ces transcrits serviraient de leurres pour la production de vsARN afin d'éviter que l'enzyme DCL1 ne produise des vsARN à partir des zones transcrites du génome viral. Les auteurs ont ainsi montré que l'inactivation de l'enzyme DCL1 produit une baisse de la production de vsARN à partir de la zone leader du virus. Il s'agit là d'une stratégie nouvelle de contre-défense virale mise en place par le virus (Blevin et al., 2011).

Suite à ces résultats, il a été proposé par Hohn et Vasquez (2011) que des mécanismes de type TGS pourraient être impliqués dans la défense antivirale du CaMV. En effet, les travaux de Blevin et al., (2006 et 2011) ont mis en évidence des vsARN de 22nt et 24nt. Les vsARN de ces tailles sont connus comme spécifiquement impliqués dans la régulation de type TGS (Zilberman et al., 2006) qui a lieu dans le noyau des cellules (figure 2-4) . Ces hypothèses sont en accord avec le mode de réplication des *Caulimoviridae* qui forment un minichromosome dans le noyau de la cellule qu'ils infectent (Jacquot et al., 1997). Dernièrement, les travaux de Raja et al, (2008) ont montré l'implication potentielle du TGS dans les mécanismes de défense antivirale spécifique aux virus à ADN ayant un passage dans le noyau sous forme de minichromosome. Le BSV pour sa part possède trois ORF. Le rôle des protéines produites par l'ORF 1 et 2 n'a pas été caractérisé, l'ORF 3 code une polyprotéine contenant la reverse transcriptase ainsi que la ribonucléase H (Jacquot et al., 1997). Les études ont montré que la région intergénique (IG) comme pour le CaMV possédait un promoteur constitutif fort de type 35S et potentiellement codait aussi au niveau de la région leader pour un transcrit 8S ayant une structure secondaire de type

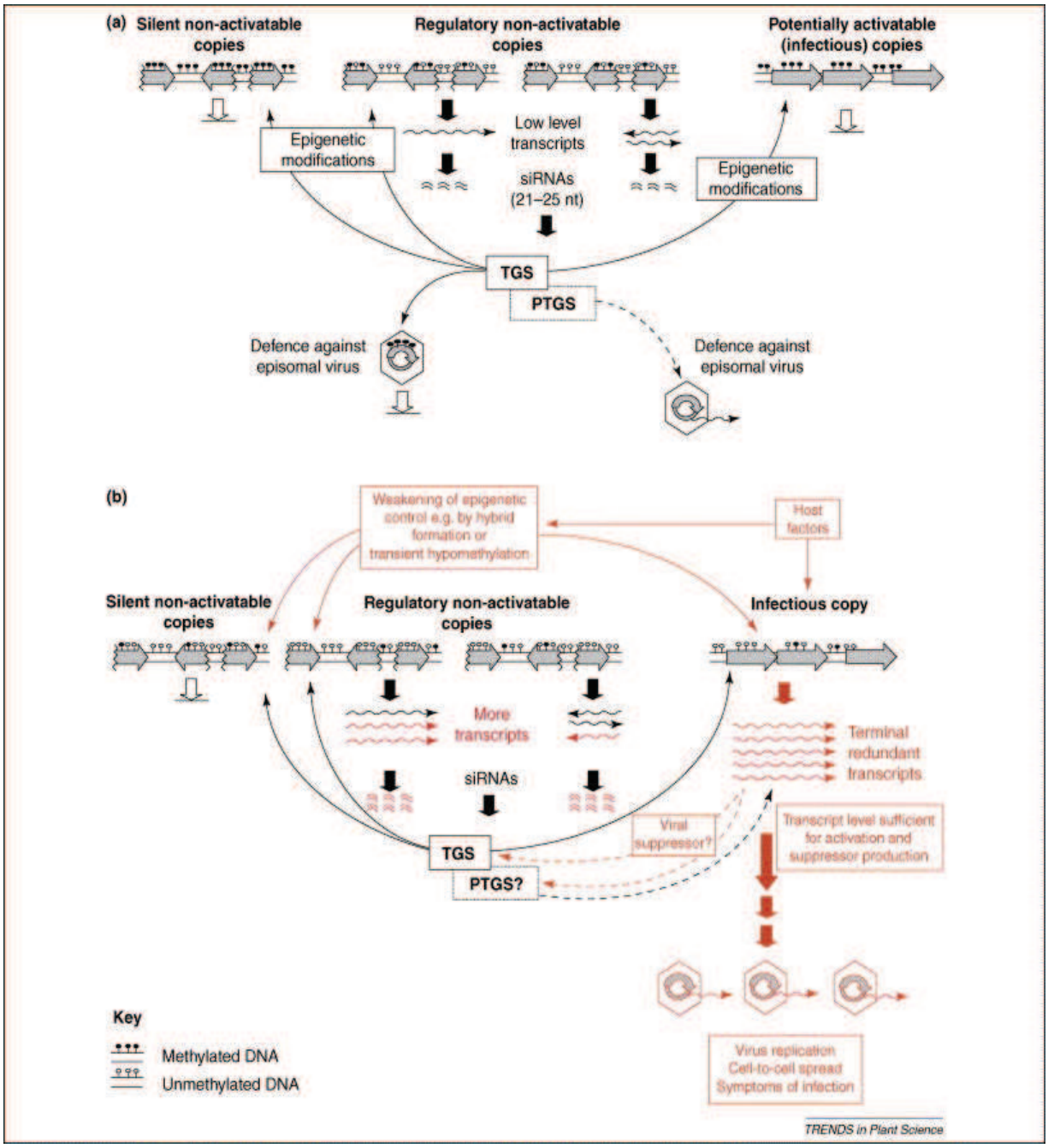


Figure 2-4 : Modèles des mécanismes de régulation des EPRV

(a) Modèle de régulation épigénétique des EPRV. Le génome des plantes contient un grand nombre de copies EPRV dites silencieuses car méthylées et associées à l'hétérochromatine (gauche), ainsi qu'un faible nombre de copies EPRV pouvant être transcrites à un faible niveau par l'ARN polymérase (milieu). Elles ont pour la plupart perdu la structure appropriée pour restituer une particule virale infectieuse. Certains génomes possèdent cependant des copies infectieuses (droite). Les transcrits produits à partir des EPRV non activables sont utilisés comme matrice pour la production de siARN qui induisent ou maintiennent les modifications épigénétiques (méthylation de l'ADN et/ou des histones) des autres locus EPRV homologues par TGS (transcriptional gene silencing). Ils permettent ainsi de contrôler l'activation des éventuelles copies EPRV infectieuses ou les virus exogènes. Des mécanismes de PTGS (post-transcriptional gene silencing – flèches en pointillées) pourraient également être impliqués.

(b) Modèle d'activation des EPRV. Le réveil des EPRV aurait lieu suite à un affaiblissement du contrôle épigénétique qui peut par exemple être due à une hypométhylation transitoire ou une d'hybridation génétique interspécifique. Cela induirait une augmentation de la transcription des transcrits viraux suffisante pour permettre la formation de particules virales. Ces virus pourraient posséder des suppresseurs de silencing pour contourner le système de défense de la plante et conduire à l'infection systémique de la plante hôte.

D'après Staginnus et Richert-Pöggeler, (2006)

hairpin (Pooggin et al., 1999). Bien que des études précises n'aient pas été réalisées, il est suggéré que la réplication du BSV pourrait être semblable à celle du CaMV (Jacquot et al., 1997).

2- Hypothèse de recherche

Des études préliminaires menées sur les eBSV apportent différents éclairages permettant de suspecter une régulation de type « silencing » pour contrôler leur expression. En effet les intégrations sont très souvent dans l'hétérochromatine du génome du bananier et dans des zones riches en éléments transposables (chapitre 1). Nous savons que ce sont ces parties du génome qui sont le plus soumises aux régulations épigénétiques. D'autre part, la structure même des eBSV peut servir de matrice à une régulation de type silencing puisque chez PKW leurs divers réarrangements pourraient conduire après transcription à la formation de structures secondaires de type « Hairpin » reconnues comme structures aberrantes. Des études précédentes réalisées au laboratoire (Lheureux et al., 2003), ont mis en évidence que le stress génomique provoqué lors de croisements interspécifiques engendrait très fréquemment le réveil des eBSV. D'après différents travaux comme ceux de Madlung et Comai (2004), il semble que le stress génomique provoque un relâchement momentané des régulations épigénétiques ce qui expliquerait le réveil des eBSV dans notre système d'étude.

Devant le faisceau d'indices recueillis nous avons proposé comme hypothèse de travail que les eBSV soient sous contrôle de l'ARNi. Cela impliquerait une régulation négative des eBSV de type TGS se traduisant, lors du relâchement du système sous l'action de stress, par la production de particules virales. Ce mécanisme aboutirait à une résistance induite de la plante aux virus libres référents. Afin de tester notre hypothèse nous nous sommes, dans un premier temps, intéressés à la régulation des eBSV infectieux en recherchant la présence de vsARN dans le génome de PKW qui est un bananier à graines, diploide *M. balbisiana*, porteur sain d'eBSV et résistant. Nous avons abordé, dans un deuxième temps, les régulations mises en jeu lors de l'infection de plantes non-porteuses d'eBSV afin de pouvoir les comparer avec celles des eBSV de PKW. Enfin, nous avons recherché si les régulations établies chez PKW étaient également présentes pour les autres bananiers de la diversité *M*. *balbisiana*.

Matériels et méthodes

1-Matériel végétal

Le matériel végétal utilisé provient des serres insecte-proof de confinement S2 de l'UMR BGPI. Le matériel végétal est constitué de plants du bananier fertile *Pisang klutuk wulung* (PKW), de plants du cultivar petite naine Cavendish sains et/ou infectés par les espèces virales BSV suivantes : *Banana streak goldfinger virus* (BSGFV), *Banana streak obino l'ewai virus* (BSOLV), *Banana streak imove virus* (BSImV) et *Banana streak mysore virus* (BSMyV).
Les bananiers représentant la diversité *Musa balbisiana* proviennent de l'International Transit Center (Bioversity international, Leuven, Belgique). Ils ont été reçus sous forme de vitro plants et sont maintenus en serre après sevrage. Ces plantes sont des diploïdes fertiles BB et des hybrides interspécifiques diploïdes et triploïdes *Musa acuminata* (noté A) *x Musa balbisiana* (noté B) (Figure 2-13B).

2- Détection du BSV par IC- PCR

La technique d'immunocapture PCR permet la détection de particules virales libres et combine une approche sérologique à une approche moléculaire. La première étape consiste en une capture des particules virales d'un échantillon grâce à des anticorps spécifiques reconnaissant la protéine de capside du BSV. La deuxième étape est l'amplification du génome viral de ces particules par PCR en utilisant des amorces spécifiques de chaque espèce BSV.
Le protocole utilisé est une variante de celui décrit par Le Provost et al., (2006) puisqu'on rajoute un traitement à la DNase pour éliminer l'ADN génomique résiduel de la plante. 0,5 g de feuille sont broyés dans 5 ml du tampon de broyage suivant (PVP 2%(w/v), Na_2SO_3 0.2% (w/v), BSA 0.2% (w/v) dans du PBS-T 1X pH 6.8) puis sont clarifiés par centrifugation. Vingt cinq microlitres d'extrait de feuille ainsi obtenu sont déposés dans un tube PCR prétraité une nuit à 4°C avec 30 µl de sérum polyclonal anti BSV (AF 1660 ou 1659) dilué au 1/8000 dans du tampon carbonate (Na_2CO_3 15 mM, $NaHCO_3$ 34 mM, qsp H_2O, pH 9.6). Le tout est incubé 3h à température ambiante. Après 5 rinçages au Phosphate Buffered Saline with Tween (PBS-T) suivi de 3 rinçages à l'eau ultra pure ou pyrogene free, chaque tube reçoit 30 µl de mix DNase constitué de 3 µl de tampon 10X (400 mM Tris-HCl [pH 8.0 à 25°C], 100 mM $MgSO_4$, 10 mM $CaCl_2$), - 3U d'enzyme RQ1 Rnase free DNase (Promega™, Madison, WI) et qsp 30µl - ultra pure ou pyrogene free. Les échantillons sont ensuite portés à 37°C pendant 1h. Le mix est éliminé par pipetage, puis le tube est rincé avec de l'eau ultra pure ou pyrogene

free. Un traitement thermique de 10 min à 95°C est ensuite appliqué afin d'inactiver l'action de la DNase.

La PCR est réalisée par addition dans les tubes 0.4µM d'amorces forward et reverse spécifiques de chaque espèce BSV, 2,5 µl de tampon 10X, -100µM de chaque dNTP, 1.5mM de MgCl2, -1U de Taq polymérase et qsp avec de l'eau jusqu'à 25µl. Le programme PCR est le suivant : 5 min à 94°C (1étape) suivi de 30 cycles (25 cycles pour la détection de BSImV) de - 30 s à 94°C, 30 s à 58°C et 30 s à 72°C - puis une étape de 10 min à 72°C. Les produits PCR sont visualisés sous lumière UV après migration sur un gel d'agarose à 1% (TAE 0.5X) et coloration au bromure d'éthidium. Les amorces utilisées pour l'étape de PCR sont indiquées dans la publication de Le Provost et al., (2006).

3-Analyse northern blot

3-1 Extraction d'ARN

Un gramme de feuilles de bananier est réduite en poudre en présence d'azote liquide dans un mortier pré-refroidi afin de limiter au maximum la dégradation des ARNs. La poudre est transférée dans un tube de 15 ml en polypropylène contenant 10 ml de tampon d'extraction (100 mM Tris HCl (pH 7.5), 500mM NaCl, 25mM EDTA (pH 8.0), 1.5% SDS, 2% PVP, 0.7% 2 β mercaptoethanol à ajouter au moment de l'utilisation). Le tout est incubé 10 min à température ambiante avec des agitations régulières. Les tubes sont centrifugés à 8000rpm pendant 30 min pour culotter les débris cellulaires. Une fois le surnageant prélevé, on ajoute 1/3 du volume en acétate de sodium (3 M pH 6.0,) pré-refroidi. Après agitation, les échantillons sont incubés 30 min dans de la glace. Cette étape permet la précipitation de l'ADN et des ARN de haut poids moléculaire qui sont par la suite culottés par une centrifugation de 15 min à 12000rpm (4°C). Le surnageant est prélevé (7-8ml) et un volume équivalent de phénol/chloroforme/alcool isoamylique saturé avec 0.1 M Tris HCl pH 8.0 (25/24/1 v/v/v) est ajouté. Après agitation par retournement, une centrifugation de 10 min à 10000 rpm à 4°C permet de séparer la phase organique de la phase aqueuse. La phase aqueuse est prélevée et retraitée une deuxième fois au phénol/chloroforme/alcool isoamylique. La phase aqueuse est prélevée et 1 volume isopropanol est ajouté. Afin de précipiter les ARN. Les échantillons sont ensuite centrifugés 30 min à 12000 rpm à 4°C, puis le surnageant est éliminé. Un dernier lavage est effectué avec 1 volume d'éthanol à 75% avant une dernière centrifugation à 10000rpm pendant 5 min. L'éthanol est éliminé et le culot est séché avant d'être remis en suspension dans 200µl d'eau DEPC.

3-2 Sélection des amorces

Un alignement des petits ARN (sARN) de PKW issus des données du séquençage profond a été fait sur les séquences eBSV, grâce au logiciel BWA (Li H. et Durbin, 2009). Ces données sous format SAM ont été traitées avec le logiciel R (R development Core Team, 2008) afin d'obtenir un tableau rapportant le nombre de sARN base à base pour chaque séquence eBSV en prenant comme référence l'emplacement du premier nucléotide de l'eBSV. Nous avons ensuite identifié, pour chacune des intégrations, les 5 séquences qui correspondaient aux productions les plus fortes de petits ARN viraux (vsARN) pour dessiner des sondes de 24 nucléotides de long.

3-3 Northern blot pour la détection de petits ARN

La technique de Northern consiste à révéler par hybridation moléculaire des ARN fixés sur des membranes de nitrocellulose qui ont été préalablement séparés selon leur taille par électrophorèse en conditions dénaturantes. L'hybridation se fait avec des sondes radioactives. Un gel de polyacrylamide à 15% dénaturant (PAGE urée) est réalisé à partir de 15 % d'acryl-bis-acryl 19 :1 à 40% , d'urée 8M dans du TBE 1X final, - est complété avec 30µl de temed et 300 µl d'APS(Amonium persulfate) 10% pour un volume final de 50ml. Trente microgrammes d'ARN par échantillon sont séchés au speed vac puis remis en suspension dans un tampon de charge composé de formamide desionisée saturée en bleu de bromophénol. Avant le dépôt des échantillons, les puits du gel sont rincés afin d'éviter la présence d'urée cristallisée. Une pré-migration est réalisée pendant 30 min à 350-400 V dans du tampon TBE 1X. Les ARNs sont ensuite dénaturés à 95°C pendant 1min et mis immédiatement dans de la glace pour être déposés rapidement sur le gel. Une migration à 350 V pendant 30 min puis à 450 V (28 mA, 9W) pendant 5 heures est réalisée à 4°c. Le transfert des ARN sur une membrane Hybond N+ (GE healthcare) est réalisé par électromigration (transfert actif ou électrique) à 10V 2h dans du TBE 1X. Puis l'application d'un rayonnement d'une intensité de 70000 µJ/cm2 (crosslinker uvilink CL-E508 (uvitec)) permet une fixation définitive des ARN sur la membrane. Elle est ensuite pré-hybridée avec 10ml de tampon ULTRAhyb® (Ambion® Life Technologies™) à 35°C pendant 30 minutes à 1h. Les sondes sont marquées radioactivement par une polynucléotide kinase pendant 30 min à 37°C 10pmoles de sonde, 1U de PNK, 2 µl de tampon PNK (10X), et 20mCi de dATP-γ-32 P dans un volume réactionnel final de 20 µl, qsp avec de l'eau stérile RNase et DNase free. La sonde marquée est ensuite purifiée sur une colonne illustra microspin G25 (GE healthcare) et dénaturée à 95°C pendant 1 min puis refroidie dans la glace pendant 2 min. La sonde est

ensuite ajoutée aux 10ml du tampon d'hybridation. L'hybridation de la membrane s'effectue la nuit à 35°C. La membrane est rincée deux fois avec un tampon 2SSC - 0.5% SDS puis lavée avec un tampon 2 SSC - 0.5% SDS durant 30 min à 35°C. La membrane est mise en cassette durant 24 h puis révélée au phosphorimager FLA9000 (GE Healthcare). Les sondes utilisées sont présentées dans en données supplémentaires dans le tableau sup 2-2.

La membrane peut être ensuite déshybridée pour une nouvelle utilisation. Un premier lavage est alors réalisé avec du tampon 0.5SSC - 0.5% SDS pendant 30 min à 80°C puis un second avec du 0.1SSC - 0.5% SDS pendant 30 min à 80°C.

4- Séquençage profond Illumina des petits ARN (sARN) et analyse bio-informatique

Une quantité de 10 µg d'ARN totaux des bananiers PKW, Cavendish sain et infectés par BSGFV, BSOLV, BSImV ou BSMyV (extrait selon le protocole ci-dessus) ont été envoyés à l'entreprise FASTERIS pour un séquençage profond. La qualité des ARN a été testée avant l'envoi grâce à un bio-analyseur (QIAGEN QIAxel™) qui permet d'obtenir l'ARN Integrity Number (RIN) qui détermine la qualité de l'ARN de chacun des échantillons. Seuls les échantillons avec un RIN supérieur à 7 ont été sélectionnés ; 1 indiquant une dégradation totale des ARN et 9-10 une intégrité parfaite de l'échantillon.

L'entreprise FASTERIS a réalisé une banque à partir des sARN de 15-30nt purifiés sur gel d'acrylamide à 15% TBE-Urée. Des adaptateurs 5' adenylés simple brin ont été positionnés en 3' des sARN grâce à la T4 RNA ligase sans ATP et d'autres adaptateurs simple brin ont été positionnés en 5' des sARN en présence d'ATP grâce à la même enzyme. Les sARN sont ensuite purifiés sur gel d'acrylamide à 10% TBE/Urée avant de réaliser la synthèse des cDNA et leur amplification par PCR. La banque a été séquencée par « Illumina Genome Analyser » suivant les recommandations du fournisseur.

Un contrôle qualité des données a été réalisé afin de vérifier la dégradation des ARN et l'homogénéité (quantitative/qualitative) des différents échantillons traités. Les données sont triées afin d'obtenir un fichier FASTQ de référence ne contenant que les sARN de 20 à 25 nt pour les 6 échantillons étudiés. Les sARN sont ensuite alignés avec des séquences de référence à l'aide du logiciel Burrows-Wheeler Aligner (BWA) (Li H. and Durbin, 2010). Les alignements avec BWA ont tous été réalisés sans mismatch. Ces alignements permettent d'obtenir des fichiers de données (fichier texte (.Txt)) indiquant le nombre de sARN débutant pour chacun des nucléotides de la séquence de référence que ces sARN soient en orientation sens (première base du sARN) ou en orientation antisens (dernière base du sARN) . Les alignements ont été réalisés soit sans répétition, c'est à dire que chaque sARN

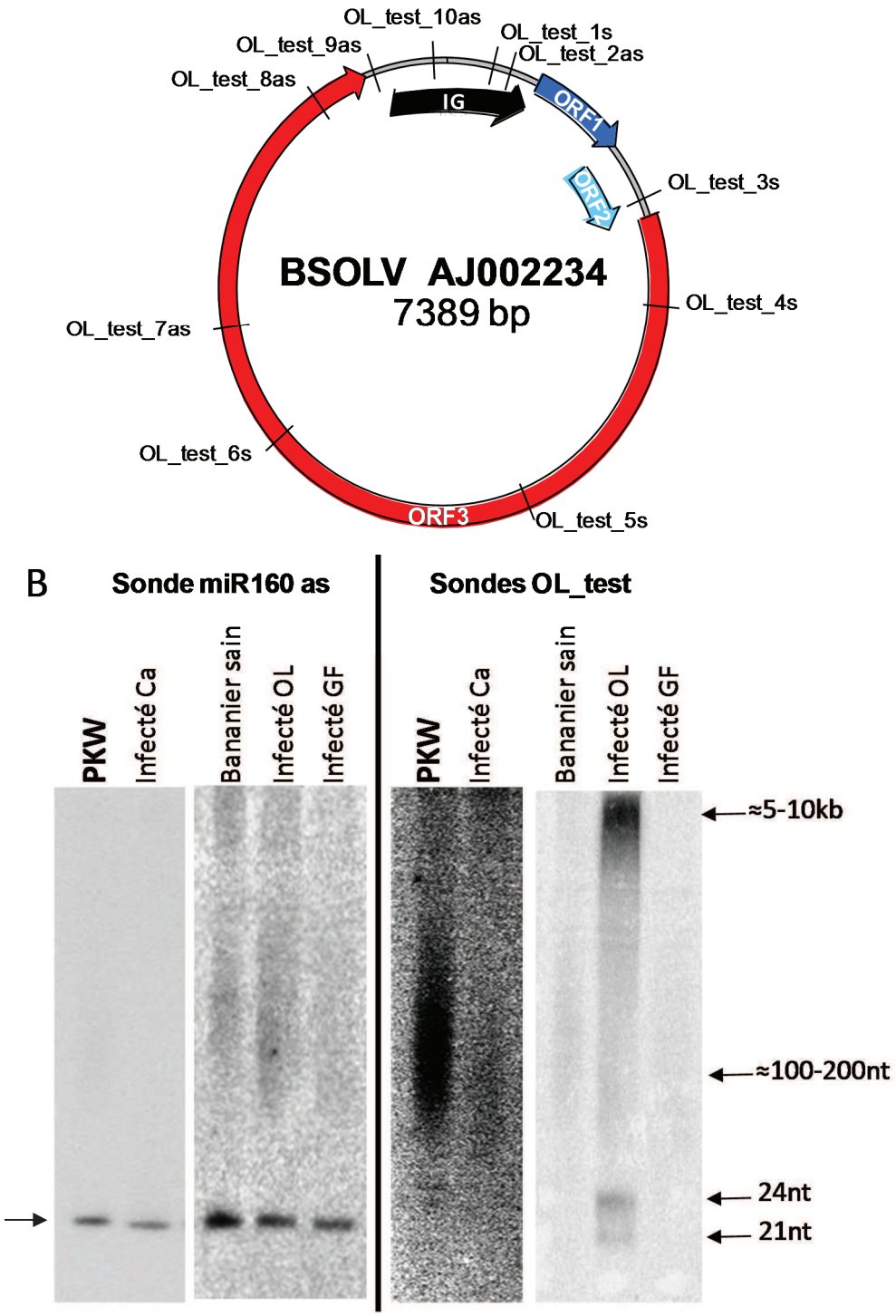


Figure 2-5: Analyse de la production de vsARN BSOLV par northern blot

génome de référence est mentionnée par les lettres (s) et (as) pour une orientation sens et antisens respectivement. Les sondes mesurent de 20 à 26nt.

B: Analyse des ARN de faible poids moléculaire extraits à partir du bananier PKW (BB) sain qui est porteur des intégrations BSOLV (eBSOLV) et de bananiers Cavendish (AAA) sain ou infectés par les virus BSOLV, BSGFV ou BSCaV. Les ARN sont séparés en fonction de leur taille par migration sur gel dénaturant sPAGE 15% puis hybridés avec la sonde miR160 antisens (partie gauche de la figure) et un mélange des 10 sondes OL_test présentées en A (partie droite de la figure).

n'est positionné qu'une seule fois sur la séquence eBSV, soit avec répétitions et dans ce cas, le sARN est positionné sur l'eBSV autant de fois que nécessaire. Le décompte des sARN est obtenu à partir des fichiers textes et la visualisation grâce au logiciel ARTEMIS implémenté du logiciel BAMview (Carver et al., 2012). Les quantités de sARN obtenues ont été ensuite transformées afin d'obtenir des nombres de « reads » (séquences de sARN) en reads par kilobase par million (RPKM). Pour cela les quantités de sARN (ou read) obtenues lors du « mapping » sont divisées par la quantité de sARN (ou read) totale obtenue pour l'échantillon et multipliées par 1 million (pour que le chiffre ne soit pas trop petit). Puis cette donnée en read par million est ensuite divisée par le nombre de kilobases de la séquence référence utilisée pour le « mapping ». Cette méthode de calcul développée par Mortazavi et al., (2008) permet de pouvoir comparer des données provenant de différents échantillons sur des séquences références de différentes tailles.

Résultats

1- Recherche ciblée de la production de vsARN BSOLV chez le bananier

Nous avons dans un premier temps recherché des petits ARN viraux (vsARN) pour l'espèce BSOLV dans des bananiers présentant des caractéristiques différentes vis à vis du BSV. Nous avons ainsi utilisé le bananier PKW (BB) qui possède des eBSOLV, qui est indemne de particules virales BSOLV et est résistant au BSOLV. Nous avons utilisé les cultivars (cv) Cavendish (AAA) qui sont dépourvus d'eBSOLV. Ces bananiers étaient soit sains soit infectés par le BSOLV. Nous avons tout d'abord vérifié par immuno-capture PCR avec des amorces spécifiques de chacune des différentes espèces BSV que le bananier Cavendish n'était infecté que par l'espèce BSOLV. Les bananiers sains se sont révélés indemnes de virus.

L'analyse des vsARN a été réalisée en deux temps. Nous avons tout d'abord vérifié la présence des ARN en utilisant une sonde miR160 as. Cette sonde est ubiquitaire dans le règne végétal et identifie un miARN de 21 nt. Elle nous renseigne également sur les quantités de sARN présents pour les différents échantillons. Les vsARN ont ensuite été recherchés par northern blot grâce à 10 sondes de 20 à 26 nt réparties sur le génome du BSOLV (figure 2-5A, tableau sup 2-2). L'hybridation des ARN totaux avec la sonde miR160 montre qu'il y a présence de ce miRNA de 21nt pour tous les échantillons analysés comme attendu (figure 2-5B). La production de sARN est relativement homogène d'un individu à l'autre. Les résultats d'hybridation avec les sondes BSOLV sont présentés sur la figure 2-5B. Nous observons pour PKW porteur d'eBSOLV un smear intense autour de 100 à 200 nt

pouvant correspondre à des transcrits ARN partiels ou dégradés des séquences eBSOLV. Nous notons également une bande de très faible intensité correspondant potentiellement à des ARN de 24nt. Le bananier *Cavendish* infecté par BSOLV (et sans eBSOLV) présente principalement des vsARN spécifiques du BSOLV de 21 et 24nt. On note un signal d'hybridation en haut de la membrane pour cet échantillon qui pourrait correspondre à des transcrits d'ARN viraux prégénomiques de 7.5kb. Aucun signal d'hybridation n'est obtenu pour les bananiers *Cavendish* sain. Ces résultats indiquent clairement que le BSOLV induit la production de vsARN chez les bananiers infectés, tout comme cela avait été observé pour le CaMV infectant *Arabidopsis thaliana* (Blevin et al., 2011). Les eBSOLV semblent induire de manière plus discrète la production de vsARN chez PKW. Nous avons alors décidé de poursuivre la caractérisation des mécanismes épigénétiques mis en place chez PKW et chez des plants Cavendish infectés par différentes espèces BSV en réalisant un séquençage profond des sARN de ces différentes plantes.

