Evolution des génomes mitochondriaux de plantes

—

Approche de génomique comparative chez *Zea mays* et *Beta vulgaris*

THÈSE

présentée et soutenue publiquement le 12 juillet 2010

pour l'obtention du

Doctorat de l'Université de Lille 1 – Sciences et Technologies

(spécialité Biologie Évolutive et Écologie)

par

Aude Darracq

Composition du jury

Président :	Joël Cuguen, Professeur	GEPV, Université de Lille 1
Rapporteurs :	Guillaume Fertin, Professeur	LINA, Université de Nantes
	Jérome Salse, Chargé de Recherche INRA	INRA, Clermont Ferrand
Examinateur :	Laurent Duret, Directeur de Recherche CNRS	PBF, Université Claude Bernard - Lyon 1
Directeurs :	Pascal Touzet, Maître de Conférences	GEPV, Université de Lille 1
	Jean-Stéphane Varré, Maître de Conférences	LIFL-INRIA, Université de Lille 1

UNIVERSITÉ DE LILLE 1 – SCIENCES ET TECHNOLOGIES
ÉCOLE DOCTORALE SCIENCES DE LA MATIÈRE, DU RAYONNEMENT ET DE L'ENVIRONNEMENT
Laboratoire de Génomique et Evolution des Populations Végétales — UMR CNRS 8016
U.F.R. de Biologie – Bât. SN2 – 59655 VILLENEUVE D'ASCQ CEDEX
Tél. : +33 (0)3 20 43 40 24 – Télécopie : +33 (0)3 20 43 69 79

Remerciements

Je tiens tout d'abord à remercier Guillaume Fertin et Jérôme Salse d'avoir accepté d'être rapporteurs de cette thèse ainsi que pour leurs remarques constructives sur ce manuscrit. Je tiens également à remercier Laurent Duret et Joël Cuguen d'avoir accepté d'être examinateurs de ma thèse.

Un très grand merci à mes directeurs de thèse, Pascal Touzet et Jean-Stéphane Varré pour m'avoir proposé un sujet aussi intéressant et qui ont réussi à me supporter pendant ces années. Merci Pascal pour ton enthousiasme débordant et Jean-Stéphane pour ta bonne humeur. Merci à tous les deux d'avoir été disponibles quand il le fallait tout en me laissant l'autonomie dont j'avais besoin ainsi que de m'avoir soutenue dans mes moments de doute, ce fût un encadrement idéal.

Au niveau de mes bases d'accueil, merci à Joël Cuguen et Hélène Touzet de m'avoir accueillie dans leurs locaux. Merci aux membres du GEPV et de SEQUOIA/BONSAI de m'avoir soutenue que ce soit dans mon projet de thèse mais également au niveau des enseignements. Vous êtes malheureusement trop nombreux pour être tous cités mais aucun d'entre vous n'est épargné. Je passe tout de même un merci spécial à Laurent N. pour YASS, quel outil merveilleux et à Fabrice pour les nombreuses discussions (scientifiques ou non).

Dans l'aventure "séquençage des génomes mitochondriaux de betterave", je tiens tout d'abord à remercier Jacky pour la construction d'étagère à betterave de Kangoo et sans qui il aurait été impossible d'amener les 80 plantes à Strasbourg. Merci également à toute l'équipe de la serre (Nathalie, Angélique, Eric et Cédric) pour avoir réagi si rapidement lorsqu'il a fallu semer en urgence des betteraves et pour s'être si bien occupés de mes plantes. Merci à Laurence pour nous avoir accueillies à l'IBMP et pour nous avoir montré LA technique d'extraction. Merci à toute l'équipe du Génoscope qui a travaillé sur le séquençage, en particulier à Valérie qui a dirigé ce travail mais également à Sophie et Patricia qui ont accepté de faire les PCR (même les plus farfelues) que j'ai proposées pour la circularisation des génomes. Bien entendu, un immense merci à Adeline pour m'avoir accompagnée (sans compter son temps) dans l'aventure d'extraction des *mitos* à ses risques et périls.

Je remercie également mes colocataires de bureaux pour tous les bons moments et leur soutien (même après leur départ). Merci ma tacounette (Sarah) d'avoir coopéré dans l'élimination de nos stocks de chocolats, merci Marta de m'avoir ressourcée et convertie aux jus de légumes, merci Docteur Fontaine (Arnaud) et Docteur Geekette (Ségolène) pour cette si bonne humeur dans le bureau 12 et pour votre humour, souvent geek, qui me plaît tant, et merci Benon (Benon) de m'avoir supportée ces derniers mois de rédaction. Merci à tous les autres thésards

pour les bons moments ainsi que pour leur soutien permanent et plus particulièrement à Isa et Jibounet pour m'avoir sortie de force (au début), Camillo pour tes débats même si très souvent tu es un incompris, Camilla pour... tes révélations ainsi que Meriem et Lucy expatriées d'un autre monde mais dont la présence est indispensable.

Merci à tous mes amis qui, en dehors du cadre de travail et malgré le peu de communication ces derniers temps ne m'ont pas oubliée. Merci Amandine, Stellou, Véro, Mickey. Merci GG et Céline d'avoir pris des nouvelles si souvent, merci Fabrice et Tomoka pour le colis "food and goodies" en provenance directe du Japon et un grand merci à Kev pour... être Kev et être venu si souvent en renfort.

Un grand merci à toute ma famille qui m'a toujours encouragée dans mes études. Plus particulièrement à mes grands-parents, Papillon, Minette et Mamie qui n'ont jamais cessé de penser à moi. Merci à Élise pour m'avoir poussée à faire du sport si souvent, ce qui m'a bien défoulée ainsi que pour les sorties au marché. Enfin un très grand merci à mon frère et à mes parents qui ont fait de moi la fille obstinée et perfectionniste que je suis et qui m'ont soutenue sans relâche tout au long de mes études. Merci pour les nombreux ravitaillements en produits du Sud-Ouest qui ont largement contribué à l'équilibre de mon moral.

Enfin un grand merci à Alex pour m'avoir suivie à Lille et aussi soutenue, supportée, épaulée (je vais m'arrêter là, la liste serait beaucoup trop longue), pendant ces années de thèse et toutes les précédentes et sans qui je n'aurais certainement pas eu la force de me lancer dans cette voie.

Table des matières

Table des matières

Introduction

La théorie d'évolution des espèces proposée par Darwin, désormais largement acceptée par la communauté scientifique, consiste à considérer que les espèces évoluent sous l'effet du hasard et sous la pression de sélection pour s'adapter à leur environnement. Cette évolution se fait par le biais de mutations créant ainsi l'apparition de nouvelles espèces. Les premières observations de cette théorie étaient basées sur les études morphologiques d'espèces qui, bien que différentes, présentaient des caractéristiques morphologiques communes. Ces études ont permis d'établir une classifications des espèces et d'en suggérer un arbre d'évolution. Avec la connaissance de la structure d'ADN (*Acide Désoxyribonucléique*), support de l'information génétique, et les méthodologies de séquençage, il a effectivement été établi que les espèces évoluent par des mutations de celui-ci. Il est alors devenu possible de retracer des arbres phylogénétiques en analysant les séquences d'ADN et leurs mutations entre différentes espèces. Désormais, les études menées sur l'évolution des espèces se font essentiellement par génomique comparative, au niveau de l'étude de ces séquences. Cependant, des observations sur les structures chromosomiques ont permis de démontrer des patrons d'évolution communs aux espèces les plus proches. Ainsi, les analyses d'évolution en terme d'étude de structure des chromosomes sont également devenues un champ de la génomique comparative. Ma thèse se place dans ce cadre, étudier l'évolution des génomes en se basant sur l'évolution de leurs structures. Certains génomes sont plus soumis aux réarrangements que d'autres : c'est le cas pour les génomes mitochondriaux de plantes, connus pour leur faible taux de mutation μ (nombre de mutations par site nucléotidique et par génération) et leur fort taux de réarrangement, celui-ci étant vraisemblablement la source principale d'évolution de ces génomes. Les réarrangements mitochondriaux chez les plantes sont décrits comme étant nombreux et complexes entre les différentes espèces, et aucune étude de ces réarrangements du point de vue phylogénétique n'a été proposée. L'objectif principal de cette thèse est de considérer plus attentivement ces réarrangements chromosomiques et de savoir, chez les plantes, s'il est possible de les analyser et d'en tirer une histoire évolutive commune. Dans cette optique, nous avons choisi d'étudier l'évolution de génomes mitochondriaux de plantes où les réarrangements sont nombreux (maïs et betterave), en nous plaçant à un niveau intraspécifique, donc sur une échelle évolutive courte pour limiter le nombre d'événements de réarrangement à analyser.

Cette thèse s'inscrit dans le cadre d'un projet Génoscope permettant le séquençage de cinq génomes mitochondriaux du groupe *Beta* (betterave). Bien que ces génomes soient de petite taille (maximum 500 kpb), l'obtention de leurs séquences a pris beaucoup de temps. En effet, le projet a commencé fin 2006 ; les séquences furent obtenues fin 2009. Plusieurs problèmes se sont posés. Tout d'abord concernant l'extraction de l'ADN mitochondrial de ces individus : la technique n'ayant pas été utilisée depuis longtemps au laboratoire, il a fallu la remettre à jour. L'extraction a finalement pu être effectuée avec l'aide de Laurence Maréchal-Drouard (IBMP, Strasbourg), qui n'a pas hésité à nous inviter dans son laboratoire afin d'acquérir la technique nécessaire. Le

second problème concernait la quantité d'ADN à fournir pour que le séquençage puisse être fait. L'extraction d'ADN exige d'avoir suffisamment de matériel biologique pour obtenir la quantité d'ADN finale nécessaire au séquençage. Afin d'optimiser la quantité d'ADN récupérée, les extractions ont été effectuées à partir de racines de plantes, ce qui implique d'avoir des plantes âgées pour obtenir une quantité de racines suffisante. De ce fait, une fois l'extraction faite, les plantes utilisées ne peuvent plus nous servir. Il nous est arrivé d'avoir une quantité d'ADN finale jugée, après quantification par le Génoscope, insuffisante pour réaliser le séquençage (via la construction d'une banque). Dans ce cas, il a fallu recommencer le processus d'extraction incluant le semis de nouvelles plantes. Le troisième problème est apparu au moment de la reconstruction des génomes mitochondriaux où de potentielles duplications ont empêché le rattachement de tous les contigs et la circularisation de ces génomes. Cette étape de ré-assemblage de contigs sera discutée dans le Chapitre 6. Le séquençage des génomes mitochondriaux de betterave ayant pris beaucoup plus de temps que prévu, nous avons commencé l'étude des réarrangements sur le maïs dont six génomes (plus deux groupes externes) ont été séquencés fin 2006. Cette étude nous a permis de détecter des motifs de duplications, parfaitement conservés entre les espèces étant donné le taux de mutation μ faible. Il s'est avéré que ces duplications étaient importantes à prendre en compte pour réaliser des analyses évolutives. Les méthodes d'analyse des réarrangements ne prenant pas encore en compte les événements de duplication, nous avons établi une méthodologie de tri et de distinction de ces duplications. Cette étude nous a également menés à établir une méthode permettant de retrouver des motifs communs dupliqués entre plusieurs génomes. Une fois les données de betterave obtenues, nous les avons annotées puis utilisé les méthodologies mises en place sur le maïs afin d'établir une phylogénie basée sur les réarrangements.

Plan de lecture

Les deux premiers chapitres de la thèse sont consacrés à l'étude bibliographique des travaux menés sur l'analyse des génomes mitochondriaux d'une part, et sur les méthodes d'étude des réarrangements génomiques d'autre part.

Dans la premier chapitre, nous dressons un état de l'art des connaissances actuelles sur les génomes mitochondriaux, chez les animaux comme chez les plantes. Nous discutons à la fois du rôle fonctionnel, de l'évolution et du contenu en gènes des mitochondries. Nous abordons également le problème de l'architecture des génomes mitochondriaux et relevons les espèces ou clades pour lesquels des données ont été produites.

Dans le second chapitre, nous introduisons les problématiques liées à l'étude des réarrangements chromosomiques ; de l'identification des marqueurs au calcul de phylogénies basées sur les événements de réarrangements génomiques. Nous n'entrons pas dans les détails des algorithmes mais présentons plutôt l'évolution des idées au cours des quinze dernières années. En particulier, nous présentons les outils utilisés pour l'étude des génomes de maïs et de betterave.

Le troisième chapitre présente un premier apport de notre travail concernant la mise en place d'une base de données de génomes mitochondriaux et chloroplastiques de plantes, interrogeable par une interface web, et possédant un outil d'annotation. Cet outil est utile à la fois lorsqu'il s'agit de comparer le contenu des génomes, lorsqu'il faut extraire les relations d'orthologie et de paralogie pour l'analyse des réarrangements ainsi que pour l'annotation automatique des nouveaux génomes séquencés.

Le quatrième chapitre décrit les travaux réalisés sur les génomes de maïs. Ces travaux étaient

portés sur l'analyse des réarrangements de huit génomes du groupe *Zea*. Nous nous sommes retrouvés confrontés au problème de régions dupliquées, pour certaines communes entre les génomes, qui furent essentielles à prendre en compte pour réaliser une étude phylogénétique. Nous en avons tiré l'hypothèse que, chez le maïs, les génomes mitochondriaux évoluent certainement par le biais de duplications en tandem, phénomène décrit dans les mitochondries animales. Cette étude nous a permis de mettre en place un protocole manuel d'identification des régions dupliquées puis de leur distinction afin de procéder aux analyses de réarrangements.

Le cinquième chapitre présente l'algorithme que nous avons mis au point, permettant la détection de régions dupliquées dans un ensemble de génomes. Cet algorithme est basé sur un algorithme paru en 2007 permettant de retrouver des suites de gènes similaires (à l'ordre près et autorisant des délétions) dans un ensemble de génomes. Nous présentons l'extension réalisée permettant de trouver des duplications, en tandem ou non qui peuvent être communes à plusieurs génomes. Cette méthode a ensuite été appliquée aux génomes mitochondriaux de maïs que nous avions analysés manuellement. Les résultats obtenus s'avèrent concluants même s'il reste des paramètres à ajuster.

Le sixième chapitre décrit le programme de séquençage des génomes mitochondriaux de betterave ainsi que les résultats issus de l'analyse de ceux-ci. L'apport ici est triple. Dans un premier temps, nous présentons les résultats du séquençage des régions chloroplastiques des mêmes génomes séquencés au niveau mitochondrial, afin d'obtenir une phylogénie supplémentaire externe au génome mitochondrial (une coévolution des génomes mitochondriaux et chloroplastiques est attendue) qui avait pour objectif d'asseoir la phylogénie de réarrangements et donc l'histoire évolutive des génomes mitochondriaux. Ensuite, nous traitons du séquençage réalisé sur cinq génomes mitochondriaux de betterave ainsi que de l'analyse de leur contenu. Enfin, nous analysons l'histoire évolutive de ces génomes en terme de réarrangements. Notre méthode de détection des duplications a été mise à profit afin de mettre en œuvre plus facilement la méthodologie proposée lors de l'étude des génomes de maïs.

Le document se termine par une discussion sur les apports de la thèse et les pistes ouvertes par ce travail.

Chapitre 1

Génomes mitochondriaux des plantes et des animaux

Dans ce chapitre, nous ferons un rappel des notions générales sur les mitochondries Eucaryotes : de leur origine, de leur fonction et leurs principales caractéristiques. Bien que cette thèse concerne les mitochondries végétales, nous parlerons également des mitochondries animales. En effet, ces deux types de mitochondries présentent des caractéristiques différentes bien qu'ils aient une origine commune. Pour finir, nous verrons que les différences entre certaines caractéristiques, telles que la taille des génomes, le code génétique utilisé ou encore les événements de réarrangement peuvent en fait s'expliquer par une différence du taux de mutation μ entre les mitochondries des animaux et des végétaux (hypothèse de pression de mutation). Le taux de mutation μ représentant ici le nombre de mutations par site nucléotidique et par génération.

1.1 Généralités

1.1.1 Origine des mitochondries

Les mitochondries sont des organites que l'on retrouve dans le cytoplasme des cellules Eucaryotes. Elles sont un des seuls organites à posséder leur propre génome. Depuis les années 70, après observation de leurs ribosomes et ARN de transfert présentant une forte ressemblance à ceux des bactéries, il est admis que les mitochondries ont une origine extra-cellulaire [Margulis, 1970]. En 1999, avec le séquençage de nombreux génomes mitochondriaux dont un génome mitochondrial proche de celui des bactéries et un génome eubactérien très proche des génomes mitochondriaux, l'hypothèse d'une origine monophylétique des mitochondries a pu être confirmée par l'analyse des séquences de leurs gènes (analyse sur les gènes *cob*, *cox1* et *cox3*) [Gray et al., 1999]. Les mitochondries proviennent donc de l'endosymbiose d'une α-proteobactérie. Cependant deux hypothèses de cette endosymbiose sont proposées. Soit les cellules Eucaryotes ont été formées en deux temps, avec, dans un premier temps la formation d'une cellule Eucaryote dépourvue de mitochondries suite à la fusion d'une archéobactérie et d'une proteobactérie puis, dans un deuxième temps l'acquisition des mitochondries par endosymbiose d'une α-proteobactérie [Zillig et al., 1989] (flèches magentas dans la Figure 1.1). Soit il y a directement eu formation de cellules Eucaryotes pourvues de mitochondries suite à une fusion d'une archeobactérie méthanogénique avec une α-proteobactérie productrice d'hydrogène [Martin and Müller, 1998] (flèches bleues dans la Figure 1.1).

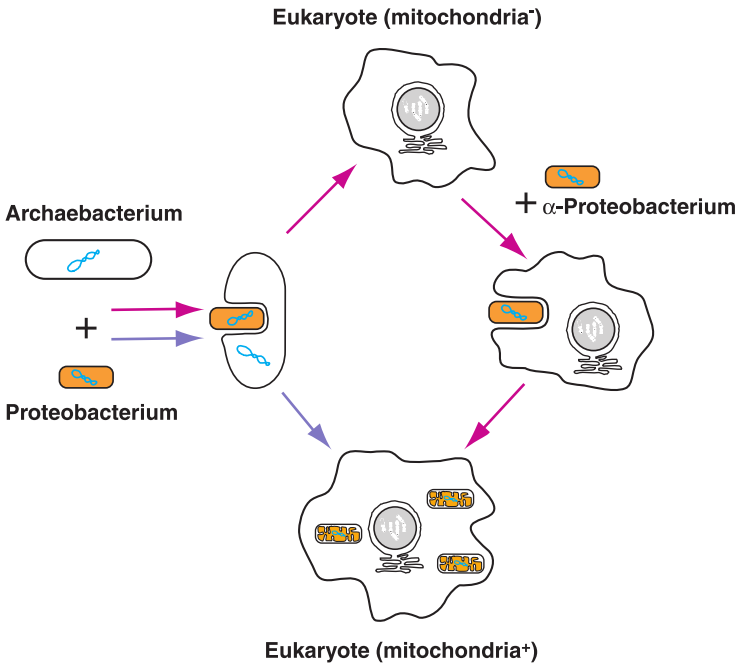


FIG. 1.1 – Hypothèses sur l'origine des mitochondries [Gray et al., 1999]. Les flèches magentas représentent l'hypothèse de la formation des cellules Eucaryotes en deux temps (formation des cellues eucaryotes sans mitochondries puis endosymbiose d'une α-proteobactérie). Les flèches bleues représentent l'hypothèse de la formation des cellules Eucaryotes en une étape (fusion d'une archeobactérie méthanogénique avec une α-proteobactérie productrice d'hydrogène).