2- Analyse de la production des sARN induits par les eBSV et BSV chez le bananier

Un séquençage profond des sARN totaux de PKW a été réalisé afin de caractériser précisément les vsARN produits par les différents eBSV présents dans son génome. En effet, PKW présente des intégrations de BSOLV, BSImV, BSGFV et BSMyV (Geering et al., 2005a). Tout comme pour eBSOLV, PKW est porteur sain de eBSImV et eBSGFV ; les intégrations BSMyV étant défectueuses. Nous savons en effet que lors de croisements interspécifiques entre PKW et le tétraploïde *M. acuminata* AAAA cultivar IDN 110-Tétraploïde, la descendance de génotype BAA stérile restitue des particules virales infectieuses à partir des allèles eBSOLV-1, eBSGFV-7 et eBSImV (cf chapitre 1).

En parallèle de PKW, un séquençage profond des sARN des bananiers Cavendish non porteurs d'eBSV infectieux mais infectés soit par BSGFV, BSImV, BSOLV ou BSMyV a été réalisé afin de préciser les mécanismes de régulation mis en jeu lors de l'infection du bananier. Ces travaux sont menés en collaboration avec l'équipe de T. Hohn et M. Pooggin de l'institut Botanique de Bâle dans le cadre du projet Cost FA0806 "Plant virus control employing RNA-based vaccines : A novel non-transgenic strategy". Outre la caractérisation du mécanisme épigénétique mis en place par la plante pour tenter de contrôler le BSV, nous souhaitions vérifier les observations faites sur le pathosystème *Arabidopsis thaliana*/CaMV (voir paragraphe 1-4 de l'introduction du chapitre) pour des virus de genre différent appartenant à la même famille des *Caulimoviridae.*

Des analyses par IC-PCR avec des amorces spécifiques de chacune des espèces BSV ont été réalisées afin de s'affranchir de co-infection ou d'infection croisée. Les sARN de ces 4 échantillons ont été séquencés en même temps que ceux du bananier Cavendish comme témoin sain et de PKW.

2-1 Vérification de la conformité des résultats de séquençage profond

Le séquençage a été réalisé par l'entreprise FASTERIS qui est localisée en Suisse et est membre du projet COST FA0806.

L'analyse préliminaire des données brutes a consisté à vérifier le niveau de dégradation des ARN, la recherche de contamination croisée, l'analyse du nombre de reads par échantillon. Le séquençage de PKW, deux fois plus profond que pour les autres échantillons, donne en moyenne 2 fois plus de « reads » que les autres séquençages. Chez PKW, les sARN de 0 à 17nt et de 27 à 44nt représentent 34% et 23% des sARN totaux respectivement. Les sARN de 0 à 17nt et de 27 à 44nt représentent pour les autres échantillons toujours moins de 50% des sARN totaux ce qui correspond à une valeur acceptable pour valider la qualité de la librairie (tableau sup 2-3). Les quantités de « reads » obtenus pour les différents échantillons sont toutes comprises entre 9,9 et 13,9 millions montrant l'uniformité/homogénéité des résultats. La profondeur de séquençage est suffisante pour l'étude à mener (tableau sup 2-3).

Les sARN les plus fréquemment impliqués dans les mécanismes d'ARNi sont de taille comprise entre 21 et 24nt. Les sARN ont donc été triés dans une gamme comprise entre 20 et 25nt. Les fichiers de séquences ainsi obtenus (format FASTQ) ont été « mappés » sur les séquences virales de référence afin d'identifier les sARN correspondants. Les séquences référence utilisées pour cette étude sont celles des différentes espèces de BSV ou d'eBSV. La stringence utilisée lors des analyses bio-informatiques est maximum puisque nous n'avons accepté aucun mismatch sauf indication contraire.

2.2 Analyse comparative des sARN avec les séquences virales référence

2-2-1 Production de vsARN chez PKW et les bananiers infectés

- <u>vsARN chez PKW</u>

Les vsARN dont la taille est comprise entre 20 et 25 nt représentent 3,7% des sARN totaux alors qu'ils ne représentent que 0,34% chez le témoin bananier Cavendish sain (figure sup1). Une étude approfondie des vsARN de ce témoin sain a montré que les profils de production étaient identiques à ceux trouvés chez PKW . Nous pouvons donc raisonnablement penser que les sARN du témoin négatif ont été contaminés par les sARN de PKW. Le re-séquençage

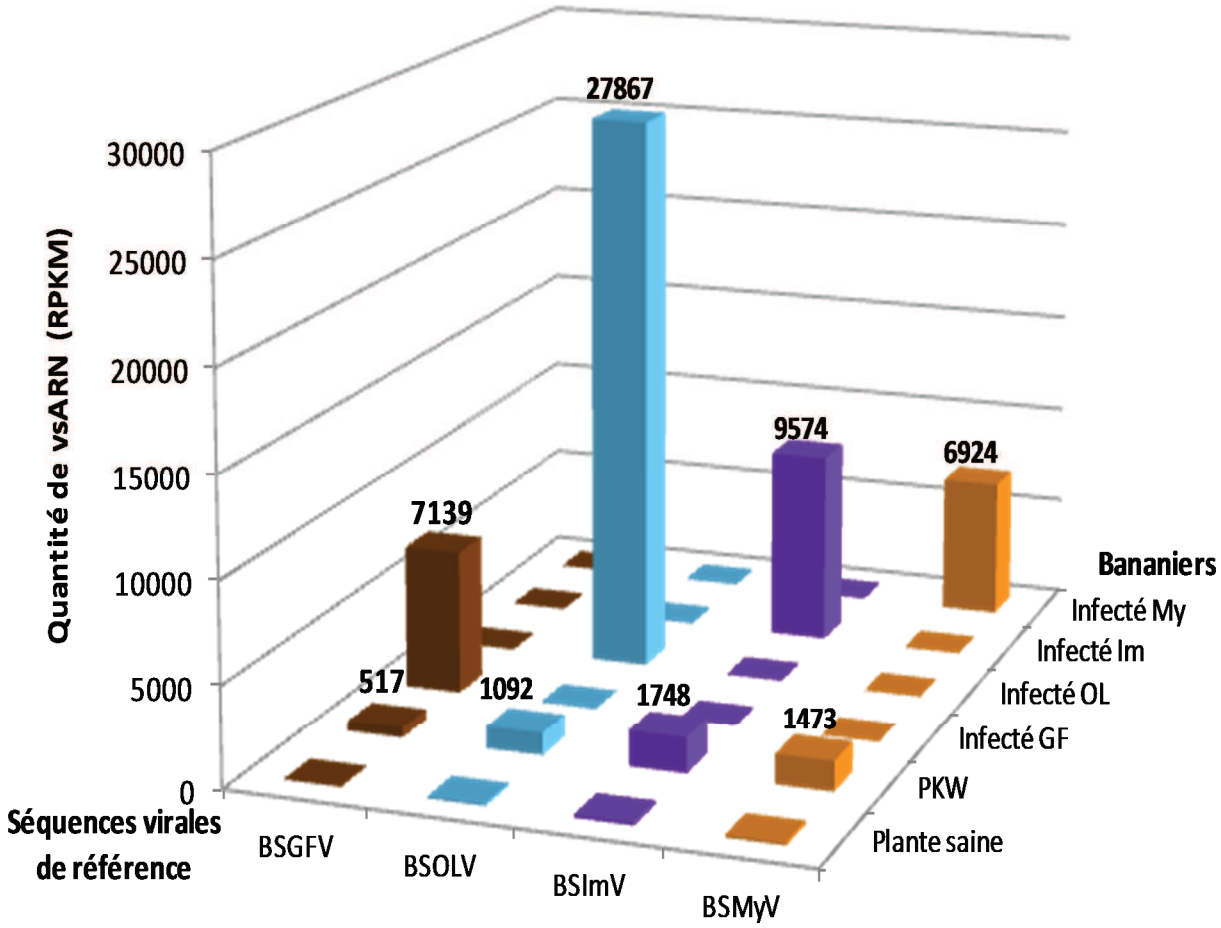


Figure 2-6 : Quantité totale des vsARN de 20 à 25 nt pour chacune des espèces BSV dans les différents bananiers analysés

Représentation des quantités de vsARN (en ordonnée) sous forme d'histogramme en 3 dimensions avec sur le premier axe des abscisses les différentes séquences références BSV utilisées pour l'analyse de mapping. Le deuxième axe des abscisses représente les différents bananiers séquencés. Les données sont exprimées en Read Par Million de Kilobase (RKPM) pour chaque espèce BSV. Seules les quantités supérieures à 50 RPKM sont affichées. Les alignements de séquences entre vsARN et séquences de référence sont réalisés dans le cadre d'identité parfaite et sans prendre en compte les répétitions de séquences observées chez les eBSV.

d'un témoin sain est en cours. De façon intéressante, on note une prédominance des vsARN de 24nt. La répartition des vsARN selon leur taille est comparable pour chacune des espèces BSV chez PKW avec systématiquement une forte production de vsARN de 24nt.

D'un point de vue quantitatif, les vsARN majoritairement produits chez PKW correspondent à la séquence référence eBSImV avec 1748 reads par millions par kilobase (RPKM), puis viennent eBSMyV et eBSOLV avec respectivement 1473 et 1092 RPKM pour finir par eBSGFV avec seulement 517 RPKM (figure 2-6). L'analyse des proportions de vsARN par catégorie de taille est présentée figure 2-7. Des profils similaires sont observés pour les quatre espèces BSV chez PKW avec des vsARN de 24nt représentant entre 75.3 et 79.7% contrairement à ceux de 21nt et 23 nt qui ne représentent qu'entre 9 et 12,4%.

L'étude s'est poursuivie par une analyse du ratio de vsARN produits en orientation sens et anti-sens par rapport au génome viral de référence (tableau sup 2-4). On note chez PKW des ratios vsARN sens/vsARN anti-sens de 0,71 pour eBSImV, de 1,21 pour eBSGFV, de 1,34 pour eBSOLV et de 1,50 pour eBSMyV. Ces résultats indiquent que l'eBSImV produit majoritairement des vsARN provenant de la séquence anti-sens alors que les autres eBSV produisent plus de vsARN provenant de la séquence sens. Néanmoins ces données doivent être utilisées avec prudence aux vues des structures fortement réarrangées observées pour eBSOLV, eBSGFV et eBSMyV chez PKW. En effet, certaines séquences internes de ces eBSV sont représentées plusieurs fois en orientation sens et anti-sens (voir chapitre 1) rendant impossible à ce stade la détermination de la (les) zone(s) de production de tel ou tel vsARN issu de ces régions répétées.

L'analyse de la répartition selon leur taille des vsARN le long du génome BSV a été conduite en prenant en compte les 3 ORF et la séquence intergénique. Ces résultats sont présentés dans les figures 2-8 et 2-9. La production de vsARN est effective pour toutes les régions du génome viral chez les quatre espèces BSV concernées à l'exception de la région intergénique et l'ORF 1 de BSImV (figure 2-8). Les vsARN correspondant à BSOLV et BSMyV se répartissent de façon similaire le long du génome viral avec une production moindre dans l'IG par comparaison aux trois ORFs où la production est assez homogène. Les vsARN correspondant à BSImV sont exclusivement concentrés sur l'ORF 2 et 3 avec une production quatre fois supérieure pour cette dernière. Concernant BSGFV, la production est maximale dans l'ORF1 et comme pour BSOLV et BSMyV une quantité plus faible apparaît dans l'IG. Enfin, on remarque que la proportion relative de chaque catégorie de vsARN (21, 22, 23 et 24nt) décrit figure 2-7 est conservée pour chacune des quatre régions (IG, ORF 1, 2 et 3) du

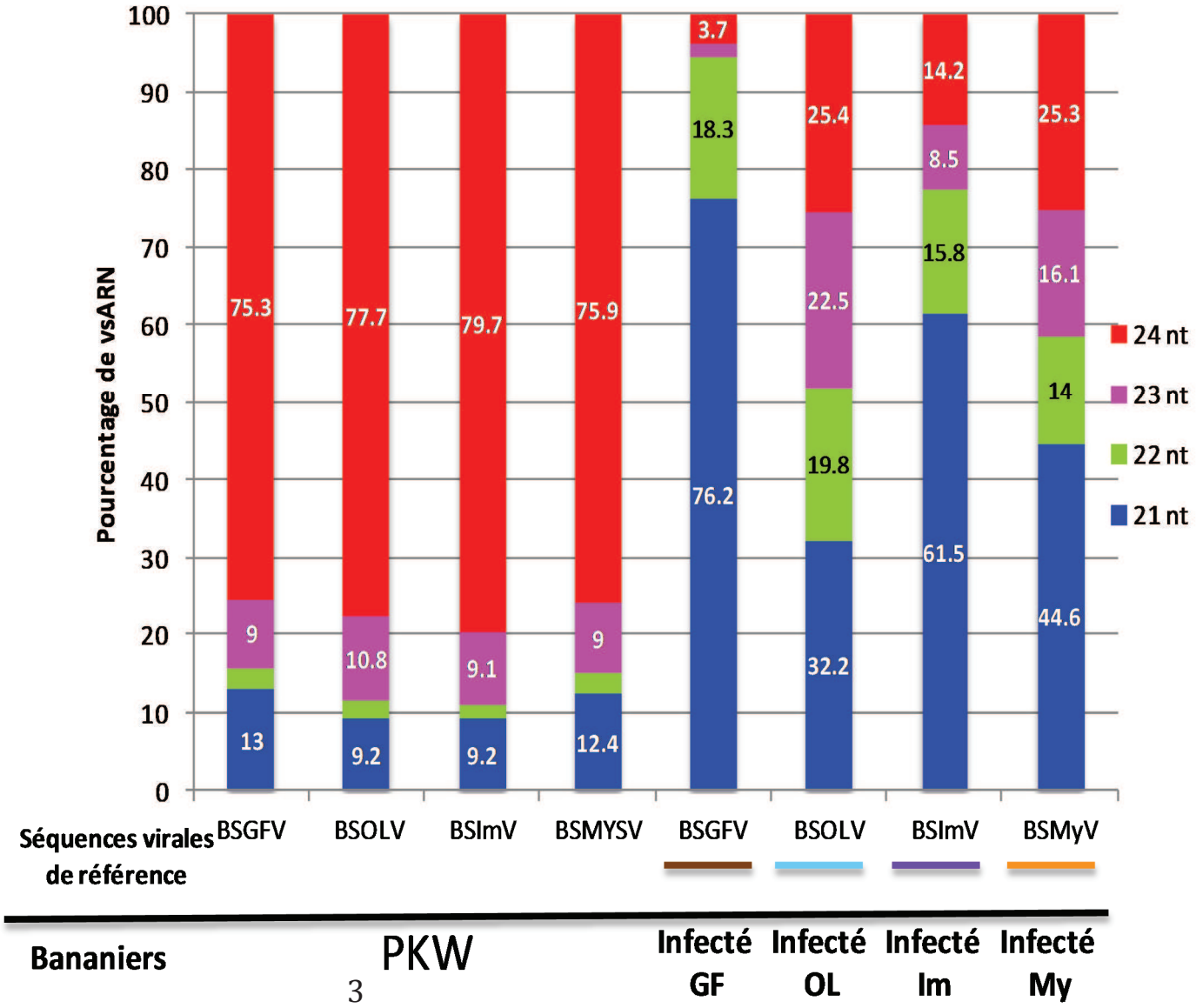


Figure 2- 7 : Répartition (%) des vsARN selon leur taille dans les bananiers analysés

Les séquences virales références utilisées pour le mapping sont BSGFV, BSOLV, BSImV et BSMyV. Les alignements

de séquences entre vsARN et séquences de référence sont réalisés dans le cadre d'identité parfaite et sans prendre en compte les répétitions de séquences observées chez les eBSV. Les chiffres sur les histogrammes représentent les pourcentages de chacune des classes de vsARN.

génome viral et ce quelle que soit l'espèce BSV considérée. En effet, aucune région du génome ne produit spécifiquement une catégorie de vsARN.

- vsARN chez les plantes infectées

Les vsARN chez les plantes infectées représentent des quantités plus importantes que ceux détectés chez PKW, allant de 5% des sARN totaux (20-25nt) pour la plante infectée par BSGFV, à 7% pour les plantes infectées par BSImV ou BSMyV pour atteindre 25% pour la plante infectée par BSOLV (tableau sup 2-3). Comme attendu, les vsARN produits correspondent presque exclusivement (>99%) à la séquence de référence BSV à l'origine de l'infection (figure 2-6). Pour chacune des plantes infectées, nous observons cependant quelques vsARN des autres espèces BSV témoignant d'identité de séquences existant entre les espèces BSV. La production de vsARN la plus importante est enregistrée pour la plante infectée par BSOLV avec 27867 RPKM soit une quantité au moins 4 fois supérieure à celles enregistrées pour les autres plantes infectées qui sont sensiblement équivalentes entre elles variant de 6924 à 9574 RPKM.

L'analyse des proportions de vsARN par catégorie de taille est présentée figure 2-7. Les résultats obtenus pour les plantes infectées montrent que les vsARN de 21nt sont majoritaires quelle que soit l'espèce BSV concernée. Nous observons cependant deux profils distincts. Un profil où les vsARN de 21nt sont largement sur-représentés. C'est le cas des plantes infectées par BSGFV (76%) et BSImV (61,5%) et un profil où les vsARN se répartissent entre deux tailles 21nt et 24nt. C'est le cas des plantes infectées par BSOLV (32,2 et 23,4 respectivement) et BSMyV (44,6 et 25,3 respectivement). Les vsARN de 21, 22 et 23nt sont en moyennes plus représentées que chez PKW.

L'étude du ratio de vsARN produits en orientation sens et anti-sens par rapport au génome viral de référence (tableau sup 2-4) montre de façon non ambiguë que les vsARN produits en sens sont équivalents ou 1,5 fois supérieurs à ceux produits en anti-sens quelque soit l'espèce BSV concernée.

L'analyse de la répartition des vsARN le long du génome viral montre comme pour PKW, qu'aucune région du génome ne produit spécifiquement une catégorie unique de vsARN et ce, quelle que soit l'espèce BSV considérée (figures 2-7 et 2-8A). Les vsARN sont produits à partir de l'ensemble du génome viral avec une production majoritaire au niveau de l'ORF 1 quelle que soit l'espèce BSV étudiée. Le profil de répartition des vsARN le long du génome est également comparable entre plantes infectées avec une production qui diminue constamment depuis l'ORF 1 jusqu'à l'IG. Seul le bananier infecté par BSOLV présente une répartition des vsARN majoritaire pour l'ORF 1 puis l'IG et des taux comparables pour les

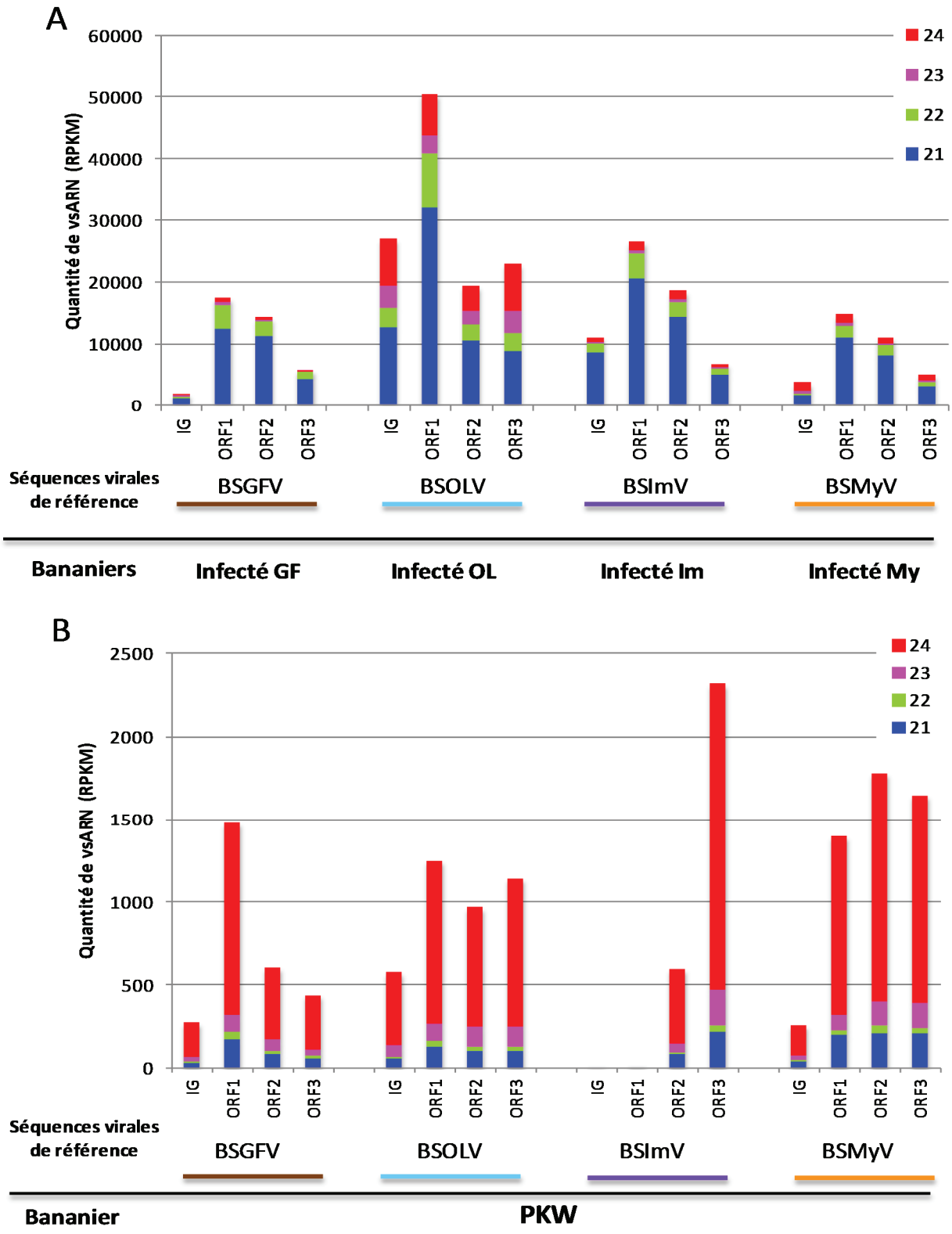


Figure 2-8 : Répartition par taille et quantité des vsARN sur le génome viral pour les bananiers analysés

Les quantités (exprimées en RPKM pour Read Par Million par Kilobase) sont décrites en fonction de la taille et de la position des vsARN sur le génome des BSV; l'IG correspondant à la zone Intergénique et les ORF aux trois séquences codantes du génome viral. Les alignements de séquences entre vsARN et séquences de référence sont réalisés dans le cadre d'identité parfaite et sans prendre en compte les répétitions de séquences observées chez les eBSV. Les résultats sont présentés pour les vsARN produits chez les différents bananiers Cavendish infectés (A) et chez PKW (B).

ORF 2 et 3. Enfin, on peut noter une production extrêmement faible au niveau de l'IG pour le bananier infecté par BSGFV.

L'ensemble de ces résultats démontrent que les zones génomiques virales impliquées dans la production de vsARN sont différentes entre PKW (cible eBSV) et les plantes infectées (cible BSV) pour une même espèce BSV et ne sont donc pas dépendante de l'organisation du génome BSV. Cela signifie que les eBSV présents dans le génome de PKW produisent des vsARN du fait de leur structure réarrangée et non en fonction du génome viral lui même.

2-2-2 Zones de production de vsARN chez PKW et les bananiers infectés

Nous avons identifié avec précision les zones de production de vsARN pour chacune des espèces BSV pour les différents bananiers analysés. Les zones produisant une quantité importante de vsARN correspondent à des régions chaudes et seront appelées «hot-spot » par opposition aux régions froides nommées « cold-spot » qui ne produisent pas ou très peu de vsARN. Les résultats sont présentés figure 2-9A.

- Zones de production de vsARN chez PKW

La figure 2-9B illustre la distribution des vsARN de PKW le long des séquences génomiques linéarisées BSV. Pour l'eBSOLV les vsARN couvrent la quasi-totalité du génome BSOLV à l'exception de la zone entre l'ORF3 et l'IG et d'une zone au milieu de l'ORF3. Il n'existe pas réellement de hot-spot très important mais beaucoup de zones productrices. Par contre, la répartition est très ciblée pour l'eBSImV où seule une zone « hot-spot » de 1887pb de l'ORF3 produit plus de 99% des vsARN. Le reste de la séquence ne produit pas ou très peu de vsARN moins de 1 RPKM pour l'IG et moins de 2 pour l'ORF 1. Les quantités de vsARN produites pour l'eBSGFV sont très largement inférieures à celles produites pour l'eBSImV. Il existe des zones « cold-spot » en particulier au niveau de l'ORF 3 ainsi que sur la zone intergénique. Par ailleurs, les hot-spots sont très ciblés mais contrairement à l'eBSImV ils sont repartis de façon plus homogène sur la séquence virale. L'eBSMyV quant à lui produit des vsARN tout le long du génome avec des zones « cold-spot » au niveau de l'IG et de la fin de l'ORF 3. Les zones hot-spot sont reparties régulièrement sur le reste de la séquence.

Les vsARN de 24nt apparaissent bien majoritaires quelle que soit l'espèce BSV intégrée. De façon intéressante, on observe que les hot-spots ne sollicitent pas les mêmes zones du génome viral entre eBSV.

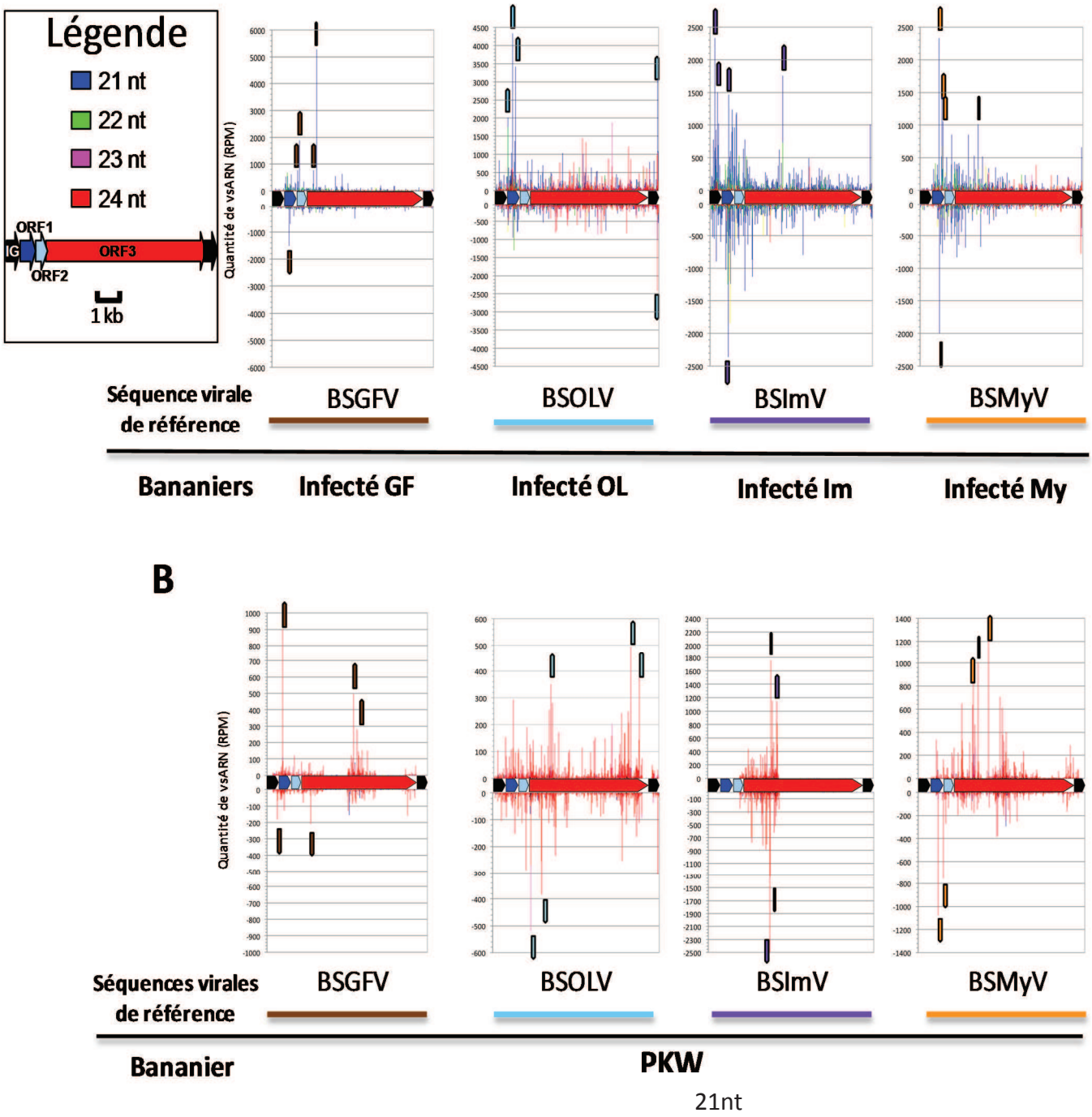


Figure 2-9 : Répartition des vsARN sur la séquence virale de référence chez les bananiers analysés

Les profils de production de vsARN de taille 21-24nt sont présentés pour les vsARN produits chez les différents bananiers Cavendish infectés (A) et chez PKW (B). Les alignements de séquences entre vsARN et séquences de référence sont réalisés dans le cadre d'identité parfaite et sans prendre en compte les répétitions de séquences observées chez les eBSV. Les données sont exprimées en RPKM (Reads par Million par Kilobase). Les histogrammes situés au dessus des séquences références représentent les vsARN en orientation sens et ceux en dessous, les vsARN en orientation antisens. Pour chaque échantillon, les 5 régions produisant le plus de vsARN sont mentionnées par les flèches de couleur. Le sens des flèches représente l'orientation des vsARN (orientation sens ou anti-sens).

- Zones de production de vsARN chez les bananiers infectés

La distribution le long du génome viral des zones de production des vsARN pour les bananiers infectés est représentée sur la figure 2-9A.

Pour la plante infectée par BSGFV, le génome est presque totalement couvert par les vsARN que ce soit en sens ou en anti-sens. Chaque ORF possède une ou deux régions chaudes présentes préférentiellement au début d'ORF. Ces hot-spots correspondent à des vsARN de 21nt et aucun d'entre eux n'est présent dans l'IG. La quantité globale des vsARN sur le génome viral est moins importante que pour les trois autres espèces BSV étudiées mais la région hot-spot située au début de l'ORF3 est la plus représentée parmi toutes celles identifiées toutes espèces confondues.

Le bananier infecté par BSOLV est celui qui produit quantitativement et qualitativement le plus de vsARN. La couverture du génome est totale en sens et en anti-sens. Les hot-spots de production de vsARN de 21nt sont très importantes et situées majoritairement dans l'ORF1. Une forte production de vsARN est toutefois notée dans l'IG et correspond à la région leader du BSV (Poogin et al., 1999). Il existe également plusieurs hot-spots de 24nt tout le long de l'ORF3.

La plante infectée par BSImV présente des vsARN de 21nt sur toute la longueur du génome viral. La région la moins productrice est au niveau du promoteur situé dans l'IG, et les hot-spots sont au niveau de la région leader de l'IG, l'ORF1, l'ORF 2 et une zone au centre de l'ORF 3.

Enfin, le bananier infecté par BSMyV, montre également une très bonne couverture de vsARN puisque seules quelques régions n'excédant pas 16nt de long ne produisent pas de vsARN. Les hot-spots sont présents au niveau des ORF1 et 2 pour les vsARN de 20nt, 21nt et 22 nt. On observe aussi une forte production en anti-sens de vsARN de 23nt et 24nt dans la région leader de l'IG.

L'ensemble de ces données révèlent que les mécanismes de défense mis en place par la plante pour lutter contre le BSV sont de même nature et font appel préférentiellement à des vsARN de 21nt quelle que soit l'espèce BSV concernée. Il apparaît également que les hot-spots sont principalement situés dans l'ORF 1 et le début de l'ORF 2. Cependant, il existe certaines régions chaudes au niveau de l'IG comme dans le cas d'une infection par BSOLV ou de l'ORF3 pour BSImV, BSGFV ou BSMyV.

Echantillon	PKW				
Référence	eBSGFV-7	eBSGFV-9	eBSOLV-1	eBSOLV-2	eBSImV
Taille en kb	**13,28**	**15,58**	**22,857**	**23,218**	**15,818**
vsARN total	**331**	**288**	**371**	**369**	**899**
vsARN spécifique eBSV (%)	**14,7**	**16,5**	**5,0**	**5,7**	**4,4**

Tableau 2-1 : Données issues du mapping des sRNA de PKW sur les séquences eBSV

Les pourcentages de vsARN spécifiques de chaque eBSV ont été calculés en soustrayant le nombre de vsARN ciblant la séquence eBSV en question du nombre de vsARN ciblant la séquence BSV correspondante. Le pourcentage de vsARN spécifique à chaque eBSV est calculé en rapportant le nombre de vsARN spécifiques au nombre de vsARN totaux ciblant l'eBSV.

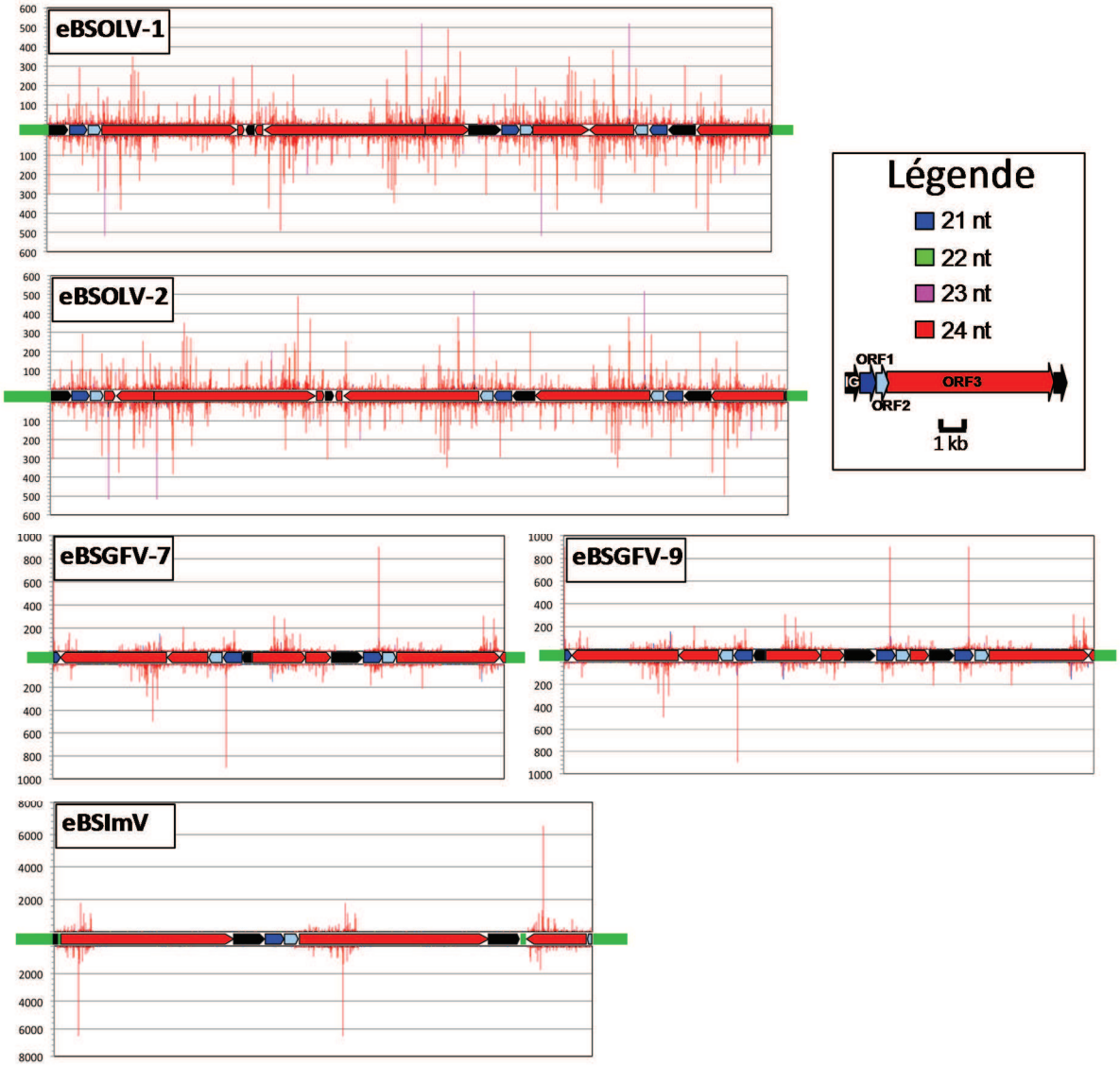


Figure 2-10 : Répartition des vsARN produits par PKW sur les séquences eBSV

La quantité et la localisation de chaque vsARN de 21 à 24 nt de long produit chez PKW sont représentées sur les séquences références eBSV. La quantité est exprimée en RPKM (Reads par Million par Kilobase). Les alignements de séquences entre vsARN et séquences de référence sont réalisés dans le cadre d'identité parfaite mais prennent en compte les répétitions de séquences observées chez les eBSV. Les histogrammes situés au dessus des séquences références représentent les vsARN en orientation sens et ceux en dessous, les vsARN en orientation antisens.

Les résultats indiquent aussi que les hot-spots de vsARN sur le génome viral sont différents entre PKW et les bananiers infectés. De plus chez ces derniers, il existe une certaine conservation des régions chaudes entre espèces BSV puisque presque toutes les zones hot-spots sont situées dans une région de 3kb entre le milieu de l'IG et le début de l'ORF3 contrairement à PKW où ces hot-spots sont repartis de façon plus aléatoire sur l'ensemble du génome viral (figure 2-9).

2-3 Analyse des vsARN produits par analyse comparative avec les séquences eBSV présentes chez PKW

2-3-1 Répartition des vsARN sur les eBSV

Après avoir analysé la distribution des vsARN produits chez PKW le long du génome viral linéarisé, nous nous sommes intéressés à leur projection sur les séquences eBSV. Les séquences et structures eBSV sont décrites dans Gayral et al., 2008 et dans l'article 1 du chapitre 1. Les structures sont schématisées sur la figure 2-10. Les analyses réalisées ont tenu compte des séquences répétées ce qui signifie qu'un même vsARN se verra assigner autant de positions que nécessaire. Elles ont aussi tenu compte des séquences présentes au niveau des zones de réarrangement. On ne peut néanmoins pas faire la distinction entre allèles à l'exception des zones présentant soit des différences structurelles soit des points de mutation puisque nos analyses sont réalisées pour une identité de séquence de 100%.

Suite à des problèmes d'assemblage non résolus à ce jour, les séquences d'eBSMyV sont toujours parcellaires. Nous avons donc choisi de ne pas inclure les deux eBSMyV dans l'analyse.

Les vsARN spécifiques de chacun des eBSV c'est-à-dire spécifiques des zones de réarrangement et des points de mutation (par comparaison avec les séquences BSV), représentent 14,7% et 16,5% des vsARN totaux pour eBSGFV-7 et 9 respectivement, alors qu'ils ne représentent que 4,5 à 5,7% pour les autres eBSV (tableau 2-1).

La figure 2-10 permet de visualiser la répartition de l'ensemble des vsARN sur les séquences eBSV en fonction de la taille des différents vsARN (21nt à 24nt).

Nous pouvons voir que la zone « hot-spot » identifiée précédemment de l'eBSImV est située dans l'ORF 3 au niveau de la seule zone répétée à trois reprises le long de l'eBSV. Le hot-spot le plus important chez BSGFV situé sur l'ORF 1est représenté 3 fois sur l'allèle eBSGFV-7 et 4 fois sur l'allèle eBSGFV-9. Nous pouvons noter l'existence d'un hot-spot présenté par une flèche sur la figure qui n'est présente qu'une fois malgré les répétitions de cette zone.

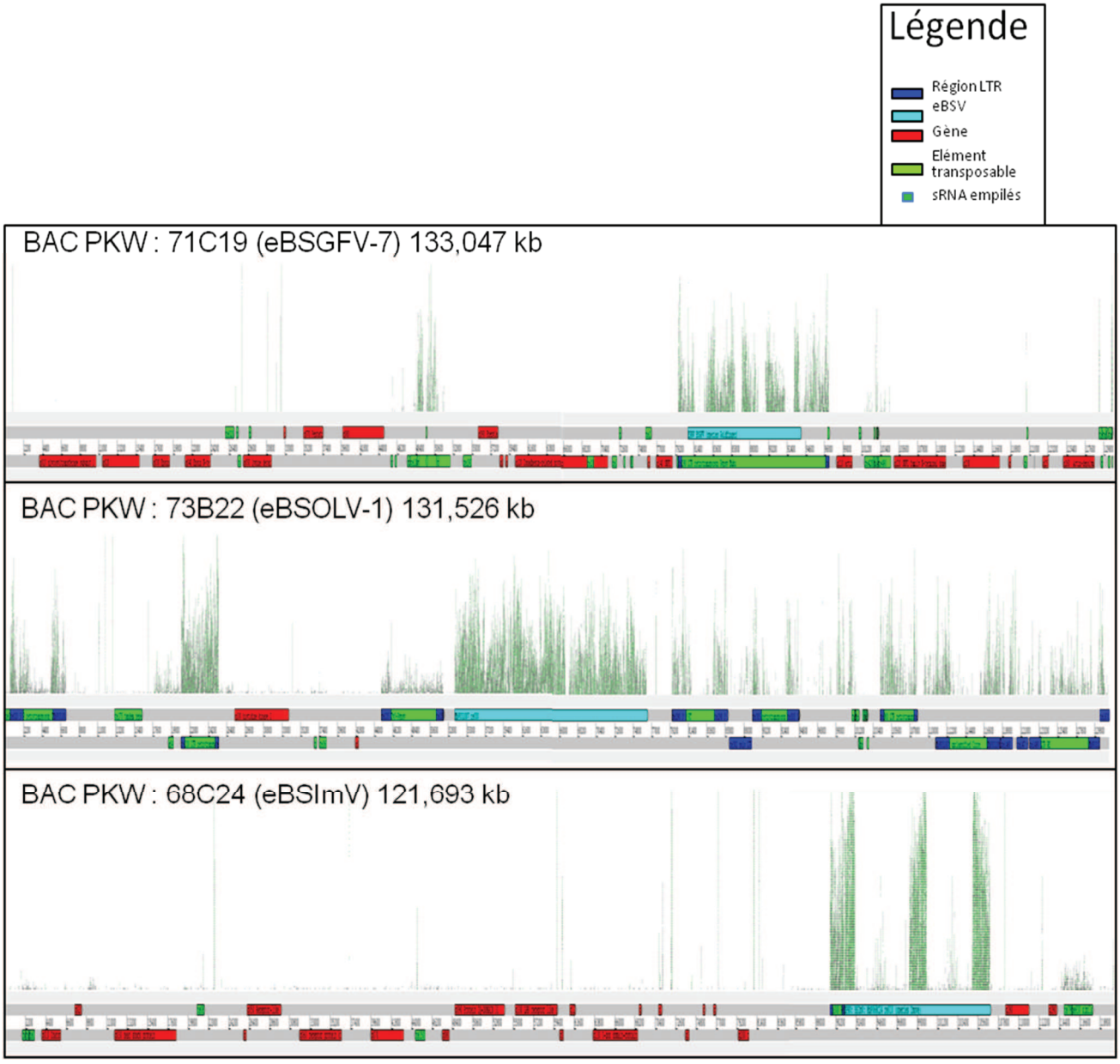


Figure 2-11 : Répartition des sARN produits par PKW sur les séquences de BAC contenant les eBSV

Représentation des sRNA de l'échantillon PKW sur la séquence des BAC porteurs d'eBSOLV-1, eBSGFV-7 ou eBSImV. Les BAC contenant eBSOLV-2 et eBSGFV-9 ne sont pas représentés car les séquences sont identiques aux séquences des BAC eBSOLV-1 et eBSGFV-7 respectivement puisque les intégrations sont alléliques (Chabannes et al., 2012). Les sARN sont représentés empilés les uns sur les autres quand ils se chevauchent ce qui donne une indication relative de la quantité de sARN produits pour une zone donnée. En rouge sont représentés les gènes putatifs, en vert les éléments transposables et en bleu les LTR 5' e 3'. Les alignements de séquences admettent jusqu'à deux mutations entre le vsARN et la séquence de référence.

La séquence correspondant à ce hot-spot sur l'eBSGFV possède une mutation qui la différencie des zones répétées de la même région sur le reste de l'eBSGFV.
Pour l'eBSOLV, le cold-spot de vsARN au niveau de l'IG correspond à la partie qui n'est présente que pour l'eBSOLV-1 et qui est la seule IG complète présente dans la séquence des eBSOLV (voir chapitre 1, article 1).
Une analyse fine au niveau des zones de jonction entre la plante et les séquences virales et des zones de réarrangement spécifiques des eBSV révèle que ces zones ne produisent pas ou très peu de vsARN (10 vsARN au maximum). C'est le cas pour les 3 eBSV infectieux (données non présentées).

De façon très intéressante, nos données indiquent que la production de vsARN se fait seulement à partir de séquences eBSV qui sont présentes au moins en deux copies et qui sont répétées inversées. Par exemple, l'eBSImV produit des vsARN seulement à partir de la région ORF3 présente en copies multiples et inversées. Inversement chez l'eBSOLV seule la zone de l'IG présente en copie unique ne produit aucun vsARN.
Ces informations confirment donc que les eBSV présents dans le génome de PKW produisent des vsARN de 24 nt du fait de leur structure réarrangée et non en fonction du génome viral lui même.

2-3-2 Répartition de la production de sARN dans le paysage génomique des eBSV

Sachant que l'environnement peut influencer les effets du silencing (Zhang et al., 2008 ; Hollister et Gaut, 2009), il nous a paru intéressant de regarder la production de sARN à proximité des eBSVs qui, comme je l'ai décrit dans le chapitre 1 (Article 1), varie considérablement d'un eBSV à l'autre. En effet, les eBSGFV et eBSImV sont présents dans des zones riches en gènes alors que l'eBSOLV est présent dans une zone riche en éléments transposables (Article 1, figure 11) ; l'eBSGFV et l'eBSOLV étant situés sur le même chromosome. Les 3 insertions eBSV sont néanmoins à proximité d'éléments transposables à LTR, voire à l'intérieur pour l'eBSGFV.
L'analyse bio-informatique a porté sur l'intégralité des séquences BAC porteurs des différents eBSV. La répartition des sARN totaux est illustrée par une représentation de type « Stacking » empilant les sARN qui se chevauchent sur la séquence référence. Cette représentation n'inclut donc pas une échelle quantitative (figure 2-11). Tout comme pour les zones virales répétées des eBSV, il n'est pas possible de savoir si les sARN proviennent des séquences représentées sur le BAC analysé ou de séquences identiques présentes dans une autre partie du génome de PKW. Contrairement aux données présentées jusqu'à présent

celles-ci acceptent jusqu'à deux mutations entre le sARN et la séquence de référence. Tous les éléments transposables complets situés à proximité des eBSV possèdent des sARN en quantité importante, la grande majorité étant de 24nt (données non présentées). Les éléments transposables à LTR produisent le plus de sARN particulièrement au niveau de leurs extrémités LTR en 3' et 5'. On note cependant quelques cas où ces ET à LTR produisent des sARN de façon plus uniforme tout le long de leur séquence (figure 2-11 BAC eBSOLV-1). Très peu de sARN correspondent aux gènes localisés à proximité des eBSV. Seuls quelques pics très ciblés de sARN de 21nt sont présents au niveau de certains correspondant à des miRNA (figure 2-11).

Les vsARN spécifiques de l'eBSOLV représentent 63% des sARN du BAC 73B22.

Pour le BAC 68C24 les vsARN de l'eBSImV représentent 56% des sARN du BAC. Il est intéressant de noter que le rétrotransposon à LTR situé à proximité représente aussi 26% des sARN totaux de ce BAC. La production de sARN est donc très concentrée autour de l'eBSV. Le rétrotransposon à LTR situé à coté de l'eBSImV possède plus en quantité relative de sARN spécifiques de sa séquence que l'eBSImV lui-même (tableau sup 2-5).

Enfin, l'eBSGFV présent sur le BAC 71C19 produit 91% des sARN du BAC soit la quasi-totalité des sARN produits. On retrouve quelques miARN ainsi que des sARN pour les deux éléments transposables complets. D'une manière générale, la quantité de sARN produits sur ce BAC est nettement inférieure aux quantités observées pour les 3 BAC analysés.

Au vu de ces résultats et de ceux présentés figure 2-6, il semble qu'il n'y ait pas de corrélation entre l'environnement génomique proche et la quantité de vsARN produits par l'eBSV. En effet, l'eBSOLV qui est situé dans une région riche en ET produit une quantité intermédiaire de vsARN par comparaison à l'eBSGFV et l'eBSImV qui sont tous les deux dans des BACs riches en gènes. Cependant, il se pourrait que ces régulations soient contrôlées sur des distances plus importantes que les 50-100 kb environnant puisqu'on observe dans le cas du BAC eBSGFV une production moindre de sARN que ce soit pour l'eBSV comme pour les ET.

3- Recherche de vsARN dans la diversité *M. balbisiana*

L'objectif de cette étude vise à déterminer le degré de conservation des régulations épigénétiques mises en évidence chez PKW au sein de l'espèce *M. balbisiana.* Les analyses ont été menées sur des génotypes représentatifs de la diversité *M. balbisiana* et dont le polymorphisme eBSV est le plus représentatif possible pour les trois espèces BSV concernées (eBSImV, eBSOLV et eBSGFV)(voir chapitre 1).

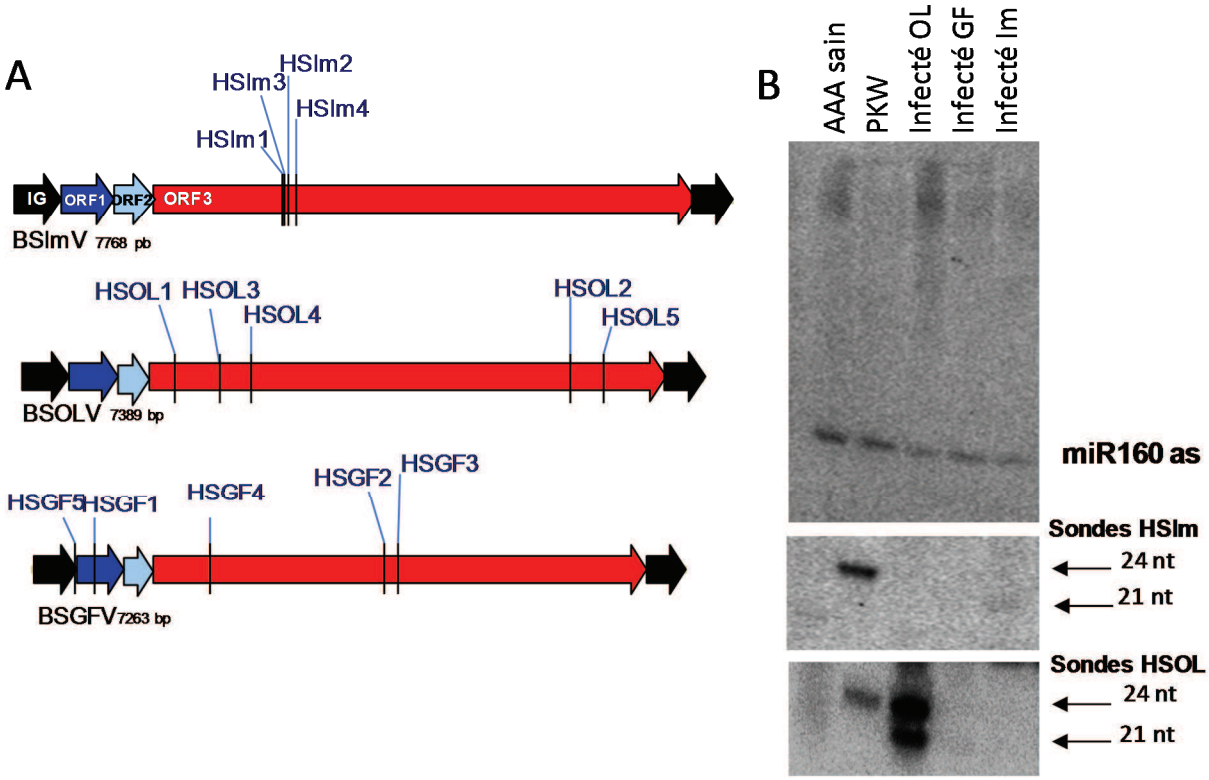


Figure 2-12 : Confirmation des régions « hot-spots » de vsARN spécifiques de eBSImV et eBSOLV chez PKW

A: Localisation des sondes de 25nt correspondantes aux régions « hot spot » de vsARN eBSImV et eBSOLV produits chez PKW sur les génomes viraux linéarisés correspondants.

B: Hybridation northern blot sPAGE 15% avec les sondes HSIm ou HSOL chacune en mélange pour rechercher les vsARN correspondants chez PKW. Les bananiers Cavendish sain et infectés par BSOLV, BSGFV et BSImV servent de contrôle. L'utilisation de la sonde mir160 antisens permet de vérifier la quantité de sARN pour chaque échantillon.

3-1 Développement de sondes spécifiques des hot-spots des eBSV.

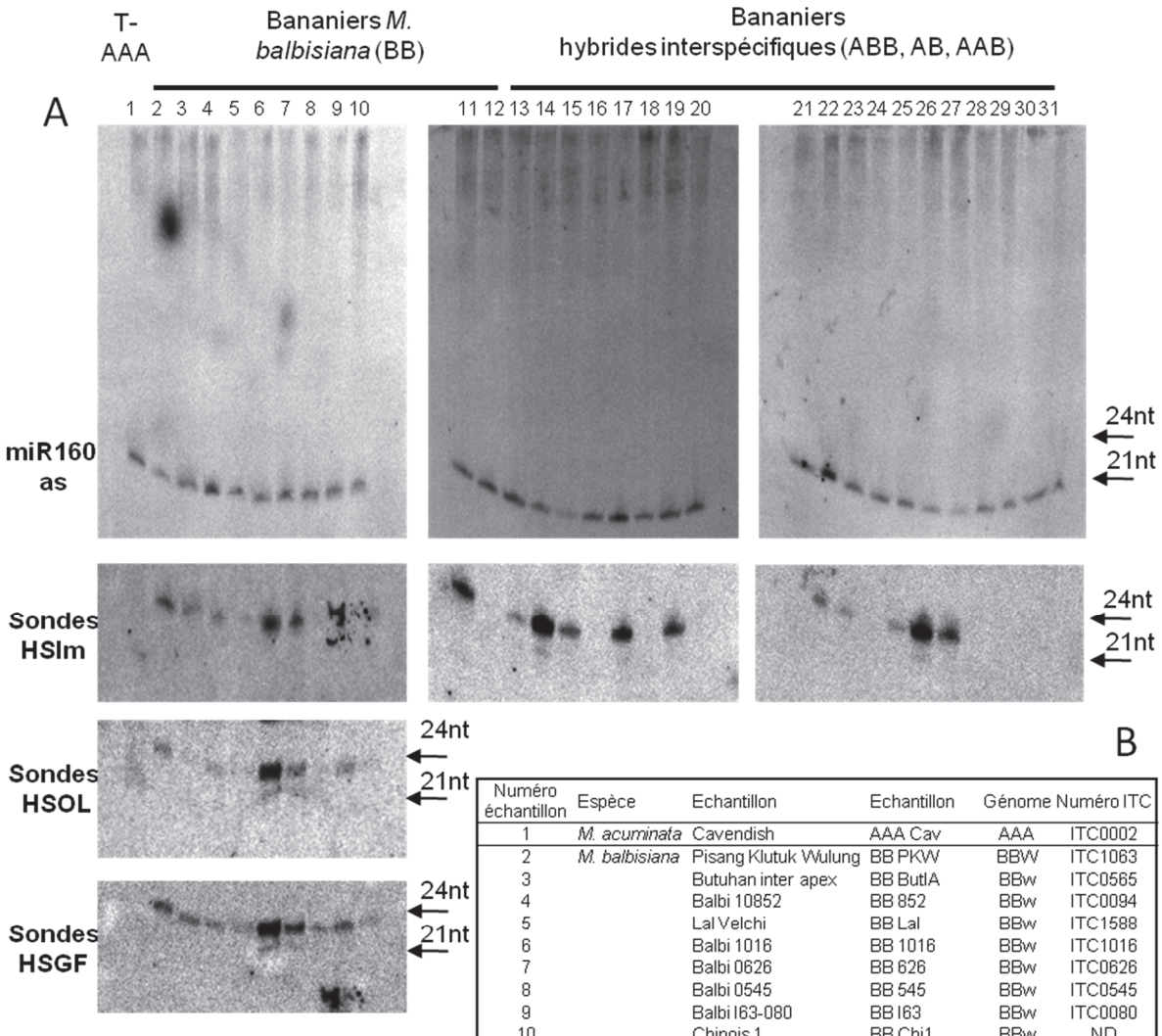
Les résultats du séquençage profond de vsARN chez PKW a permis d'identifier des hot-spots de vsARN pour chacune des espèces eBSV. Différentes sondes de 24 nt en orientation inverse du vsARN identifié ont ainsi été synthétisées à partir de ces zones clés, puis testées en northern Blot.

Nous avons pu, dans un premier temps, vérifier par l'intermédiaire de cette technique la validité des données de séquençage obtenues pour PKW (figure 2-12). Nous avons utilisé 4 et 5 sondes de 24nt pour BSImV et BSOLV respectivement; leur position sur la séquence génomique virale de référence est représentée sur la figure 2-12A. Une première hybridation des sARN de PKW et des bananiers Cavendish sain et infectés par BSOLV, BSGFV ou BSImV avec la sonde miRNA160 antisens utilisée comme témoin, a permis de vérifier la quantité et la qualité des sARN extraits pour les différents échantillons et d'avoir accessoirement un marqueur de taille de 21nt sur la membrane. Les hybridations ont été faites en mélangeant les 4 ou 5 sondes produites pour chacun des eBSV étudiés. Les analyses sont en cours de réalisation pour l'espèce Goldfinger (GF), les 5 sondes que nous avons dessinées, correspondent comme pour les autres eBSV aux hot-spots les plus importants mis en évidence chez PKW (figure 2-9B).

Comme attendu, une forte production de vsARN de 24nt est détectée avec les sondes des espèces BSImV et BSOLV chez PKW (figure 2-12B). Ces sondes mettent aussi en évidence des vsARN chez les bananiers Cavendish infectés par BSImV ou BSOLV. . L'intensité de la bande pour le bananier infecté par BSImV est cependant faible corroborant les données de séquençage qui montrent une production faible de vsARN pour cette zone là (figure 2-9A). La taille des vsARN produits pour les bananiers infectés par BSImV et BSOLV est de 21nt et 21 et 24nt respectivement. Le bananier sain et le bananier infecté par BSGFV ne présentent aucun signal d'hybridation comme attendu.

3-2 Mise en évidence de la production de vsARN chez des plantes représentant la diversité *Musa balbisiana*.