1.1.2 Fonctions

Les mitochondries, situées dans le cytoplasme des cellules Eucaryotes, sont pourvues d'une double membrane et possèdent leur propre génome. On trouve plusieurs copies de l'ADN mitochondrial dans une seule mitochondrie. Bien que les génomes mitochondriaux puissent varier, en terme de structure (taille du génome, nombre de gènes, ordre des gènes) entre les espèces, toutes les mitochondries ont la même fonction principale ayant lieu au niveau de la membrane interne : la phosphorylation oxydative (Figure 1.2). Cette fonction dépend de cinq complexes enzymatiques, nommés complexe I à V plus l'ubiquinone et le cytochrome *c*. Le complexe I (NADH :ubiquinone oxydoreductase) est codé par un ensemble de gènes appelés *nad*. Il permet la réduction du NADH en NAD^+ générant ainsi des protons (H^+) et des électrons. Le complexe II (succinate :ubiquinone oxydoreductase), encodé par un ensemble de gènes appelées *sdh*, permet la réduction du succinate (généré lors du cycle de Krebs) en fumarate, produisant également la libération d'électrons. L'ubiquinone sert à transférer les électrons provenant des oxydoréductions au complexe III. Le complexe III (ubiquinone :cytochrome c oxydoreductase), codé par un ensemble de gènes appelés *cob*, permet l'oxydoréduction de la coenzyme Q libérant ainsi des protons. Les électrons produits passent alors par le cytochrome *c* qui les transfèrent au complexe IV. Le complexe IV (ferrocytochrome c :oxygen oxidoreductase), encodé par un ensemble de gènes appelés *cox*, utilise les électrons pour la réduction d'oxygène en eau. Au cours de cette réaction, des protons sont également produits. Trois des complexes (I, III et IV) aboutissent à la formation de protons. Ces protons sont expulsés, par chacun des complexes, dans l'espace intermembranaire. Cela crée un gradient de protons fournissant l'énergie nécessaire au fonction-

nement du complexe V ou encore à l'importation de protéines. Le complexe V (ATP synthase), codé par un ensemble de gènes appelé *atp* permet la synthèse de molécule d'ATP et d'eau à partir d'ADP, d'oxygène et de protons pompés depuis l'espace intermembranaire. Les composés utilisés lors de la phosphorylation oxydative (NADH, $FADH_2$ fumarate, ADP) proviennent du cycle de Krebs.

Chez les végétaux, on trouve d'autres enzymes, telles que des NAD(P)H déshydrogénase, jouant un rôle similaire au complexe I mais sans expulsion de protons. On trouve également l'AOX (oxydase alternative), qui se substitue aux complexes III et IV puisqu'elle permet la réduction d'oxygène en eau à partir des électrons fournis par l'ubiquinone. Cependant, il n'y a pas d'expulsion de protons avec cette enzyme (et donc une moindre production d'ATP).

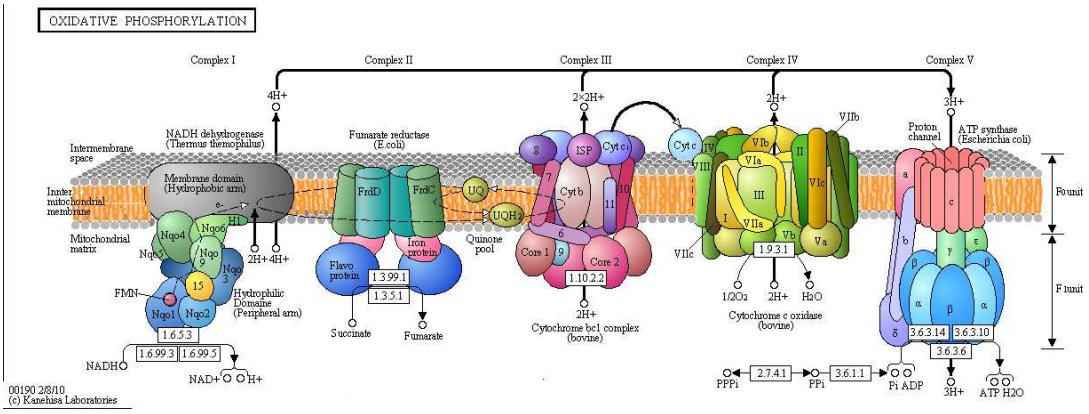


FIG. 1.2 – Phosphorylation oxydative. Extrait de KEGG pathway [Kanehisa and Goto, 2000]. Les différents complexes intervenant dans les phosphorylation oxydative sont ancrés dans la membrane interne de la mitochondrie. Les protons, générés au cours des différentes réactions, sont expulsés dans l'espace intermembranaire et les électrons sont transférés de complexes en complexes. L'ATP est formé dans la mitochondrie lorsque les protons sont pompés de l'espace intermembranaire vers la matrice mitochondriale par le complexe V.

Les mitochondries jouent également un rôle dans la synthèse des hormones stéroïdes (via des cytochromes dans la matrice mitochondriale chez les animaux), dans la régulation de la concentration intracellulaire en calcium ou encore dans l'apoptose des cellules (transport cytosolique du cytochrome *c* entrainant l'activation de protéases).

1.1.3 Hérédité

L'hérédité des mitochondries est très souvent décrite comme uniparentale et le plus souvent de type maternel. Les génomes mitochondriaux sont donc homoplasmiques à cause de cette hérédité uniparentale ce qui induit une évolution de ces génomes basée sur des mutations et des recombinaisons intragénomiques, puisqu'il n'y aura pas de mélange d'ADN mitochondrial. Différents processus de dégradation de l'ADN mitochondrial ont été établis au niveau de la reproduction sexuée, menant ainsi à une hérédité uniparentale [Birky, 2001]. La dégradation des mitochondries peut se faire dans les gamètes au moment de la fécondation, l'ADN mitochondrial n'entrant pas dans l'œuf ou une dégradation sélective peut avoir lieu dans le zygote. Chez les mammifères, le sperme est marqué par une protéine (ubiquitine) aboutissant à sa dégradation par des protéosomes et lysozomes. Même si dans la majorité des espèces, l'hérédité est uniparentale

et maternelle, il existe des exceptions. Chez les moules, par exemple, on observe un double héritage : l'ADN mitochondrial des femelles est transmis aux femelles et aux mâles (mais serait éliminé au bout de 24 heures) et l'ADN mitochondrial mâle est transmis aux mâles [Rand, 2001]. Il ne s'agit cependant pas du seul cas de fuite paternelle détecté chez les animaux, le phénomène ayant déjà été vu ponctuellement chez les drosophiles, les souris, les humains et les oiseaux [Kvist et al., 2003]. Il s'agirait d'un problème de dégradation de l'ADN mitochondrial au moment de la fécondation. Cependant dans ces cas de problèmes de dégradation, l'ADN mitochondrial d'origine paternelle est en quantité infime par rapport à celui d'origine maternelle.

Ces phénomènes d'hérédité biparentale ont également été décrits chez les plantes. Notamment, chez *Silene vulgaris*, il a été montré que certains gènes (*atpA* et *coxI*) sont hérités paternellement [Welch et al., 2006]. Les premières observations d'hétéroplasmie étaient considérées comme des duplications avec de forts taux de mutations μ [Rand, 2001].

1.1.4 Contenu en gènes

Nous l'avons vu, les mitochondries ont une origine commune entre les différentes espèces. Cependant, au cours de l'évolution, le contenu en gènes s'est diversifié. Si, chez les vertébrés, on dénombre généralement douze gènes codant pour des protéines, il en existe seulement quatre chez les plasmodium et plus d'une trentaine chez certaines plantes. Malgré cette différence en nombre de gènes codés par le génome mitochondrial, les mitochondries de ces organismes assurent toutes les mêmes fonctions. Les protéines manquantes nécessaires au bon fonctionnement des mitochondries sont codées par le noyau et sont ensuite importées dans les mitochondries.

1.2 Fonctionnement des mitochondries

Dans cette partie, nous synthétisons les différentes connaissances concernant la composition, la structure, la réplication et la transcription des mitochondries animales et végétales. La bibliographie concernant ces connaissances au niveau des animaux est impressionnante, notamment les mécanismes de réplication et transcription qui ont été très étudiés. Au contraire, chez les végétaux, ces mécanismes sont beaucoup moins connus et on trouve difficilement des études les concernant.

1.2.1 Mitochondries animales

Composition

Chez les mammifères on trouve 50 à 75 mitochondries dans les spermatozoïdes et plus de 10^5 dans les ovocytes. En général on aura entre 10^3 et 10^4 copies d'ADN mitochondrial par cellule. D'un point de vue général, sans tenir compte des exceptions, l'ordre des gènes et la taille du génome sont bien conservés, tandis que la séquence et les éléments intervenant dans la transcription et la réplication varient considérablement entre les espèces. Le génome mitochondrial chez les animaux est, généralement, circulaire et sa taille varie entre 16 et 18 kpb. Les mitochondries animales contiennent 13 polypeptides de la chaîne respiratoire, 22 ARN de transfert (ARNt) et 2 ARN ribosomiques (ARNr). Les protéines ribosomiques sont codées et synthétisées en dehors de la mitochondrie et les enzymes intervenant dans les voies métaboliques de la mitochondrie sont codées par le noyau.

Structure

L'information génétique des mitochondries animales a été observée puis décrite comme étant contenue sur une molécule circulaire. Cependant, des exceptions ont été trouvées chez certains protozoaires *Tetrahymena pyriformis* et *Paramecium aurelia* avec des molécules linéaires [Bendich, 1993]. De plus, des événements de recombinaison intramoléculaire ont été observés chez certaines espèces, comme c'est le cas pour le phytonematode *Meloidogyne javanica* où deux molécules d'ADN mitochondrial ont été vues. Cela proviendrait d'une recombinaison au niveau de la région de contrôle, entraînant la formation d'une petite sous-molécule contenant une partie de la région de contrôle et une grande sous-molécule contenant tout le génome ainsi que l'autre partie de la région de contrôle [Lunt and Hyman, 1997]. Comme nous l'avons décrit dans la Section 1.1.3, les mitochondries sont généralement observées comme étant uniparentalement héritées, à l'exception par exemple des moules présentant une biparentalité. Des recombinaisons entre ces molécules ont pu être observées. En effet, chez *Mytilus galloprovinvialis*, des molécules recombinantes coexistent avec les molécules parentales, suggérant que des recombinaisons peuvent apparaître à l'intérieur d'un même individu [Ladoukakis and Zouros, 2001]. De plus, chez les moules *Mytilus trossulus*, Burzynski et ses collaborateurs suggèrent que les molécules recombinantes peuvent être transmises aux générations suivantes [Burzynski et al., 2003]. Il ne faut pas oublier que l'on peut avoir des recombinaisons non-homologues ou des mésappariements conduisant à des insertions ou délétions, les mitochondries animales possèdant les enzymes nécessaires à la recombinaison [Rokas et al., 2003]. Cependant, l'hétéroplasmie chez les animaux reste rare et difficilement détectable lorsque les variants diffèrent de quelques bases [Barr et al., 2005].

Dans le génome mitochondrial des animaux, les gènes ne possèdent pas d'introns et, sauf pour les zones de régulation, les espaces intergéniques n'existent pas ou sont composés de quelques bases. Le génome est tellement compact que certains gènes protéiques se superposent [Taanman, 1999].

Réplication

Le génome mitochondrial des animaux possède deux brins appelés brin lourd (brin H) et brin léger (brin L). Les gènes sont répartis sur les deux brins. Par exemple, chez l'humain, les gènes mitochondriaux sont répartis de la façon suivante : douze polypeptides, quatorze ARNt et deux ARNr sur le brin H et un polypeptide et huit ARNt sur le brin L (Figure 1.3) [Taanman, 1999].

Chez les mammifères, la réplication se fait de manière unidirectionnelle et sera spécifique de chaque brin. La réplication des brins se distingue spatialement et temporellement. L'origine de réplication du brin H (O_H) se trouve dans la boucle D (chez les vertébrés, cette boucle est encadrée des gènes *ARNt-phe* et *ARNt-pro*) tandis que celle du brin L (L_H) se trouve aux deux tiers du génome. La réplication commence par le brin H, au niveau de O_H, le long du brin L. Lorsque le brin fils H synthétisé passe le niveau de O_L, la réplication du brin L commence dans le sens inverse de celui du brin H. La région O_H est composée de blocs de séquences conservés, appelés CSB (CSBI, CSBII et CSBIII). La réplication du brin H est donc initiée au niveau de la boucle D et nécessite l'interaction entre les régions CSB et l'ARN contenu au niveau de la boucle D. Cette interaction génère une liaison stable ADN/ARN, on appelle alors cette région la boucle R (qui est alors sous la forme d'un triple brin). Il existe des régions appelées *TAS* sur l'ADN mitochondrial arrêtant la réplication. Le nombre de régions *TAS* varie selon les espèces. Chez l'humain, il n'existe qu'une seule région *TAS* et il s'agit du site majeur de terminaison du

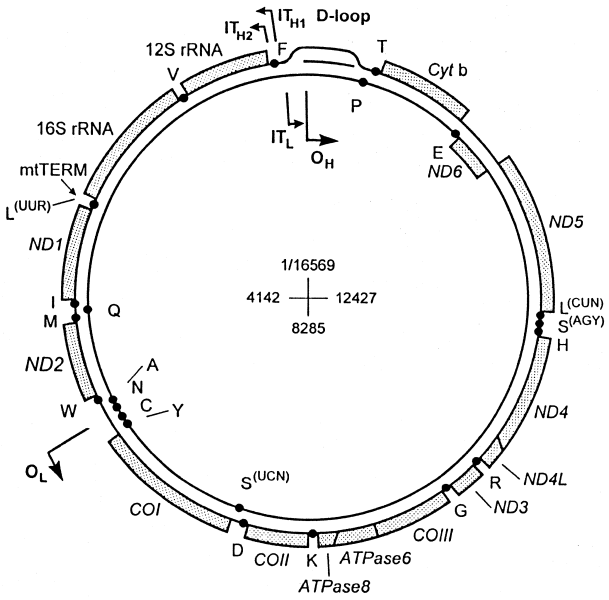


FIG. 1.3 – Schéma d'un génome mitochondrial humain [Taanman, 1999]. Le cercle externe représente le brin H et le cercle interne le brin L. La région boucle D est représentée comme une structure trois brins. Les origines de réplication de chaque brin sont indiquées (O_H et 0_L) ainsi que les origines de transcription (IT_H et IT_L). Le contenu en gènes, ARNt et ARNr est représenté sur chaque brin.

brin H. Il y a deux sites mineurs de terminaison adjacent au site majeur. On ne connaît pas ce qui détermine l'arrêt ou non de la réplication, mais cet arrêt permet de réguler le nombre de copies d'ADN mitochondrial dans les cellules. Des facteurs nucléaires interagiraient avec les *TAS*. L'initiation de la réplication du brin L commence lors du passage de la réplication du brin H au niveau de O_L. Une *ADN primase* synthétise une amorce ARN sur le site O_L dans une région riche en T pour ensuite laisser la place à la synthèse d'ADN. L'élongation se fait à l'aide de la *polymérase* γ qui est la seule polymérase présente dans la mitochondrie. Bien évidemment, l'élongation nécessite d'autres enzymes telles que des *hélicases* ou encore de *topoisomérases*, qui seront importées du cytoplasme.

Transcription et traduction

Plusieurs études menées sur l'ADN mitochondrial humain, utilisant différents procédés, ont montré l'existence de deux sites majeurs d'initiation de la transcription au niveau de la boucle D [Shadel and Clayton, 1997, Taanman, 1999]. Ces régions sont appelées IT_{H1} et IT_L. La transcription du brin H se fait au niveau de IT_{H1} et celle du brin L au niveau de IT_L. Des études *in vitro*, confirmées *in vivo* sur des patients, démontrent que ces deux promoteurs fonctionnement indépendemment. Un autre site d'initiation de la transcription a été trouvé (IT_{H2}). Celui-ci serait utilisé moins fréquemment que IT_{H1} pour la transcription du brin H.

Une fois initié au niveau de IT_L, le brin L est transcrit en un ARN précurseur polycistronique contenant quasiment toute l'information génétique codée sur le brin. Dans certaines cellules, les ARN du brin H ont un taux de transcription supérieur aux autres gènes. Cela proviendrait des deux sites d'initiation. Généralement, la transcription commencerait au site IT_{H1} pour finir

après l'ARNr 16S. Cette synthèse produirait la majeure partie des ARNr. En revanche, une transcription commençant au site IT_{H2}, moins fréquente, entraînerait la transcription entière du brin H. Un facteur (mtTERM) induit une torsion de l'hélice et stoppe la transcription. *In vitro*, ce facteur se lie à son site de fixation de façon bidirectionnelle, arrêtant la transcription du brin H et du brin L : cette transcription des brins H et L se fait donc d'une traite puisqu'il n'y a pas de séquence intronique chez les vertébrés avec un minimum de séquences intergéniques. On obtient donc deux longs messagers polycistroniques (un pour le brin H et un pour le brin L).

La synthèse des protéines est initiée avec ARNt-fMet. La traduction semble difficile étant donné qu'il n'y a pas beaucoup de nucléotides en 5' des gènes mitochondriaux. Ceci pourrait expliquer pourquoi beaucoup d'ARN messagers sont transcrits tandis que beaucoup moins sont traduits. Expérimentalement, il a été montré que la sous-unité ribosomique 28S peut se fixer sur un site de 30 à 80 pb mais une bonne liaison exige une région d'environ 400 pb. Ce serait peut-être une raison pour laquelle on a des gènes qui se chevauchent.

Le code génétique utilisé dans les mitochondries animales est différent du code génétique standard du génome nucléaire. Dans ce dernier par exemple, le codon TGA correspond à un codon stop ; dans les mitochondries animales il correspond à un tryptophane. De plus, en fonction des espèces, les codons utilisés peuvent être différents. Par exemple, AGA et AGG correspondent à un codon stop chez les vertébrés, une sérine chez les echinodermes alors qu'il s'agit d'une arginine dans le code génétique standard. Chez les animaux, les 24 ARNt sont nécessaires pour coder les gènes mitochondriaux.

Taux de substitutions

La majorité des points de mutation sont non-synonymes mais certaines délétions vont interrompre la phase de lecture [Rand, 2001]. Chez les mammifères et les oiseaux, le taux de mutation μ est plus élevé dans la lignée germinale mâle que femelle [Whittle and Johnston, 2002]. Des analyses chez les souris [Ballard and Dean, 2001] montrent que le taux d'évolution de l'ADN mitochondrial est lié à celui de l'ADN nucléaire : les gènes mitochondriaux en interaction avec les gènes nucléaires vont alors évoluer plus vite (phénomène de co-évolution) [Ballard and Dean, 2001],[Ballard and Rand, 2005]. Le taux de mutation μ intervient également dans l'élimination des gènes dupliqués. En effet, lors de duplications de gènes, si le taux de mutation μ est élevé, les gènes dupliqués seront rapidement mutés et donc transformés en pseudogènes. Les pseudogènes ne seront pas maintenus et seront éliminés. Chez l'homme, le taux de substitution mitochondrial a été évalué à $1,70 \times 10^{-8}$ substitutions par site et par année [Ingman et al., 2000].

1.2.2 Mitochondries végétales

La transcription et surtout la réplication chez les plantes sont beaucoup moins connues que chez les animaux.

Composition

La taille des génomes mitochondriaux des plantes est très variable, elle peut aller de 15 kpb (*Chlamydomonas reinhardtii*) à plus de 1 Mpb (*Cucurbita pepo*). La majorité du génome mitochondrial, dans les espèces ayant une grande taille, est composé de séquences non codantes. De plus, la composition en gènes dans ces génomes est également variable. Contrairement aux

génomes mitochondriaux animaux, des gènes dupliqués existent dans les génomes mitochondriaux végétaux. De plus, contrairement aux mitochondries animales, des gènes codant pour des protéines ribosomiques sont retrouvés dans le génome mitochondrial des végétaux. Certains gènes mitochondriaux végétaux comporteront des introns qui peuvent être épissés en cis (les exons sont séparés par des introns) ou en trans (les exons sont répartis sur le génome, séparés par d'autres gènes). Par exemple, *Marchantia polymorpha* [Schuster and Brennicke, 1994] a un génome d'environ 186 kpb et possède seize gènes codant pour des protéines, trente ARNt et trois ARNr. Cette espèce possède également des introns dont la proportion est évaluées à environ 20% du génome. *Marchantia polymorpha* est cité ici à titre d'exemple mais les tailles des génomes, leurs contenus en gènes, ARN et introns sont très variables d'une espèce à une autre. Par exemple, *Marchantia polymorpha* possède des introns sur *cox1* qui n'existent pas chez les autres plantes. Ces introns ont probablement été perdus par une *reverse transcription* des transcrits puis leur réinsertion dans le génome. Certains introns peuvent également être dupliqués puis déplacés ailleurs dans le génome. Entre les différentes espèces végétales, l'ordre des gènes n'est quasiment pas conservé contrairement aux génomes mitochondriaux des animaux.

Le code génétique utilisé dans les mitochondries végétales est le code génétique standard.

Structure

La structure des génomes mitochondriaux des plantes semble plus complexe que celles des animaux (souvent sous forme circulaire). En effet, les génomes mitochondriaux de certaines espèces ont été vus sous forme linéaire (par exemple chez *Chlamydomonas reinhardtii*) ou encore sous forme de sous-cercles (tout le contenu du génome est réparti sur plusieurs molécules circulaires) [Bendich, 1993]. Il est même possible de trouver des molécules circulaires et linéaires pour un même génome (par exemple chez le pois). Ces formes seraient dues à des recombinaisons de l'ADN. Ainsi, le génome complet construit sous la forme d'un unique cercle est appelé *cercle maître*. Dans les cultures de cellules, il a été observé que les molécules circulaires sont plus courantes mais la plupart sont des sous-cercles du cercle maître. Au contraire, chez *Brassica hirta* et *Marchantia polymorpha*, aucune recombinaison interne n'a été trouvée, seulement le cercle maître (confirmé *in vivo* par microscopie électronique). Les cercles maîtres peuvent comporter des régions dupliquées, codantes ou non, qui peuvent être impliquées ou non dans des recombinaisons [Marienfeld et al., 1997]. Ces duplications ne sont pas conservées entre les espèces : par exemple, les répétitions trouvées chez le tabac ne présentent aucune homologie avec celles des quatre autres angiospermes séquencés, ce phénomène de duplication est donc indépendant de l'évolution des angiospermes [Sugiyama et al., 2005]. Théoriquement, lorsqu'un cercle maître possède des duplications, on peut aboutir à des recombinaisons dont deux types sont envisageables : les recombinaisons intramoléculaires et les recombinaisons intermoléculaires. Si deux répétitions dans un cercle maître sont dans le même sens et que l'on a une recombinaison intramoléculaire, on aboutit alors à la formation de deux sous-cercles. Au contraire, si les duplications sont dans des sens contraire, on aura la formation d'un nouveau cercle maître dont la région entre les duplications est inversée (Figure 1.4). Sachant qu'il existe plusieurs molécules d'ADN par mitochondrie, on peut également avoir des recombinaisons intermoléculaires. Bien entendu, ces réarrangements ne restent pas forcément sans conséquences, la formation de nouvelles molécules pouvant avoir des impacts sur les gènes. On peut par exemple obtenir des pseudogènes, des fusions de gènes, un changement dans l'épissage des gènes ou encore déconnecter un gène de son promoteur (ce qui agira sur son expression) [Marienfeld et al., 1997]. Les recombinaisons peuvent également causer l'apparition de nouvelles ORF (Open Reading Frame). Les ORF induisent

plus fréquemment des effets négatifs que positifs, mais les gènes nucléaires peuvent compenser les effets négatifs [Marienfeld et al., 1997].

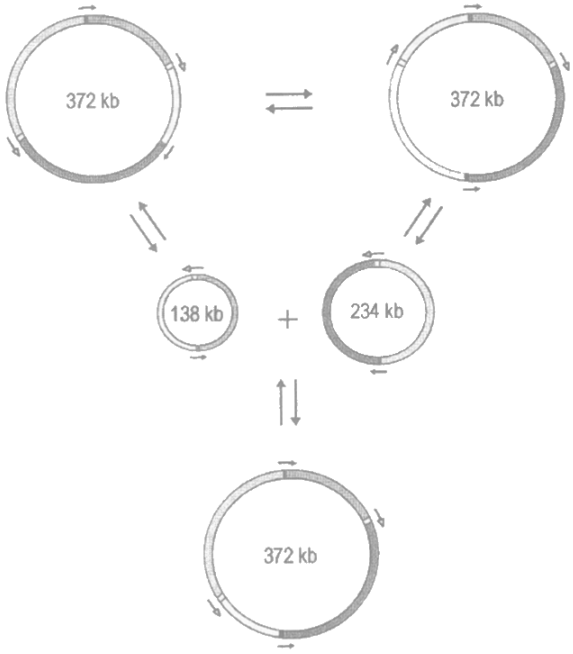


FIG. 1.4 – Exemple d'organisation de génomes mitochondriaux en cercle maître (deux cercles du haut et celui du bas) et sous-cercles (deux cercles du milieu) [Backert et al., 1997]. Une recombinaison entre deux répétitions en sens inversé va aboutir à une inversion d'une portion du génome (première ligne). Si la recombinaison se fait entre des répétitions dans un même sens on aura la formation de sous-cercles. Ces sous-molecules peuvent se recombiner si elles possèdent une répétition commune pour aboutir à un cercle maître.

Réplication

Le système de réplication des mitochondries végétales est beaucoup moins connu que celui des animaux. Il faut noter que, si les génomes mitochondriaux des plantes sont organisés en sous-cercles, pour parvenir à se répliquer, le sous-cercle doit forcément comporter une origine de réplication. Le cercle maître semble être le plus apte pour la réplication mais une sous-molécule contenant l'origine de réplication doit aussi pouvoir le faire.

En 1991, une étude menée sur le pétunia (*Petunia hybrida*) a réussi à montrer l'existence d'une origine de réplication *in vitro* [de Haas et al., 1991]. Quatre origines de réplication potentielles ont été identifiées (pPMY-I, pPMY-II, pPMY-III et pPMY-IV).
pPMY-III et pPMY-IV montrent des homologies structurales avec les origines de réplication des levures et des mammifères. pPMY-IV ressemblerait plutôt au site sur le brin L (région riche en T, site de reconnaissance de l'*ADN primase* pour l'initiation de l'amorce servant à la réplication) alors que pPMY-III ressemble à celui sur le brin H. Ces sites de reconnaissance vont donc constituer les origines de réplication O_A (pPMY-III) et O_B (pPMY-IV) et sont des sites de fixations pour les enzymes. Les analyses ont montré que l'initiation de la réplication se fait

uniquement à partir de pPMY-III. Des expériences *in vitro* ont montré qu'une *ADN-polymérase* γ, une *ARN-polymérase* ainsi que des activités *primase* et *gyrase* seraient présentes dans la mitochondrie de *Petunia hybrida*. *In vitro*, une boucle D a été vue par microscopie électronique au niveau de pPMY-III.

Après l'initiation au site O_A, l'élongation se fait de façon unidirectionnelle dans le sens des aiguilles d'une montre. Une fois le site O_B atteint, l'initiation du deuxième brin commence en direction de O_A (c'est-à-dire dans le sens inverse).
Plusieurs copies de O_B sont trouvées sur le génome (le génome étant plus grand que celui des animaux, ces multiples copies pourraient permettre que les brins soient synthétisés en même temps). Un schéma de la réplication est montré en Figure 1.5.

La localisation des origines A et B1 laisserait penser que seul le cercle maître peut se répliquer chez les plantes supérieures [de Haas et al., 1991].

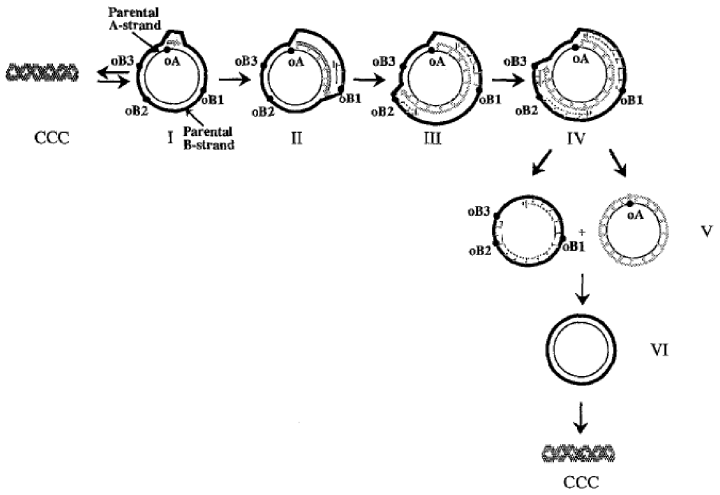


FIG. 1.5 – Représentation de la réplication chez *Petunia hybrida* [de Haas et al., 1991]. "ccc" représente plusieurs molécules d'ADN mitochondrial liées. O_A et O_B représentent les origines de réplication des brins A et B. (I) Initiation de la réplication du brin parental, (II) élongation, (III/IV) initiation et élongation du brin fils, (V) séparation des brins et complétion du brin B synthétisé et (VI) nouvelle molécule fille terminée.

Transcription et traduction

La transcription des gènes dans le génome mitochondrial des plantes est plus complexe que chez les animaux. En effet, les études menées sur les promoteurs de mitochondries de plantes monocotylédones et dicotylédones n'ont pas permis de trouver une séquence consensus communes entre les différentes lignées de plantes [Mackenzie and McIntosh, 1999]. De plus, dans un même organisme, il existe de nombreux sites d'initiation et de terminaison ainsi que des mécanismes post-transcriptionnels d'épissage et de clivage pour une même région codante du génome, aboutissant à des transcrits de tailles différentes pour un même gène [Gray et al., 1992]. Par exemple, il a été dénombré quinze sites d'initiation de la transcription chez *Oenothera berteriana* [Schuster and Brennicke, 1994]. Plusieurs ARN polymérases codées par le noyau semblent

nécessaires pour la transcription des mitochondries de plantes, étant donné que l'on trouve différents motifs d'initiation.

Dans les génomes mitochondriaux de plantes, il existe des gènes possédant des introns. Le nombre d'introns pour un même gène peut différer entre les espèces. Par exemple, chez *Zea mays* (monocotylédone), le gène *rps3* possède un intron alors qu'il n'y en a aucun pour ce même gène, chez *Beta vulgaris* (dicotylédones). Les plantes ne tendent pas forcément à perdre leurs introns au cours de l'évolution par *reverse transcription*. En effet, il a été montré que, sur trois introns (introns 3 et 4 de *cox2* et intron 4 de *nad1*) étudiés dans différentes plantes, ceux-ci seraient apparus au cours de l'évolution. Chez les plantes monocotylédones et les dicotylédones, les gènes *nad1*, *nad2* et *nad5* sont épissés en trans. La séparation des introns de ces gènes se serait donc produite par recombinaison avant la séparation de ces deux lignées [Schuster and Brennicke, 1994].

Une des particularités des génomes mitochondriaux des plantes est l'édition d'ARN. Il s'agit d'une modification post-transcriptionnelle, indépendante de la traduction, d'une cytosine (C) en uracile (U) (Figure 1.6). Plus rarement on peut avoir la transformation d'un U en C. Ce phénomène n'a pas été observé au niveau des algues vertes, ni des Bryophytes, cependant on le retrouve dans toutes les autres plantes terrestres. Plusieurs centaines de sites d'édition ont été détectés dans certaines espèces, notamment chez *Oenothera berteriana* et chez le blé. L'édition d'ARN joue plusieurs rôles [Bowe and dePamphilis, 1996]. Tout d'abord, elle permet de pallier les éventuelles mutations : un gène codant pour une protéine non fonctionnelle codera pour une protéine fonctionnelle si l'ARN messager est édité. Il a été montré expérimentalement que l'édition d'ARN produit des protéines fonctionnelles par rapport aux mêmes protéines non éditées. Elle permet également de préserver les acides aminés très conservés ainsi que les structures en boucles dans les séquences codantes pour des protéines. Elle joue également un rôle au niveau de la régulation des protéines. En effet, l'édition d'ARN crée le codon start de certains gènes. C'est le cas des gènes *atp6*, *nad1* et *nad4L* chez *Beta vulgaris* [Mower and Palmer, 2006]. L'édition peut également agir au niveau des codons stop (les gènes *atp6* et *atp9* de *Beta vulgaris* ont leurs codons stop créés par édition d'ARN). Les positions les plus visées par cette édition sont les premier et deuxième nucléotides d'un codon, c'est-à-dire les positions sur lesquelles on risque d'avoir des mutations synonymes [Bowe and dePamphilis, 1996]. L'édition d'ARN peut provoquer l'apparition de paralogues transformés. En effet, après transcription, on obtient un ARN messager, Celui-ci pouvant être édité puis réinséré dans le génome par *reverse transcription*. Les paralogues transformés peuvent également être insérés dans le génome d'un autre compartiment. Il est possible de distinguer un paralogue transformé de la copie originale si cette dernière possède des introns, le paralogue n'en aura pas puisqu'il a été réinséré dans le génome à partir de l'ARN messager. L'édition d'ARN n'aura donc pas d'effet sur les phylogénies si celles-ci sont effectuées à partir de l'ADN, à condition de rester attentif aux paralogues transformés [Bowe and dePamphilis, 1996].

1.2.3 Echanges génétiques entre les organites

Dans les cellules végétales, il existe des flux de matériel génétique entre les différents compartiments (noyau, chloroplaste, mitochondrie). La vision que nous avons des compartiments cellulaires bien délimités est erronée puisqu'en effet ceux-ci sont plus ou moins liés entre eux. Les flux de nucléotides entre les différents compartiments sont quasiment inexistants chez les animaux. Une des raisons principales pourrait s'expliquer par la différence de code génétique en-

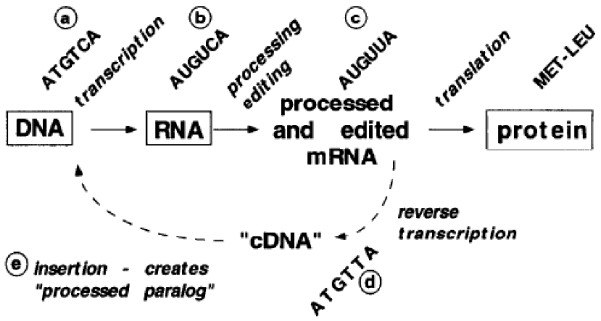


FIG. 1.6 – Edition d'ARN [Bowe and dePamphilis, 1996]. Après édition, la cytosine en (b) devient une uracile en (c). L'ARN messager édité peut alors être traduit en protéines ou réinséré dans le génome, créant un paralogue transformé, suite à une *reverse transcription*.

tre le noyau et les mitochondries chez les animaux, alors que les codes génétiques nucléaires et mitochondriaux sont identiques chez les végétaux. En effet, chez les animaux, si l'on a des échanges génétiques entre le noyau et la mitochondrie, les séquences codantes nucléaires intégrées dans le génome mitochondrial ne seront plus codantes à cause de la différence de code génétique. Ces séquences ne seront donc pas maintenues dans les mitochondries et seront éliminées. Dans les cellules végétales, les échanges de matériel génétique peuvent se faire entre le noyau, les chloroplastes et les mitochondries. Il existe des flux allant de la mitochondrie au noyau. L'intégration des gènes mitochondriaux dans le noyau se ferait à partir d'ARN messagers, dès lors que l'épissage et l'édition ont eu lieu, l'insertion de la séquence peut se faire dans le génome nucléaire [Schuster and Brennicke, 1994]. Lorsqu'un gène est inséré dans le noyau, la copie mitochondriale subit alors un fort taux de mutation μ et peut devenir un pseudogène [Blanchard and Lynch, 2000]. On peut également avoir des flux du noyau vers les mitochondries. La plupart des séquences nucléiques insérées dans les mitochondries ne sont pas fonctionnelles. Ces séquences intégrées sont majoritairement sous forme de rétrotransposons. Il existe également des insertions de séquences chloroplastiques dans les mitochondries chez les plantes terrestres. Exceptionnellement, aucune insertion chloroplastique n'est retrouvée dans le génome de *Marchantia polymorpha* [Schuster and Brennicke, 1994]. En général, les séquences de gènes chloroplastiques insérées ne sont pas codantes. Une étude des échanges génétiques a été menée chez le riz [Notsu et al., 2002], en comparant les génomes mitochondriaux, nucléaire et chloroplastiques à ceux de *Marchantia polymorpha*. Les transferts suivants ont été observés (Figure 1.7) :

- 5 gènes codant pour des protéines ribosomiques, 6 ARNt et 6 fragments de taille supérieure à 300 pb ont été transférés des mitochondries au noyau (1),
- 19 séquences montrant des homologies avec des rétrotransposons et transposons ont été transférés du noyau vers les mitochondries, signes de transfert du noyau vers les mitochondries (2),
- 37 fragments de plus de 300 pb ont été transférés entre le noyau et les mitochondries (3),
- 3 fragments chloroplastiques ont été transférés dans le noyau et les mitochondries (4),
- 3 fragments chloroplastiques ont été transférés dans les mitochondries puis dans le noyau (5),
- 17 fragments chloroplastiques ont été transférés vers les mitochondries (6).

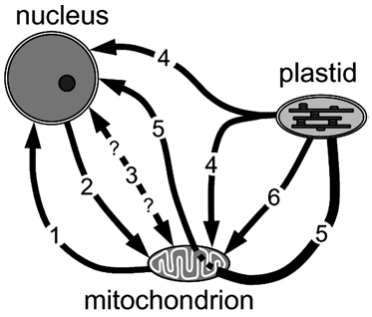


FIG. 1.7 – Exemple de transferts de gènes entre organites observés chez le riz [Notsu et al., 2002].

Taux de substitutions

Le taux de substitutions synonymes est trois fois plus élevé dans le génome chloroplastique que dans le génome mitochondrial et l'est encore plus dans le génome nucléaire. Une estimation du taux de substitutions synonymes chez les plantes a montré un taux de substitution de 0.2 à 1 $\times 10^{-9}$, 1 à 3 $\times 10^{-9}$ et 5 à 30 $\times 10^{-9}$ substitutions par site et par année dans, respectivement, les mitochondries, les chloroplastes et le noyau [Wolfe et al., 1987].

1.3 Comparaison des génomes animaux et végétaux

Nous venons de le voir, les mitochondries animales et végétales ont des caractéristiques assez différentes. Par exemple, la taille des génomes des mitochondries animales est en général inférieure à 20 kpb. Chez les végétaux, la taille est en général au moins dix fois supérieure. Le code génétique utilisé diffère également entre ces deux groupes. Chez les animaux il est modifié par rapport au code génétique standard qui est celui utilisé dans les mitochondries végétales. Nous pouvons également noter des différences au niveau des taux de mutation μ. En effet, une estimation du taux de mutation μ, basé sur les mutations silencieuses, montre des différences dans les groupes phylogénétiques majeurs [Lynch, 2007]. Chez les mammifères, on obtient un taux de 34 $\times 10^{-9}$ par site par année alors qu'il n'est que de 0.36 $\times$ 10^{-9} par site par année chez les plantes. Le taux de mutations silencieuses est ainsi 19 fois plus élevé que dans le noyau pour les vertébrés et 8 fois plus élevé que dans le noyau pour les invertébrés (sauf coraux) alors qu'il ne correspond qu'à 5% de celui du noyau pour les plantes. Le taux de substitution dans les mitochondries végétales serait cent fois inférieur à celui des animaux [Albert et al., 1996]. Nous avons également vu que chez les plantes, les ARN messagers sont édités et qu'il existe des transferts de matériel génétique entre les différents compartiments de la cellule. Les génomes mitochondriaux de plantes possèdent également des gènes codant pour des protéines ribosomiques et des introns, ce que l'on ne retrouve pas du côté des animaux. Les principales différences entre ces génomes sont résumées dans le Tableau 1.1.

Ces différences peuvent s'expliquer par l'hypothèse de pression de mutation [Lynch et al., 2006]. En effet, il a été montré qu'il existe une corrélation négative entre la taille des génomes mitochondriaux et leur taux de substitution. Plus le taux de substitution est élevé, plus les séquences dupliquées ou insérées seront rapidement mutées puis éliminées par la sélection. Au contraire, un génome avec un faible taux de substitution est plus facilement envahi de séquences qui s'accumulent de manière neutre. Dans ces génomes, parmi les séquences insérées, on peut avoir des séquences dupliquées, qui pourront causer des événements de

	Animaux	Végétaux
Taille du génome	compacte	large
Introns	non connu	oui
Séquences intergéniques	très petites	très grandes
Code génétique	modifié	standard
Gènes codant pour des protéines ribosomiques	aucun	plusieurs
ORF	non connu	oui
Taux de mutation μ	élevé	faible
Ordre des gènes	conservé	très modifié
Edition d'ARN	non connu	existant

TAB. 1.1 – Comparaison entre les différentes caractéristiques des génomes mitochondriaux animaux et végétaux.

réarrangements.

Cependant, au fil des années, de plus en plus de génomes mitochondriaux sont séquencés et, chez certaines espèces, on trouve des génomes mitochondriaux montrant des spécificités propres aux génomes animaux et végétaux, confirmant l'origine commune des mitochondries. Quelques espèces sont répertoriées dans le Tableau 1.2.