Les sondes présentées au paragraphe précédent ont été utilisées pour rechercher la présence de vsARN chez des plantes représentatives de la diversité *M. balbisiana* et du polymorphisme eBSV rencontré pour les BSV étudiés. L'échantillonnage testé est constitué de 11 plantes *M. balbisiana* diploïdes (BB) dont PKW, 4 hybrides interspécifiques *M. acuminata* et *M. balbisiana* diploïdes pour le génome B (ABB), 13 hybrides interspécifiques



B

Numéro échantillon	Espèce	Echantillon	Echantillon	Génome	Numéro ITC
1	*M. acuminata*	Cavendish	AAA Cav	AAA	ITC0002
2	*M. balbisiana*	Pisang Klutuk Wulung	BB PKW	BBW	ITC1063
3		Butuhan inter apex	BB ButIA	BBw	ITC0565
4		Balbi 10852	BB 852	BBw	ITC0094
5		Lal Velchi	BB Lal	BBw	ITC1588
6		Balbi 1016	BB 1016	BBw	ITC1016
7		Balbi 0626	BB 626	BBw	ITC0626
8		Balbi 0545	BB 545	BBw	ITC0545
9		Balbi I63-080	BB I63	BBw	ITC0080
10		Chinois 1	BB Chi1	BBw	ND
11		Lep Chang Kut	BB LCK	BBBcv	ITC0647
12		Eti Kehel	BB EK	BBw	ITC0271
13	Hybride interspécifique	Fougamou 1	ABB Foug	ABB	ITC0101
14		ND	ND	ND	ND
15		Yangambi n°2	AAB Yang2	AAB	ITC1275
16		Safet Velchi	ABSafet	AB	ITC0245
17		Pelipita	ABB Peli	ABB	ITC0396
18		Blue Java	ABB Blue	ABB	ITC0361
19		Daru	ABB Daru	ABB	ITC0795
20		Corne 1	AAB Corn	AAB	ITC0754
21		Boung Fu	AAB Bong	AAB	ITC0940
22		Chuoi Mat	AAB ChuM	AAB	ITC1381
23		Bengani	AAB Beng	ABB	ITC0855
24		Kunaimp	AAB Kuna	AAB	ITC0836
25		Mai'a Popo'ulu Moa	AAB Moa	AAB	ITC1169
26		Tigua	AAB Tig	AAB	ITC0875
27		Gamaha	AAB Gama	AAB	ITC1006
28		kluai roi wii	AAB Kluai	AAB	ITC0535
29		tang tang	AAB Tang	AAB	ITC0901
30		Dimaemamosi	AAB Dimae	AAB	ITC0920
31		Pisang Nangka	AAB Nang	AAB	ITC1062

Figure 2-13 : Mise en évidence de vsARN dans la diversité *M. balbisiana*

A : Analyse des vsARN extraits à partir de bananiers *M. balbisiana* diploïdes (génotype BB) et d'hybrides bananiers interspécifiques (AAB, ABB et AB). La plante *M. acuminata* (AAA) sert de témoin négatif. Les membranes sont hybridées successivement avec mir160 (antisens) puis avec un mélange de sondes correspondantes aux 4 régions chaudes de production des vsARN les plus importantes de l'eBSImV de PKW (figure 12A). Les accessions diploïdes *M. balbisiana* de 2 à 10 ont été hybridées avec un mélange de sondes des régions « hot spot » de l'eBSOLV (sondes HSOL, figure 12A) ou de l'eBSGFV (sonde HSGF, figure 12A) identifiées chez PKW.

B : Tableau récapitulatif des accessions utilisées pour cette analyse

haploïdes pour le génome B (AAB) et une plante hybride interspécifique AB. Toutes ces plantes ont été testées préalablement en IC-PCR pour les 3 espèces BSV étudiées et sont indemnes de particules virales. Elles proviennent de zones géographiques différentes et présentent différents niveaux de ploïdie vis à vis du génome B en lien avec le réveil ou pas des eBSV et la libération de génomes viraux (voir article 2). Les différentes caractéristiques de ces accessions sont rapportées dans le tableau de la figure 2-13B. Leurs allèles ainsi que leurs degrés de modification par rapport aux eBSV présents chez PKW sont présentés dans les dendrogrammes qui ont été construits à partir des résultats obtenus dans l'article 2 (chapitre 1, article 2) (figure 2-14).

Les résultats des hybridations sont présentés dans la figure 2-13. Trois membranes différentes ont été réalisées afin d'analyser tous les individus sélectionnés. L'accession 14 s'est révélée de génotype non confirmé et ne sera pas exploitée. Les accessions 28/29 (Kluai Roi Wii et Tang Tang), ne sont pas présentées dans les arbres car elles ne faisaient pas partie de l'échantillonnage utilisé pour l'article 2.

L'hybridation réalisée avec la sonde miRNA160 antisens indique que les accessions ont des sARN de bonne qualité et en quantité comparable.

Les hybridations virales ont été réalisées comme précédemment avec les sondes utilisées en mélange pour chacune des trois espèces eBSV. Les sondes spécifiques de l'eBSImV ont été utilisées avec succès sur les trois membranes. Seule la membrane présentant les accessions de type *M. balbisiana* diploïdes (BB) est présentée pour les hybridations avec les sondes spécifiques de l'eBSGFV et l'eBSOLV. En effet, les résultats apparaissent plus difficiles à obtenir, il est possible que les quantités faibles de vsARN produits chez ces individus hybrides interspécifiques rendent complexe la détection par cette technique.

L'hybridation avec les sondes eBSImV (sonde HSIm) ne montre aucun signal pour le témoin négatif et les accessions non porteuses d'eBSImV (échantillons 1, 16, 18, 20, 21, 24, 31). Les accessions possédant l'intégration telle que décrite chez PKW produisent systématiquement des vsARN de 24nt à part l'accession 8 (BB-0545). Les accessions avec des eBSImV peu modifiées produisent des vsARN de 24nt comme PKW à l'exception de l'accession 30 (Dimaemamosi). Concernant les accessions ayant des intégrations fortement modifiées, on observe deux cas de figure : les bananiers 5 et 12 diploïdes BB ne présentent aucun signal d'hybridation contrairement aux accessions 15 et 26 AAB qui enregistrent un signal à 24nt et 21 et 24nt respectivement.

Les résultats d'hybridations avec les sondes eBSGFV (sonde HSGF) montrent que les bananiers *M. balbisiana* diploïdes BB produisent tous des vsARN de 24nt tels que ceux mis

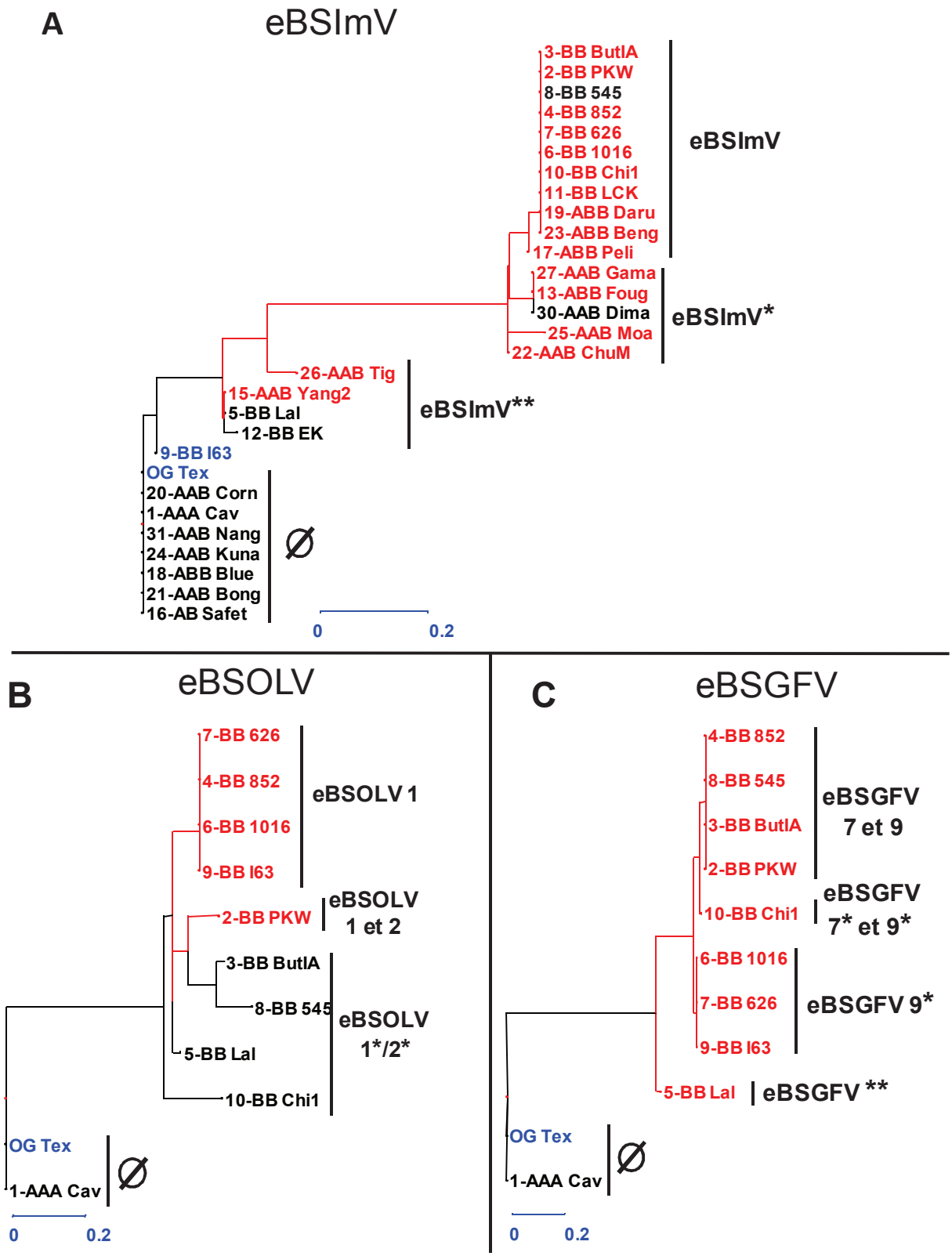


Figure 2-14 : Relation entre production de vsARN et structure des eBSV au sein de la diversité *M. balbisiana*

Arbres de diversité construits selon la méthode du Neighbour joining à partir des données obtenues en Southern Blot et PCR pour eBSImV (A), eBSOLV (B) et BSGFV (C) (Article 2, Chapitre 1). Ces arbres sont construits à partir des accessions renseignées figure13. Ils sont enracinés avec les accessions ne possédant pas d'eBSV (AAA Cav et l'out-group *M. Textilis*). Les échantillons en rouge produisent les vsARN de 24nt attendus contrairement à ceux en noir qui n'en produisent pas. En bleu sont représentés les échantillons non déterminés. Une étoile (*) indique des mutations légères par rapport à l'eBSV présent dans le génome de PKW et deux étoiles (**) des différences importantes de structure.

en évidence chez PKW. Ces résultats sont conformes à ce que l'on pouvait attendre puisque tous les bananiers sont porteurs de l'intégration eBSGFV telle que décrite chez PKW ou très légèrement modifiée. Aucune différence allélique n'a pu être observée car les accessions sont toutes porteuses des deux allèles ou du seul allèle eBSGFV-9.
Les résultats d'hybridations avec les sondes spécifiques eBSOLV (sonde HSOL) montrent que les accessions diploïdes BB ayant au moins l'allèle eBSOLV-1 tel que décrit chez PKW produisent en plus ou moins grande quantité les vsARN à 24nt comme attendu (accessions 2, 4, 6, 7, et 9). Par contre les accessions porteuses d'eBSOLV modifié ne produisent pas de vsARN (figure 2-14).

Ces résultats montrent que les bananiers possédant le même génotype que PKW ainsi que les mêmes intégrations eBSV produisent des vsARN à partir des mêmes hot-spots. A l'inverse, des différences structurelles importantes au niveau des eBSV ne permettent plus de mettre en évidence la production de vsARN à partir des zones ciblées par nos sondes. Cela ne signifie en rien que ces eBSV ne produisent plus de vsARN mais seulement que les zones de production identifiées chez PKW ne sont plus présentes et/ou que les zones de production sont différentes.

Discussion

L'existence et le maintien de séquences virales fonctionnelles et infectieuses dans le génome du bananier PKW interroge sur le mode de régulation de telles séquences ainsi que sur les effets potentiellement bénéfiques pour les bananiers. Ces questions nous ont amenés dans un premier temps à préciser le mécanisme de défense mis en place par PKW pour éviter une infection et à préciser le degré de spécificité vis à vis du virus. Sur la base des différents travaux menés sur les séquences virales endogènes (EVE) (Noreen et al., 2007) ainsi que sur les éléments transposables (ET) (Rigal et Matthieu, 2011) et après avoir montré leurs similitudes avec les eBSV, notre étude a porté sur les régulations épigénétiques en lien avec la défense des plantes vis à vis des séquences étrangères invasives.

Les séquençages haut débit des sARN de bananiers ont mis en évidence que les séquences eBSV, fonctionnelles ou non, présentes dans le génome du bananier PKW étaient productrices de vsARN. C'est également le cas de bananiers non porteurs d'eBSV mais infectés par le virus du BSV qui produisent des vsARN à partir du génome viral. Ces données

indiquent que les mécanismes de régulation des séquences BSV au sens large (BSV + eBSV) font appel au processus générique de l'ARNi.

1-Nature des mécanismes de régulation des eBSV chez PKW

Nature de la régulation

Nous avons montré que les vsARN produits chez PKW pour les 4 eBSV sont en majorité des vsARN de 24nt. Les différentes études menées montrent que les sARN de cette taille sont associés à un contrôle de type TGS (Transcriptional Gene Silencing) ayant pour conséquence principale la méthylation de l'ADN ou la modification des histones (Zilberman et al., 2006 ; Lisch, 2009 ; Bucher et al., 2012). Dans notre étude, ces vsARN seraient la résultante d'un mécanisme épigénétique qui empêcherait l'expression massive de la séquence incriminée, ici les eBSV. Ce résultat est en accord avec ceux publiés par Noreen et al., (2007) pour les ePVCV chez le pétunia qui montrent que les séquences virales intégrées sont associées à une production de vsARN et une modification des histones au niveau H3K9 réduisant la transcription de l'ADN. La production majeure de vsARN de 24nt dans notre cas indique donc un mécanisme de type TGS et laissent penser à la mise en place d'une méthylation de l'ADN ou de modification des histones soutenant une faible production de transcrits. La présence de vsARN de 21nt en plus des 24nt majoritaires dans les données issues du séquençage profond laisse penser que l'ARN polymérase 2 produirait quand même des transcrits cibles qui seraient ensuite dégradés par le PTGS pour produire des vsARN de 21nt en faible quantité. Cette double régulation TGS-PTGS a d'ailleurs déjà été démontrée pour les rétroéléments (Rigal et Matthieu, 2011). Cette production de 21 nt est effective et homogène pour les 4 eBSV présents dans le génome de PKW quel que soit leur potentiel infectieux (figure 2-7).

Les connaissances récentes acquises sur les ET indiquent que la production des sARN de 24nt est répartie de manière homogène sur ces éléments transposables (Zilberman et al., 2006 ; Mirouze et al., 2009). Cependant les études, menées sur les rétrotransposons, n'ont pas encore mis en évidence les cibles privilégiées de l'ARN polymérase IV qui est impliquée dans la synthèse des sARN de 24nt (Huettel et al., 2006, 2007). Et il n'existe pas actuellement de données traitant de la mise en place du processus de TGS à la suite d'une néo-insertion de rétroélément. Il est donc difficile de connaître de manière dynamique la mise en place et le fonctionnement précis du TGS suite à cela. Des données préliminaires semblent néanmoins indiquer que la mise en place de l'ARNi est rapide afin de lutter immédiatement contre

l'expression de séquences parasites. Les premières étapes impliqueraient le processus de PTGS qui glisserait vers un processus de TGS au bout de plusieurs générations stabilisant la mise sous contrôle de la séquence par le mécanisme de l'ARNi. La dynamique du système est en passe d'être démontrée pour des rétroéléments présents dans le génome d'*Arabidopsis thaliana* par Voinnet et al., et Tanurdzic et al., (données non publiées). Nous pouvons néanmoins formuler des hypothèses vis à vis de ce que nous avons obtenu sur les eBSV. Les résultats suggèrent que la production de vsARN évolue dans le temps une fois l'intégration réalisée certainement en lien avec les réarrangements que subit l'eBSV. Dans ce contexte là, les données obtenues pour l'eBSImV pourraient indiquer que le passage du PTGS au TGS est récent pour cette intégration chez PKW. On peut aussi penser qu'une fois la structure de l'eBSV fortement réarrangée, la quantité de vsARN produit diminuerait avec le temps puisqu'ils ne serviraient plus qu'à entretenir le système. Cette hypothèse prend tout son sens au vu des données collectées avec une production forte pour l'eBSImV qui est une intégration récente et peu réarrangée versus eBSOLV ou eBSGFV qui produisent moins de vsARN car plus anciennes et plus réarrangées. On ne peut cependant pas exclure que la quantité de vsARN soit aussi liée à une régulation dépendante de l'environnement des zones d'intégration des eBSV (voir partie résultat).

Rôle de la structure pour la régulation

L'analyse des zones productrices de vsARN chez PKW le long des eBSV a montré qu'elles étaient différentes pour chacune des 4 espèces BSV intégrées (figure 2-9). Cette production est ainsi répartie de manière homogène sur toute la séquence eBSOLV, ne concerne que certaines parties d'eBSGFV et d'eBSMyV et est focalisée sur une seule zone de l'ORF3 pour l'eBSImV (figure 2-10). L'analyse des structures eBSV montre clairement que seules les zones répétées inversées donnent lieu à une production significative de vsARN quelle que soit l'espèce eBSV analysée. Ceci est facilement observable pour l'eBSImV qui est peu réarrangée et qui produit donc des vsARN uniquement pour la zone dupliquée inversée de l'ORF3 (figure 2-10). Des analyses plus fines pour l'eBSGFV et l'eBSOLV montrent exactement la même chose (données non présentées). Il apparaît également que l'eBSMyV, bien que non infectieuse, soit soumise au même type de régulation que les autres eBSV. Il faut cependant noter que ces analyses devront être confirmées une fois les séquences finales de l'eBSMyV disponibles.

Ces structures répétées /inversées peuvent donner lieu, lors de la transcription des eBSV, à des appariements aberrants plus ou moins complexes de type ARNdb facilement reconnus

par la plante comme ARN étranger. Ces structures sont le support indispensable à la mise en place de l'ARNi. La production de vsARN semble donc être structure eBSV-dépendante et non organisation génomique viral-dépendant.

Rôle du contexte génomique dans la régulation

Les 4 eBSV apparaissent intégrés dans des régions génomiques différentes. Ces régions sont riches en ET pour BSOLV et BSMyV et riches en gènes pour eBSGFV et eBSImV. Par contre, tous les eBSV sont soit à proximité soit intégrés dans des rétroéléments. Ces différents contextes ne modifient en rien le mode de régulation des différents eBSV qui sont tous contrôlés par du TGS principalement. Seules les quantités de vsARN produites sont différentes. Cette variation pourrait traduire une régulation TGS plus ou moins forte en fonction du lieu d'insertion et/ou d'une ancienneté d'intégration qui nécessiterait moins de vsARN pour maintenir le système actif (cf §1-Nature des mécanismes de régulation des eBSV chez PKW). La quantité de vsARN produits à partir des eBSV est maximale pour l'eBSImV puis décroît progressivement chez eBSMyV puis chez eBSOLV pour atteindre une production la plus faible pour l'eBSGFV (figure 2-6). Cela pourrait indiquer que la plante a mis en place un niveau élevé de silencing pour contrôler l'eBSImV qui ne possède qu'une seule séquence répétée inversée donc une seule cible potentielle pour produire des ARNdb. Par contre, une fois les eBSV très réarrangés comme c'est le cas pour les trois autres eBSV, la production de vsARN diminue du fait d'un plus grand nombre de zones répétées inversées offrant un choix plus important aux TGS pour les contrôler. De plus, la reconstitution d'un génome viral fonctionnel à partir des structures fortement réarrangées nécessite plusieurs étapes de recombinaison homologue (Iskra-Caruana et al., 2010 et Chabannes, communication personnelle) et est plus complexe que pour l'eBSImV où un génome viral circulaire peut être obtenu par simple transcription. Ce dernier argument renforce l'idée qu'une production moindre de vsARN pour contrôler ces eBSV est suffisante car le risque de restitution virale est plus faible. Le cas de l'eBSMyV est difficile à statuer car la structure moléculaire définitive n'est pas connue. D'autre part, les données quantitatives de vsARN semblent indiquer que l'environnement proche c'est-à-dire à l'échelle du BAC n'influence pas la production de vsARN car chez eBSOLV et eBSMyV qui sont dans des zones riches en TE la production est supérieure à celle de l'eBSGFV qui est dans une zone riche en gènes mais inférieure à celle de l'eBSImV qui est, elle aussi, entourée de gènes. Il n'y a donc pas de corrélation évidente entre quantité de vsARN et environnement de proximité. Par contre, on peut envisager qu'une régulation à plus grande échelle soit mise en place par la plante car

l'eBSGFV et l'eBSOLV qui sont tous deux portés par le chromosome 1 produisent nettement moins de vsARN que l'eBSImV qui est situé sur le chromosome 2.

L'analyse des régions génomiques encadrant les eBSV chez PKW a montré que tous les éléments transposables situés dans ces régions étaient la cible de vsARN majoritairement de 24 nt ; confirmant que la régulation observée pour les eBSV était étendue à la zone d'intégration. Ces études épigénétiques qui sont les premières sur bananier semblent confirmer les données obtenues sur d'autres plantes concernant le fonctionnement du TGS. On note par exemple que les séquences Long Terminal Repeat (LTR) des rétrotransposons sont une cible privilégiée des sARN comme chez *Arabidopsis thaliana* ou le maïs (Zilberman et al., 2006 ; Cantu et al., 2010). D'après les auteurs de ces travaux, la sur-représentation de ces séquences LTR par rapport à celles situées au niveau de la zone centrale des RE expliquerait la sur-production de sARN (Cantu et al., 2010). Cependant ces arguments ne semblent pas se vérifier si l'on regarde à une échelle plus large que le simple TE. En effet, d'une manière générale, la production de sARN correspondant aux séquences des rétroéléments présents sur les BAC porteurs des eBSV est inférieure ou similaire à celle des eBSV à l'exception du TE copia proche de l'eBSImV. Il faut aussi garder à l'esprit que ces sARN dirigés contre les TE peuvent provenir de TE identiques situés dans une zone différente du génome ce qui tend à minorer davantage les quantités de sARN observées pour ces TE sur les BACs. Ces résultats montrent que le nombre de copies n'est pas un argument suffisant pour expliquer les différences de quantité de sARN car les eBSV bien que présents de façon unique dans le génome produisent autant, voir plus, de sARN que les ET présents à proximité. Il est possible par contre que l'âge des intégrations soit important pour expliquer la production de sARN. En effet, l'article 1 (Chapitre 1) montre au travers d'études de synténie entre les BACs appartenant à l'espèce *M. balbisiana* et le génome de l'espèce *M. acuminata* qui vient d'être séquencé (D'Hont et al., 2012) que les rétroéléments des BACs contenant l'eBSImV et eBSGFV étaient présents sur les deux génomes *Musa* alors que les eBSV ne sont présents que chez les bananiers *M. balbisiana.* Ces données montrent que les eBSV se sont intégrés après les ET. Cette différence d'âge d'intégration pourrait expliquer les différences de quantités de sARN produites entre ces séquences.

La présence de rétroélements à proximité immédiate des eBSV pourrait, au-delà d'un rôle potentiel dans les mécanismes d'intégration (introduction §4-1), faciliter aussi la mise en place de la régulation de type TGS au niveau des eBSV. En effet, (Zhang et al., 2008) ont montré chez *Arabidospsis thaliana* que la présence des éléments transposables avait un

effet sur les gènes situés à proximité en provoquant une baisse de leur expression. D'après eux, le TGS mis en place pour les éléments transposables serait transmis aux séquences environnantes (Zhang et al., 2008 ; Hollister et Gaut, 2009). Ces données pourraient expliquer la mise en place du TGS chez les eBSV qui sont tous très proches d'éléments transposables sous contrainte de TGS. Cette observation est exacte même dans le cas des eBSGFV et eBSImV qui sont pourtant dans des zones riches en gènes car ils sont situés dans un ET pour l'eBSGFV et à proximité directe (moins de 30pb) pour l'eBSImV. De manière surprenante, l'eBSOLV et les eBSMyV qui sont dans des zones riches en ET, sont plus éloignés du premier ET que l'eBSImV. Cette proximité différente avec des ET pourrait s'expliquer par le fait que les zones riches en gènes sont moins sujettes au TGS et doivent donc « regrouper » toutes les séquences à risque au même endroit, alors que dans les zones riches en ET, le TGS agit de manière plus globale afin de méthyler la zone dans sa totalité.

Les différents arguments développés dans ce paragraphe tendent à montrer l'importance des réarrangements et de l'âge des intégrations dans la production des sARN en particulier de 24nt. Ainsi seules les zones inversées répétées produisent des sARN et il semble que plus l'eBSV est ancien, plus il accumule ce genre de zones inductrices. Il semblerait également que le contexte génomique immédiat (quelques centaines de pb autour de l'eBSV) jouerait un rôle dans la mise en place du TGS au niveau des eBSV, et que l'environnement à plus large échelle (centaines de Mb) influencerait le niveau de régulation nécessaire.

2- Mécanismes de défense antivirale mis en place par le bananier contre le BSV

L'étude de la production de sARN chez des bananiers infectés par le BSV montre que les bananiers possèdent des mécanismes de défense afin de lutter contre l'infection du BSV. Ce mécanisme connu sous le nom de PTGS a déjà été mis en évidence dans de très nombreuses interactions virus/hôtes. Il génère une production de sARN de 21nt qui ont pour cible séquence spécifique les ARN viraux prégénomiques et induisent leur dégradation avant traduction (Voinnet, 2005). Dans le cas du bananier, nous avons pu mettre en évidence une production de vsARN de 21 nt pour les 4 espèces BSV testées lors d'infections de bananiers Cavendish, qui sont dépourvus d'eBSV, correspondant à ces 4 espèces virales. De façon intéressante, nous avons montré que les zones virales génératrices de vsARN sont différentes de celles impliquées dans la production des vsARN de 24nt des eBSV de PKW. D'autre part, nos résultats montrent que la région située entre la fin de l'IG et le début de l'ORF 3 est une zone préférentielle pour la production de vsARN quelle que soit l'espèce BSV

considérée (figure 2-8). Cette région pourrait produire des vsARN ayant des conformations spécifiques de type « hairpin » permettant une production facilitée d'ARNdb nécessaire au PTGS.

Il est intéressant de noter que le BSOLV produit une quantité de vsARN 3 à 4 fois supérieure aux autres BSV ce qui peut correspondre à une charge virale différente dans le bananier infecté sélectionné et/ou à une phase de multiplication du virus dans la plante différente. Le BSOLV est l'espèce virale dont la prévalence est la plus grande, qui fait de graves dégâts sur les bananiers *M. acuminata*, et qui semble la plus adaptée à son hôte. Outre une quantité de vsARN supérieure, on note aussi que la répartition entre les sARN de 21nt à 24nt est plus homogène par rapport aux autres bananiers infectés (figure 2-7). Il est donc envisageable qu'au delà d'un certain seuil de virus, la plante, se sentant en danger, rééquilibre les mécanismes de défense de type TGS et PTGS pour réguler plus efficacement la multiplication virale. Comme cela a été décrit par Raja et al. (2008), le TGS viendrait dans ce cas méthyler l'ADN viral présent dans le noyau des cellules alors que le PTGS ciblerait les ARN présents dans le cytoplasme. Ce qui est vrai pour BSOLV l'est aussi pour les trois autres espèces BSV mais avec une quantité de vsARN de 24nt moindre. De façon surprenante on observe que ces vsARN de 24nt présents chez les bananiers infectés par BSImV, BSMyV et BSGFV sont en quantité similaire ou légèrement inférieure à ce qu'ils sont chez PKW au niveau des eBSV de même espèce. Bien que les hot-spots soient différents de ceux des eBSV. Ces résultats confirment que le TGS joue un rôle non négligeable dans la défense antivirale contre les BSV. Cette possibilité a d'ailleurs été envisagée par Hohn et Vasquez en 2011. Ils supposent que les virus à ADN et les *Caulimoviridae* en particulier seraient régulés par ce type de mécanisme au cours de la phase de formation du minichromosome dans le noyau de la cellule hôte.

La présence limitée de vsARN au niveau de la zone intergénique indique que la région leader du BSV initiant la transcription de l'ADN viral mis en évidence par Pooggin et al., (1999) n'est pas une source importante de production de vsARN comme c'est le cas chez le CaMV qui appartient lui aussi à la famille *Caulimoviridae*. La stratégie de leurre décrite par Blevins et al., 2011 ne semble donc pas universelle au sein de cette famille mais est peut-être propre au genre *caulimovirus*. Il est important de préciser que cette séquence leader est extrêmement courte chez le BSV contrairement au CaMV ce qui pourrait expliquer les différences observées entre ces deux virus. En effet, les positions du primer binding site (PBS) et du signal Poly (A) dans la région leader du BSV sont plus proches que pour le CaMV

et proposent une boucle qui serait créée par le transcrit juste après le PBS. L'intérêt d'utiliser cette zone pour servir de leurres, serait alors beaucoup plus limité car elle concerne une zone importante pour la réplication du BSV.