Organisme	Taille (pb)	Nb de gènes	Nb d' ARNt	Nb d' ARNr	Nb de protéines ribosomiques	Autre	Référence
Chondrus crispus	25836	13	23	3	0	6 ORF potentielles 4 ORF communes aux plantes 2 gènes nucléaires 4 gènes chloroplastiques 1 intron	[Leblanc et al., 1995]
Trichoplax adhaerens	43079	12	24	2	0	3 ORF introns (*cox1* et *rrnL*) épissage trans (*cox1*)	[Signorovitch et al., 2007]
Acanthamoeba castellanii	41591	17	16	2	16	8 ORF gènes superposés	[Smith et al., 2010]
Dunaliella salina	28300	7	3	2	0	18 introns séquences intergéniques	[Signorovitch et al., 2007]
Reclinomonas americana	69034	26	26	3	18	introns séquences nucléaires dans les autres espèces	[Lang et al., 1997]

TAB. 1.2 – Contenu de quelques génomes mitochondriaux présentant des caractéristiques des génomes mitochondriaux animaux et végétaux. Par exemple, une taille supérieure à 30 Kpb, la présence de protéines ribosomiques et d'introns sont des caractéristiques végétagles, tandis qu'une taille plus petite, la présence de gènes chevauchants ainsi que l'absence de protéines ribosomiques sont des caractéristiques animales.

Par exemple, chez l'algue rouge *Chondrus crispus*, organisme proche de la base des Eucaryotes, la taille du génome (25 kpb) et le code génétique (modifié) correspondent aux caractéristiques mitochondriales des animaux. Cependant, ses ARNr et ses protéines sont phylogénétiquement plus proches des plantes. Cette espèce possède également des ORF et un intron (caractéristiques des plantes). De plus, sur le génome de cet organisme, les gènes et ORF sont regroupés en fonction du brin sur lequel ils se trouvent, on a donc eu ici des événements de recombinaison pour rassembler les gènes. Ce regroupement suggère l'existence de deux unités de transcription [Leblanc et al., 1995].

Le génome mitochondrial de *Trichoplax adhaerens*, espèce appartenant aux Placozoaires (animaux qui présentent le plan d'organisation le plus simple, il s'agit en quelque sorte de couches de cellules empilées, capables de se mouvoir), a également été séquencé [Signorovitch et al., 2007]. Comme pour les algues rouges, cette espèce montre des caractéristiques animales car son génome ne contient pas de gènes de protéines ribosomiques et son code génétique est modifié. Cependant, elle possède des caractéristiques des végétaux avec un génome de 43079 pb et surtout des régions espacées intergéniques, des ORF et des introns. *Trichoplax adhaerens* a été comparé à trois autres Placozoaires, l'analyse phylogénétique sur douze gènes de la chaîne respiratoire supportant la monophylie des Placozoaires. De ce fait, une hypothèse est émise selon laquelle la réduction de taille du génome mitochondrial serait apparue après l'émergence des animaux [Signorovitch et al., 2007]. Les quatre espèces de Placozoaires possèdent le même nombre de gènes, ARNt, ARNr et introns, mais la structure de leur génome est différente : on peut observer des inversions et des translocations (un fragment du génome est excisé et réinséré à une autre position).

Acanthamoeba castellanii, organisme positionné à la base des animaux, fungis et plantes, possède également les doubles caractéristiques des mitochondries animales et végétales [Burger et al., 1995]. En effet, le code génétique de cette espèce est modifié mais il est différent de celui des animaux (UGA = tryptophane) et on trouve des gènes codant pour des protéines ribosomiques. Dans ce génome, des événements de réarrangements ont été observés puisque les gènes et ORF sont compactés sur le même brin. Cette espèce a été comparée à deux autres espèces (*Prototheca wickerhamii* et *Chlamydomonas reinhardtii* ayant respectivement des génomes de 55,326 pb et 15,758 pb) appartenant au groupe des algues (Chlorophyta). *Acanthamoeba castellanii* et *Prototheca wickerhamii* ont le même contenu en protéines ribosomiques et en protéines de la chaîne respiratoire que les plantes. Au contraire, *Chlamydomonas reinhardtii* a des gènes de la chaîne respiratoire manquants et n'a pas de gène codant pour des protéines ribosomiques. De plus, seule *Acanthamoeba castellanii* contient le gène *atp9* présent chez les fungis. Par contre, elle ne possède pas le gène *atp8* présent chez les fungis et vertébrés. Comme chez les animaux, cette espèce contient des gènes qui se superposent.

Toujours dans le clade des Chlorophyta, nous pouvons ajouter *Dunaliella salina* dont le génome mitochondrial a récemment été séquencé [Smith et al., 2010]. La taille du génome reste plus proche de celle des animaux (28 kpb) et son contenu en gènes est restreint. Cependant, son génome contient 29% de séquences intergéniques et dix-huit introns potentiels. On peut également noter que tous les gènes codants sont regroupés sur le même brin.

Enfin, le génome mitochondrial de *Reclinomonas americana*, décrit comme étant un des génomes les plus proches du génome mitochondrial ancestral des Eucaryotes, a été étudié [Lang et al., 1997]. *Reclinomonas americana* contient des gènes non retrouvés dans d'autres génomes, suggérant que certains gènes codant pour des protéines n'ont pas encore disparu de la mitochondrie : le gène *atp3*, un gène intervenant dans complexe III, 9 gènes codant pour des protéines ribosomiques, un facteur de traduction, une protéine de sécrétion, une protéine

putative d'assemblage des cytochromes oxydase et quatre composants de l'*ARN-polymérase*. Retrouver cette *ARN-polymérase*, de type eubactérienne et codée par le noyau dans les autres espèces, supporte l'origine endosymbiotique des mitochondries. La perte de cette polymérase se serait donc faite très tôt lors de l'évolution des mitochondries. Le génome mitochondrial de *Reclinomonas americana* possède 8% de séquences intergéniques et un intron. On y trouve également l'ARNr 5S retrouvé, pour le moment, chez les plantes et *Prototheca wickerhamii*. Le code génétique utilisé est le code génétique standard. L'étude de l'organisation des protéines ribosomiques montre une structure identique chez *Reclinomonas americana*, *Marchantia polymorpha*, *Acanthamoeba castellanii* et l'opéron α d'*Escherichia coli*. De plus, *Reclinomonas americana* a des gènes codant pour des protéines ribosomiques que n'ont plus *Marchantia polymorpha* (division des Streptophytes) et *Acanthamoeba castellanii*.

L'étude de génomes mitochondriaux proches de la base des Eucaryotes confirme donc cette hypothèse d'origine monophylétique des mitochondries, malgré les caractéristiques spécifiques des mitochondries animales et végétales. Les codes génétiques utilisés entre ces mitochondries sont différents. Il a été suggéré que le code génétique standard est vraisemblablement le code génétique des mitochondries ancestrales, mais celui-ci évolue sous l'influence d'un biais de mutations généré pendant la réplication [Osawa et al., 1992]. Chez les animaux, les mitochondries ont évoluées vers une réduction du génome (génome de petite taille ne contenant quasiment que des séquences codantes). Le génome mitochondrial des plantes a une tendance à accumuler des séquences non codantes augmentant ainsi considérablement sa taille par rapport à celui des animaux. Avec cette accumulation se séquences, on trouve des séquences dupliquées dans ces génomes engendrant des réarrangements chromosomiques.

1.4 Réarrangements

Si les réarrangements dans les génomes mitochondriaux végétaux sont nombreux, pour le moment, aucun modèle de réarrangement du génome mitochondrial n'a été proposé, les réarrangements étant décrits comme très nombreux et difficiles à analyser (les réarrangements sont visibles au niveau intraspécifique). Pendant plusieurs années, il a été considéré que les génomes animaux ne se réarrangeaient quasiment pas, c'est pourquoi les génomes mitochondriaux sont souvent utilisés pour des analyses phylogénétiques. En effet, Boore et Brown ont montré que généralement, chez les animaux, l'ordre des gènes est le même dans un phylum mais sera différent entre les phylums [Boore and Brown, 1998]. Certains nœuds des phylogénies peuvent alors être résolus à l'aide des réarrangements établis entre différentes espèces (par exemple la construction du phylum des Lophotrochozoan [Vallès and Boore, 2005]). Cependant, chez certains animaux, on peut trouver de très forts taux de réarrangements, souvent décrits comme étant des inversions ou des transpositions [Inoue et al., 2003]. Pour Dowton et Campbell, les causes des réarrangements chez les parasites sont dues au stress oxydatif imposé par la réponse immunitaire de l'hôte [Dowton and Campbell, 2001]. La Figure 1.8 répertorie les différentes structures de génomes observées chez les animaux.

Dans cette partie, nous feront une synthèse non exhaustive des différents types de réarrangement proposés et observés chez les animaux. Les types de réarrangement décrits sont souvent des duplications avec pertes de gènes mais on peut également trouver des inversions de séquence ou encore des translocations.

Certains réarrangements observés dans les génomes mitochondriaux des animaux ont été

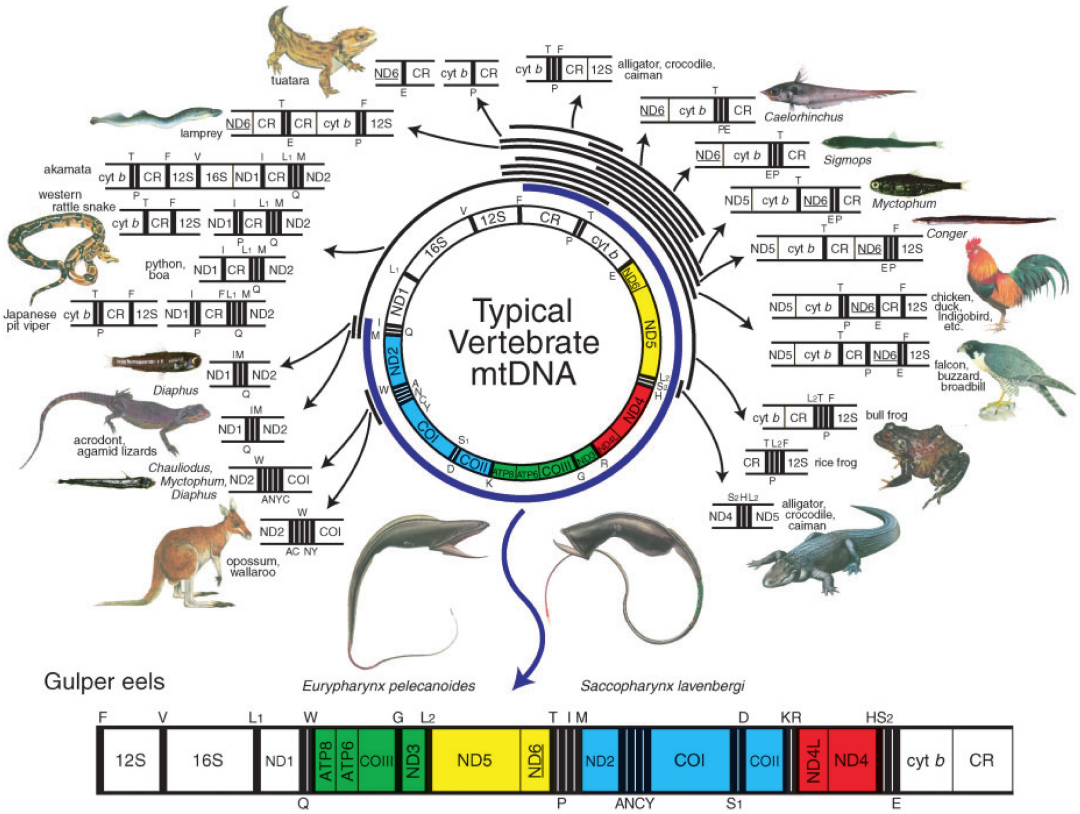


FIG. 1.8 – Différents réarrangements connus chez les animaux [Inoue et al., 2003]. Les génomes sont circulaires mais, pour chaque espèce, la région dont la position des gènes et ARN varie est représentée linéairement. La différence de structure entre le génome mitochondrial typique des vertébrés et celui des *Gulper eels* est expliqué par une duplication avec perte de gènes. Les groupes de gènes conservés entre ces deux génomes ont été colorés.

décrits comme étant une duplication totale du génome avec des pertes non aléatoires des gènes. Par exemple, chez les Arthropodes, les génomes de deux Millipedes ont été séquencés [Lavrov et al., 2002]. *Narceus annularus* (14868 pb) et *Thyropygus sp.* (15133 pb) contiennent 37 gènes et présentent tous les deux le même arrangement. Par contre, cet arrangement est différent de celui trouvé chez les autres arthropodes séquencés. Sur ces deux génomes, les gènes dans le même sens de transcription sont tous rassemblés. On a ainsi deux groupes de gènes, chacun contenant tous ses gènes dans le même sens de transcription, à l'exception d'un ARNt qui est dans le sens opposé aux autres gènes de son groupe. Les deux groupes de gènes sont séparés par une région non codante. Cette disposition des gènes peut s'expliquer par une duplication totale du génome ancestral, ce qui équivaut à une liaison de deux molécules d'ADN. De telles structures ont déjà été observées dans des tissus anormaux mais aussi dans des tissus normaux (par exemple chez le rat). Les promoteurs de la transcription se retrouvent dans les régions non codantes séparant les deux groupes de gènes. Des mutations sur une des copies des promoteurs entraînera la perte progressive des gènes sous l'influence de ce promoteur. Le perte des gènes suite à cette duplication en tandem est donc influencée par la polarité de ces gènes. La seule exception est l'ARNt resté dans le mauvais sens dans un des groupes. Il a été montré que cet ARNt est produit en faible quantité et utilisé à moins de 1% dans les acides aminés. Le

séquençage d'un autre Millipede, *Antrokoreana gracilipes* (14,747 pb) confirme cette duplication totale du génome ancestrale de Millipedes [Woo et al., 2007]. Cette structure des génomes en deux groupes de gènes rassemblés selon leur sens de transcription avait également été observée chez *Chondrus crispus* [Leblanc et al., 1995].

Dans un certain nombre d'espèces, des duplications en tandem avec des pertes aléatoires des gènes ont été décrites. Une étude portant sur la comparaison de trois génomes de mollusques (*Crassostrea honkongensis*, *Crassostrea gigas* et *Crassostrea virginica*) explique les réarrangements des gènes entre ces génomes par une duplication en tandem couplée à des recombinaisons intramoléculaires et des transpositions [Yu et al., 2008]. Chez les Geckos, l'analyse de onze génomes a montré que les différences dans l'ordre des gènes provenaient de larges duplications en tandem (environ 10 kpb). De plus, dans ces génomes, les évènements de duplication en tandem se situent dans les mêmes régions mais sont indépendants entre les génomes. Il a été observé que les gènes dupliqués sont très vite éliminés, par contre certaines copies d'ARNt et d'ARNr sont conservées [Fujita et al., 2007]. Chez les poissons, l'analyse de deux génomes, *Eurypharynx pelecanoides* (Eurypharyngidae) et *Saccopharynx lavenbergi* (Saccopharyngidae), révèle la duplication d'une partie du génome (environ 12 kpb) avec pertes de gènes. Même si l'arrangement des gènes dans ces deux espèces est identique, il reste différent de celui des autres poissons, laissant penser à une duplication ancestrale pour ces deux espèces [Inoue et al., 2003]. L'analyse de six génomes de Salamandres a également montré l'existence de duplications en tandem encore visibles car il reste des traces de pseudogènes [Mueller and Boore, 2005]. Des réarrangements par duplication en tandem ont également été décrits chez les gastéropodes [Grande et al., 2008], les insectes [Carapelli et al., 2006], les échinodermes [Perseke et al., 2008], les amphibiens [Mauro et al., 2006], les oiseaux [Cho et al., 2009] et également les crabes [Sun et al., 2005].

D'autres types de réarrangement ont été décrits chez les animaux tels que des duplications ([Vallès and Boore, 2005, Shao and Barker, 2003]), des transpositions ([Vallès and Boore, 2005, Campbell and Barker, 1999, Shao and Barker, 2003]) et des inversions ([Sun et al., 2005, Shao and Barker, 2003]).

1.5 Conclusion

Dans ce chapitre, nous avons vu que les mitochondries des Eucaryotes ont une origine commune. Des organismes ayant une place ancestrale entre les animaux et les végétaux présentent des caractéristiques propres à ces deux groupes. La Figure 1.9 rassemble les caractéristiques mitochondriales de quelques espèces citées dans ce chapitre. Au cours de l'évolution, des différences sont apparues entre les mitochondries animales et végétales. Une des principales différences entre ces génomes est le taux de mutation μ, plus élevé chez les animaux. Une étude ayant montré que le code génétique évoluait sous l'influence des mutations au cours de la réplication [Osawa et al., 1992], il semble alors logique que le code génétique chez les animaux soit modifié. De plus, le taux de réarrangement est beaucoup plus élevé chez les plantes. Même si nous avons décrit plusieurs études montrant des réarrangements chez les animaux, la majorité des groupes d'espèces conservent une même structure de leur génome, tandis que chez les végétaux, on trouvera des réarrangements à un niveau intraspécifique (par exemple chez la betterave [Satoh et al., 2004]). Le taux de mutation μ pourrait expliquer cette différence. En effet, Lynch et ses collaborateurs montrent une relation entre le taux de mutation μ et la compaction des génomes [Lynch et al., 2006] : le taux de mutation μ est inversement proportionnel à la taille du

génome, sous l'hypothèse d'une taille efficace de population égale. La différence de taille entre les génomes ainsi que le taux de réarrangement plus élevé peuvent alors s'expliquer. En effet, si une duplication ou une insertion de séquence apparaît dans les génomes des animaux celle-ci sera très vite éliminée car le taux de mutation μ est élevé. Ceci explique également pourquoi lors de duplications chez les animaux, il ne reste généralement qu'une copie fonctionnelle des gènes dupliqués. Chez les plantes, le taux de mutation μ étant très faible, les séquences insérées dans le génome ne seront pas éliminées. Ces insertions de séquences peuvent entraîner l'apparition de séquences dupliquées pouvant mener à des réarrangements. Les réarrangements décrits chez les animaux sont généralement la conséquence de mésappariements lors de la réplication ou de recombinaisons intra-moléculaires ([Campbell and Barker, 1999, Mueller and Boore, 2005]). Chez les animaux, l'évolution des génomes dépendra principalement des mutations. Au contraire, chez les végétaux, l'évolution des génomes dépendra essentiellement des réarrangements. Si l'on veut étudier l'évolution des génomes mitochondriaux des végétaux, il faut avant tout s'intéresser aux réarrangements qui s'y produisent.

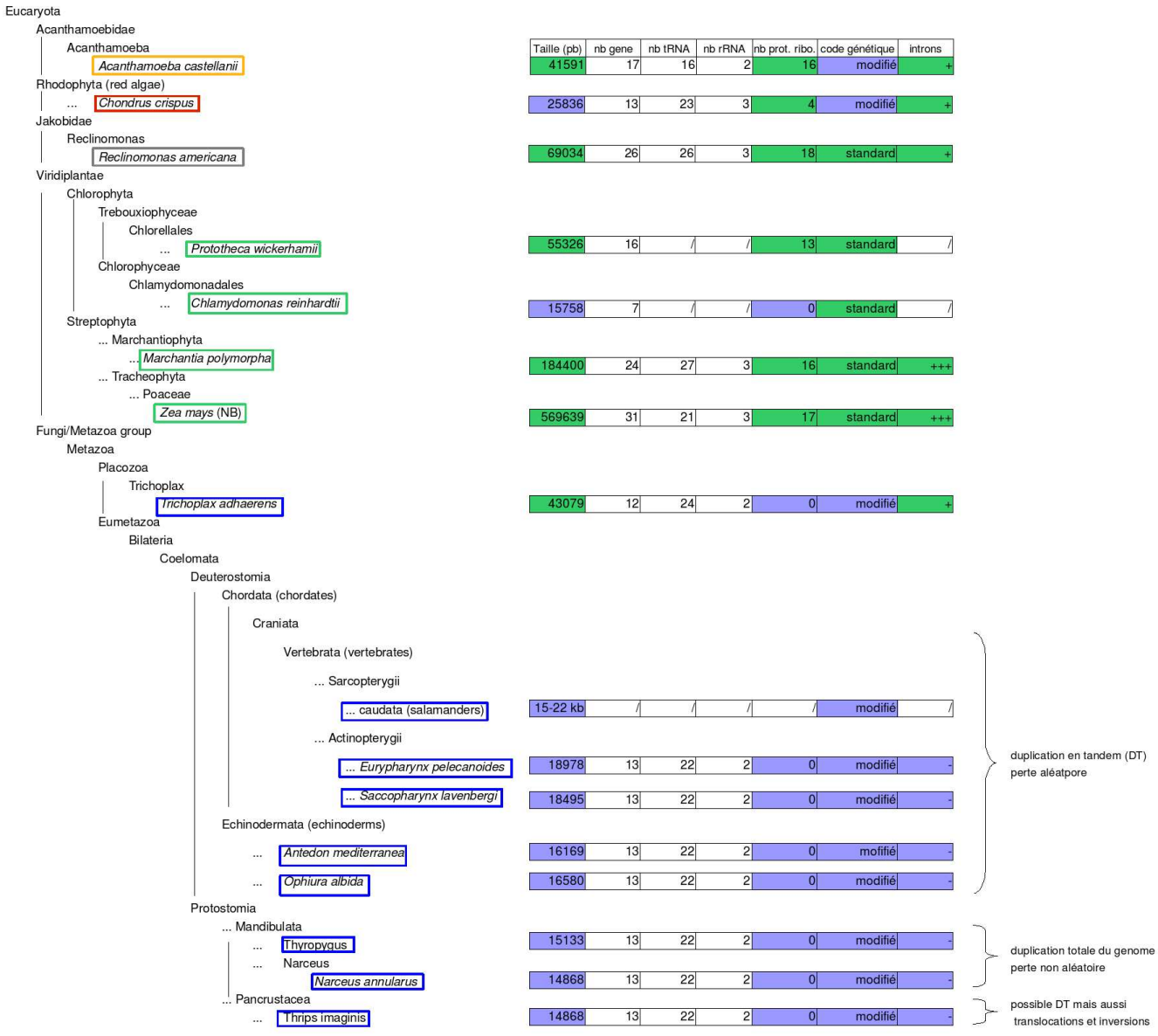


FIG. 1.9 – Résumé de données de génomes mitochondriaux. Sont Encadrées en bleu, les espèces appartenant au groupe des animaux, en vert celles appartenant au groupe des végétaux, en orange une espèce de protozoaire pathogène, en rouge une espèce d'algue rouge et en gris une espèce de protozoaire se nourrissant de bactéries. Les caractéristiques des mitochondries animales sont surlignées en bleu, celles des mitochondries végétales en vert.

Chapitre 2

Aperçu des méthodes d'analyse de réarrangements génomiques

Les réarrangements génomiques désignent l'ensemble des modifications s'opérant sur l'ordre des gènes d'un génome. Ces remaniements sont connus depuis longtemps [Dobzhansky and Sturtevant, 1938] et ont été observés chez tous les types d'organismes, aussi bien dans les génomes nucléaires que les génomes des organites comme nous l'avons vu au chapitre précédent. L'étude de ces remaniements permet de répondre à plusieurs questions.

Tout d'abord, la distance de réarrangements (le nombre de remaniements qui se sont déroulés au cours de l'évolution) fournit une méthode supplémentaire pour reconstruire une phylogénie des espèces [Boore and Brown, 1998, Blanchette et al., 1999, Snel et al., 1999, Suyama and Bork, 2001, Wang et al., 2002]. Cela peut s'avérer particulièrement utile pour les cas où la conservation des gènes est très importante comme c'est le cas pour les génomes mitochondriaux des plantes.

L'obtention d'une histoire des réarrangements, des scénarios évolutifs, donne également de précieuses informations sur la coopération entre gènes, ce qui est important du point de vue fonctionnel.

Dans ce chapitre, nous débuterons par une description des outils permettant d'identifier des régions dupliquées dans les génomes et d'étudier les remaniements chromosomiques. Puis nous retracerons une brève histoire des méthodes développées au cours des quinze dernières années, permettant la reconstruction de scénarios de réarrangements et la reconstruction phylogénétique. Nous nous concentrerons plus particulièrement sur les méthodes utilisées dans le cadre de cette thèse.

2.1 La comparaison de génomes complets

La comparaison de deux génomes n'est pas aussi simple à appréhender que la comparaison de deux séquences homologues. En effet, les génomes ne peuvent pas être considérés comme des suites linéaires de symboles et ce pour plusieurs raisons. Les réarrangements réorganisent la séquence. Les transferts horizontaux sont aussi un facteur de réorganisation. Les duplications ou les répétitions en sont un autre. Ceci fait qu'aligner des génomes revient à aligner un nombre maximum de segments similaires des génomes. On peut interpréter cela comme une combinaison d'alignements locaux.

2.1.1 Aligner des génomes complets

L'intérêt pour la comparaison de très longues séquences ou de génomes complet est apparu à la fin des années 1990 ou au début des années 2000 avec des méthodes telles que MUMmer [Delcher et al., 1999] ou LAGAN [Brudno et al., 2003a]. Ces méthodes sont basées sur l'idée que pour comparer des séquences très longues, il ne faut pas essayer d'obtenir un alignement global complet. En effet les temps de calcul très importants des techniques usuelles (type Needleman et Wunsh [Needleman and Wunsch, 1970]) sont rédhibitoires et les mutations accumulées au cours du temps, sur les segments soumis à moins de contraintes, empêcheraient d'obtenir un alignement satisfaisant. Ces méthodes proposent alors de travailler à partir d'ancres, correspondant à des segments communs très bien conservés, puis de les combiner en proposant ou non d'aligner les parties entre les ancres. Si ce type de méthode fonctionne bien pour la comparaison de génomes dont l'architecture n'a pas été modifiée, elles ne sont plus du tout adaptées dans le cas de génomes réarrangés.

Les méthodes permettant de comparer des génomes complets ayant subi des réarrangements sont peu nombreuses. Une évolution de LAGAN, nommé Shuffle-LAGAN [Brudno et al., 2003b], a été proposée dans ce sens. Shuffle-LAGAN utilise un algorithme d'alignement local (CHAOS) des séquences afin de retrouver des similarités entre les régions des séquences couplé à un algorithme d'alignement global (LAGAN) permettant d'identifier les transformations entre deux séquences. Un problème de Shuffle-LAGAN est qu'il ne permet d'effectuer des alignements de génomes complets que par paires de génomes. Un autre outil, Mauve, discuté ci-dessous, a été proposé en 2004 et est toujours maintenu et actualisé aujourd'hui. On peut également citer MAGIC [Swidan et al., 2006], apparu en 2006, qui est un concurrent de Mauve. MAGIC permet de réaliser des alignements de génomes complets réarrangés. Cependant, il ne fait que des comparaisons par paires de génomes. Or nous cherchons un outil capable de traiter plus de deux génomes simultanément. De plus MAGIC requiert un fichier d'ancres (généralement des gènes) pour aider à la recherche de régions similaires. Dans une récente revue [Swidan and Shamir, 2009], la comparaisons de ces outils dans le cadre de l'analyse de génomes bactériens est discutée. L'étude consiste à comparer les performances de Mauve et MAGIC selon deux mesures : une mesure basée sur le nombre de gènes disruptés lorsque les régions communes sont définies et une mesure sur le nombre de segments créés et leur conservation. Les génomes de 41 espèces bactériennes ont été étudiés. Les résultats montrent que MAGIC est plus performant que Mauve sur ces deux mesures. Afin d'éviter de favoriser MAGIC sur la mesure de disruption de gènes, les ancres données en entrée étaient celles calculées par Mauve et non par des annotations de gènes. Par contre, le temps de calcul de MAGIC est plus long que celui de Mauve et MAGIC ne permet pas la comparaison multiple.

GRIMM-synteny [Pevzner and Tesler, 2003a] propose une stratégie un peu différente permettant de comparer des génomes ayant des similarités moins marquées car il prend en compte des micro-réarrangements. L'idée de la méthode est d'être capable d'extraire des blocs synténiques. Un bloc synténique est constitué d'un ensemble d'ancres réarrangées les unes par rapport aux autres mais tout en restant dans une fenêtre de faible taille sur chaque génome. L'outil offre ainsi à la fois une vision microscopique en observant les arrangements des ancres formant chaque bloc synténique et une vision macroscopique en observant les arrangements des blocs synténiques eux-mêmes. Pour utiliser GRIMM-Synteny il faut déjà disposer des ancres et il est donc nécessaire d'utiliser des programmes tels que ceux discutés ci-dessus pour les produire. Il est possible de donner des marqueurs dupliqués en entrée de GRIMM-synteny. Cependant, celui-ci génère au final des permutations sans duplicats en intégrant les duplicats aux autres marqueurs sur des

critères de synténie conservées entre les génomes. Ces regroupements sont peu efficaces lorsque les régions autours des duplicats ne sont pas bien conservées entre les génomes.

2.1.2 L'outil Mauve

Mauve [Darling et al., 2004] est un outil conçu pour réaliser l'alignement multiple de génomes en autorisant des opérations de réarrangements. L'outil est capable de retrouver un ensemble de segments conservés communs à tout ou partie de l'ensemble des génomes donnés en entrée. Les segments identifiés ne sont pas nécessairement dans le même ordre dans tous les génomes.

La stratégie d'alignement développée par Mauve est basée sur l'identification des segments conservés entre les génomes, les multi-MUMs (multiple Maximal Unique Matches), et sur l'utilisation d'un arbre guide calculé par Mauve pour trouver une combinaison non chevauchante de ces segments.

Les MUMs sont des segments communs à au moins deux génomes de l'ensemble de départ. Ils ont la propriété d'être uniques, c'est-à-dire qu'il n'existe qu'une copie de ce segment dans chaque génome, et sont maximaux, c'est-à-dire que les nucléotides en amont et en aval du MUM sont différents entre tous les génomes. Ces multi-MUMs servent de base à la construction d'ancres qui sont constituées de suites colinéaires de multi-MUMs, conservées entre tous les génomes. Pour finaliser l'alignement, un calcul d'alignement multiple progressif classique, utilisant un arbre guide construit à partir des MUMs, est réalisé sur les régions entre les suites colinéaires de multi-MUMs.

Le principal paramètre de la méthode est la taille minimale des MUMs à détecter.

Adaptation au cas des génomes mitochondriaux de plantes. Mauve est un logiciel très pratique et très utilisé. Cependant, on aura remarqué que l'utilisation des MUMs empêche l'étude de génomes comportant des duplications. Cette difficulté peut être contournée lorsque les duplications sont extrêmement bien conservées.

Il est assez facile dans les génomes que nous avons étudiés de masquer toutes sauf l'une des occurrences de chacune des régions dupliquées. On peut alors utiliser Mauve pour extraire les régions conservées puis réaligner ces régions avec les parties masquées.

Une illustration du résultat du calcul de segments conservés par Mauve sur les données de betterave est donnée Figure 2.1.

2.1.3 Étudier les remaniements chomosomiques

Un autre point de vue consiste à s'abstraire de la séquence pour comparer les génomes à un niveau plus élevé. On n'observe plus la ressemblance de séquence mais on observe l'arrangement des séquences, identifiées comme similaires, entre les génomes.

Détecter les marqueurs. Avant d'entreprendre la comparaison des réarrangements dans les génomes, il est nécessaire d'identifier les séquences similaires. De manière générale, nous parlerons de *marqueur* pour les désigner. Effectivement, les études de remaniements chromosomiques se font sur des génomes représentés sous forme de suite de marqueurs. Un marqueur est donc défini comme un segment d'un génome. Deux marqueurs seront dit orthologues si on considère qu'une occurrence était présente dans l'ancêtre commun. La plupart du temps, les marqueurs sont les gènes. Parfois cela est suffisant si l'étude porte par exemple sur des génomes possédant peu de séquences inter-géniques. Parfois cela n'est pas suffisant, ou plutôt incomplet, comme nous

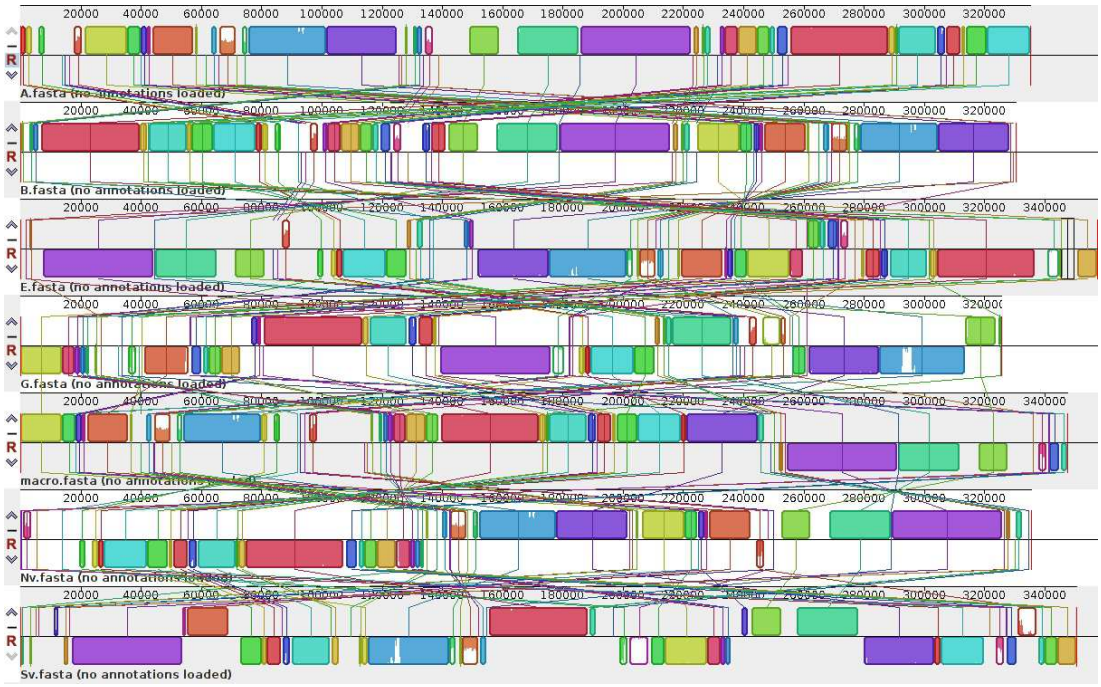


FIG. 2.1 – Alignement produit par Mauve sur les huit génomes de *Beta vulgaris* (voir Chapitre 6). Les parties dupliquées ont été supprimées avant de soumettre les séquences à Mauve. Les régions colorées entre chaque génomes (un génome par ligne) représentent les fragments homologues.

pourrons le voir dans le Chapitre 4. On est alors amené à considérer des segments de génomes plus longs, portant également sur des séquences inter-géniques.

De plus, si des suites identiques de marqueurs identifiés entre plusieurs génomes forment des régions synténiques, on pourra alors les réduire à un seul marqueur étant donné que cette suite de marqueurs est ordonnée de la même manière dans tous les génomes étudiés.

Dans nos analyses, un marqueur représentera donc la position d'une séquence (codante ou non) conservée entre les génomes comparés. Un exemple de transformation de génomes en suites de marqueurs est présenté Figure 2.2. Il faudra donc dans un premier temps retrouver les marqueurs homologues. Un numéro sera alors attribué à chaque groupe d'homologues, on obtiendra ainsi des suites de marqueurs.

Selon les méthodes, les génomes seront vus soit comme des *permutations*, soit comme des *séquences*. Selon la formalisation mathématique, les permutations sont des suites de marqueurs ne contenant pas de marqueur dupliqué et dont l'ensemble des marqueurs est identique entre plusieurs permutations. Lorsque l'on compare deux génomes, une des deux permutations est représentée sous forme de permutation identité, c'est-à-dire que la suite de marqueurs va de 1 à n avec n le nombre de marqueurs composant les permutations. Par exemple, dans la Figure 2.2, le génome A est la permutation identité. Des signes (+ ou -) en fonction de l'orientation des marqueurs sur le génome peuvent être attribués. Dans ce cas on parlera de permutation signée. Lorsque l'on regarde des permutations à partir de génomes contenant des marqueurs dupliqués, il est possible d'éliminer les dupliqués de deux façons : soit toutes les copies des dupliqués sont éliminées dans tous les génomes, soit on garde une copie de chaque dupliqué dans chacun des génomes. Les séquences quant à elles permettent d'inclure des marqueurs dupliqués.

L'identification des marqueurs, lorsqu'il ne s'agit pas des gènes, peut être réalisée grâce

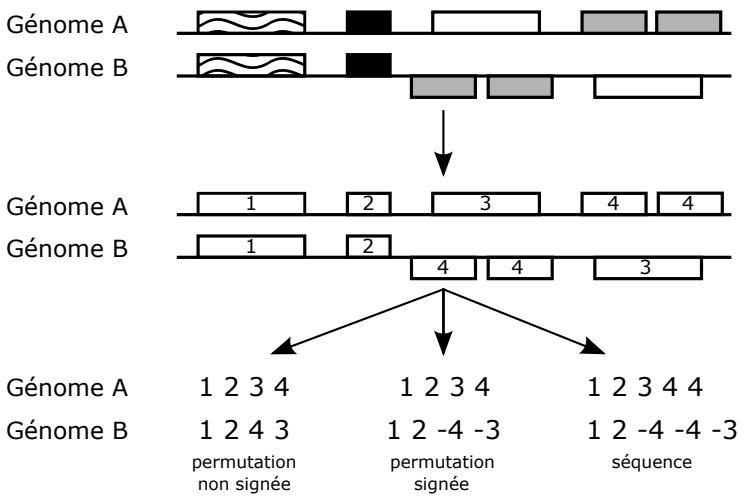


FIG. 2.2 – Transformation de génomes en suite de gènes. Les gènes homologues sont recherchés entre plusieurs génomes et sont ensuite transformés en suites de numéros (correspondant aux homologies), représentant soit des permutations (une copie des dupliqués), soit des séquences (tous les dupliqués).

aux méthodes que nous avons vu ci-dessus. Nous précisons ici que nous sommes dans un cadre où nous cherchons des marqueurs orthologues pour résoudre une histoire de réarrangement. D'autres méthodes utilisent la détection de réarrangement pour identifier des séquences orthologues [Fu et al., 2007, Ma et al., 2006].

De nombreux événements possibles. Une fois les marqueurs identifiés, la comparaison de deux génomes est réalisée grâce à l'identification d'une suite d'événements permettant la transformation de l'un à l'autre. Les événements de réarrangement qui ont été envisagés sont les suivants :

- l'inversion : fragment inversé,
- la transposition : fragment déplacé dans le chromosome,
- la délétion : fragment délété du chromosome,
- l'insertion : fragment inséré dans le chromosome,
- la translocation : fragments échangés entre deux chromosomes,
- la fusion : fusion de deux chromosomes,
- la fission : un chromosome scindé en deux,

Une représentation schématique de ces événements est donnée Figure 2.3.