Comme évoqué plus haut, la région située entre la fin de l'IG et le début de l'ORF 3 est une zone préférentielle pour la production de vsARN quelle que soit l'espèce BSV considérée. Le rôle des ORF1 et 2 n'a pas encore été déterminé. Ces protéines pourraient avoir une ou des fonction(s) essentielle(s) dans l'interaction avec la plante justifiant d'être la cible de très nombreux vsARN. Un rôle dans le contournement des défenses antivirales de la plante, comme c'est le cas de nombreuses protéines virales identifiées à ce jour, peut dès lors être proposé. Depuis quelques années plusieurs protéines « suppresseur de silencing » ont été mises en évidence. Comme par exemple la protéine P6 du CaMV qui a été identifiée comme interagissant avec une protéine impliquée dans le PTGS et qui jouerait un rôle dans la capacité du virus à pouvoir infecter son hôte (Haas et al., 2008 ; Love et al., 2007). Il est donc indispensable de déterminer si le BSV possède aussi de telles protéines et si elles peuvent être impliquées dans la détermination des zones productrices de vsARN.

3- Conservation des mécanismes de régulation des eBSV dans la diversité *Musa*

Les études s'intéressant à la transmission des mécanismes épigénétiques entre individus d'une même espèce ont montré que la méthylation des ET ainsi que la production de sARN associés étaient conservés, que ce soit au cours du temps (Becker et al., 2011) ou entre les individus (Vaughn et al., 2007). Ces études montrent l'importance pour les plantes de garder les séquences parasites non exprimées lorsque les conditions de culture sont stables. Par contre en cas de stress, comme ceux qui peuvent exister lors d'un croisement génétique interspécifique (Lockton et Gaut, 2010) ou lors de changements de température (Ito et al., 2010) il a été montré que les ET pouvaient être activés, suite à un relâchement temporaire du système de régulation, et devenir une menace pour le génome de leur hôte.

L'étude que nous avons réalisée a montré qu'au sein de l'espèce *M. balbisiana* les zones impliquées dans la production de vsARN chez PKW pour réguler l'expression des eBSV semblaient conservés, en particulier quand les eBSV présentaient des structures identiques et représentaient donc le même risque pour le bananier. Lorsque les structures des eBSV changent, les hot-spots de production de sARN semblent différents. Les individus porteurs de l'eBSImV très incomplet ou les porteurs de l'eBSOLV muté présentent une absence de vsARN principalement en raison de l'absence de la zone incriminée. Nous pouvons aussi penser que l'absence de risque que représentent ces eBSV très dégénérés peut influencer la

production de vsARN au cours du temps. Nous n'avons malheureusement pas pu vérifier pour l'eBSGFV si la production de vsARN était corrélée ou non à la présence de l'allèle infectieux car nous n'avons pas pu identifier d'accession BB avec l'allèle eBSGFV-7 uniquement.

Nos données concernant la conservation du TGS chez les bananiers interspécifiques sont plus limitées. Il semble cependant que pour l'eBSImV les hot-spots de vsARN identifiés chez PKW soient conservés même pour des bananiers possédant des intégrations dégénérées. Les difficultés rencontrées pour produire des résultats pour les deux autres eBSV semblent indiquer que les limites de sensibilité de détection des vsARN par la technique de northern blot ont été atteintes. En effet, si l'on cumule les valeurs absolues des vsARN reconnus par les 4 ou 5 sondes servant aux hybridations on observe que les vsARN de l'eBSImV sont 5 à 6 fois plus nombreux (environ 11000 reads) que ceux des sondes eBSGFV (2100 reads) et eBSOLV (2200 reads) d'après les données obtenues chez PKW (figure 2-9). Ceci est d'ailleurs confirmé par les hybridations réalisées sur PKW (figure 2-12) qui montrent un signal d'hybridation plus faible avec les sondes HSOL qu'avec les sondes HSIm. Dans le cas des hybrides, les intégrations étant à l'état haploïde, on peut logiquement envisager un effet dose qui expliquerait que les vsARN ciblant eBSGFV et eBSOLV soient non détectés. Il est d'ailleurs intéressant de noter que ce sont les eBSOLV et eBSGFV qui sont le plus souvent observés comme produisant des particules virales chez les bananiers hybrides (Galzi et Caruana, communication personnelle). Ces observations sont certainement à corréler avec le fait que ces eBSV produisent moins de vsARN et que le TGS mis en place soit plus facilement relâché.

L'étude des mécanismes de régulation des eBSV chez PKW a permis de montrer que les eBSV étaient soumis au processus de régulation de l'ARNi de type TGS pour ne pas permettre la restitution de particules virales. Les hot-spots impliqués dans la production de vsARN sont différents entre les différents eBSV et sont dépendants de la structure de l'eBSV. Ceci est renforcé par les résultats obtenus sur les diploïdes *M. balbisiana* qui possèdent des eBSV identiques ou très proches et pour lesquels le mécanisme semble très conservé.

Données supplémentaires

Abbreviation	Nom	Séquence	Taille en nt
OL_test_1s	OlqPCR9	GAATGGGGGAAAAATTGGTT	20
OL_test_2as	OlqPCR10	TCACGATGCCCATGTTTCTA	20
OL_test_3s	OlqPCR11	TTCAAGGGAGTCAACCAAGG	20
OL_test_4s	Cl 1	ATGGCCTTAATAGTCTTTCGTGAT	24
OL_test_5s	sig1 eBSOLV R	CCTGGTCTGCACAGAGATGA	20
OL_test_6s	sig1 eBSOLV F	TTCGAGGAGTCAACGGAGTC	20
OL_test_7as	Delanoy 5445F	TTGGTGTTTAACTATAAGAGGCTGAA	26
OL_test_8as	Delanoy 6688R	TTATTGCATCCACATTTGAAAAC	23
OL_test_9as	OIR	GCTCACTCCGCATCTTATCAGTC	23
OL_test_10as	Marker2-BSOLV(2) F	ACTCGCACAAAGTGAACTCG	20
miRNA160as	miR160a	TGGCATACAGGGAGCCAGGCA	21
HSGF1	GF_sRNA1_712-689	GCTCCTGCAATTTAGTTCTAAGAT	24
HSGF2	GF_sRNA2_3931-3908	ACTATCTTCTTTGTCAAATTTCCA	24
HSGF3	GF_sRNA3_4097-4074	CAGCTAGTTCTCTCCAGTCATGCT	24
HSGF4	GF_sRNA4_1965-1988	AAGATTTCTATAGACACATCCAGA	24
HSGF5	GF_sRNA5_462-485	AATTACTTCTACCTAACCTTGTAT	24
HSIm1	Im_sRNA1_2878-2901	TCTCCACCAATGAGTCTGCAATCA	24
HSIm2	Im_sRNA2_2968-2945	CCTCCTGTCGCAGGTGCATATCCT	24
HSIm3	Im_sRNA3_2902-2925	GCAGCTACAACCCTGGAGATATTG	24
HSIm4	Im_sRNA4_3052-3029	TTGAACCATTCCGAACTATTTTCC	24
HSOL1	OL_sRNA1_1680-1703	TTGCCCACAGAGCAAGAACAGTAC	24
HSOL2	OL_sRNA2_6149-6126	GCAGACTATCAGTTTTCCACTTAC	24
HSOL3	OL_sRNA3_2181-2204	TCGGATTCGCCTATAATATTCAAA	24
HSOL4	OL_sRNA4_2565-2542	TCTCCCACCATGTTTTCAAGAGGA	24
HSOL5	OL_sRNA5_6528-6505	TCCACTGCATTGATCAACACACCT	24

Tableau sup 2-2 : Liste des primers utilisés pour les analyses northern blot

Taille des sARN	PKW reads	PKW % reads	Plante saine reads	Plante saine % reads	Infecté GF reads	Infecté GF % reads
1-17	36278	0.260%	36848	0.309%	13926	0.140%
18-26	4710475	33.705%	4317053	36.177%	4825284	48.642%
27-44	8384702	59.996%	6473115	54.245%	4541805	45.784%
Restant	172324	1.233%	587288	4.921%	98671	0.995%
	671770	4.807%	518897	4.348%	440287	4.438%

Taille des sARN	Infecté Im reads	Infecté Im % reads	Infecté My reads	Infecté My % reads	Infecté OL reads	Infecté OL % reads
1-17	30767	0.227%	24229	0.238%	18435	0.153%
18-26	5889111	43.356%	3393883	33.362%	4635104	38.545%
27-44	6848222	50.417%	6212632	61.070%	6730928	55.974%
Restant	79402	0.585%	99183	0.975%	156061	1.298%
	735653	5.416%	443105	4.356%	484549	4.029%

Tableau sup 2-3 : Vérification de la qualité du séquençage haut débit

A : Nom et nombre de reads des différents échantillons utilisés lors du séquençage

B : Tableau indiquant les pourcentages de chacune des différentes catégories de sARN pour les différents échantillons étudiés dans ce chapitre.

Ces données proviennent de l'équipe de bio-informatique de l'entreprise FASTERIS.

Echantillon	PKW				Infecté GF	Infecté OL	Infecté Im	Infecté My
Séquence Référence	BSGFV	BSOLV	BSImV	BSMysV	BSGFV	BSOLV	BSImV	BSMyV
Nombre de sARN de 20-25 nt	6423019				2337407	2429208	4651195	3010584
vsARN total	24098	51814	87198	75355	121194	619900	345897	226429
vsARN total (RPKM)	517	1092	1748	1473	7139	27867	9721	6819
Ratio sens/antisens	1,21	1,33	0,71	1,49	1,19	1,51	1.05	1,23
vsARN ORF 1 (RPKM)	1524,8	1281,7	1,8	1432,4	20129	56503	31152	18010
vsARN ORF 2 (RPKM)	624,6	986,6	617,2	1833,2	15431	21538	20994	13422
vsARN ORF 3 (RPKM)	452,3	1171,1	2372,0	1695,2	6247	25348	7359	5948
vsARN IG (RPKM)	285,1	597,7	0,9	262,7	2011	29937	12133	4339
% vsARN 21nt	13,0	9,2	9,2	12,4	76,2	32,2	61,5	44,6
% vsARN 22	2,7	2,3	2,0	2,6	18,3	19,8	15,8	14,0
% vsARN 23	9,0	10,8	9,1	9,0	1,8	22,5	8,5	16,1
% vsARN 24	75,3	77,7	79,7	75,9	3,7	25,4	14,2	25,3

Tableau sup 2-4 : Données globales obtenues suite au séquençage haut débit
Ces données ont été calculées à partir d'analyse de type « mapping ». Les alignements de séquences entre vsARN et séquences de référence sont réalisés dans le cadre d'identité parfaite. Les ratios sens/ antisens ont été obtenus par division du nombre total de vsARN de 20 à 25nt en sens par celui en antisens sur la séquence référence. Les échantillons analysés concernent la plante PKW qui possèdent des eBSV et des plantes Cavendish dépourvues d'eBSV mais infectées soit par BSImV, BSMyV, BSGFV ou BSOLV.

Echantillon	PKW
eBSGFV-7	389,50
BAC 71C19 sans l'eBSGFV	40,05
Rétrotransposon Muku BAC 71C19 (GF)	136,99
eBSOLV-1	416,88
BAC 73B22 sans l'eBSOLV	123,86
Rétrotransposon Circe BAC 73B22 (OL)	156,47
Gene histidine Kinase BAC 73B22 (OL)	2,54
eBSImV	1000,56
BAC 68C24 sans l'eBSImV	83,13
Rétrotransposon Clio BAC 68C24 (Imv)	1527,85

Tableau sup 2-5 : Représentativité des sARN sur les différents BAC PKW porteurs des eBSV
Ces résultats ont été calculés à partir d'analyse de type « mapping ». Les alignements de séquences admettent jusqu'à deux mutations entre le vsARN et la séquence de référence. Toute les données sont exprimées en reads par kilobase et par millions (RPKM). Les BACs analysés sont les suivants: BAC 71C19 contenant l'eBSGFV, BAC 68C24 contenant l'eBSImV et BAC 73B22 contenant l'eBSOLV.

Points clés du chapitre 2

Le chapitre 2 a permis de mettre en évidence des mécanismes de type ARN interférent impliqués dans la régulation des eBSV et en cas d'infection par le BSV.

Régulation des eBSV :

- Les eBSV présents dans le génome de PKW sont la cible de vsARN majoritairement **de 24 nt**. La régulation des eBSV semble donc de type **TGS « transcriptionnal gene silencing ».**

- Les séquences **inversées/répétées** des eBSV sont les zones cibles des vsARN témoignant certainement de la restitution et l'implication d'une structure aberrante en épingle à cheveux suite à une transcription initiatrice du processus.

- **L'âge** des intégrations et **les zones d'intégrations** semblent être deux facteurs clés expliquant les différences de quantité de vsARN produit.

- Les bananiers ***M. balbisiana*** **fertiles** possédant des eBSV très proches de ceux de PKW en terme de structure produisent des vsARN à partir **des mêmes hot-spots que PKW**. Les bananiers ayant des eBSV différents structurellement ne produisent pas de vsARN à partir des mêmes zones. Ce qui confirme que la production de vsARN est **structure dépendante** pour les eBSV.

Mécanismes de défenses en cas d'infection par le BSV :

- Les bananiers **non-porteurs d'eBSV** et **infectés** par du BSV exogène mettent en place des mécanismes de défenses anti-viraux impliquant le processus de **PTGS « post-transcriptionnal gene silencing »** par l'intermédiaire d'une **production de vsARN de 21nt**. La production de vsARN de 24nt observée aussi, semble indiquer une pondération du système par du « transcriptional gene silencing » impliqué dans ces mécanismes de défense mais dans une moindre mesure.

- Des zones spécifiques du génome BSV sont la cible de vsARN chez les bananiers infectés. Les zones semblent partagées par les différentes espèces de BSV. **La séquence** des BSV semble donc jouer un rôle dans la **production de vsARN**.

- Les **zones génératrices de vsARN** ne sont pas les mêmes pour une même espèce entre **BSV et eBSV**.

- Les séquences du BSV productrices de vsARN ne sont pas les mêmes que chez le **CaMV**. Et donc les hypothèses développées pour le pathosystème *Arabidopsis thaliana*/CaMV expliquant ces productions ciblées (Blevin et al, 2011) ne sont pas partagées avec notre pathosystème.

DISCUSSION GENERALE

Ces travaux de thèse visent la compréhension évolutive de l'interaction entre le BSV et sa plante hôte le bananier. Nous nous sommes pour cela plus particulièrement focalisés sur l'explication en termes de coût /bénéfices du maintien de telles séquences dans le génome de cette plante. En effet les eBSV sont à l'origine d'un paradoxe évolutif pour le bananier qui possède dans son génome des séquences ayant des effets délétères sur sa fitness. Nous avons évalué tout d'abord dans le chapitre 1 l'étendue de la conservation des eBSV du bananier *M. balbisiana* PKW pour trois espèces BSV, puis analysé cette signature eBSV dans la diversité *M. balbisiana*. Et dans un second temps, nous avons souhaité comprendre les mécanismes mis en place chez le bananier afin de contourner les effets délétères imposés par la présence de ces séquences dans le génome. Nous avons choisi de préciser la nature des mécanismes basaux existant chez le bananier pour lutter contre les infections exogènes BSV et de les comparer à ceux qui régissent le contrôle des séquences endogènes telles que les eBSV.

1- Le contrôle des eBSV par les mécanismes de silencing

1-1 Schéma de régultaion des eBSV et des BSV

Un schéma global évolutif expliquant les différentes interactions possibles entre les endogenous pararetrovirus et leur plante hôte avait été proposé par Staginnus et Richert-Pöggeler (2006) sur la base des connaissances des ePVCV principalement et des intégrations virales chez les solanacées. Ce travail avait mis en lumière l'importance des mécanismes épigénétiques dans la régulation des EPRV. Les données obtenues dans le chapitre 2 de cette thèse confirment cette étude en montrant que les eBSV sont eux aussi régulés par un mécanisme de type ARNi.

Ainsi, sur la base des résultats obtenus et au travers de l'analyse de deux contextes d'interactions BSV-bananier, nous proposons un schéma récapitulatif des interactions existantes actuelles résultant d'une histoire évolutive commune entre le bananier et le virus (figure 3-1). Les deux contextes d'interactions que nous avons choisi d'illustrer concernent les bananiers non porteurs d'eBSV qui sont infectés par le BSV (ici Cavendish AAA) et les bananiers porteurs d'eBSV, résistants à la multiplication virale quelle que soit l'origine de la production virale eBSV et/ou BSV (ici PKW BB). Chez les bananiers infectés BSV, nous avons observé que la plante réagissait en utilisant les voies classiques de défense contre les virus cependant sans que cela n'aboutisse à l'élimination du virus.

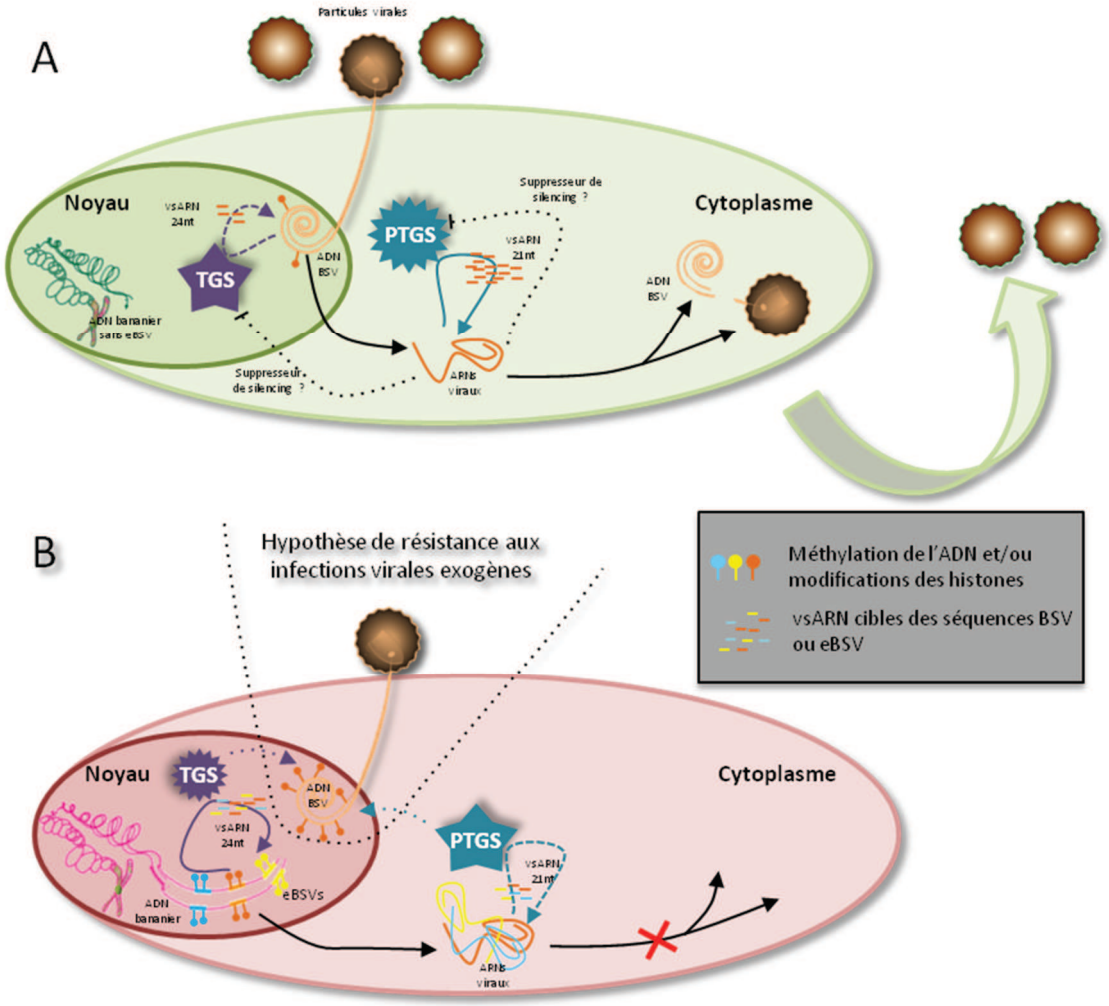


Figure 3-1 : Schéma synthétisant la régulation liée aux BSV ou aux eBSV

Les flèches pleines représentent les mécanismes majoritaires et les flèches en pointillées les mécanismes minoritaires qui ont lieu lors de la régulation. Les traits en pointillés très fins représentent des mécanismes qui sont supposés faire partie de la régulation. Les étoiles sont une représentation simplifiée des mécanismes de Transcriptional Gene Silencing (TGS) et de Post Transcriptional Gene Silencing (PTGS), plus l'étoile a de pointes et plus le mécanisme est important.

A : Cellule de bananier *M. acuminata* infectée par le BSV Ce schéma représente une infection par le virus du BSV d'un bananier non porteur d'eBSV. Lors d'infections, l'ADN viral va pénétrer dans le noyau de la cellule où il détournera la machinerie cellulaire pour produire des ARN pré-génomiques qui vont transiter jusqu'au cytoplasme et permettre la synthèse des protéines nécessaires à la néo-production de particules virales. Le TGS agit dans le noyau pour diminuer la transcription de l'ADN viral en méthylant ce dernier – minichromosomes viraux. Le PTGS agit dans le cytoplasme pour diminuer la traduction des ARN viraux. Ces deux mécanismes agissent par l'intermédiaire de petits ARN (vsARN car il s'agit de séquences virales) que nous avons mis en évidence dans notre étude. La taille des sARN est informative sur le type de mécanisme qui est mis en place, 21nt pour le PTGS et 24nt pour le TGS. Dans ce type d'infection le PTGS est plus important que le TGS. Nous faisons l'hypothèse de la présence de protéines anti- silencing pour expliquer la production de particules virales malgré la mise en place des mécanismes de défense.

B : Cellule de bananier *M. balbisiana* contenant des eBSV (PKW) Ce schéma représente une cellule du bananier PKW, qui est porteur de 3 eBSV fonctionnels. Le contrôle des eBSV pour éviter qu'elles ne produisent des particules virales est assuré essentiellement par le TGS. Ce mécanisme induit la production de sARN de 24nt et la méthylation de l'ADN et/ou une modification des histones spécifiquement au niveau de ces séquences afin d'empêcher leur transcription. D'après nos données, le PTGS joue aussi un rôle dans la régulation des eBSV mais de manière moins importante que le TGS. D'autre part, le contrôle des eBSV via le TGS et PTGS permettrait de manière très précoce la mise en place de mécanismes de défense contre des infections exogènes de BSV. Aucune donnée ne nous permet de trancher quant à l'importance relative de l'un et de l'autre dans ce mécanisme de résistance c'est pourquoi sur le schéma nous accordons la même importance au TGS et PTGS.

Ces mécanismes basés sur l'ARNi agissent principalement au niveau post-transcriptionnel (PTGS) en dégradant une partie des ARN produits par le virus. Cela nécessite la formation de structures double-brins ou simple brin aberrantes issues certainement des phases de la multiplication virale (au moment de la transcription inverse). Il est aussi probable au vu de nos résultats que le contrôle épigénétique agisse en même temps au niveau transcriptionnel (TGS) en méthylant l'ADN viral présent en tant que minichromosome dans le noyau. Ce deuxième mécanisme de défense se mettrait en place ou s'amplifierait considérablement lorsque l'infection virale deviendrait trop importante et à risque pour la fitness de la plante. Malgré la régulation en place, le virus se multiplie dans les cellules du bananier. Suite à ces observations et sur la base des travaux réalisés par Blevin et al., (2006 et 2011), Raja et al., (2008) et Hohn et Vasquez (2011), nous proposons que le virus utilise soit des protéines suppresseurs de silencing et/ou la mise en place de leurres qui serviraient à détourner la machinerie PTGS de la plante comme cela a été démontré pour le CaMV. Quelle que soit la nature de la stratégie mise en place, le résultat permet au virus de continuer à se multiplier et à induire une infection systémique.

Dans le contexte d'un bananier diploïde *M. balbisiana* tel que PKW, nous avons démontré que les eBSV étaient soumis à une régulation de type TGS via la production de vsARN de 24 nt. D'après nos données, la proximité entre eBSV et éléments transposables est essentielle à la mise en place du TGS. L'âge, l'environnement à large échelle et la structure des eBSV sont autant de données essentielles pour cibler les régions productrices de vsARN et déterminer la quantité de vsARN produits. Une fois méthylés, les eBSV continueraient tout de même à induire une production basale de vsARN essentiellement de taille 24nt pour entretenir le système, mais pas exclusivement puisque nos résultats montrent chez PKW 10% environ de vsARN de 21nt. Ces productions de vsARN pourraient permettre à la plante de résister à une infection exogène de BSV en mettant en place du PTGS ou du TGS extrêmement rapidement.

Ces observations sont plus en faveur de la mise en place/existence d'un mécanisme basal de défense réversible basé sur un mécanisme séquence spécifique, qu'une résistance réelle au BSV.

1-2 L'impact du silencing sur l'évolution des eBSV

La mise en place des mécanismes épigénétiques de régulation constitue une étape essentielle de la conservation dans le génome de séquences parasites (Feschotte et Gilbert, 2012). Les études récentes sur le silencing tendent à montrer que ces mécanismes seraient

opérationnels rapidement après l'intégration (Voinnet, données non publiées) et empêcheraient la transcription des séquences parasites que constituent les éléments transposables ou bien les EPRV. Il est donc possible que ce mécanisme s'applique rapidement aux eBSV chez le bananier au moment de leur fixation et permette leur inactivation. Les mécanismes de reconnaissance des séquences parasites sont encore inconnus mais la présence à proximité des éléments transposables ainsi que les similarités de séquences avec les rétrotransposons peuvent être des facteurs influençant la mise en place rapide du TGS (Hollister et al., 2011).

Cependant nous savons que chez les hybrides interspécifiques, la production de particules virales peut avoir lieu, lorsqu'ils sont porteurs d'allèle fonctionnel, sous l'action de stress. Cela informe sur la fragilité du système face aux stress, comme cela a été mis en évidence pour les éléments transposables (Lockton et al., 2010). Ces données sont autant d'arguments nous indiquant que la ploïdie des eBSV est un facteur clé dans la mise en place du silencing. Nous savons que les intégrations sont intégrées à l'état hémizygote. Et en cas de stress tels que des changements d'environnement, les eBSV ont pu être une contrainte pour les bananiers *M. balbisiana* fertiles du fait de leur état hémizygote, et ont entrainé différentes modifications génomiques (duplications et statut homozygote, réarrangements permettant des structure en hairpin etc…). Ces modifications ont abouti à l'inactivation des eBSV dans la plante et sont ainsi autant d'étapes traduisant l'évolution de structure des eBSV jusqu'à la diversité que nous observons aujourd'hui. Par exemple, l'intégration eBSMyV qui est présente chez le bananier PKW à l'état hémizygote est la seule ne pouvant produire de particules virale dans la descendance hybrides de PKW du fait d'une structure fortement réarrangée certainement non fonctionnelle (Chabannes et Duroy, communication personnelle).

L'inactivation des eBSV fonctionnels par cette régulation a permis aux eBSV de ne plus être sujet à une forte pression de sélection et peut donc expliquer en partie leur conservation dans le génome des bananiers. Cependant il est difficile de croire que seule cette inactivation par le silencing a contribué à leur fixation et à leur conservation dans les génomes. La résistance induite et acquise aux BSV constituerait dès lors l'hypothèse majeure pour pouvoir expliquer leur présence dans le génome des bananiers mais elle demande encore à être vérifiée.

1-3 Perspectives

L'étude que j'ai menée au cours de ma thèse apporte des données préliminaires importantes quant aux régulations mises en place par le bananier vis-à-vis des eBSV et BSV. Ces régulations sont différentes et feraient appel au TGS dans le cas des eBSV et au PTGS dans le cas des particules virales libres.

Il me semble important dans un premier temps de confirmer l'existence du TGS comme processus de résistance/régulation négative aux eBSV infectieux en recherchant des marques épigénétiques chez PKW. Pour cela, il faudrait tout d'abord caractériser le niveau de méthylation de l'ADN de ces eBSV avant de s'intéresser aux modifications qui pourraient affecter les histones. Une approche basée sur la technique utilisant l'enzyme de restriction McrBC sera retenue. En effet, elle permet d'évaluer l'état de méthylation des cytosines grâce à des coupures qui ont lieu entre deux cytosines méthylées lorsqu'elles sont précédées d'une base purine (Lippman et al., 2005). Cela permettra d'étudier tous les types de méthylation (CG, CHG et CHH). D'autres approches, comme le séquençage au bisulfite ou des Southern blots utilisant des enzymes de restriction ayant des sensibilités différentes pour la méthylation, pourront par la suite être privilégiés selon les résultats obtenus pour confirmation. Ces autres techniques s'avèrent néanmoins beaucoup plus longues à mettre en place notamment sur un hôte tel que le bananier. La modification des histones pourra être étudiée également par immunoprécipitation de la chromatine (ChIP) associée à de la qPCR ou du séquençage. Cette étude portera sur l'ensemble des eBSV même si les régions hot-spots seront dans un premier temps, privilégiées.

Pour compléter les données obtenues dans le cadre de cette thèse, je pense que l'étude des mécanismes de régulation chez les hybrides interspécifiques ne possédant qu'un seul génome B est essentielle. Elle devrait permettre une compréhension globale des interactions moléculaires existantes entre le BSV et le bananier grâce aux trois situations complémentaires rencontrées qui sont : bananier sain et résistant (PKW) possédant des eBSV, hybrides interspécifiques producteurs ou non de BSV (hybrides AAB) et plantes sans eBSV mais infectées par le BSV.

Dans cette optique, le séquençage haut débit des sARN de bananiers hybrides issus de la descendance de PKW (BB) par le bananier tétraploïde IDN-T 110 (AAAA) a été initié. Chaque individu de cette descendance qui se compose de bananiers AAB stériles produisant ou non des particules virales, et a un contexte génétique se rapprochant de celui de PKW pour 1/3 de son génome. Les individus rassemblant les diverses combinaisons possibles des allèles

eBSV existent à un état haploïde et montrent des niveaux divers d'infection par les BSV. Cette population a été génotypée vis à vis des allèles eBSV et son statut sanitaire caractérisé (Chapitre 1, article 1). Nous avons sélectionné 4 bananiers hybrides tous porteurs des eBSV pour les trois espèces BSV et présentant les caractéristiques suivantes : un bananier BSV free possédant les deux allèles eBSOLV-2 et eBSGFV-9 non infectieux et un des 2 allèles eBSImV et trois bananiers infectés, porteurs chacun d'un des 3 allèles infectieux (eBSOLV1, eBSGVF-7 ou eBSImV) et d'allèles non infectieux pour les autres espèces BSV. Les analyses des vsARN produits devraient nous permettre d'étudier l'impact du changement de ploïdie sur la régulation des eBSV (effet dose entre génome haploïde versus génome diploïde), de voir si les deux allèles d'une même espèce eBSV produisent les mêmes catégories et les mêmes quantités de vsARN, mais aussi de voir si le réveil d'une espèce eBSV influence le pattern (quantité et qualité) des vsARN des autres espèces eBSV. Les 4 échantillons en cours de séquençage ne représentent cependant pas l'ensemble des combinaisons alléliques possibles. Il me parait donc important, pour compléter ces analyses, d'envisager le séquençage de nouveaux échantillons. Il faudrait notamment inclure des plantes ayant un des allèles infectieux (eBSOLV-1 ou eBSGFV-7) mais exemptes de particules virales correspondantes si un tel individu existe.

Une fois la compréhension des mécanismes inhérents à la régulation des eBSV décryptée, il va être possible d'aborder comment la résistance des bananiers porteurs des eBSV se met en place vis à vis des BSV exogènes. Pour cela nous pouvons envisager d'étudier la production de sARN chez PKW lors d'une tentative d'infection par cochenille par exemple ou lors de bombardement par des virus. Cette étude permettrait de mieux comprendre la pondération de la régulation et comment les eBSV peuvent participer à la mise en place de la résistance et par là même de connaître quels sont les éléments génétiques essentiels de ce mécanisme.

Finalement et d'un point de vue plus général, je pense qu'il serait intéressant d'observer la régulations des séquences BSV-like *Badnavirus* récemment mises en évidence dans le génome du bananier *M. acuminata* cv Pahang. Ces séquences, qui pour certaines sont de taille et de structure comparables aux eBSV des génomes B, ont été découvertes lors du séquençage du génome de ce bananier Pahang (D'Hont et al., 2012). Contrairement aux eBSV du génome B qui sont en faibles copies, ces BSV-like *Badnavirus* sont réparties sur la quasi-totalité des chromosomes du génome A et seraient le résultat de plusieurs évènements d'intégration (Baurens et Chabannes, communication personnelle) d'espèces

virales différentes (au moins 4). Ces séquences ne semblent pas fonctionnelles c'est-à-dire qu'elles sont incapables de restituer un génome viral fonctionnel permettant la production de particules virales. De façon intéressante, les études en cours montrent que certains de ces BSV-like *Badnavirus* sont uniques à l'espèce *M. acuminata* ou l'espèce *M. balbisiana* alors que d'autres sont au contraire ubiquitaires du genre *Musa* puisque présents dans les deux génomes (Muller et Chabannes, communication personnelle). Il apparait donc évident que certains d'entre eux se sont intégrés dans l'ancêtre commun des *M. balbisiana* et des *M. acuminata* et sont par conséquent bien plus anciens que les eBSV caractérisés au cours de cette thèse qui sont restreints à l'espèce *M. balbisiana*. Comprendre leur rôle et caractériser leur régulation devraient apporter des données complémentaires à celles qui sont décrites dans ce manuscrit en terme d'évolution des eBSV infectieux du génome B. C'est également une étape essentielle pour comprendre l'évolution des mécanismes de régulation du bananier.

2- Evolution des eBSV

Ce travail de thèse constitue également une avancée significative dans la compréhension des forces évolutives qui ont pesé sur les eBSV, et qui ont contribué à leur conservation dans le génome des bananiers *M. balbisiana*.

2-1 L'évolution moléculaire des eBSV

Nous avons en premier étudié la diversité nucléotidique au sein des eBSV de PKW. Comme l'avait montré Gayral et al. (2008) pour l'eBSGFV, l'évolution moléculaire mesurée entre les eBSV et les séquences correspondantes du BSV épisomal est plus faible que celle attendue théoriquement et ce pour les 3 eBSV. Ce résultat suggère que les eBSV participent activement à la restitution des virus identifiés lors d'infections et modifient significativement les paramètres évolutifs des BSV, voire seraient la source majoritaire de production des BSV vue l'inexistence d'épidémie à BSV à travers le monde. L'absence de connaissances sur les séquences virales ancestrales s'étant intégrées dans les génomes à l'origine nous empêche de pouvoir quantifier de façon précise l'évolution des eBSV par ce type de méthode. Nous avons par ailleurs montré que l'évolution moléculaire entre les séquences répétées ainsi qu'entre allèles d'un eBSV est très faible ce qui peut indiquer soit que l'intégration serait très récente ou qu'il existe des pressions de sélection positive favorisant la conservation de séquences pour les eBSV.

Néanmoins, nous avons pu mettre en évidence que des mutations présentes sur les eBSV empêchent la production de particules virales fonctionnelles. Par exemple l'eBSImV possède deux mutations conduisant à des codons stop le long de sa séquence bien que ces allèles soient impliqués dans la restitution de virions. L'eBSGFV possède au niveau de l'allèle-9 une quantité de mutations synonymes plus importante que sur l'autre l'allèle expliquant en partie son potentiel non infectieux même dans un contexte bananier hybride. L'eBSOLV, quant à lui a, une séquence très proche de celle du virus présentant très peu de mutations non-synonymes, l'allèle eBSOLV-2 présente une délétion qui correspond à la fin de l'ORF et le début de l'IR. La présence de ces mutations parmi un niveau très faible de divergence semble en faveur de pressions de sélection appliquées de façon très ponctuelle sur les séquences eBSV et de manière différente suivant les eBSV.

2-2 La structure des eBSV

Nous avons par la suite vérifié les variations de structure des eBSV au sein de la diversité des bananiers porteurs de génome *M. balbisiana*. En effet le « post-insertional gene rearrangement » qui génère des différences de structure entre les ERV a été identifié comme étant un des mécanismes partagé et sélectionné chez les ERV (Hughe et al., 2001). L'étude de Gayral et al (2010) avait mis en évidence qu'il existait effectivement des divergences de structures pour un même eBSV dans la diversité des bananiers *M. balbisiana* notamment en ce qui concerne l'allèle non infectieux. La poursuite de cette étude en augmentant la finesse et l'étendue de l'analyse (Article 2) a révélé que ces réarrangements post-insertionnels sont observés chez tous les eBSV fonctionnels et qu'ils représentent un facteur clé de leur diversité. Nous avons, pour tous les eBSV, mis en évidence de nouveaux allèles, l'eBSGFV est néanmoins très conservé au sein de la diversité alors que l'eBSOLV avec l'eBSImV présentent des changements de structures majeurs. L'étude n'ayant pas été au niveau de la séquence même, nous ne pouvons pas dire si ces changements alléliques se traduisent par la production d'un génome viral fonctionnel. Il semblerait néanmoins que les allèles fonctionnels de l'eBSOLV comme de l'eBSGFV soit conservés, voire sélectionnés, et que l'apparition de nouveaux allèles résultent d'une pseudogénisation de l'intégration dans certains génotypes bananiers. Cette hypothèse apparaît plus clairement pour l'eBSImV qui présente soit une conservation de la structure décrite chez PKW soit des disparitions internes significatives des fragments laissant des traces fossiles de l'intégration par le biais de la présence des zones d'intégration dans le génome bananier.

La position génomique des eBSV dans des zones riches en éléments transposables ou à proximité de ces éléments est supposée jouer un rôle important dans la mise en place de tels mécanismes. Chez les plantes, ces régions possèdent des taux de recombinaisons supérieures au reste du génome. De plus, les répétitions de séquences sont des facteurs favorisant la recombinaison (Mezard et al., 2006). L'absence totale de l'eBSGFV de certains hybrides interspécifiques peut par exemple s'expliquer par son positionnement au sein d'un rétrotransposon à LTR pour lequel il a été montré que la recombinaison entre ces deux séquences répétées favorisait l'excision complète de la séquence (Devos et al., 2002). Il serait donc intéressant d'aller rechercher chez ces hybrides si le rétroélement est encore présent, ce qui permettrait de valider le fait que cet eBSV a bien été intégré chez tous les bananiers *M. balbisiana*.

Ces réarrangements sont à l'évidence un facteur limitant l'expression des eBSV. Comme mis en évidence par Iskra-Caruana et al. (2010), l'eBSGFV-7 doit subir deux étapes de recombinaison homologue pour produire un génome viral fonctionnel et produire des particules virales. Les analyses menées sur l'eBSOLV-1 indiquent également que cet eBSV doit faire l'objet de recombinaisons homologues pour pouvoir exciser un génome viral fonctionnel (Chabannes, communication personnelle). Il est d'ailleurs intéressant de noter que la zone supposée responsable de la production des particules virales à partir de l'eBSVOLV-1 est celle qui présente le plus de diversité pour l'eBSOLV chez les bananiers étudiés et à l'origine des nombreux nouveaux allèles identifiés. Des pressions de sélections spécifiques sur cette partie de l'eBSV ont pu être appliquées et induire cette variabilité afin de sélectionner les eBSV ne pouvant pas ou difficilement produire des particules virales.

Les études menées dans le chapitre 2 ont révélé que les réarrangements de structures sont impliqués dans le contrôle des eBSV et sont le siège de production de vsARN. En effet ces molécules résultent, de façon spécifique, des zones inversées répétées issues des réarrangements. La production facilitée d'ARN double brin type hairpin, grâce à ce type de séquence, semble être l'explication la plus probable pour expliquer cette répartition des vsARN sur la séquence des eBSV (Huettel et al., 2007).

Ces informations montrent que les réarrangements post-insertionnels sont des évènements fréquents chez les eBSV, et qu'ils semblent avoir été sélectionnés durant l'évolution du bananier pour empêcher et/ou complexifier la production de particules virales.

2-3 Histoire évolutive des eBSV

L'eBSOLV qui est proposé comme étant l'intégration la plus ancienne a été soumise à un grand nombre de réarrangements en particulier dans la zone, selon nos propositions, recrutée pour la restitution d'un génome viral, mais a très peu de mutations. Son intégration dans une zone riche en éléments transposables peut être une explication à cette observation et peut avoir orienté son contrôle au travers d'un grand nombre de recombinaisons plutôt que de mutations.

L'eBSImV qui semble être le plus récent eBSV fixé car peu réarrangé, a été soumis pour le moment à très peu de recombinaisons observables, certaines ayant pu aboutir à la disparition d'une partie de l'intégration comme constaté pour certaines accessions. Le seul réarrangement observé sur l'allèle disponible et supposé fonctionnel propose des séquences inversées/répétées que nous savons indispensables à la mise en place et au maintien du mécanisme de type TGS. Néanmoins, l'eBSImV que nous avons analysé de PKW, possède deux mutations induisant des codons stop qui rendent très complexe la production de particules virales à partir de telles séquences. Aucune hypothèse n'a d'ailleurs été proposée jusqu'à présent. Nous savons que cet eBSV est dans une zone riche en gène où les recombinaisons ont sûrement un taux faible, l'évolution a donc dû préférentiellement se traduire par l'apparition de mutations. Quant à l'eBSGFV, il constitue un intermédiaire entre les deux autres eBSV, car il possède le taux d'évolution moléculaire le plus important des eBSV infectieux (Gayral et al., 2010) ce qui ne s'est pas traduit pas une structure plus réarrangée que celle de l'eBSOLV. Nous pouvons donc nous demander, bien qu'ayant un grand nombre de gènes à ses côtés, si sa présence au sein d'un rétrotransposon n'a pas pu influencer son évolution.

Les marques de l'évolution observées chez les eBSV présents dans le génome des bananiers *M. balbisiana* ne sont pas en accord les unes avec les autres comme nous venons de le voir pour l'eBSGFV et l'eBSOLV. Ces informations nous indiquent que les voies évolutives prises par les différents eBSV sont probablement liées aux zones d'intégrations et de fixation dans le génome et à la structure des eBSV lors de leur intégration.

Les études menées sur les différents organismes ont révélé que presque tous les EVE présents de manière sauvage dans le génome de leurs hôtes sont fortement mutés ou bien réarrangés menant à l'impossibilité de produire des particules virales. Les pressions de sélections sur la fitness de l'hôte, induites par le potentiel infectieux des EVE sont trop importantes pour que des séquences virales infectieuses soient conservées dans les

génomes. Le caractère non infectieux des EVE constitue donc un prérequis essentiel à leur présence dans les génomes hôtes (Holmes, 2011). Il semble que chez le bananier, cette pression ait aussi eu lieu pour les eBSV. En effet, aucune expression des eBSV n'a jamais été observée chez les bananiers fertiles de l'espèce *M. balbisiana* qui sont porteurs de séquences d'eBSV. Les travaux de cette thèse ont ainsi permis de discuter les différents processus évolutifs conduisant à la non-expression ou à l'expression complexifiée des eBSV.

En quoi le contrôle eBSV pourrait servir la résistance au BSV.

Nous avons vu que les réarrangements et les mutations subis ne suffisent pas à empêcher l'activation spontanée des eBSV chez les hybrides interspécifiques, et le maintien de séquences virales fonctionnelles dans le génome pose question. Nous pouvons donc nous demander si la conservation de la fonctionnalité des intégrations a été soumise ou est encore soumise à une pression de sélection positive ou bien, si ce n'est qu'une étape du processus d'évolution conduisant *in fine* à la pseudogénisation des eBSV. L'hypothèse la plus développée dans la littérature pour d'autres plantes (Mette et al., 2002 ; RTBV et GRD dans le génome du tabac) ainsi que pour les animaux (Aswad et Katzourakis, 2012), est que cette conservation est associée aux potentiels effets bénéfiques des eBSV sur la fitness du bananier en particulier en apportant une résistance face aux BSV exogènes.

Cette résistance a déjà été observée pour les deux autres pathosystèmes impliquant des EPRV fonctionnels chez des hybrides interspécifiques (Norreen et al., 2007 ; Mette et al., 2002) alors qu'aucune résistance n'est décrite chez les plantes possédant d'autres types d'EPRV. Dans le cas du bananier, nous pouvons alors envisager que le maintien d'eBSV fonctionnels est nécessaire pour apporter une résistance constitutive à la plante. Les mécanismes impliquant l'ARNi seraient dans ce cas là des candidats forts de la mise en place de cette résistance. Une course à l'armement telle que décrite dans l'introduction entre le virus exogène et la plante au travers des eBSV devrait être observée. Les données montrent qu'aucune divergence réelle n'existe entre eBSV et BSV pour chacune des espèces BSV étudiées et que les fixations BSV ont eu lieu bien avant la diversification de l'espèce *M. balbisiana*. On peut dès lors s'interroger sur l'origine de la diversité des espèces BSV étudiées ; reflète-t-elle la diversité ancestrale ou est-elle la résultante d'une évolution dans le génome hôte ? Force est de constater aujourd'hui que la seule diversité observée pour le BSV issu d'épidémie est celle rapportée en Ouganda. Les BSV épisomaux se rassemblent presque exclusivement dans le groupe 3 de la phylogénie des badnavirus, groupe bien

distinct de celui des BSV étudiés. Ces virus ne présentent aujourd'hui aucune correspondance eBSV dans les génomes bananiers quels qu'ils soient.

La compréhension des mécanismes impliqués dans la résistance des bananiers aux BSV devrait permettre de répondre à ces différentes questions. En effet selon le niveau de régulation (TGS ou le PTGS) le maintien de la fonctionnalité des eBSV ne se justifie pas de la même manière. Une régulation de type PTGS nécessite la transcription des séquences intégrées pour pouvoir produire des sARN de manière efficace dès les prémices de l'infection. La transcription basale permettant la synthèse d'ARN pré-génomique nécessite l'action de l'ARN Pol II qui a besoin que les séquences eBSV soient fonctionnelles pour pouvoir les transcrire. Alors que la régulation de type TGS nécessite l'action de l'ARN Pol IV qui n'a pas besoin que les séquences eBSV aient des ORF fonctionnelles pour pouvoir produire des ARN, qui seront ensuite dégradés en sARN. Nous pouvons donc tendre vers une action du PTGS dans la mise en place de la résistance dans un premier temps. Cependant, le peu de connaissances disponibles aujourd'hui sur le fonctionnement exact de l'ARN PolIV limite nos propositions.

Une dernière hypothèse pouvant expliquer le maintien des eBSV dans le génome bananier serait qu'à un moment donné de l'évolution des bananiers diploïdes *M. balbisiana*, elle ait servi le bananier pour résister ou tolérer un contexte environnemental d'épidémie BSV important. Dès lors que les pressions environnementales ont changé, elles n'apporteraient plus aucun intérêt et constituerait une étape du processus qui tend vers de la pseudogénisation de la séquence.

En quoi le contrôle eBSV pourrait servir le BSV.

Inversement, nous pouvons nous demander alors si l'eBSV ne constitue pas un nouveau type de parasitisme en ayant investi le génome de son hôte. L'ARNi, qui est le mécanisme régissant la régulation des eBSV, a deux caractéristiques, il est réversible en cas de stress et il est séquences spécifiques. Cependant nous avons observé qu'il n'induisait pas de résistance constitutive totale dans les plantes. Les eBSV seraient alors une forme latente du virus dans son hôte. Leur contrôle par l'ARNi obligerait à une conservation de la séquence virale, donc du virus, et la réversibilité du système permettrait au virus, en cas de stress environnementaux ou génomique, d'être libéré et peut-être de coloniser un autre hôte. Ce processus constituerait un cycle vertueux pour le virus lui permettant de rester présent de manière exogène dans l'environnement malgré l'intégration. (Malgré tout), comme nous

pouvons le voir aujourd'hui, lors de périodes de grande stabilité, la dérive génétique fait que les séquences virales évoluent au cours du temps et tendent à la pseudogénisation comme c'est le cas pour les BSV-like badnavirus amenant le virus petit à petit à disparaître. Les eBSV infectieux dans le génome B ayant été fixés récemment ne seraient qu'une image/étape du processus d'évolution, les BSV-like badnavirus observés dans les génomes A et B en étant la fin.

2-4 Perspective

Application pour l'amélioration du bananier

Chez le bananier l'étude des EPRV prend une dimension particulière, en effet cette plante représente un poids économique majeur dans l'agriculture moderne. La production spontanée de BSV à partir d'eBSV infectieux chez cette plante a conduit à l'arrêt des programmes d'amélioration génétique impliquant les génomes *M. balbisiana* (porteur des intégrations infectieuses) alors qu'ils étaient garants d'avancées significatives pour l'amélioration variétale du bananier. Il semblerait que ce soit les pratiques culturales inhérentes à la culture, telles que la micro propagation de masse, qui seraient l'origine indirecte de la résurgence de la maladie. Ce sont ces raisons qui font de l'étude des eBSV chez le bananier une priorité agronomique pour le Cirad.

Les avancées que nous avons réalisées ont donc un intérêt tout particulier pour l'amélioration du bananier avec tout d'abord la restitution des marqueurs moléculaires de type PCR développés pour détecter les eBSV infectieux dans le génome de PKW. Ces marqueurs sont en cours d'utilisation afin de sélectionner les bananiers non-porteurs des eBSV, quand cela est possible (comme l'accession Honduras n'ayant pas l'eBSImV), ou des plantes qui ne portent que les allèles désarmés non-infectieux.

Les caractérisations complètes des structures des eBSV des bananiers *M. balbisiana* fertile ont ainsi permis d'obtenir des informations essentielles quant au potentiel pouvoir infectieux de certains allèles pour lesquels les marqueurs PCR n'apportaient qu'une information partielle. Mais des études supplémentaires vont être nécessaires afin de valider que les allèles nouvellement mis en évidence (en particulier pour l'eBSOLV) sont bien non-infectieux. Le but ultime étant de pouvoir réintégrer ces génotypes comme géniteurs dans les programmes d'améliorations.

Une étude globale de tous les bananiers possédant du génome B présents dans la collection mondiale de l'ITC (plus de 300 bananiers) est d'ailleurs en cours au sein de notre équipe

pour vérifier grâce aux marqueurs que nous avons développé la répartition des 3 eBSV infectieux dans ces plantes et en particulier les allèles infectieux. Cette étude va être d'une aide précieuse pour les améliorateurs bananiers qui pourront enfin connaître la signature eBSV de tout le germplasm *Musa* disponible. Elle va de plus être d'une grande valeur scientifique afin de valider l'analyse que nous avons réalisée dans l'article 2. Nous allons pouvoir vérifier à plus grande échelle si la répartition des eBSV respecte bien les groupes phylogénétiques établis, l'origine phylogénétique de tous ces bananiers ayant déjà été obtenue lors de l'étude de Hippolyte et al, (2012).

Les relations phylogénétiques mise en évidence constituent enfin, une avancée majeure pour l'amélioration des bananiers car pour le moment les bananiers *M. balbisiana* fertiles n'étaient pas reliés phylogénétiquement aux bananiers hybrides. Or ce sont très majoritairement les hybrides qui sont cultivés dans le monde. L'intérêt est tout particulier pour les plantains où l'origine B était inconnue jusqu'à présent. L'utilisation des BB proches va pouvoir être envisagée pour participer à l'amélioration génétique et à la création de nouveaux bananiers hybrides possédant les mêmes caractéristiques agronomiques et gustatives.

La découverte des BSV-like *Badnavirus* dans le génome de *M. acuminata* (D'Hont et al., 2012) ouvre la voie à des études à plus large échelle d'utilisation des virus intégrés comme marqueurs phylogénétiques. Ces virus ayant des origines plus anciennes que les eBSV, ils vont sûrement pouvoir révéler une diversité supérieure entre les différents individus à celle connue, et donc pouvoir apporter une vision plus fine et complète des liens phylogénétiques existant entre les bananiers.

Enfin, bien que les *Badnavirus* soient un genre viral émergent, que le BSV affecte une culture économiquement importante, et que les EPRV infectieux constituent une singularité dans le monde viral, la biologie générale du BSV reste à ce jour très peu connue. Des études du cycle de multiplication, de la localisation tissulaire et cellulaire du BSV lors de l'infection, ou de génomique (fonction des protéines virales) sont en effet nécessaires à la compréhension de la complexité de ce modèle biologique, quelle que soit la discipline, les questions de recherche et les champs d'application. Ces études permettrons notamment de pouvoir vérifier la présence de protéines suppresseur de silencing dans le génome des BSV comme cela a été suspecté dans le §1.

Evolution moléculaire des eBSV

L'étude que nous avons menée sur l'évolution moléculaire chez PKW nous a montré que les taux d'évolutions sont faibles au sein d'un même individu par rapport au BSV libre et entre les allèles. Ces informations ont donné assez peu d'explications quant à l'évolution des eBSV qui semble s'effectuer de manière plus active par d'autres voies comme nous l'avons développé par ailleurs. Afin d'obtenir une diversité plus importante nous avons débuté durant ma thèse l'étude de l'évolution des séquences eBSV de manière plus large en utilisant les bananiers que nous avons étudiés dans l'article 2. Nous avons sélectionné un échantillonnage de bananiers représentatifs de la diversité des génomes *M. balbisiana* et possédant un large spectre d'eBSV et de génotypes. Nous avons ensuite développé des amorces permettant d'amplifier les séquences de manière spécifique sur les eBSV. Nous avons également séquencé le gène proche de chacun des eBSV au niveau d'introns et d'exons afin d'obtenir des séquences sous pressions de sélection et d'autres sous contraintes sélectives neutres. Le but étant d'avoir des référentiels évolutifs au plus près des eBSV pour pouvoir comparer l'évolution de ces séquences avec celles du bananier.

Cette étude est en cours et va permettre une étude fine des pressions de sélection appliquées sur les eBSV en fonction des connaissances que nous avons déjà acquises sur ces séquences et dans des génotypes permettant ou non la production de particules virales. Nous avons bien évidemment choisi d'amplifier des séquences au niveau des zones fortement productrices de sARN afin de vérifier si la mise en place de la régulation TGS peut avoir une influence sur les pressions exercées sur les séquences eBSV. Toutes les séquences pour cette étude ont déjà été obtenues et vont être analysées prochainement. Ces analyses vont mettre en lumière le pourquoi du maintien de la séquence eBSV dans le génome des bananiers et son importance pour la conservation des mécanismes de régulation et de défense.

Elargissement de l'analyse de la diversité des eBSV

Les études que nous avons réalisées se sont concentrées sur les eBSV infectieux, l'obtention de la séquence complète de l'eBSMyV chez PKW va permettre d'ouvrir notre étude à un eBSV qui est non-fonctionnel chez PKW et pour lequel le virus épisomal est connu. Nous allons pouvoir analyser si les pressions de sélection appliquées sur ces séquences ont été plus importantes et auraient pu influencer sa dégénérescence par rapport aux autres eBSV.

Nous allons également, dès connaissance de la structure finale, pouvoir étudier leur diversité parmi les bananiers *M. balbisiana* afin de voir si sa présence à l'état hémizygote chez PKW à deux sites d'insertion indépendants est conservée pour les autres bananiers et si cet état a pu influencer son évolution de structure au sein de ces bananiers. De façon plus large la découverte des séquences BSV-like *Badnavirus* dans le génome de *M. acuminata* a ouvert la voie à des études plus larges sur l'évolution des eBSV dans le génome des bananiers. Ces intégrations étant, semblent-il, plus anciennes, vont permettre de connaître à plus long terme le devenir de ces séquences dans les génomes hôtes. D'autre part le séquençage du génome B est en train d'être discuté et devrait avoir lieu dans les prochaines années. Il permettra de vérifier si le goulot d'étranglement qui est suspecté comme ayant contribué à la faible diversité de l'espèce *M. balbisiana*, a réellement existé. La connaissance complète de ce génome va donner accès à toutes les séquences virales de type BSV ou BSV-like intégrées. Des études comparatives d'évolution entre les séquences virales du génome A et B vont pouvoir être conduites.

L'activation et la résistance virale induite par les eBSV

Les connaissances relatives à l'activation et à la résistance virale induite par les eBSV infectieux sont indispensables pour pouvoir comprendre l'évolution des eBSV dans le génome des bananiers. Ils constituent les deux paramètres du paradoxe évolutif déjà évoqué précédemment pour expliquer la présence des eBSV dans le génome hôte. C'est pourquoi, une fois le fonctionnement des mécanismes totalement décrypté (§1) , leur mise en évidence au sein des *Musa* va constituer une étape importante pour pouvoir mettre en regard l'évolution des eBSV telle que nous l'avons décrite dans cette thèse avec les effets positifs ou négatifs que les eBSV induisent sur leur hôte.

Pour cela nous allons pouvoir utiliser des croisements génétiques faisant varier la ploïdie du génome B donc celle allélique des eBSV dans des contextes génotypiques différents afin d'étudier l'effet de cette ploïdie sur la production des particules virales. Une étude portant sur les bananiers utilisés dans les études précédentes vis à vis de leur capacité à résister aux BSV ou bien à produire des particules virales en cas de stress apparaît alors nécessaire.

3- Histoire évolutive bananier-BSV

Afin de répondre à la question de ma thèse qui porte sur les enjeux évolutifs ayant conduit au maintien des eBSV dans le génome du bananier, j'ai réalisé un schéma reprenant les étapes de la co-évolution entre le bananier et le BSV conduisant au maintien des eBSV tel qu'on l'observe aujourd'hui. Ce paragraphe permettra de comprendre de façon optimale ce schéma.

La présence de BSV-like *Badnavirus* dans le génome des bananiers *M. acuminata* et *M. balbisiana* témoigne que l'interaction entre les deux protagonistes de ce pathosystème est très ancienne. Les intégrations et plus particulièrement leurs fixations dans la plante hôte, bien que ne faisant pas partie du cycle de multiplication des badnavirus, constituent apparemment une pratique récurrente et habituelle de ce genre viral avec son hôte. L'hypothèse d'un virus ancêtre fixé dans le génome bananier ancêtre de *M. acuminata* et *M. balbisiana* dans un contexte épidémique fort peut être faite. Cette fixation aurait donné un avantage sélectif aux bananiers porteurs de BSV-like. Ces bananiers étant devenus tolérants voir résistants, le virus a été contraint de s'adapter en trouvant un autre hôte, ou d'évoluer. Aucune donnée ne nous permet de privilégier une hypothèse plutôt qu'une autre. Petit à petit, la divergence entre séquences virales endogènes et virus exogènes aurait contribué à la diminution de la pression de sélection positive sur les séquences endogènes et favorisé leur pseudogénisation jusqu'à ce que l'on observe aujourd'hui.