D'un point de vue biologique, les mécanismes entraînant de telles modifications du génome ont été très étudiés, notamment sur les génomes nucléaires des mammifères. Ces réarrangements sont dûs à des coupures des deux brins de l'ADN. L'origine de ces cassures peut être induite par des facteurs externes (rayons X, radicaux libres) ou induites par la cellule elle-même, grâce à des enzymes spécifiques, soit pour corriger des problèmes de structure durant la réplication [Aguilera and Gómez-González, 2008], soit pour modifier la structure de l'ADN en réponse à un stress (par exemple lors de la réponse immunitaire)[Pfeiffer et al., 2000]. Les brins cassés sont ensuite réparés. Deux mécanismes sont alors possibles. Soit une réparation rapide des fragments cassés par leur réintégration sans avoir besoin de l'information génétique [Pfeiffer et al., 2000] (appelé recombinaisons non homologue) consistant simplement à recoller les fragments après di-

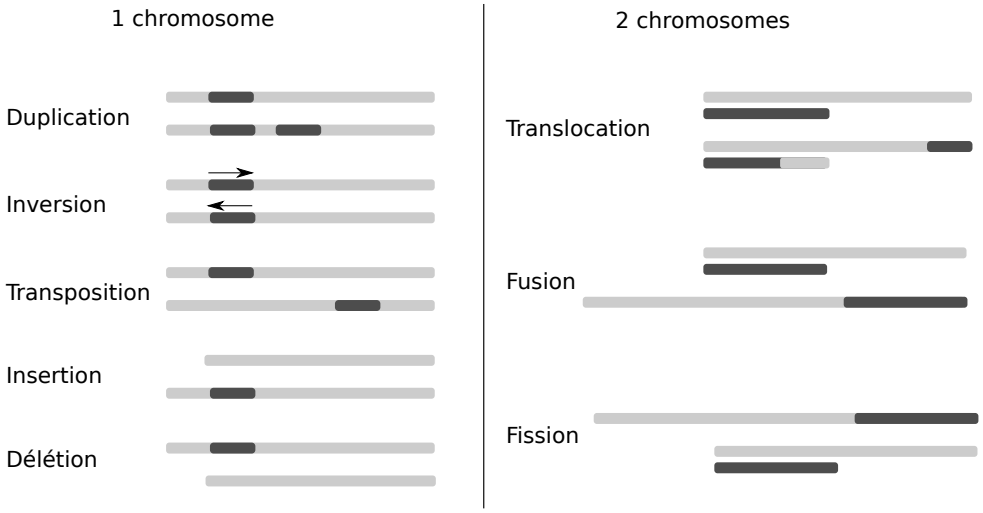


FIG. 2.3 – Illustration des différents événements de réarrangements. A gauche, les événements affectant un chromosome, à droite, les événements affectant plusieurs chromosomes.

gestion des extrémités cassées (entraîne une perte d'information), soit par recombinaison homologue, consistant à s'appuyer sur l'information génétique. La recombinaison homologue consiste à rassembler les fragments cassés en prenant comme matrice le chromosome homologue non cassé. Au moment de la recombinaison homologue, il est possible d'avoir des recombinaisons entre le brin matrice et le nouveau fragment synthétisé. Ces mécanismes de réparation de l'ADN, qu'ils soient homologues ou non, peuvent entraîner des recombinaisons d'ADN. Dans le cas de recombinaisons non homologues, si plusieurs cassures sont produites sur un même chromosome, l'assemblage des extrémités peut alors être différent de la séquence de départ, ce qui conduira à des inversions. Dans le cas de recombinaisons homologues, si il existe dans le génome une région très similaire mais non homologue à la région cassée et qu'elle est utilisée comme matrice, une recombinaison entre ces deux régions, lors de la réparation, mènera alors à différents réarrangements tels que des inversions ou des translocations [Bailey and Eichler, 2006]. L'induction d'inversions et de translocations à partir de régions très similaires est montrée Figure 2.4.

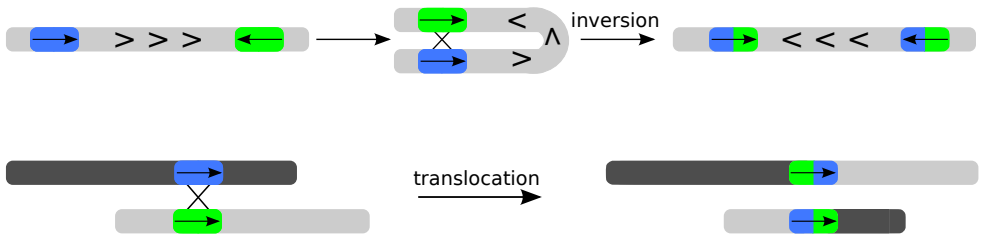


FIG. 2.4 – Exemples de réarrangements à partir de recombinaisons non alléliques. Deux segments similaires (bleu et vert) mais à des positions différentes du génome peuvent induire des recombinaisons.

Les événements d'inversion, de transposition, d'insertion et de délétion se réalisent à l'intérieur d'un même chromosome tandis que les événements de translocation (échange de deux régions), fusion et fission s'opèrent entre deux chromosomes. La translocation correspond à la transposition d'un fragment chromosomique sur un chromosome non homologue.

De plus, les événements d'insertion, de délétion, de fusion et de fission sont des réarrangements dit déséquilibrés puisqu'ils entraînent une modification de l'information génétique contenue à l'échelle d'un chromosome.

Etant donné deux suites de marqueurs et un ensemble d'événements possibles, on pourra alors s'intéresser à la recherche d'une histoire, un scénario, permettant de transformer une suite de marqueurs en une autre. Une illustration est donnée Figure 2.5.

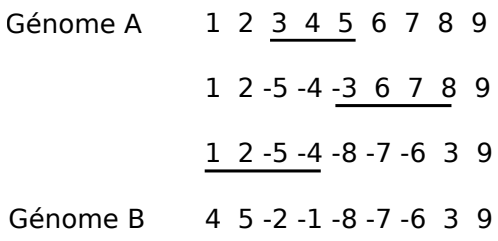


FIG. 2.5 – Illustration d'un scénario de réarrangement permettant de passer d'un génome A au génome B en utilisant trois inversions (marqueurs soulignés).

Il est également possible de calculer des phylogénies, soit en se basant sur les distances obtenues à partir de scénarios calculés pour chaque paire de séquences, soit en ayant une approche plus globale. Ces problèmes sont discutés dans les deux sections suivantes.

2.2 Reconstruction de scénarios

Nous nous intéressons maintenant aux méthodes permettant le calcul de scénarios de réarrangements entre deux génomes. Les génomes sont donnés, comme expliqué ci-dessus, comme des suites de marqueurs.

Le nombre de scénarios de réarrangements entre deux génomes est bien entendu très grand. Ainsi, le problème posé est celui de la recherche d'un scénario parcimonieux, c'est-à-dire utilisant un nombre minimum d'événements. Toutefois, il peut exister plusieurs scénarios parcimonieux. Le nombre d'événements d'un tel scénario définit une distance qui permet de quantifier la ressemblance entre les deux génomes d'intérêt. Si π est une permutation, on notera $d(\pi)$ la longueur d'un scénario parcimonieux permettant d'aboutir à la permutation identité.

2.2.1 Le problème de la distance d'inversion

L'étude des algorithmes de calcul de scénarios commence au début des années 1990 avec les articles de Kececioglu et Sankoff [Kececioglu and Sankoff, 1993, Kececioglu and Sankoff, 1995]. A l'époque, on se concentrait surtout sur l'étude d'une suite minimale d'inversions. Par convention, on supposera toujours que le génome auquel on souhaite aboutir est la permutation identité, c'est-à-dire la suite croissante des éléments de 1 à n. C'est pourquoi on utilise le vocabulaire « trier une permutation » pour rechercher une suite d'inversions.

Une première borne pour la distance d'inversions a été démontrée :

$$d(\pi) \geq \frac{b(\pi)}{2}$$

où π était la permutation initiale et $b(\pi)$ le nombre de points de cassures. Un point de cassure (*breakpoint* en anglais) est la région entre deux marqueurs formants une suite adjacente dans

l'une des deux permutations mais pas dans l'autre. Dans le cas signé, on considérera en plus un point de cassure lorsque deux marqueurs, adjacents dans les deux permutations, n'ont pas leur signe conservé. La Figure 2.6 donne une illustration de cette notion dans le cas signé. Par exemple, on trouve un point de cassure entre 6 et -5 car leur signe n'est pas conservé par rapport à la permutation identité (5 6 ou -6 -5).

Génome		2		9		6		-5	-4	-3		7	8		1				
Génome linéaire	•	2	•	9	•	6	•	-5	-4	-3	•	7	8	•	1	•			
Génome circulaire		2	•	9	•	6	•	-5	-4	-3	•	7	8	•	1				

FIG. 2.6 – Points de cassure dans les permutations (par rapport à la permutation identité). Les points de cassure sont représentés par des •. Suivant que le génome soit circulaire ou non, le nombre de points de cassures n'est pas le même.

Le premier algorithme, polynomial en temps, pour le calcul de la distance entre deux permutations signées, fut proposé par Hannennalli et Pevzner [Hannenhalli and Pevzner, 1995]. Il utilisait une modélisation intelligente des génomes sous la forme d'un graphe appelé « graphe des points de cassures » [Bafna and Pevzner, 1993]. Ce graphe représente les adjacences entre marqueurs dans les deux génomes (voir Figure 2.7). L'identification de structures dans ce graphe

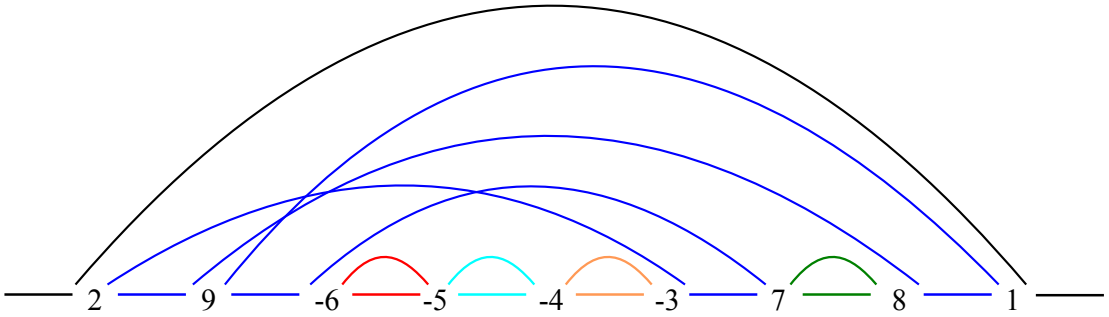


FIG. 2.7 – Le graphe des points de cassure sur le génome 2 9 6 -5 -4 -3 7 8 1. Si ce génome est circulaire, on compte 6 cycles (chaque cycle a une couleur différente) et la distance d'inversion sera ici $9 + 1 - 6 = 4$.

a permis l'obtention d'une formule exacte pour la distance d'inversions :

$$d(\pi) = n + 1 - c(\pi) + h(\pi) + f(\pi)$$

où n est le nombre de marqueurs, $c(\pi)$ le nombre de cycles dans le graphe et h et f deux fonctions décrivant des structures plus complexes du graphe qui sont rarement observées sur des données réelles. L'algorithme permettant d'obtenir un scénario correspondant à cette distance repose sur le choix d'inversions dites « triantes » en fonction des structures observées dans le graphe. On cherche à maximiser le nombre de cycles puisque, lorsqu'il est maximum, la permutation est triée. Comme ce n'est pas notre propos de décrire en détails ces algorithmes, nous renvoyons le lecteur à [Setubal and Meidanis, 1997] par exemple pour une description exhaustive. Des présentations différentes et simplifiées du même résultat furent ensuite proposées [Bergeron, 2005]. Depuis 2001, le calcul de la distance d'inversions entre deux permutations signées s'exécute en temps linéaire [Bader et al., 2001] et le meilleur algorithme de tri a été conçu en 2004 [Tannier and Sagot, 2004].

Le calcul pour le cas non signé a quant à lui été montré NP-dur [Caprara, 1997]. Il existe

cependant un algorithme polynomial avec ratio d'approximation pour réaliser le calcul (ration de 1,375) [Berman et al., 2001].

2.2.2 Avec d'autres événements

Le calcul de scénarios utilisant la transposition a aussi été bien étudié. Malheureusement la complexité du problème reste aujourd'hui ouverte et on dispose actuellement d'algorithmes avec ratio d'approximation (ratio de 1,375) [Elias and Hartman, 2006]. Dans le même type d'opération, il a été introduit la notion d'échange de blocs (*block interchange* en anglais) qui correspond à un échange de suites de marqueurs distants et non nécessairement voisins comme dans la transposition. On dispose d'un algorithme exact polynomial [Christie, 1996].

Dans le cas où le génome est composé de plusieurs chromosomes, il faut introduire les opérations citées plus haut permettant de « transférer » des marqueurs d'un chromosome à un autre. Le cas de la translocation se résout grâce à un algorithme linéaire [Bergeron et al., 2006]. L'utilisation combinée de la translocation, de la fusion et de la fission aboutit à un algorithme polynomial [Dias and Meidanis, 2001, Lu et al., 2006].

Une opération de réarrangement plus générale, appelée DCJ pour *Double-Cut-and-Join*, a été proposée en 2005 [Yancopoulos et al., 2005] (voir Figure 2.8 pour une illustration). Cette

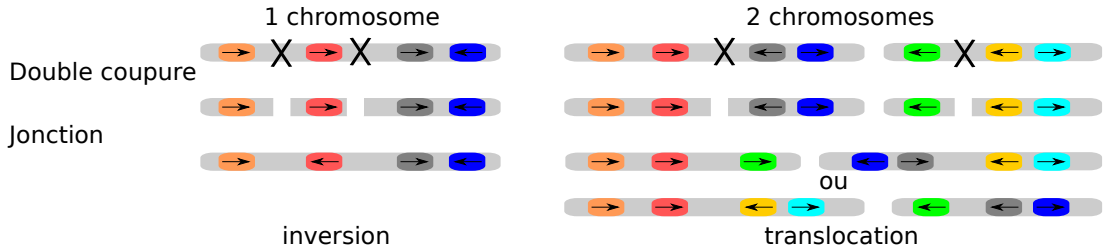


FIG. 2.8 – Illustration des opérations DCJ. A gauche, opération de DCJ sur un chromosome aboutissant à une inversion, à droite opération de DCJ sur deux chromosomes aboutissant à deux configurations possibles de translocation.

opération, qui ne correspond pas réellement à un événement évolutif, permet de modéliser chacune des opérations citées avant en combinant plusieurs DCJ. Le problème du calcul de scénarios par DCJ a l'avantage d'être résolu de manière exacte grâce un algorithme linéaire.

2.2.3 Avec un contenu en marqueurs différent

Dans le cas où le contenu en marqueurs est différent entre les deux génomes, il est possible d'ajouter les opérations de délétion et d'insertion de marqueurs. Le problème du tri par inversions (avec signe) avec délétion/insertion peut être résolu par un algorithme polynomial [El-Mabrouk, 2001].

Le problème des marqueurs dupliqués. Pour le moment, aucun outil existant, pour l'étude des réarrangements de génomes, ne prend réellement en compte les événements de duplication. Biologiquement, les duplications de gènes sont des événement courants, nous pouvons également noter que les génomes mitochondriaux de plantes sur lesquels nous travaillerons comportent de nombreux marqueurs dupliqués. Nous discuterons des méthodes existantes permettant de retrouver des régions dupliquées dans des permutations, dans la Section 5.

L'ajout de marqueurs dupliqués pose le problème de l'identification des relations paralogues/orthologues. Si celles-ci ne sont pas connues alors on comprend qu'il faut tester toutes les possibilités d'orthologie possibles, ce qui aboutit à une combinatoire très importante.

Pour tenter de résoudre ce problème, deux approches furent proposées : construire des génomes exemplaires (*exemplar genomes*) [Sankoff, 1999] ou *maximum matching* [Tang and Moret, 2003]. L'objectif de la technique de génome exemplaire est de ne considérer plus qu'une seule copie de chaque duplicat dans les permutations. Cette méthode permet donc d'obtenir des génomes sans dupliqués. Son principal problème est qu'il faut choisir une des copies des dupliqués. Si les marqueurs paralogues sont parfaitement conservés, il n'est pas possible de savoir si l'on sélectionne le vrai orthologue ou non par rapport aux marqueurs des autres génomes. De plus, si l'on veut prendre en compte toutes les possibilités, si un génome contient n marqueurs dupliqués (les dupliqués sont en deux copies), il y aura n^2 possibilités de génomes exemplaires. Plus généralement, si n marqueurs sont en k copies et m marqueurs en l copies, il y aura $n^k \times m^l$ possibilités. Dans l'exemple illustré Figure 2.9, on obtient deux et quatre génomes exemplaires possibles pour respectivement le génome A et le génome B. De ce fait, calculer une phylogénie entre A et B revient à tester huite arbres (2×4) pour trouver celui qui paraît, au niveau évolutif, le plus probable. Quand le nombre de génomes ainsi que le nombre de marqueurs dupliqués augmentent, il devient impossible d'analyser tous ces arbres dont, pour la plupart, les relations d'orthlologie et de paralogie des marqueurs est biaisée.

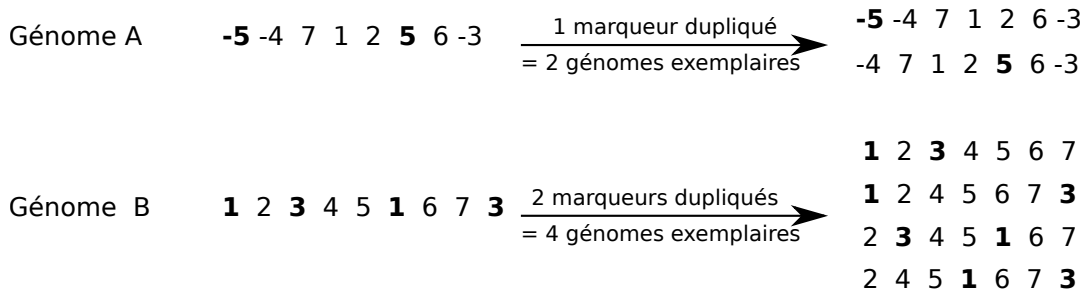


FIG. 2.9 – Construction de génomes exemplaires pour les génomes A et B. Les génomes A et B contiennent, respectivement, un et deux marqueurs dupliqués. Les marqueurs 5, 1 et 3 dupliqués sont indiqués en gras.

La méthode de *maximum matching model* consiste, quant à elle, à garder un maximum de copies des dupliqués. Par exemple, lorsque l'on compare deux génomes, si on trouve deux copies dans un génome et trois copies dans l'autre génome d'un même marqueur, on conservera le nombre maximal de marqueurs communs, c'est-à-dire deux copies.

Pour les deux méthode que nous venons de décrire, le choix des copies conservées se fait à l'aide d'une fonction d'optimisation. Cette fonction, basée sur la parcimonie, consiste à minimiser les distances d'évolution entre deux génomes.

Le problème de ce type de méthodes est qu'elles restent applicables lorsque le nombre de marqueurs dupliqués dans les génomes est faible si l'on veut pouvoir comparer les résultats de différents matching. De plus elles possèdent l'inconvénient de ne pas proposer des événements de duplication à proprement dit dans les scénarios. Très récemment, [Bader, 2009, Bader, 2010] ont proposé une méthode permettant de passer d'un génome avec dupliqués à un génome, supposé ancestral, contenant exactement une copie de chaque marqueur en incluant notamment des événements d'échange de blocs, d'inversions et de duplications en tandem. Malheureusement, la méthode ne permet pas de comparer plusieurs génomes dont chacun possède des marqueurs

dupliqués.

2.2.4 Spécialisation et adaptation au contexte biologique

Des méthodes ont été proposées essayant de se rapprocher de réalités biologiques. Par exemple, des méthodes ont été proposées pour prendre en compte un éventuel coût par rapport aux événements d'inversion, transposition et translocations [Sankoff et al., 1997, Sankoff et al., 2000]. En effet, en fonction des organismes et organites observés, certains événements sont plus fréquents que d'autres. Par exemple, les translocations sont très courantes dans les génomes de mammifères et inexistantes dans les génomes de Drosophilles. Des inversions existent dans les génomes mitochondriaux de mammifères et sont beaucoup plus rares dans ceux des Fungi [Blanchette et al., 1996]. Chez les primates les réarrangements sont surtout des inversions alors que chez les lémuriens ce sont des translocations. On peut également noter que les taux de réarrangements au sein des espèces sont différents. Par exemple le taux de tanslocation est plus élevé chez les mammifères que chez les insectes [Coghlan et al., 2005]. D'autres méthodes, ont proposé de prendre en compte la taille des inversions [Dalevi et al., 2002, McLysaght et al., 2002]. En effet, une étude sur des génomes bactériens laisse supposer que les petites inversions seraient plus fréquentes que les grandes inversions [Lefebvre et al., 2003]. En effet, les petites inversions touchant uniquement un gène, ne déconnectent pas ce gène du cluster de gènes dans lequel il est impliqué et jouent donc uniquement sur la régulation de ce gène. Enfin, il a été proposé de tenir compte des régions du génome dans lesquelles sont effectués les réarrangements. Il a notamment été observé que des régions de génomes semblent plus fragiles au cassures et donc que ces cassures ne se feraient pas aléatoirement [Pevzner and Tesler, 2003b].

2.2.5 Les outils disponibles

GRIMM (Genome Rearrangements In Man and Mouse) [Tesler, 2002]

GRIMM est un des rares outils disponibles à la communauté permettant le calcul de scénarios et de distances d'inversions entre deux génomes. Si on lui donne en entrée un ensemble de génomes dont les marqueurs sont signés ou non, il calcule la matrice de distance entre ces génomes. Les génomes peuvent être multi-chromosomes, auquel cas les opérations de translocation, fusion et fission sont utilisées en plus des inversions. Les génomes uni-chromosome peuvent être linéaires ou circulaires.