Etape 1- L'intégration et la fixation du virus dans les génomes bananiers

Après la spéciation *M. acuminata/M. balbisiana*, la présence et la fixation des BSV ne peuvent être expliquées que par un contexte épidémique fort et endémique à la zone d'origine de l'espèce *Musa balbisiana*. L'avantage évolutif de l'acquisition d'une résistance pour les bananiers diploïdes sauvages aux BSV aurait contribué à la fixation de ces intégrations dans le génome des bananiers et à leur sélection (Lheureux, 2002). Ainsi tous les bananiers diploïdes *M. balbisiana* auraient des intégrations BSV, comme nous l'avons observé.

Les études de Gayral et Iskra-Caruana (2009) et D'Hont et al., (2012) ont permis de mettre en évidence les liens phylogénétiques existants entre les BSV-like badnavirus et les BSV. Il est possible que les BSV tels que nous les connaissons aujourdhui aient émergés à partir des BSV-like badnavirus. Leur intégration et la pseudogénisation qui en a découlé ont pu contri-

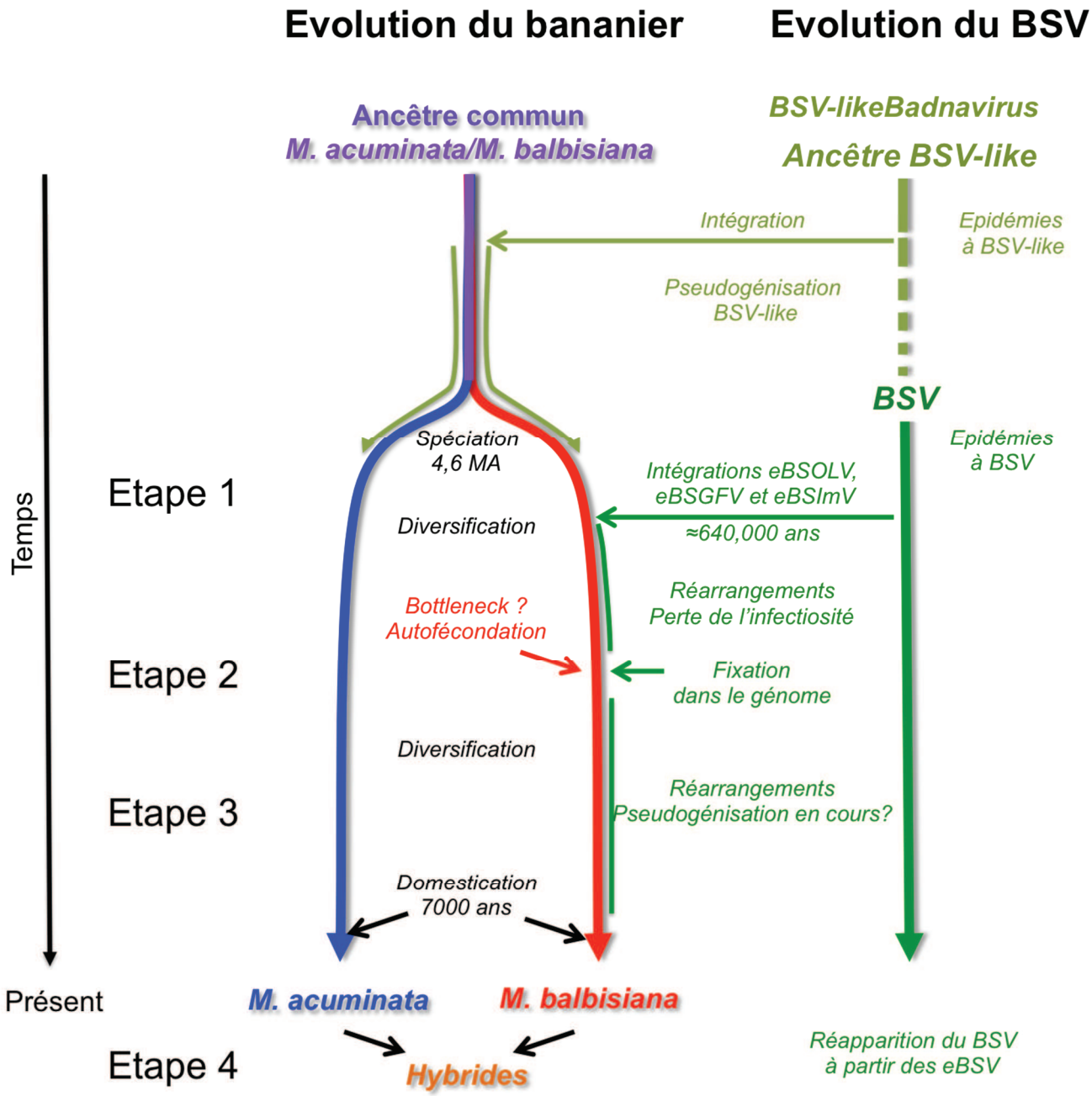


Figure 3-2 : Etapes clés de l'évolution BSV/Bananier

Schéma représentant la co-évolution entre le BSV et le bananier et les principaux enjeux subit par ce couple durant leur évolution commune. Les étapes présentées à gauche correspondent aux explications données sur ce schéma dans la thèse de Pierre-Olivier Duroy (2012).Les indications en vert clair correspondent aux BSV-related badnavirus, celles en vert foncé au BSV, celles en violet à l'ancêtre commun aux deux espèces *M. balbisiana* et *M. acuminata* , celle en rouge seulement à *M. balbisiana* et celles en bleu seulement à *M. acuminata.* La date de la spéciation *M. acuminata* et *M. balbisiana* a été obtenue par Lescot et al., (2008) et la date de l'intégration du BSV correspond à la date la plus ancienne possible pour l'intégration de l'eBSGFV (Gayral et al. 2008), la date de la domestication du bananier a été obtenue lors de l'étude de Perrier et al., (2009).

-buer à l'émergence de nouveaux badnavirus contre lesquels les bananiers n'étaient pas résistants.

Etape 2 - Maintien de 3 eBSV fonctionnels

Nous avons montré l'ubiquité de la présence des 3 séquences eBSV dans le génome de tous les bananiers *M. balbisiana* fertiles et en aucun cas dans le génome des bananiers de l'espèce *M. acuminata*. Ce que nous interprétons comme une intégration et fixation des 3 eBSV étudiés dans cette thèse chez l'ancêtre commun de l'espèce *M. balbisiana* avant la diversification de cette espèce et après la spéciation avec *M. acuminata*. Il n'est pas à exclure que, la présence des eBSV et la résistance qu'elle semble engendrer pour les bananiers, puissent avoir pris part à la divergence entre ces deux espèces en obligeant à une séparation spatiale entre ces espèces, en permettant aux bananiers *M. balbisiana* de pouvoir rester dans la zone d'origine alors que l'autre espèce ait dû se déplacer en dehors de la zone épidémique du BSV.

Une fois l'intégration réalisée, différents phénomènes non-exclusifs peuvent être envisagés pour expliquer la fixation et la présence des 3 eBSV infectieux dans le génome de tous les bananiers *M. balbisiana*. Dans un premier temps il est possible que la population originelle *M. balbisiana* ait subi un goulot d'étranglement rétrécissant de façon drastique la diversité au sein de cette espèce. Cette hypothèse est renforcée par la faible diversité génétique des bananiers *M. balbisiana* que nous avons mise en évidence lors de l'analyse phylogénétique utilisant des marqueurs microsatellites. Ce goulot d'étranglement a pu être accentué par l'avantage évolutif qu'à constituée une tolérance/résistance contre les BSV exogènes en cas d'épidémie. Enfin nous savons que les eBSV lors de l'intégration sont à l'état hémizygote ; cet état permet le relargage de particules virales chez les hybrides interspécifiques. Il est cependant difficile de savoir si cet état chez les bananiers de type BB a les mêmes effets car aucun bananier présentant un eBSV fonctionnel à l'état hémizygote n'a été identifié à ce jour. Nous pouvons tout de même penser que cet état était non stable et pouvait poser problème en particulier en cas de changements environnementaux. En conséquence, le passage au stade homozygote lors d'autofécondation a pu être favorisé et, de concert, a favorisé le maintien des eBSV dans les génomes bananier (cf figure 9 article 1 chapitre 1).

Une fois la fixation BSV réalisée dans les génomes B ancestraux, et la régulation de ces séquences en place, la plante a acquis une résistance au BSV jusqu'à peser sur l'épidémie

environnante. Comme dans le cas des BSV-like, la résistance/tolérance des bananiers porteurs d'eBSV aurait contribué à diminuer la pression épidémique contraignant le virus à s'adapter ou disparaître. Il semblerait que les épidémies aient disparus ainsi que le virus. Les eBSV ont alors cessé d'être un avantage évolutif pour le bananier et sont devenus même, en cas de relargage de particule virale, un désavantage évolutif. Ne connaissant pas les bases réelles de la résistance aux BSV , ni même si elle existe, il est difficile de savoir si la capacité fonctionnelle des eBSV est vraiment nécessaire pour qu'un bananier soit résistant. Il est donc possible que déjà lors de l'extinction d'épidémies les BSV n'étaient plus fonctionnels, et soient devenus des séquences sous pression de sélection neutre soumise à dérive génétique. Cette absence de pression de sélection positive aurait favorisé les modifications de structure ou les mutations moléculaires comme évoqué dans le §2.

Etape 3 - La pseudogénisation des eBSV ?

Les divergences observées au sein des bananiers BB, pour une même intégration ancestrale, nous indiquent que les bananiers n'ont pas évolué de la même façon avec, pour certains des réarrangements massifs de l'eBSV témoignant d'une forte pression de sélection purificatrice, et pour d'autres comme PKW une conservation plus importante permettant la production de particules virales dans une descendance hybrides interspécifiques. Une première explication peut être que tous les bananiers *M. balbisiana* n'aient pas subi d'autofécondation. Une partie d'entre eux ont pu se retrouver en dehors des contextes environnementaux d'origine, ce qui a pu favoriser la production de particules virales. La sélection naturelle aurait alors favorisé des réarrangements importants des eBSV afin de préserver les nouveaux bananiers BB. Il semble que cela ait été le cas pour les bananiers présents en Inde qui possèdent une absence totale ou des dégénérescences importantes au niveau des eBSV. Malgré tout l'absence de d'épidémie de BSV va contribuer à la dérive génétique de ces séquences. L'exemple des séquences BSV-like est révélateur de ce qui peut arriver aux séquences BSV que nous avons étudiées et pourrait traduire à plus grande échelle ce que nous sommes en train d'observer pour les eBSV du génome B.

Etape 4 - Le retour des eBSV grâce aux hybrides interspécifiques.

La diversité des eBSV que nous avons observée au sein des bananiers *M. balbisiana* fertiles reprend une part importante de celle des hybrides interspécifiques et pour certains eBSV, ces modifications sont très importantes par rapport aux eBSV de PKW. Nous pouvons donc penser qu'une grande partie de cette diversité ait eu lieu dans le contexte *M. balbisiana* et non chez les bananiers hybrides. Cet argument est renforcé par les temps d'évolution qui sont beaucoup plus importants pour les eBSV au sein des bananiers *M. balbisiana* qu'au sein des hybrides. En effet nous savons que l'intégration de l'eBGFSV a pu avoir lieu à partir de 640,000 ans (Gayral et al., 2008) alors que les hybrides ont été créés seulement il y a 7000 ans pour les plus anciens (Perrier et al., 2011).

L'étude des eBSV chez les hybrides AAB nous a révélé que l'impact des eBSV sur l'évolution des génomes peut être très important car la grande majorité de ces plantes ne possèdent pas ou de façon dégénérée les eBSV, ce qui semble indiquer qu'une sélection a eu lieu au moment de leur création où seules les plantes non productrices de particules virales ont survécu. Ce qui est renforcé par le fait que les bananiers ABB, qui ne peuvent pas produire de particules virales, possèdent une diversité d'eBSV similaire aux plantes BB alors qu'ils sont présents dans les même zones que les hybrides AAB et ont été créés aux mêmes moments (Perrier et al., 2009 et 2011).

Cependant la proximité des séquences BSV endogènes et exogènes qui sont observées aujourd'hui tend à indiquer que des échanges génétiques ont eu lieu entre les eBSV et les BSV (Gayral et al, 2008). De même une partie des foyers d'infections que nous observons aujourd'hui est le résultat de relargage de particules virales à partir des séquences eBSV (Harper et al., 2005 ; Gayral et Iskra-Caruana, 2009). Ces données signifient que la création des hybrides interspécifiques, largement suppléée par la main de l'Homme (Perrier et al., 2009 ; 2011), a pu contribuer au retour du BSV alors que de manière naturelle ce virus ne constituait plus un problème pour cette plante. La création des hybrides interspéfiques par l'Homme a donc permis de justifier de manière fortuite le choix évolutif du BSV à s'intégrer dans le génome de son hôte.

« Tout à une fin sauf la banane qui en a deux »

Expression africaine

REFERENCES BIBLIOGRAPHIQUES

- A -

Argent, G.C.G. (1976). The wild bananas of Papua New Guinea. Notes Royal Botanical Garden Edinburgh.

Ashby, M.K., Warry, A., Bejarano, E.R., Khashoggi, A., Burrell, M., and Lichtenstein, C.P. (1997). Analysis of multiple copies of geminiviral DNA in the genome of four closely related Nicotiana species suggest a unique integration event. Plant Molecular Biology **35,** 313-321.

Aswad, A., and Katzourakis, A. (2012). Paleovirology and virally derived immunity. Trends in Ecology & Evolution **27,** 627-636.

- B -

Bakry, F., Carreel, F., Jenny, C., and Horry, J.P. (2009). Genetic improvement of banana. Breeding plantation tree crops: Tropical species, 3-50.

Banks, D.J.D.J., Beres, S.B.S.B., and Musser, J.M.J.M. (2002). The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. Trends in Microbiology **10,** 515-521.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science **315,** 1709-1712.

Becker, C., Hagmann, J., Müller, J., Koenig, D., Stegle, O., Borgwardt, K., and Weigel, D. (2011). Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature **480,** 245-249.

Bejarano, E.R., Khashoggi, A., Witty, M., and Lichtenstein, C. (1996). Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. PNAS **93,** 759-764.

Bergh, O., Børsheim, K.Y., Bratbak, G., and Heldal, M. (1989). High abundance of viruses found in aquatic environments. Nature **340,** 467-468.

Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature **409,** 363-365.

Bertsch, C., Beuve, M., Dolja, V.V., Wirth, M., Pelsy, F., Herrbach, E., and Lemaire, O. (2009). Retention of the virus-derived sequences in the nuclear genome of grapevine as a potential pathway to virus resistance. Biology Direct **4,** 21.

Betley, M.J., and Mekalanos, J.J. (1985). Staphylococcal enterotoxin A is encoded by phage. Science **229,** 185-187.

Bezier, A., Annaheim, M., Herbiniere, J., Wetterwald, C., Gyapay, G., Bernard-Samain, S., Wincker, P., Roditi, I., Heller, M., Belghazi, M., Pfister-Wilhem, R., Periquet, G., Dupuy, C., Huguet, E., Volkoff, A.N., Lanzrein, B., and Drezen, J.M. (2009). Polydnaviruses of Braconid Wasps Derive from an Ancestral Nudivirus. Science **323,** 926-930.

Biémont, C. (2009). Are transposable elements simply silenced or are they under house arrest? Trends in genetics : TIG **25,** 333-334.

Bill, C.A., and Summers, J. (2004). Genomic DNA double-strand breaks are targets for hepadnaviral DNA integration. PNAS **101,** 11135-11140.

Bioversity, Musa Germplasm Information System (MGIS) http://www.crop-diversity.org/banana/.

Blevins, T., Rajeswaran, R., Aregger, M., Borah, B.K., Schepetilnikov, M., Baerlocher, L., Farinelli, L., Meins, F., Hohn, T., and Pooggin, M.M. (2011). Massive production of small RNAs from a non-coding region of Cauliflower mosaic virus in plant defense and viral counter-defense.

Nucleic Acids Research **39,** 5003-5014.

Blevins, T., Rajeswaran, R., Shivaprasad, P.V., Beknazariants, D., Si-Ammour, A., Park, H.S., Vazquez, F., Robertson, D., Meins, F., Hohn, T., and Pooggin, M.M. (2006). Four plant Dicers mediate viral small RNA biogenesis and DNA virus induced silencing. Nucleic Acids Research **34,** 6233-6246.

Bourc'his, D., and Voinnet, O. (2010). A Small-RNA Perspective on Gametogenesis, Fertilization, and Early Zygotic Development. Science **330,** 617-622.

Bousalem, M., Douzery, E.J.P., and Seal, S.E. (2008). Taxonomy, molecular phylogeny and evolution of plant reverse transcribing viruses (family Caulimoviridae) inferred from full-length genome and reverse transcriptase sequences. Archives of Virology **153,** 1085-1102.

Breitbart, M., and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? Trends in Microbiology **13,** 278-284.

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. (2008). Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. Science **321,** 960-964.

Brüssow, H., Canchaya, C., and Hardt, W.-D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiology and Molecular Biology Reviews **68,** 560-602.

Bucher, E., Reinders, J., and mirouze, M. (2012). Epigenetic control of transposon transcription and mobility in Arabidopsis. Current Opinion in Plant Biology **15,** 503-510.

Burke, G.R., and Strand, M.R. (2012). Polydnaviruses of Parasitic Wasps: Domestication of Viruses To Act as Gene Delivery Vectors. Insects **3,** 91-119.

- C -

Cantu, D., Vanzetti, L.S., Sumner, A., Dubcovsky, M., Matvienko, M., Distelfeld, A., Michelmore, R.W., and Dubcovsky, J. (2010). Small RNAs, DNA methylation and transposable elements in wheat. BMC Genomics **11,** 408.

Capy, P., Gasperi, G., Biémont, C., and Bazin, C. (2000). Stress and transposable elements: co-

evolution or useful parasites? Heredity **85,** 101-106.

Carreel, F., Fauré, S., González de León, D., Lagoda, P.J.L., Perrier, X., Bakry, F., Tezenas du Montcel, H., Lanaud, C., and Horry, J.P. (1994). Evaluation de la diversité génétique chez les bananiers diploïdes (Musa sp). Genetics Selection Evolution **26,** 125-136.

Carver, T., Harris, S.R., Berriman, M., Parkhill, J., and McQuillan, J.A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics **28,** 464-469.

Chao, L., Vargas, C., Spear, B.B., and Cox, E.C. (1983). Transposable elements as mutator genes in evolution. Nature **303,** 633-635.

Cheng, C.P., Lockhart, B.E., and Olszewski, N.E. (1996). The ORF I and II proteins of Commelina yellow mottle virus are virion-associated. Virology **223,** 263-271.

Cheesman, E. E. (1947). Classification of the banana. *Kew bulletin* **2**, 97-117

Côte, F.X., Galzi, S., Folliot, M., Lamagnère, Y., Teycheney, P.-Y., and Iskra-Caruana, M.-L. (2010). Micropropagation by tissue culture triggers differential expression of infectious endogenous Banana streak virus sequences (eBSV) present in the B genome of natural and synthetic interspecific banana plantains. Molecular plant pathology **11,** 137-144.

Crouch, H.K., Crouch, J.H., Jarret, R.L., Cregan, P.B., and Ortiz, R. (1998). Segregation at microsatellite loci in haploid and diploid gametes of Musa. Crop Science **38,** 211-217.

- D -

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M., Da Silva, C., Jabbari, K., Cardi, C., Poulain, J., Souquet, M., Labadie, K., Jourda, C., Lengellé, J., Rodier-Goud, M., Alberti, A., Bernard, M., Correa, M., Ayyampalayam, S., McKain, M.R., Leebens-Mack, J., Burgess, D., Freeling, M., Mbéguié-A-Mbéguié, D., Chabannes, M., Wicker, T., Panaud, O., Barbosa, J., Hřibová, E., Heslop-Harrison, P., Habas, R., Rivallan, R., Francois, P., Poiron, C., Kilian, A., Burthia, D., Jenny, C., Bakry, F., Brown, S., Guignon, V., Kema, G., Dita, M., Waalwijk, C., Joseph, S., Dievart, A., Jaillon, O., Leclercq, J., Argout, X., Lyons, E., Almeida, A., Jeridi, M., Doležel, J., Roux, N.,

Risterucci, A.-M., Weissenbach, J., Ruiz, M., Glaszmann, J.-C., Quétier, F., Yahiaoui, N., and Wincker, P. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature **488,** 213-217.

D'Hont, A., Paget-Goy, A., Escoute, J., and Carreel, F. (2000). The interspecific genome structure of cultivated banana, Musa spp. revealed by genomic DNA in situ hybridization. Theoretical and Applied Genetics **100,** 177-183.

Dahal, G., Hughes, J., Gauhl, F., and Pasberg-Gauhl, C. (2000). Symptomatology and development of banana streak, a disease caused by banana streak badnavirus, under natural conditions in Ibadan, Nigeria. ISHS Acta Horticulturae **540,** International conference on banana and plantain for Africa.

Dahal, G., Hughes, J.A., Thottappilly, G., and Lockhart, B.E.L. (1998). Effect of temperature on symptom expression and reliability of banana streak badnavirus detection in naturally infected plantain and banana (Musa spp.). Plant Disease **82,** 16-21.

Dallot, S., Acuña, P., Rivera, C., Ramírez, P., Côte, F., Lockhart, B.E.L., and Caruana, M.L. (2000). Evidence that the proliferation stage of micropropagation procedure is determinant in the expression of banana streak virus integrated into the genome of the FHIA 21 hybrid (Musa AAAB). Archives of Virology **146,** 2179-2190.

Daniells, J.W., Geering, A.D.W., Bryde, N.J., and Thomas, J.E. (2001). The effect of Banana streak virus on the growth and yield of dessert bananas in tropical Australia. Annals of applied biology **139,** 51-60.

De Langhe, E., Vrydaghs, L., de Maret, P., Perrier, X., and Denham, T. (2009). Why bananas matter: an introduction to the history of banana domestication. Ethnobotany Research & Applications **7,** 165-177.

Devos, K.M., Brown, J.K.M., and Bennetzen, J.L. (2002). Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in Arabidopsis. Genome Research **12,** 1075-1079.

Dewannieux, M., Harper, F., Richaud, A., Letzelter, C., Ribet, D., Pierron, G., and Heidmann, T. (2006). Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Research **16,** 1548-1556.

Drezen, J.M., Provost, B., Espagne, E., Cattolico, L., Dupuy, C., Poirié, M., Periquet, G., and Huguet, E. (2003). Polydnavirus genome: integrated vs. free virus. Journal of Insect Physiology **49,**

407-417.

Duffy, S., Shackelton, L.A., and Holmes, E.C. (2008). Rates of evolutionary change in viruses: patterns and determinants. Nature Reviews Genetics **9,** 267-276.

Dunoyer, P. (2009). La bataille du silence. Med Sci (Paris) **25,** 505-512.

Dupuy, C., Huguet, E., and Drezen, J.-M. (2006). Unfolding the evolutionary story of polydnaviruses. Virus Research **117,** 81-89.

- E -

Eickbush, T.H., and Jamburuthugoda, V.K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Research **134,** 221-234.

Ekwall, K. (2004). The RITS complex-A direct link between small RNA and heterochromatin. Molecular cell **13,** 304-305.

Elbashir, S.M.S., Lendeckel, W.W., and Tuschl, T.T. (2001). RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes & Development **15,** 188-200.

Espagne, E., Dupuy, C., Huguet, E., Cattolico, L., Provost, B., Martins, N., Poirié, M., Periquet, G., and Drezen, J.-M. (2004). Genome sequence of a polydnavirus: insights into symbiotic virus evolution. Science **306,** 286-289.

- F -

Fauquet, C.M. (2005). Virus Taxonomy: VIIIth Report of the International Committee on Taxonomy of Viruses. (Academic Press).

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics **9,** 397-405.

Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on

host biology. Nature Publishing Group **13,** 283-296.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature **391,** 806-811.

Fischer, A., Hofmann, I., Naumann, K., and Reuter, G. (2006). Heterochromatin proteins and the control of heterochromatic gene silencing in Arabidopsis. Journal of Plant Physiology **163,** 358-368.

Forterre, P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. Virus Research **117,** 5-16.

- G -

Gago, S., Elena, S.F., Flores, R., and Sanjuan, R. (2009). Extremely High Mutation Rate of a Hammerhead Viroid. Science **323,** 1308-1308.

Gambley, C.F., Geering, A.D.W., Steele, V., and Thomas, J.E. (2008). Identification of viral and non-viral reverse transcribing elements in pineapple (Ananas comosus), including members of two new badnavirus species. Archives of Virology **153,** 1599-1604.

García, P., Rodríguez, L., Rodríguez, A., and Martínez, B. (2010). Food biopreservation: promising strategies using bacteriocins, bacteriophages and endolysins. Trends in Food Science & Technology **21,** 373-382.

Garcia-Ruiz, H., Takeda, A., Chapman, E.J., Sullivan, C.M., Fahlgren, N., Brempelis, K.J., and Carrington, J.C. (2010). Arabidopsis RNA-Dependent RNA Polymerases and Dicer-Like Proteins in Antiviral Defense and Small Interfering RNA Biogenesis during Turnip Mosaic Virus Infection. The Plant Cell **22,** 481-496.

Gayral, P., Blondin, L., Guidolin, O., Carreel, F., Hippolyte, I., Perrier, X., and Iskra-Caruana, M.L. (2010). Evolution of Endogenous Sequences of Banana Streak Virus: What Can We Learn from Banana (Musa sp.) Evolution? Journal of Virology **84,** 7346-7359.

Gayral, P., and Iskra-Caruana, M.-L. (2009). Phylogeny of Banana Streak Virus Reveals Recent and Repetitive Endogenization in the Genome of Its Banana Host (Musa sp.). Journal of Molecular Evolution **69,** 65-80.

Gayral, P., Noa-Carrazana, J.C., Lescot, M., Lheureux, F., Lockhart, B.E.L., Matsumoto, T., Piffanelli, P., and Iskra-Caruana, M.L. (2008). A Single Banana Streak Virus Integration Event in the Banana Genome as the Origin of Infectious Endogenous Pararetrovirus. Journal of Virology **82,** 6697-6710.

Gazzani, S. (2004). A Link Between mRNA Turnover and RNA Interference in Arabidopsis. Science **306,** 1046-1048.

Geering, A.D.W., McMichael, L.A., Dietzgen, R.G., and Thomas, J.E. (2000). Genetic Diversity Among Banana streak virus Isolates from Australia. Annual Review of Phytopathology **90,** 921-927.

Geering, A.D.W. (2005). Banana contains a diverse array of endogenous badnaviruses. Journal of General Virology **86,** 511-520.

Geering, A.D.W., Parry, J.N., and Thomas, J.E. (2011). Complete genome sequence of a novel badnavirus, banana streak IM virus. Archives of Virology **156,** 733-737.

Geering, A.D.W., Pooggin, M.M., Olszewski, N.E., Lockhart, B.E.L., and Thomas, J.E. (2005). Characterisation of Banana streak Mysore virus and evidence that its DNA is integrated in the B genome of cultivated Musa. Archives of Virology **150,** 787-796.

Gifford, R., and Tristem, M. (2003). The evolution, distribution and diversity of endogenous retroviruses. Virus Genes **26,** 291-315.

Grandbastien, M.A. (2008). Retrotransposons in Plants. Encyclopedia of Plants**,** 428-436.

Gregor, W., Mette, M.F., Staginnus, C., Matzke, M.A., and Matzke, A.J.M. (2004). A distinct endogenous pararetrovirus family in Nicotiana tomentosiformis, a diploid progenitor of polyploid tobacco. Plant physiology **134,** 1191-1199.

- H -

Haas, G., Azevedo, J., Moissiard, G., Geldreich, A., Himber, C., Bureau, M., Fukuhara, T., Keller, M., and Voinnet, O. (2008). Nuclear import of CaMV P6 is required for infection and suppression of the RNA silencing factor DRB4. The EMBO Journal **27,** 2102-2112.

Hamilton, A.J. (1999). A Species of Small Antisense RNA in Posttranscriptional Gene Silencing in Plants. Science **286,** 950-952.

Hammond, M.S., Bernstein, E., Beach, D., and Hannon, J.G. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. Nature **404,** 293-296.

Hansen, C., and Heslop-Harrison, J.S. (2004). Sequences and phylogenies of plant pararetroviruses, viruses, and transposable elements. Advance in Botanical Research **41,** 165-193.

Hansen, C., Harper, G., and Heslop-Harrison, J.S. (2005). Characterisation of pararetrovirus-like sequences in the genome of potato (Solanum tuberosum). Cytogenetic and Genome Research **110,** 559-565.

Harper, G., Hart, D., Moult, S., and Hull, R. (2004). Banana streak virus is very diverse in Uganda. Virus Research **100,** 51-56.

Harper, G., Hart, D., Moult, S., Hull, R., Geering, A., and Thomas, J. (2005). The diversity of Banana streak virus isolates in Uganda. Archives of Virology **150,** 2407-2420.

Harper, G., Hull, R., Lockhart, B., and Olszewski, N. (2002). Viral sequences integrated into plant genomes. Annual Review of Phytopathology **40,** 119-136.

Harper, G., Osuji, J.O., Heslop-Harrison, J.S., and Hull, R. (1999). Integration of banana streak badnavirus into the Musa genome: molecular and cytogenetic evidence. Virology **255,** 207-213.

Harper, G., and Hull, R. (1998). Cloning and sequence analysis of banana streak virus DNA. Virus Genes **17,** 271-278.

Heslop-Harrison, J.S., and Schwarzacher, T. (2007). Domestication, Genomics and the Future for Banana. Annals of Botany **100,** 1073-1084.

Hippolyte, I., Bakry, F., Seguin, M., Gardes, L., Rivallan, R., Risterucci, A.-M., Jenny, C., Perrier, X., Carreel, F., Argout, X., Piffanelli, P., Khan, I.A., Miller, R.N.G., Pappas, G.J., Mbéguié-A-

Mbéguié, D., Matsumoto, T., De Bernardinis, V., Huttner, E., Kilian, A., Baurens, F.-C., D'Hont, A., Cote, F., Courtois, B., and Glaszmann, J.-C. (2010). A saturated SSR/DArT linkage map of Musa acuminata addressing genome rearrangements among bananas. BMC Plant Biology **10,** 65.

Hippolyte, I., Jenny, C., Gardes, L., Bakry, F., Rivallan, R., Pomies, V., Cubry, P., Tomekpe, K., Risterucci, A.M., Roux, N., Rouard, M., Arnaud, E., Kolesnikova-Allen, M., and Perrier, X. (2012). Foundation characteristics of edible Musa triploids revealed from allelic distribution of SSR markers. Annals of Botany **109,** 937-951.

Hohn, T., Richert-Pöggeler, K.R., Staginnus, C., Harper, G., Schwarzacher, T., Teo, C.H., Teycheney, P.Y., Iskra-Caruana, M.L., and Hull, R. (2008). Evolution of integrated plant viruses. Plant Virus Evolution. Berlin: Springer**,** 53-81.

Hohn, T., and Vazquez, F. (2011). RNA silencing pathways of plants: Silencing and its suppression by plant DNA viruses. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms **1809,** 588-600.

Hollister, J.D., and Gaut, B.S. (2009). Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Research **19,** 1419-1428.

Holmes, E.C. (2011). The Evolution of Endogenous Viral Elements. Cell Host and Microbe **10,** 368-377.

Horie, M., and Tomonaga, K. (2011). Non-Retroviral Fossils in Vertebrate Genomes. Viruses **3,** 1836-1848.

Hudson, M.E., Lisch, D.R., and Quail, P.H. (2003). The FHY3 and FAR1 genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. Plant Journal **34,** 453-471.

Huettel, B., Kanno, T., Daxinger, L., Bucher, E., van der Winden, J., Matzke, A.J.M., and Matzke, M.A. (2007). RNA-directed DNA methylation mediated by DRD1 and Pol IVb: A versatile pathway for transcriptional gene silencing in plants. Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression **1769,** 358-374.

Huettel, B., Kanno, T., Daxinger, L., Aufsatz, W., Matzke, A.J., and Matzke, M.A. (2006). Endogenous targets of RNA-directed DNA methylation and Pol IV in Arabidopsis. EMBO Journal **25,** 2828-2836.

Hughes, J.F., and Coffin, J.M. (2001). Evidence for genomic rearrangements mediate by human endogenous retroviruses during primate evolution. Nature genetics **29,** 487-489.

Hull, R. (2001). Classifying reverse transcribing elements: a proposal and a challenge to the ICTV. International Committee on Taxonomy of Viruses, pp. 2255-2261.

Hull, R., Harper, G., and Lockhart, B. (2000). Viral sequences integrated into plant genomes. Trends in Plant Science **5,** 362-365.

- I -

Iskra-Caruana, M.-L., Baurens, F.-C., Gayral, P., and Chabannes, M. (2010). A four-partner plant–virus interaction: enemies can also come from within. Molecular Plant-Microbe Interactions **23,** 1394-1402.

Iskra-Caruana, M.L., Lheureux, F., and Teycheney, P.Y. (2003). Endogenous pararetroviruses (EPRV), an alternative mode of propagation in plants. Virologie **7,** 255-265.

Ito, H., Gaubert, H., Bucher, E., mirouze, M., Vaillant, I., and Paszkowski, J. (2011). An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature **472,** 115-119.

- J -

Jacquot, E., Dautel, S., Leh, V., Geldreich, A., and Yot, P. (1997). Les pararétrovirus de plantes. Virologie **1,** 11-20.

Jacquot, E., Hagen, L.S., Jacquemond, M., and Yot, P. (1996). The open reading frame 2 product of cacao swollen shoot badnavirus is a nucleic acid-binding protein. Virology **225,** 191-195.

Jakowitsch, J., Mette, M.F., Van der Winden, J., Matzke, M.A., and Matzke, A.J. (1999). Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. PNAS **96,** 13241-13246.

James, A.P., Geijskes, R.J., Dale, J.L., and Harding, R.M. (2011). Molecular characterisation of six badnavirus species associated with leaf streak disease of banana in East Africa. Annals of applied biology **158,** 346-353.

Jaufeerally-Fakim, Y., Khorugdharry, A., and Harper, G. (2006). Genetic variants of Banana streak virus in Mauritius. Virus Research **115,** 91-98.

Jeridi, M., Bakry, F., Escoute, J., Fondi, E., and Carreel, F. (2011). Homoeologous chromosome pairing between the A and B genomes of Musa spp. revealed by genomic in situ hybridization. Annals of Botany **108,** 975-981.

Jern, P., and Coffin, J.M. (2008). Effects of retroviruses on host genome function. Annual Review of Genetics **42,** 709-732.

Johnson, L., Cao, X., and Jacobsen, S. (2002). Interplay between Two Epigenetic Marks - DNA Methylation and Histone H3 Lysine 9 Methylation. Current Biology **12,** 8-8.

Johnson, E.W., and Coffin, M.J. (1999). Constructing primate phylogenies from ancient retrovirus sequences. PNAS **96,** 10254-10260.

Jones, D. R. (2000). *Diseases of banana, abacá, and enset*. Wallingford, Oxon, UK ; New-York

- K -

Katzourakis, A., and Gifford, R.J. (2010). Endogenous Viral Elements in Animal Genomes. PLoS Genetics **6,** e1001191.

Kenyon, L., Lebas, B.S.M., and Seal, S.E. (2008). Yams (Dioscorea spp.) from the South Pacific Islands contain many novel badnaviruses: implications for international movement of yam germplasm. Archives of Virology **153,** 877-889.

Kristensen, D.M., Mushegian, A.R., Dolja, V.V., and Koonin, E.V. (2010). New dimensions of the virus world discovered through metagenomics. Trends in Microbiology **18,** 11-19.

Kubiriba, J., Legg, J.P., Tushemereirwe, W., and Adipala, E. (2001). Vector transmission of Banana streak virus in the screenhouse in Uganda. Annals of applied biology **139,** 37-43.

Kumar, A., and Bennetzen, J.L. (1999). Plant retrotransposons. Annual Review of Genetics **33,** 479-532.

Kunii, M., Kanda, M., Nagano, H., Uyeda, I., Kishima, Y., and Sano, Y. (2004). Reconstruction of putative DNA virus from endogenous rice tungro bacilliform virus-like sequences in the rice genome: implications for integration and evolution. BMC Genomics **5,** 80.

- L -

Lafleur, D.A., Lockhart, B.E.L. & Olszewski, N.E. (1996). Portions of *Banana streak badnavirus* genome are integrated in the genome of its host *Musa* sp. *Phytopathology (supplement)* **86**, 100.

Lassoudière, A. (1974). La mosaïque dite "à tirets" du bananier 'Poyo' en Côte d'Ivoire. *Fruits* **29,** 349-357.

Lagoda, P.J.P., Noyer, J.L.J., Dambier, D.D., Baurens, F.C.F., Grapin, A.A., and Lanaud, C.C. (1998). Sequence tagged microsatellite site (STMS) markers in the Musaceae. Molecular Ecology **7,** 659-663.

Le Provost, G., Iskra-Caruana, M.-L., Acina, I., and Teycheney, P.-Y. (2006). Improved detection of episomal Banana streak viruses by multiplex immunocapture PCR. Journal of Virological Methods **137,** 7-13.

Levin, H.L., and Moran, J.V. (2011). Dynamic interactions between transposable elements and their hosts. Nature Reviews genetics **12,** 615-627.

Lheureux, F. (2002). Etude des mécanismes génétiques impliqués dans l'expression des séquences EPRVs pathogènes des bananiers au cours de croisements génétiques interspécifiques. Thèse de doctorat de l'université Montpellier II**,** 1-116.

Lheureux, F., Carreel, F., Jenny, C., Lockhart, B.E.L., and Iskra-Caruana, M.L. (2003). Identification of genetic markers linked to banana streak disease expression in inter-specific Musa hybrids. TAG Theoretical and Applied Genetics **106,** 594-598.

Lheureux, F., Laboureau, N., Muller, E., Lockhart, B.E.L., and Iskra-Caruana, M.L. (2007). Molecular characterization of banana streak acuminata Vietnam virus isolated from Musa acuminata siamea (banana cultivar). Archives of Virology **152,** 1409-1416.

Li, H.H., and Durbin, R.R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Audio, Transactions of the IRE Professional Group on **26,** 589-595.

Li, L.-F., Häkkinen, M., Yuan, Y.-M., Hao, G., and Ge, X.-J. (2010). Molecular phylogeny and systematics of the banana family (Musaceae) inferred from multiple nuclear and chloroplast DNA fragments, with a special reference to the genus Musa. Molecular Phylogenetics and Evolution **57,** 1-10.

Lin, R., Ding, L., Casola, C., Ripoll, D.R., Feschotte, C., and Wang, H. (2007). Transposase-Derived Transcription Factors Regulate Light Signaling in Arabidopsis. Science **318,** 1302-1305.

Lippman, Z., Gendrel, A.V., Colot, V., and Martienssen, R. (2005). Profiling DNA methylation patterns using genomic tiling microarrays. Nature Methods **2,** 219-224.

Lisch, D. (2009). Epigenetic Regulation of Transposable Elements in Plants. Annual Review of Plant Biology **60,** 43-66.

Llave, C. (2010). Virus-derived small interfering RNAs at the core of plant-virus interactions. Trends in Plant Science **15,** 7-7.

Lockhart, B.E. (1995). Banana streak badnavirus infection in Musa : Epidemiology, Diagnosis and Control. Food Fert. Tech. Bull, 143.

Lockhart, B.E.B., Menke, J.J., Dahal, G.G., and Olszewski, N.E.N. (2000). Characterization and genomic analysis of tobacco vein clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. Journal of General Virology **81,** 1579-1585.

Lockhart, B.E.L., and Olszewski, N.E. (1993). Serological and genomic heterogeneity of banana streak badnavirus: implications for virus detection in *Musa* germplasm. Breeding Banana and Plantain for Resistance to diseases and Pests. J. Ganry, ed. International Network for the Improvement of Bananas and Plantain, Montpellier, France.

Lockhart, B.E.L. and Jones, D.R. (2000a). Diseased caused by virus: banana mosaic. In *Diseases of banana, abacà and enset*, pp. 256-263. Ed D. R. Jones. New york: CABI Publishing

Lockton, S., and Gaut, B.S. (2010). The evolution of transposable elements in natural populations of self-fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata. BMC Evolutionary Biology **10,** 10.

Love, A.J., Laird, J., Holt, J., Hamilton, A.J., Sadanandom, A., and Milner, J.J. (2007). Cauliflower mosaic virus protein P6 is a suppressor of RNA silencing. Journal of General Virology **88,** 3439-3444.

- M -

Madlung, A., and Comai, L. (2004). The Effect of Stress on Genome Regulation and Structure. Annals of Botany **94,** 481-495.

Magiorkinis, G., Gifford, R.J., Katzourakis, A., De Ranter, J., and Belshaw, R. (2012). Env-less endogenous retroviruses are genomic superspreaders. Proceedings of the National Academy of Sciences **109,** 7385-7390.

Malik, H.S. (2001). Phylogenetic Analysis of Ribonuclease H Domains Suggests a Late, Chimeric Origin of LTR Retrotransposable Elements and Retroviruses. Genome Research **11,** 1187-1197.

Mallory, A.C. and Bouché, N. (2008). MicroRNA-directed regulation: to cleave or not to cleave. Trends in Plant Science **13,** 9-9.

Matzke, M., Gregor, W., Mette, M.F., Aufsatz, W., Kanno, T., Jakowitsch, J., and Matzke, A.J.M. (2004). Endogenous pararetroviruses of allotetraploid Nicotiana tabacum and its diploid progenitors, N. sylvestris and N. tomentosiformis. Biological journal of the Linnean Society **82,** 627-638.

McCue, A.D.A.D., Nuthikattu, S.S., Reeder, S.H.S.H., and Slotkin, R.K.R.K. (2012). Gene expression and stress response mediated by the epigenetic regulation of a transposable element small RNA. PLoS Genetics **8**.

Mette, M.F., Aufsatz, W., Van der Winden, J., Matzke, M.A., and Matzke, A.J.M. (2000). Transcriptional silencing and promoter methylation triggered by double-stranded RNA. The EMBO Journal **19,** 5194-5201.

Mette, M.F.M., Kanno, T.T., Aufsatz, W.W., Jakowitsch, J.J., van der Winden, J.J., Matzke, M.A.M., and Matzke, A.J.M.A. (2002). Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. EMBO Journal **21,** 461-469.

Meyer, J.B., Kasdorf, G.G.F., Nel, L.H., and Pietersen, G. (2008). Transmission of activated-episomal banana streak ol (badna) virus (bsolv) to cv. Williams banana (musa sp.) by three mealybug species. Plant Disease **92,** 1158-1163.

Mézard, C.C. (2006). Meiotic recombination hotspots in plants. Biochemical Society Transactions **34,** 531-534.

Mi, S., Lee, X., Li, X., Veldman, G.M., Finnerty, H., Racie, L., LaVallie, E., Tang, X.Y., Edouard, P., Howes, S., Keith, J.C., and McCoy, J.M. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature **403,** 785-789.

Mirouze, M., Reinders, J., Bucher, E., Nishimura, T., Schneeberger, K., Ossowski, S., Cao, J., Weigel, D., Paszkowski, J., and Mathieu, O. (2009). Selective epigenetic control of retrotransposition in Arabidopsis. Nature **461,** 427-430.

Moissiard, G., and Voinnet, O. (2006). RNA silencing of host transcripts by cauliflower mosaic virus requires coordinated action of the four Arabidopsis Dicer-like proteins. PNAS **103,** 19593-19598.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods **5,** 621-628.

Mura, M., Murcia, P., Caporale, M., Spencer, T.E., Nagashima, K., Rein, A., and Palmarini, M. (2004). Late viral interference induced by transdominant Gag of an endogenous retrovirus. PNAS **101,** 11117-11122.

Murad, L., Bielawski, J.P., Matyasek, R., Kovarik, A., Nichols, R.A., Leitch, A.R., and Lichtenstein,

C.P. (2004). The origin and evolution of geminivirus-related DNA sequences in Nicotiana. Heredity **92,** 352-358.

Ndowora, T., Dahal, G., LaFleur, D., Harper, G., Hull, R., Olszewski, N.E., and Lockhart, B. (1999). Evidence that badnavirus infection in Musa can originate from integrated pararetroviral sequences. Virology **255,** 214-220.

Niu, Q.-W., Lin, S.-S., Reyes, J.L., Chen, K.-C., Wu, H.-W., Yeh, S.-D., and Chua, N.-H. (2006). Expression of artificial microRNAs in transgenic Arabidopsis thaliana confers virus resistance. Nature Biotechnology **24,** 1420-1428.

- N -

Noreen, F., Akbergenov, R., Hohn, T., and Richert-Pöggeler, K.R. (2007). Distinct expression of endogenous Petunia vein clearing virus and the DNA transposon dTph1 in two Petunia hybrida lines is correlated with differences in histone modification and siRNA production. The Plant Journal **50,** 219-229.

- P-

Pahalawatta, V., Druffel, K., and Pappu, H. (2008). A new and distinct species in the genus Caulimovirus exists as an endogenous plant pararetroviral sequence in its host, Dahlia variabilis. Virology **376,** 253-257.

Paszkowski, J., Shillito, R.D., Saul, M., Mandák, V., Hohn, T., Hohn, B., and Potrykus, I. (1984). Direct gene transfer to plants. EMBO Journal **3,** 2717-2722.

Perrier, X., Bakry, F., Carreel, F., Jenny C., Horry, J.-P., Lebot, V., and Hippolyte, I., (2009). Combining biological approaches to shed light on the evolution of edible bananas. Ethnobotany Research & Applications **7,** 199-216.

Perrier, X., De Langhe, E., Donohue, M., Lentfer, C., Vrydaghs, L., Bakry, F., Carreel, F., Hippolyte, I., Horry, J.P., Jenny, C., Lebot, V., Risterucci, A.M., Tomekpe, K., Doutrelepont, H., Ball, T., Manwaring, J., de Maret, P., and Denham, T. (2011). Multidisciplinary perspectives on banana (Musa spp.) domestication. Proceedings of the National Academy of Sciences **108,** 11311-11318.

Perron, H., and Lang, A. (2009). The Human Endogenous Retrovirus Link between Genes and Environment in Multiple Sclerosis and in Multifactorial Diseases Associating Neuroinflammation. Clinical Reviews in Allergy and Immunology **39,** 51-61.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Sanyal, A., Saniyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S., Wing, R.A., and Panaud, O. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genes and Development **16,** 1262-1269.

Pooggin, M.M., Futterer, J., Skryabin, G.K., and Hohn, T. (1999). A short open reading frame terminating in front of a stable hairpin is the conserved feature in pregenomic RNA leaders of plant pararetroviruses. Journal of General Virology **80,** 2217-2228.

Prangishvili, D., Forterre, P., and Garrett, R.A. (2006). Viruses of the Archaea: a unifying view. Nature Review Microbiology **4,** 837-848.

Puchta, H. (2005). The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. Journal of Experimental Botany **56,** 1-14.

- Q -

Qu, J., Ye, J., and Fang, R. (2007). Artificial MicroRNA-Mediated Virus Resistance in Plants. Journal of Virology **81,** 6690-6699.

- R -

Raja, P., Sanville, B.C., Buchmann, R.C., and Bisaro, D.M. (2008). Viral Genome Methylation as an Epigenetic Defense against Geminiviruses. Journal of Virology **82,** 8997-9007.

Rajeswaran, R., and Pooggin, M.M. (2012). Role of Virus-Derived Small RNAs in Plant Antiviral Defense: Insights from DNA Viruses. MicroRNAs in Plant Development and Stress Responses**,** 261-289.

Ribet, D., Louvet-Vallée, S., Harper, F., de Parseval, N., Dewannieux, M., Heidmann, O., Pierron, G., Maro, B., and Heidmann, T. (2008). Murine endogenous retrovirus MuERV-L is the progenitor of the "orphan" epsilon viruslike particles of the early mouse embryo. Journal of Virology **82,** 1622-1625.

Richert-Pöggeler, K.R., Noreen, F., Schwarzacher, T., Harper, G., and Hohn, T. (2003). Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. EMBO Journal **22,** 4836-4845.

Rigal, M., and Mathieu, O. (2011). A “mille-feuille” of silencing: Epigenetic control of transposable elements. BBA - Gene Regulatory Mechanisms **1809,** 452-458.

Roossinck, M.J. (2012). Plant Virus Metagenomics: Biodiversity and Ecology. Annual Review of Genetics **46,** 359-369.

Rowe, H.M., and Trono, D. (2011). Dynamic control of endogenous retroviruses during development. Virology **411,** 273-287.

- S -

Šafář, J., Noa-Carrazana, J.C., Vrána, J., Bartoš, J., Alkhimova, O., Sabau, X., Šimková, H., Lheureux, F., Caruana, M.-L., Doležel, J., and Piffanelli, P. (2004). Creation of a BAC resource to study

the structure and evolution of the banana (Musa balbisiana) genome. Genome **47,** 1182-1191.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. (1998). The paleontology of intergene retrotransposons of maize. Nature genetics **20,** 43-45.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A. et *al.* (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science **326,** 1112-1115.

Schwind, N., Zwiebel, M., Itaya, A., Ding, B., Wang, M.-B., Krczal, G., and Wassenegger, M. (2009). RNAi-mediated resistance to Potato spindle tuber viroidin transgenic tomato expressing a viroid hairpin RNA construct. Molecular plant pathology **10,** 459-469.

Shepherd, K.R. (1999). Cytogenetics of the genus Musa. International Network for the improvement of banana and plantain, Montpellier, France.

Simmonds, N. W., and Shepherd, K. (1955). The taxonomy and origins of the cultivated bananas. *Journ. Linn. Soc. Bot.* LV, 302-312

Simmonds, N.W. (1962). The evolution of the bananas.

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics **8,** 272-285.

Springer, N.M., and Stupar, R.M. (2007). Allelic variation and heterosis in maize: how do two halves make more than a whole? Genome Research **17,** 264-275.

Staginnus, C., Gregor, W., Mette, M.F., Teo, C., Borroto-Fernández, E., Machado, M.L., Matzke, M., and Schwarzacher, T. (2007). Endogenous pararetroviral sequences in tomato (Solanum lycopersicum) and related species. BMC Plant Biology **7,** 24.

Staginnus, C., and Richert-Pöggeler, K.R. (2006). Endogenous pararetroviruses: two-faced travelers in the plant genome. Trends in Plant Science **11,** 485-491.

Stoye, J.P. (2012). Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nature Publishing Group **10,** 395-406.

Suttle, C.A. (2005). Viruses in the sea. Nature **437,** 356-361.

- T -

Tarlinton, R.E., Meers, J., and Young, P.R. (2006). Retroviral invasion of the koala genome. Nature **442,** 79-81.

Team, R.D.C. (2008). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Teycheney, P.Y., and Geering, A.D.W. (2011). Endogenous Viral Sequences in Plant Genomes. In Recent advances in plant virology, C. Caranta, M. Tepfer, and J.J. Lopez-Moya, 343-361.

Thézé, J., Bézier, A., Periquet, G., Drezen, J.-M., and Herniou, E.A. (2011). Paleozoic origin of insect large dsDNA viruses. PNAS **108,** 15931-15935.

Thomas, J.H., and Schneider, S. (2011). Coevolution of retroelements and tandem zinc finger genes. Genome Research **21,** 1800-1812.

Tomlinson, P.B. (1969). Anatomy of the Monocotyledons. III. Commelinales-Zingiberales. Ed Metcalfe C.R., XX-446

- V -

Vaucheret, H. (2006). Post-transcriptional small RNA pathways in plants: mechanisms and regulations. Genes & Development **20,** 759-771.

Vaucheret, H., Béclin, C., Elmayan, T., Feuerbach, F., Godon, C., Morel, J.B., Mourrain, P., Palauqui, J.C., and Vernhettes, S. (1998). Transgene-induced gene silencing in plants. Plant Journal **16,** 651-659.

Vaughn, Ić, T., Lippman, Jiang, Carrasquillo, Rabinowicz, Dedhia, McCombie, Agier, Bulski, Colot, Doerge, and Martienssen. (2007). Epigenetic natural variation in Arabidopsis thaliana. PLoS Biology **5,** e174.

Voinnet, O. (2005). Induction and suppression of RNA silencing: insights from viral infections.

Nature Reviews Genetics **6,** 206-220.

- W -

Wagner, P.L., and Waldor, M.K. (2002). Bacteriophage control of bacterial virulence. Infection and Immunity **70,** 3985-3993.

Wang, and Maule. (1994). A Model for Seed Transmission of a Plant Virus: Genetic and Structural Analyses of Pea Embryo Invasion by Pea Seed-Borne Mosaic Virus. Plant Cell **6,** 777-787.

Wang, J., Huang, B., Chen, Y., Feng, S., and Wu, Y. (2011). Identification and characterization of microsatellite markers from Musa balbisiana. Plant Breeding **130,** 584-590.

Wang, Q., and Dooner, H.K. (2006). Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. PNAS **103,** 17644-17649.

Weber, B., and Schmidt, T. (2009). Nested Ty3-gypsy retrotransposons of a single Beta procumbens centromere contain a putative chromodomain. Chromosome Research **17,** 379-396.

White, K.A., and Nagy, P.D. (2004). Advances in the molecular biology of tombusviruses: gene expression, genome replication, and recombination. Progress in nucleic acid research and molecular biology **78,** 187-226.

Wierzbicki, A.T. (2012). The role of long non-coding RNA in transcriptional gene silencing. Current Opinion in Plant Biology **15,** 517-522.

- X -

Xiong, Y., and Eickbush, T.H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. The EMBO Journal **9,** 3353-3362.

- Y -

Yot-Dauthy, D., and Bové, J. M. (1966). Mosaïque du bananier: Identification et purification de diverses souches du virus. *Fruit* **21**(9)**,** 449-465.

- Z -

Zhang, X., Shiu, S., Cal, A., and Borevitz, J.O. (2008). Global Analysis of Genetic, Epigenetic and Transcriptional Polymorphisms in Arabidopsis thaliana Using Whole Genome Tiling Arrays. PLoS Genetics **4,** e1000032.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T., and Henikoff, S. (2006). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nature genetics **39,** 61-69.

Résumé :

Le génome du bananier (*Musa sp.*) est envahi par un nombre important de séquences de *Banana streak virus* (BSV), virus à ADN double brin de la famille *Caulimoviridae* qui n'a aucune étape d'intégration au génome hôte au cours de son cycle de multiplication. La majorité de ces intégrations eBSV (endogenous BSV) est défective mais certaines sont restées fonctionnelles et peuvent être à l'origine de particules virales suite à des stress. L'objectif de ce travail de thèse est de préciser si les eBSV sont maintenus ou non dans le génome *Musa balbisiana* des bananiers et d'étudier les conséquences évolutives que cela engendre. Nous avons tout d'abord caractérisé les eBSV fonctionnelles pour trois espèces BSV (*Banana streak goldfinger virus* (BSGFV), *Banana streak obino l'ewai virus* (BSOLV), *Banana streak imove virus* (BSImV) présentes dans le génome du bananier modèle *M. balbisiana cv* Pisang Klutuk Wulung (PKW). Nous avons montré que les intégrations eBSGFV et eBSOLV étaient di-alléliques avec un seul allèle fonctionnel à chaque fois, contrairement à eBSImV qui est mono-allélique et pour lequel nous n'avons pas pu identifier l'allèle à l'origine de l'infection. Leur contexte génomique d'intégration diffère avec une co-localisation d'eBSGFV et d'eBSOLV sur le chromosome 1 et d'eBSImV sur le chromosome 2. Ces résultats nous ont permis de développer les outils moléculaires nécessaires à la caractérisation de ces trois eBSV dans la diversité de *M. balbisiana*. Cette caractérisation a révélé la diversité de structures des eBSV et éclairé une partie encore inconnue de la phylogénie de l'espèce *M. balbisiana*. Dans un second temps nous avons abordé les mécanismes de régulation des eBSV. Ce travail a porté sur les mécanismes d'ARN interférent pouvant expliquer le maintien des eBSV dans le génome des bananiers. Cette analyse révèle que les eBSV sont effectivement sous contrôle d'un mécanisme de type ARNi et la forte production de petits ARNs de 24nt ciblant les eBSV suggère qu'il s'agit d'un silencing au niveau transcriptionnel (TGS). En parallèle, nous avons aussi recherché les mécanismes mis en place par les bananiers non-porteurs d'eBSV en cas d'infection afin de connaître les défenses constitutives des bananiers face à une attaque virale BSV. Nous avons, sur la base de ces résultats, proposé un modèle de régulation des eBSV et des BSV et discuté de l'impact que ces mécanismes auraient pu avoir sur l'évolution des eBSV. L'ensemble des données de ce travail ont permis de préciser les étapes évolutives qu'ont connues les eBSV dans le génome du bananier, expliquant le maintien que l'on observe aujourd'hui.

Mots-clés : Bananier (*Musa sp.*), *Banana streak virus* (BSV), Silencing, Phylogénie, Endogenous Pararetrovirus (EPRV).

ABSTRACT:

The nuclear genome of banana plants is invaded by numerous viral sequences of *banana streak virus* (BSV), a DNA virus belonging to the family *Caulimoviridae,* which does not require integration for its replication. These endogenous BSV (eBSV) are mostly defective; however, some can release a functional viral genome following activating stresses. The objectives of this work were to identify whether the eBSV are maintained or not in the *M. balbisiana* genome and to study the impacts of this on the evolution of banana plants. First, we characterized three functional eBSV sequences present within the *Musa balbisiana* cv PKW genome: (*Banana streak goldfinger virus* (BSGFV); *Banana streak obino l'ewai virus* (BSOLV) ; and, *Banana streak imove virus* (BSImV). We show that eBSOLV and eBSGFV are di-allelic with just one functional allele contrary to eBSImV which are mono-allelic and for which we cannot identified the functional allele. Their genomic areas of integration are different and we also observe that eBSOLV and eBSVGFV are both on chromosome 2 whereas eBSImV is on chromosome 1. These results allowed us to develop the molecular tools required for the characterization of these 3 functional eBSVs within the diversity of *M. balbisiana*. This characterization has revealed the structural diversity of eBSV and has thus clarified previously unresolved details of *M. balbisiana* phylogeny. Secondly, we studied the regulatory mechanism of eBSV expression. This work investigated if RNA interference (RNAi) mechanisms could explain the maintenance of eBSV in the *Musa* genome. Our analyses have shown that, as expected, eBSV was under the control of RNAi mechanisms and the strong production of 24nt small RNAs that target eBSV suggests that Transcriptional Gene Silencing (TGS) was involved in this control. In parallel, we investigated the mechanisms implicated in the anti-viral defense during a BSV infection on a banana plant without eBSV in order to understand the constitutive defense of banana plants. On the basis of these results we have proposed a regulation model of eBSV and BSV and we discuss the impact of silencing regulation on eBSV evolution. Data accumulated during this work have clarified several steps in the co-evolutionary history of *Musa sp*. and eBSV and explain the maintenance of eBSVs in *Musa* genomes that we observe today.

Key words: Banana Plant (*Musa* sp.), *Banana streak virus* (BSV), Silencing, Phylogeny, Endogenous Pararetrovirus (EPRV).