SPRING (Sorting Permutation by Reversals and block-INterchanGes) [Lin et al., 2006]

SPRING est un outil de calcul de distances de réarrangements utilisant les inversions et les échanges de blocs sur des marqueurs signés. Les génomes doivent être uni-chromosome et il est possible de définir si ils sont circulaires ou linéaires. Une particularité de SPRING est d'accepter, en entrée, des suites de marqueurs mais également des séquences. Dans ce cas SPRING calcule des blocs de synténie conservés grâce à Mauve dont nous avons discuté plus haut. Cette option ne nous intéressait pas dans notre étude car les segments dupliqués ne sont pas gérés.

Il reste alors que SPRING, limité aux inversions fournit le même service que GRIMM. Nous avons choisi d'utiliser GRIMM dans nos analyses.

2.3 Reconstruction phylogénétique

La reconstruction phylogénétique basée sur les réarrangements peut s'entendre de plusieurs manières, tout comme lorsqu'elle est faite sur la base des données de séquences.

Une première approche consiste à prendre la distance d'inversion, de transposition, *etc.* comme estimateur de la distance évolutive et d'appliquer des algorithmes de reconstruction d'arbres à partir d'une matrice de distance. Dans ce cadre, il peut être nécessaire de réaliser une correction des distances de réarrangements qui sous-estiment généralement la vraie distance évolutive. Cette idée a été largement explorée par Moret et ses collègues au début des années 2000 [Moret et al., 2001, Moret et al., 2002]. Cette première approche est donc relativement simple à mettre en place puisqu'il suffit de calculer les distances entre les paires de génomes avec des outils tels que GRIMM ou SPRING. On obtient alors une matrice de distances sur laquelle on peut, par exemple, construire des phylogénies de Neighbor-Joining. Cependant, les méthodes basées sur des matrices de distances, ne permettent pas de reconstruire les séquences ancestrales.

Une autre approche consiste à explorer les méthodes par parcimonie et par maximum de vraisemblance, bien connues en reconstruction phylogénétique. Elles offrent l'avantage de fournir des séquences ancestrales putatives en plus d'un arbre de réarrangements.

2.3.1 L'analyse simultanée de plusieurs génomes

Le problème de la reconstruction d'un arbre de réarrangements, accompagné des scénarios et des séquences ancestrales, a été introduit dès 1996 [Sankoff et al., 1996]. Le problème limité à trois génomes, dont le but est de retrouver une séquence ancestrale à partir de trois séquences de marqueurs, est nommé problème du médian. Ce problème est en fait un sous problème du problème de réarrangements de génomes multiples. Pour un jeu de génomes de taille n, le problème MGPR est de retrouver un arbre dont les feuilles sont les séquences de marqueurs données, les nœuds correspondent à des permutations de taille n et ces séquences minimisent les distances entre les nœuds. Le problème du médian ayant été montré NP-dur pour les inversions sur des marqueurs signés [Caprara, 1999], la communauté s'est attachée à trouver des solutions soit heuristiques, soit suffisamment efficaces en pratique pour traiter un nombre raisonnable de génomes.

Nous citerons ici trois outils qui résolvent le problème de réarrangements de génomes multiples. Le premier outil, BPAnalysis, était basé sur l'analyse des points de cassure [Sankoff and Blanchette, 1998]. Un autre outil, GRAPPA, a été proposé permettant d'utiliser la distance d'inversion [Moret et al., 2002] et a ensuite été amélioré en intégrant les nouveaux résultats algorithmiques. Enfin, il existe MGR [Bourque and Pevzner, 2002], un outil également basé sur la distance d'inversion fournissant un des arbres les plus parcimonieux ainsi que les séquences ancestrales putatives.

2.3.2 Les outils GRAPPA et MGR

GRAPPA et MGR sont deux outils permettant de résoudre le problème de réarrangements de génomes multiples grâce à des algorithmes résolvant le problème de médian. Ce sont des approches basées sur la parcimonie permettant une reconstruction de l'ordre des gènes entre plusieurs génomes en un temps d'exécution raisonnable.

GRAPPA [Moret et al., 2002]

GRAPPA est en quelque sorte une amélioration de BPAnalysis. Cependant, GRAPPA utilise des distances d'inversions contrairement à BPAnalysis qui était basé sur les distances de points de cassure entre les génomes. De plus, l'algorithme utilisé dans GRAPPA est plus rapide. En fait, GRAPPA est basé sur deux algorithmes différents permettant de résoudre le problème de médian de manière exacte. Pour résumer son fonctionnement, à partir de plusieurs permutations données en entrée, GRAPPA va tester toutes les topologies possibles d'arbres de ces permutations. Pour chaque arbre, les nœuds internes seront établis et un médian est alors calculée pour ces nœuds.

MGR [Bourque and Pevzner, 2002]

MGR utilise une heuristique pour répondre au problème de médian, c'est-à-dire que pour trois permutations, MGR essaie de trouver des inversions qui rapprochent une permutation des deux autres. Pour résumer son fonctionnement, MGR commence par identifier les inversions rapprochant une permutation de toutes les autres. Si pour une permutation donnée aucune inversion pouvant le rapprocher des autres génomes n'est trouvée, celle-ci est alors non-résolue. L'algorithme connecte alors les permutations non résolues en identifiant les triplets de permutations qui minimisent les distances d'inversions lorsque l'on ajoute la permutation non résolue. Pour ces triplets, le problème de médian est alors facile à résoudre.

Bien qu'il n'y ait pas eu d'étude montrant que GRAPPA ou MGR fournit les meilleurs résultats, MGR qui est basé sur une heuristique est plus rapide que GRAPPA. De plus, MGR retourne un des arbres les plus parcimonieux tandis que GRAPPA fournit toutes les topologies d'arbres possible que l'on peut obtenir avec les permutations entrées.

2.3.3 Conclusions

Nous avons fait ici un rappel non exhaustif des méthodes existantes pour analyser les réarrangements entre plusieurs génomes. Nous avons surtout insisté sur les méthodes que nous avons choisies pour l'analyse de nos génomes. Dans un premier temps, nous avons choisi de conserver le logiciel Mauve pour nous aider à détecter les marqueurs communs existants entre plusieurs génomes. Nous possédons des génomes mitochondriaux composés d'un seul chromosome circulaire. Les méthodes correspondant le plus à ce type de génomes sont donc les méthodes basées sur les distances d'inversion. Les méthodes d'étude de réarrangements que nous avons choisies sont donc GRIMM pour l'étude des distances d'inversion et MGR pour l'analyse d'un arbre phylogénétique parcimonieux et les séquences ancestrales associées.

Chapitre 3

Mise en place d'une base de données dédiée à l'analyse des génomes mitochondriaux de plantes

Nous le savons, pour comparer l'évolution des structures de différents génomes mitochondriaux de plantes, il faut commencer par chercher des séquences homologues entre ces génomes afin de créer les fichiers d'entrée nécessaires à l'utilisation des outils de réarrangement. Nous avons donc développé Plamidb (PLAnt MItochondrial DataBase), une base de données dédiée à la comparaison des contenus en séquences codantes dans différents génomes mitochondriaux. Si, dans un premier temps, Plamidb a essentiellement servi à comparer les gènes communs entre différentes espèces, nous l'avons amélioré dans l'optique de pouvoir annoter facilement et rapidement des génomes mitochondriaux.

Dans ce chapitre nous présenterons donc la base de données et son contenu, nous décrirons ensuite sa fonctionnalité d'annotation de génomes. Enfin nous décrirons son interface web.

3.1 Base de données

Plamidb est une base de données conçue pour l'analyse des génomes mitochondriaux de plantes. Elle contient les informations relatives aux génomes mitochondriaux complets séquencés que l'on peut trouver dans les bases de données publiques. Nous verrons dans cette partie comment sont extraites et stockées les informations de ces génomes.

3.1.1 Contenu

Cette base de données, gérée par le système de gestion de bases de données relationnelles (SGBDR) PostgreSQL, contient un maximum d'informations extraites de fichiers GenBank. Les informations retenues pour les comparaisons de génomes sont, ce que nous appelerons de manière générique, les éléments (*feature* en anglais, par exemple les gènes). De plus, la base de données visant à la comparaison rapide de génomes mitochondriaux, un processus a été mis en place afin de stocker, au moment de l'ajout d'une nouvelle espèce, toutes les relations d'homologie entre les différents éléments des génomes. Plamidb va donc s'articuler autour de deux modules qui sont les éléments d'un génome et les relations entre ces éléments. La Figure 3.1 montre le schéma relationnel de Plamidb. Bien que cette base de données soit dédiée à la comparaison de

génomes mitochondriaux, et sachant qu'il existe des transferts de séquences entre les génomes mitochondriaux et les génomes chloroplastiques, nous avons également intégré ces derniers afin de repérer les éléments identiques entre ces deux organites. Tous les génomes contenus dans la base sont montrés dans le Tableau 3.1.

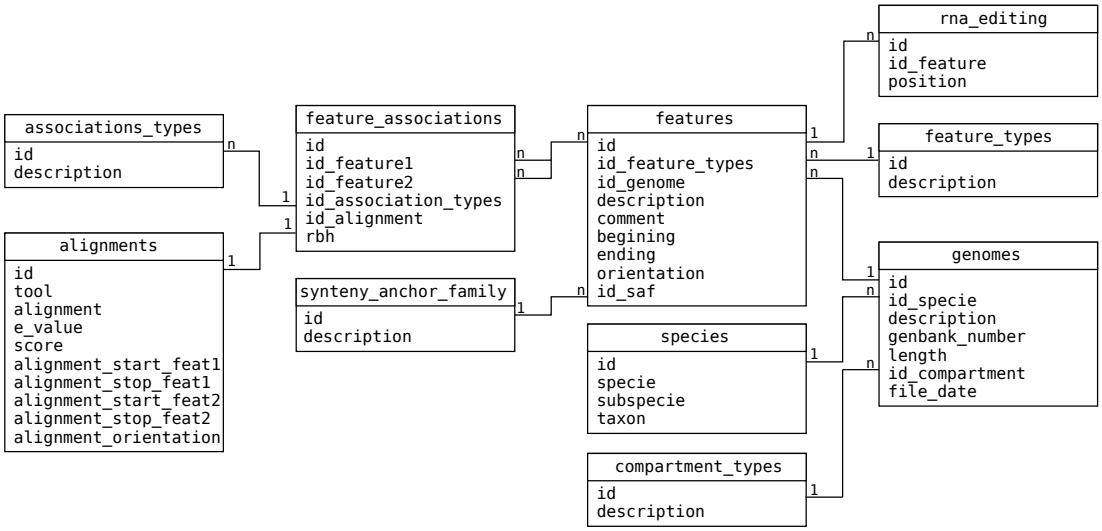


FIG. 3.1 – Schéma relationnel de PLAMIDB.

Éléments

Les éléments qui nous intéressent sont les gènes, ORF et ARN contenus dans les fichiers GENBANK des génomes mitochondriaux. Toutes les informations nécessaires au remplissage de la base de données, pour être stockées dans les tables prévues à cet effet, sont extraites à l'aide de scripts Python en utilisant la librairie bioPython. En effet, cette librairie contient des modules parfaitement adaptés à la lecture et l'extraction de données présentes dans les fichiers de type GENBANK. Pour chaque génome, il nous est alors facile de récupérer les caractéristiques de chaque élément c'est-à-dire leurs noms, leurs positions dans le génome ainsi que les positions des sites édités. Un génome, appartenant à une espèce donnée, sera constitué de plusieurs éléments.

Relations entre les éléments

La première fonction de PLAMIDB est d'identifier rapidement les relations d'homologie entre les différents éléments de plusieurs génomes. Afin d'optimiser la vitesse des requêtes, ces relations sont établies et stockées dans la base de données au moment de l'ajout d'un génome. Plusieurs relations sont analysées : les éléments paralogues, les éléments orthologues ainsi que tous les éléments présentant une homologie, même partielle avec d'autres éléments. Quand un génome est ajouté, les comparaisons se font donc en deux temps : dans un premier temps, entre paires d'éléments dans un génome pour détecter les éléments paralogues et, dans un deuxième temps, entre génomes pour détecter les éléments orthologues ou partiellement homologues.

Nom de l'espèce	N° d'accession	Organite	Taille (bp)
Arabidopsis thaliana	Y08501	mitochondrie	366924
Beta vulgaris (TK81-MS)	BA000024	mitochondrie	501020
Beta vulgaris (TK81-O)	BA000009	mitochondrie	368801
Brassica napus	AP006444	mitochondrie	221853
Carica papaya	NC_012116	mitochondrie	476890
Nicotiana tabacum	BA000042	mitochondrie	430597
Vitis vinifera	NC_012119	mitochondrie	773279
Zea mays mays (CMS-C)	DQ645536	mitochondrie	739719
Zea mays mays (CMS-S)	DQ490951	mitochondrie	557162
Zea mays mays (CMS-T)	DQ490953	mitochondrie	535825
Zea mays mays (NA)	DQ490952	mitochondrie	701046
Zea mays mays (NB)	AY506529	mitochondrie	569630
Zea mays parviglumis	DQ645539	mitochondrie	680603
Zea luxurians	NC_008333	mitochondrie	539368
Zea perennis	NC_008331	mitochondrie	570354
Nicotiana tabacum	Z00044	chloroplaste	155943
Spinacia oleracea	NC_002202	chloroplaste	150725

Tab. 3.1 – Espèces contenues dans Plamidb.

3.1.2 Ajout de données

Comme nous l'avons vu, Plamidb contiendra deux types de données : les éléments contenus dans les génomes et les relations d'homologie entre ces éléments. Les données concernant le contenu des génomes sont directement extraites des fichiers GenBank. Une étape cruciale avant l'ajout de données dans la base est la vérification de ces fichiers. En effet, les champs de ces fichiers sont remplis au libre arbitre des auteurs et il peut arriver que les renseignements ne se retrouvent pas forcément au même endroit en fonction des génomes. Par exemple, dans les génomes mitochondriaux de plantes, nous savons que certains gènes sont épissés en trans. De ce fait nous considérerons les différents exons comme des éléments à part entière. Il n'est pas obligatoire de renseigner les champs correspondants aux exons dans les fichier GenBank si bien qu'ils sont parfois inexistants : il faut donc les rajouter manuellement dans le fichier afin que ces données soient prises en compte.

Lorsqu'un nouveau génome est ajouté, le processus de remplissage est le suivant : dans un premier temps toutes les informations relatives au génome, telles que l'espèce, la séquence, la date du fichier entré sont extraites et ajoutées à la base. Tous les éléments et leurs caractéristiques sont également extraits et insérés dans la base, et un fichier *multifasta* avec les séquences de tous les éléments du génome est créé afin de procéder à leur comparaison. Dans un deuxième temps, les comparaisons des éléments sont effectuées. Toutes les comparaisons sont réalisées à l'aide de l'outil d'alignement de séquences YASS [Noe and Kucherov, 2005] (matrice choisie : +1 pour les identités, -3 pour les substitutions).

Les relations de paralogie sont définies à partir de l'alignement de chaque élément du génome contre les autres éléments que contient le génome. On établira une relation de paralogie entre deux éléments s'ils sont retrouvés comme *RBH* (Reciprocal Best Hit), c'est-à-dire que, pour chacun des deux éléments, le meilleur alignement retrouvé est avec l'autre élément. De plus, il

faut que la taille de l'alignement des deux séquences soit au pire inférieure à 8% de la taille des éléments, quant à la *E-value*, elle doit être inférieure à 1×10^{-170} si la taille des éléments est supérieure à 100 pb, ou inférieure à 1×10^{-26} si leur taille est inférieure à 100 pb. La *E-value* indique la probabilité que l'alignement trouvé soit dû au hasard. Cette valeur dépendra de la taille des séquences alignées, c'est pourquoi nous autorisons une valeur plus faible pour les courtes séquences (ARNt et petits exons). Bien entendu, nous ne considérerons pas comme paralogues les éléments qui se chevauchent.

Les relations d'orthologie sont établies en comparant le nouveau génome avec les génomes déjà contenus dans la base de données. Nous effectuons les comparaisons par paires de génomes. Pour chaque génome contenu dans PLAMIDB, un fichier *multifasta*, contenant toutes les séquences de ses éléments est créé. Ce fichier est ensuite comparé à l'aide de YASS au fichier *multifasta* créé pour le nouveau génome. Pour chaque alignement de chaque élément, les orthologues sont définis selon les mêmes critères que les paralogues (*RBH*, *E-value* et taille de l'alignement). Quand il existe un alignement entre un élément du nouveau génome et un élément d'un des génomes de la base, dont la *E-value* est supérieure à 1×10^{-26} (peu importe la taille des séquences), mais que les autres critères ne permettent pas d'attribuer une relation d'orthologie, les deux éléments sont alors considérés comme des homologues partiels.

Que ce soit pour la définition de paralogues ou d'orthologues, les critères choisis sont assez restrictifs. Cependant, dans le cas des génomes mitochondriaux des plantes, nous savons que le taux de mutation μ est très faible. De ce fait, nous devons chercher et déterminer comme paralogues et orthologues, des éléments dont les séquences sont très conservées. De plus, lorsque des paralogues et orthologues sont déterminés, ceux-ci sont définis comme appartenant à une même famille (*synteny_anchor_family* dans la Figure 3.1). Par exemple, si pour un élément du nouveau génome, on trouve un orthologue qui appartient déjà à une famille de gène, l'élément du nouveau génome sera alors associé à cette famille de gène. A chaque famille de gène sera donné un identifiant unique avec lequel on retrouvera facilement tous les orthologues et paralogues la constituant. Ceci facilitera la transformation des génomes en permutations lorsque nous voudrons analyser les réarrangements entre deux ou plusieurs génomes.

3.2 Interface web

Une interface Web a été développée afin que tout utilisateur puisse effectuer des recherches de comparaison sur les génomes de la base de données et visualiser les résultats correspondants.

3.2.1 Organisation

L'organisation de PLAMIDB est basée sur une architecture à trois niveaux appelée architecture trois tiers (Figure 3.2). Dans ce cadre, l'utilisateur (appelé client) enverra des requêtes à un serveur Web qui effectuera les requêtes sur la base de données et les renverra à l'utilisateur. Le transfert de requêtes entre le client est le serveur Web est assuré par le protocole HTTP (*HyperText Transfert Protocol*). Ces requêtes seront alors interprétées par le serveur Web Apache qui va alors effectuer les requêtes sur la base de données. Le langage utilisé par les SGBDR pour effectuer des requêtes sur une base de données est le langage SQL (*Structure Query Language*).

Le langage HTML (*HyperText Markup Language*) est la base de tout site Web. Il s'agit d'un langage balisé qui est interprété par les navigateurs Web afin d'en assurer sa bonne visualisation. Le problème du langage HTML est qu'il ne fournit que des pages statiques, alors qu'ici, les résultats affichés changeront en fonction des requêtes effectuées. Nous avons donc utilisé, côté

FIG. 3.2 – Architecture trois tiers de PLAMIDB. Les requêtes de l'utilisateur sont envoyées au serveur Web qui interroge la base de données et les résultats obtenus sont affichés à l'utilisateur.

serveur, le langage PHP (*PHP : Hypertext Preprocessor*). Ce langage permet d'interpréter et de récupérer les informations d'une page HTML, telles que celles contenues dans l'interface où l'utilisateur soumet ses requêtes. Il va aussi permettre de générer des pages dynamiques (contenant du HTML) correspondant aux résultats obtenus. De plus, le langage PHP est particulièrement adapté pour la communication avec les bases de données en effectuant les requêtes SQL au niveau du SGBDR. Nous avons également mis en place un système de vérification de remplissage des champs au niveau de l'interface utilisateur grâce au langage JavaScript.

L'interface générale sur laquelle arrive l'utilisateur est présentée en Figure 3.3. L'utilisateur peut alors accéder à l'interface de requêtes ou encore l'interface d'annotation (dont nous parlerons en Section 3.3). Pour les administrateurs, une mise à jour est possible par l'intermédiaire de l'onglet dédié.

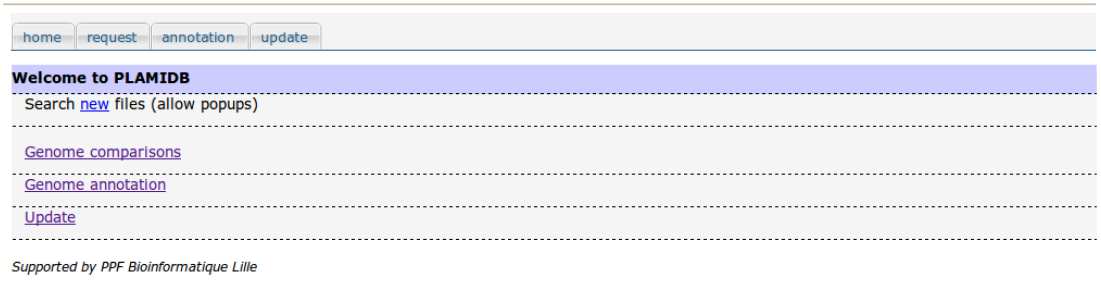


FIG. 3.3 – Page d'accueil de PLAMIDB.

3.2.2 Comparaison de génomes

Interface de requêtes

L'interface de requêtes proposée à un utilisateur voulant comparer des génomes mitochondriaux est présentée en Figure 3.4(a). Les requêtes se déroulent de la façon suivante : l'utilisateur choisit un génome que nous appellerons génome requête (*Beta vulgaris* TK81-MS dans l'exemple) pour lequel il choisit les différents éléments (ARNt dans l'exemple) qu'il veut comparer avec les autres génomes de la base. Dans l'exemple nous avons choisi de le comparer au génome mitochondrial de *Beta vulgaris* TK81-O et au génome chloroplastique de *Spinacia oleracea*. L'utilisateur sélectionne également le type d'homologie recherchée. Ici, nous avons choisi des homologies avec tous les génomes de la base afin de comparer génomes mitochondriaux et chloroplastiques. Il est possible de choisir des relations d'orthologie avec les génomes mitochondriaux ou encore des relations de paralogie dans le génome requête. L'utilisateur peut également rechercher les éléments

du génome requête par leurs noms ou dans une région donnée du génome ainsi qu'opter pour la visualisation, dans les résultats, des séquences des éléments, de leurs alignements ou encore de leurs positions. Un fois le formulaire rempli, il n'y a plus qu'à le soumettre.

Résultats

Les résultats obtenus pour l'exemple de requête vu précédemment sont présentés dans la Figure 3.4(b). La première colonne correspond au génome requête, les autres aux génomes comparés. Nous pouvons voir par exemple que l'ARN de transfert Pro, appartenant au génome de *Beta vulgaris* TK81-MS est également retrouvé dans le génome mitochondrial de *Beta vulgaris* TK81-O et chloroplastique de *Spinacia oleracea*. Les génomes mitochondriaux ont un fond gris et les génomes chloroplastiques un fond vert. L'utilisateur peut connaître les scores et E-value de l'alignement entre la séquence du génome requête et la séquence du génome comparé. S'il le souhaite, il peut accéder aux séquences (des éléments et des alignements) en cliquant sur les liens *sequence* et *alignment*. Lorsqu'un élément n'est pas retrouvé dans un des génomes, la case correspondante sera vide. C'est ici le cas de l'ARN de transfert Glu qui n'est pas retrouvé dans le génome chloroplastique de *Spinacia oleracea*. Dans le cas des comparaisons d'ARNt, l'utilisateur peut donc avoir rapidement une idée sur les ARN d'origine mitochondriale ou chloroplastique.

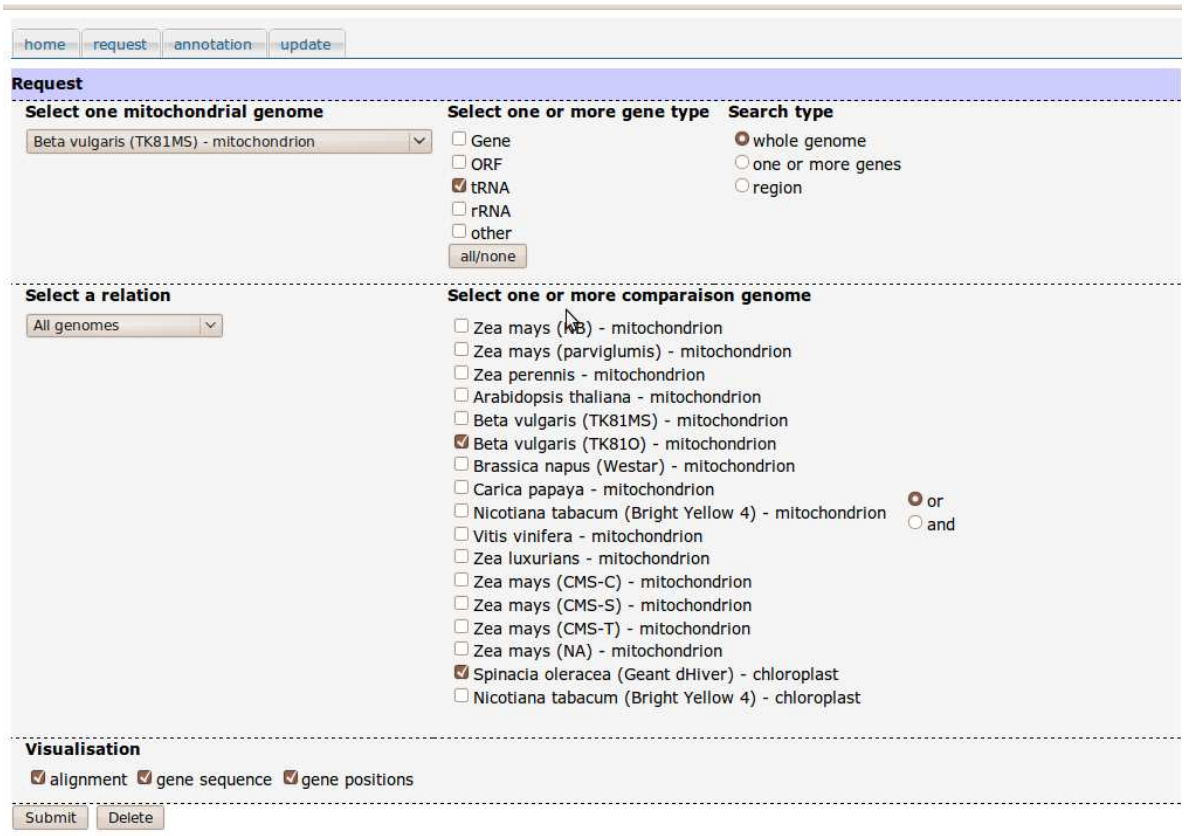
Plamidb s'est avéré un outil particulièrement utile lors de l'analyse des génomes mitochondriaux de *Beta*. Nous avons ainsi utilisé les fonctionnalités de comparaison de gènes, ORF et ARN de Plamidb, pour obtenir les familles de gènes contenues dans les génomes ainsi que leur nombre d'occurrences (Chapitre 6, Tableau 6.2). Au niveau des ARNt, l'utilisation de la base nous a également permis de retrouver les ARNt homologues entre les génomes mais également de prédire leur origine en comparant ces génomes à des génomes chloroplastiques (Chapitre 6, Tableau 6.2). Enfin, nous avons utilisé plamidb pour extraire les ORF homologues entre tous ces génomes (Annexe Chapitre 6, page 196).

3.2.3 Mise à jour

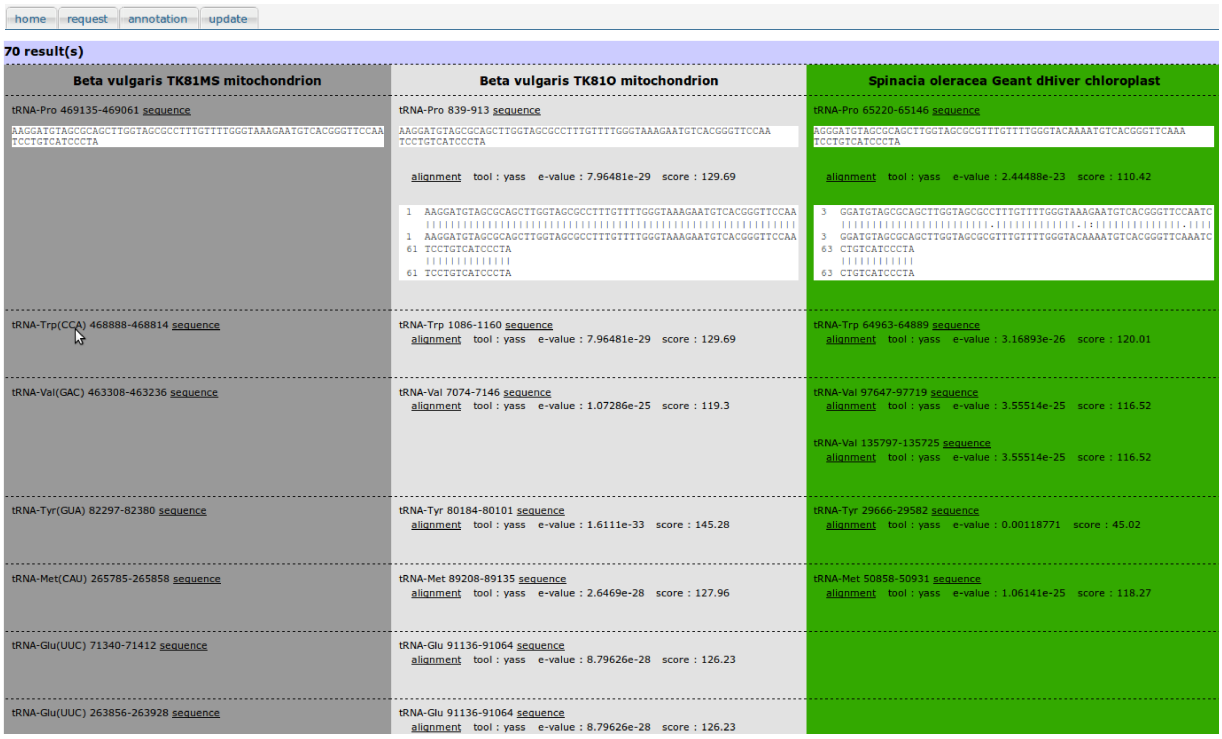
La mise à jour se fait également par une interface, réservée aux administrateurs (Figure 3.5). Trois choix sont proposés : mettre à jour toute la base de données, l'ajout et la suppression de génomes. Dans le cas de la mise à jour, une vérification automatique de la date de publication des génomes contenus dans Plamidb par rapport à GenBank est effectuée. Si la date dans Plamidb est inférieure, le génome est alors remplacé. L'ajout de génomes peut se faire par téléchargement direct de GenBank si l'utilisateur connaît le numéro d'accession du génome ou simplement en entrant un fichier de format GenBank. Enfin, la suppression se fait par sélection d'un ou plusieurs génome(s).

3.3 Outil d'annotation

Nous avons ajouté une fonction d'annotation des génomes mitochondriaux à Plamidb. Cette fonction nous sera très utile lorsque nous devrons annoter les génomes mitochondriaux de betterave.



(a) Interface de requête.



(b) Résultats de requête.

FIG. 3.4 – Exemple de requête dans PLAMIDB. Recherche des ARNt, chez *Beta vulgaris* TK81-O (mt) et *Spinacia oleracea* (cp), homologues à ceux de *Beta vulgaris* TK81-MS (mt).

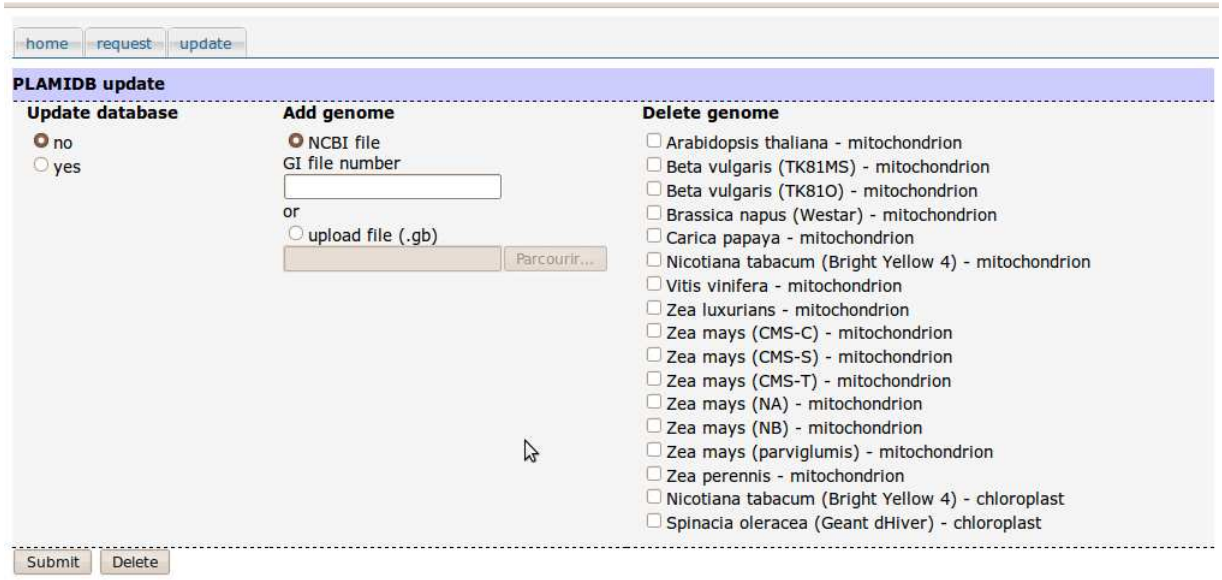


FIG. 3.5 – Interface de mise à jour de PLAMIDB.

3.3.1 Principe

Le but de cette fonctionnalité est de permettre à un utilisateur d'entrer un fichier de séquence de type *fasta* pour lequel tout gène, ORF, ARN et duplication sera recherché. Il pourra ensuite voir les similarités retrouvées avec les éléments contenus dans la base, les sélectionner et obtenir un fichier d'annotations. Le fichier d'annotations créé ici est de type *embl*, un choix qui s'explique par la nécessité, en Europe, de soumettre les fichiers dans EMBL. L'annotation se déroulera donc en cinq phases à partir de la soumission d'un génome sous forme d'un fichier de type *fasta*. Tout d'abord, les gènes seront recherchés par homologie de séquence avec tous les gènes contenus dans la base. La recherche s'effectuera ensuite sur les ORF, puis les ARNt et les ARNr. Enfin, les duplications au sein de ce génome seront retrouvées. Comme nous l'avons déjà expliqué, les taux de mutation μ entre les génomes mitochondriaux sont très faibles, il est ainsi possible de retrouver les gènes conservés entre plusieurs espèces, par alignement des séquences. Même si le contenu en gène peut varier entre différentes espèces, si la base de données contient suffisamment d'espèces de plantes différentes, cela suffira à retrouver les gènes contenus dans un nouveau génome.

Gènes

Pour évaluer les gènes contenus dans le génome soumis, tous les éléments de type gène sont extraits de la base de données. Les gènes appartenant au même groupe d'orthologues et paralogues (retrouvés par le numéro de famille qui leur avait été attribué) sont rassemblés dans un fichier *multifasta*. Quand un gène n'a ni orthologue, ni paralogue, il est placé seul dans un fichier *fasta*. Pour chaque fichier, un alignement avec YASS est réalisé. Les critères d'attribution d'orthologie entre les gènes contenus dans la base de données et le génome à annoter sont les mêmes que ceux choisis pour la construction de la base de données. Si une relation d'orthologie est retrouvée, un fichier d'alignement multiple entre la région concernée du génome à annoter et la famille de gène est réalisé. Si des gènes appartenant à cette famille ont des sites d'édition

d'ARN connus, leurs positions sont également repérées.

ORF

La recherche d'ORF se fait en deux phases. Dans une première phase, nous recherchons les ORF orthologues aux éléments de type ORF contenus dans PLAMIDB (les critères utilisés sont toujours les mêmes). Dans un deuxième temps, nous procédons à la recherche de toute région du génome commençant par un codon start (ATG) et finissant par un codon stop (TAA, TAG ou TGA) d'au moins 300 pb. Les ORF trouvées dont les bornes correspondent exactement aux bornes des régions pour lesquelles on a établi la présence d'un gène, sont éliminées.

ARNt

Les ARN de transfert sont également recherchés en deux temps. D'abord par comparaison avec les éléments de type ARNt de la base de données, par application de tRNAScan-SE version 1.23 [Lowe and Eddy, 1997], un logiciel d'annotation d'ARNt. Le fait d'avoir des génomes de type chloroplastiques peut ici aider à déterminer l'origine des ARNt qui seront annotés.

ARNr

Les ARN ribosomiques sont détectés par comparaison avec les éléments de type ARNr contenus dans la base de données.

Duplications

Les duplications sont déterminées par l'alignement du génome face à lui même grâce à YASS. Tous les fragments d'au moins 500 pb retrouvés sont alors considérés comme dupliqués.

3.3.2 Interface

Sur la page d'annotation (Figure 3.6), l'utilisateur entre son fichier *fasta* dans le champ prévu à cet effet, sélectionne le nom de l'espèce qu'il veut annoter et le type d'organite. A titre d'exemple, nous avons effectué l'annotation d'un des génomes mitochondriaux de betterave, séquencé par le Génoscope. Pour simplifier l'affichage des résultats, nous avons soumis ce génome uniquement contre les deux autres génomes de *Beta vulgaris* présents dans la base.

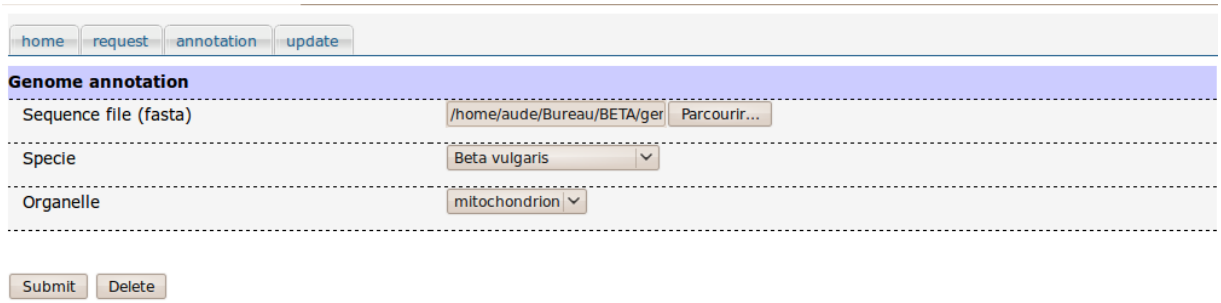
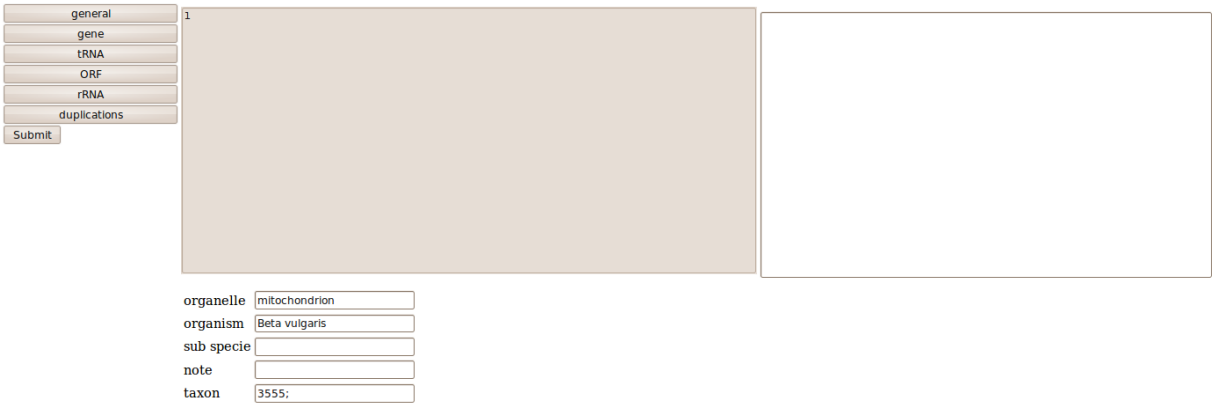


FIG. 3.6 – Interface d'annotation de PLAMIDB, première phase.

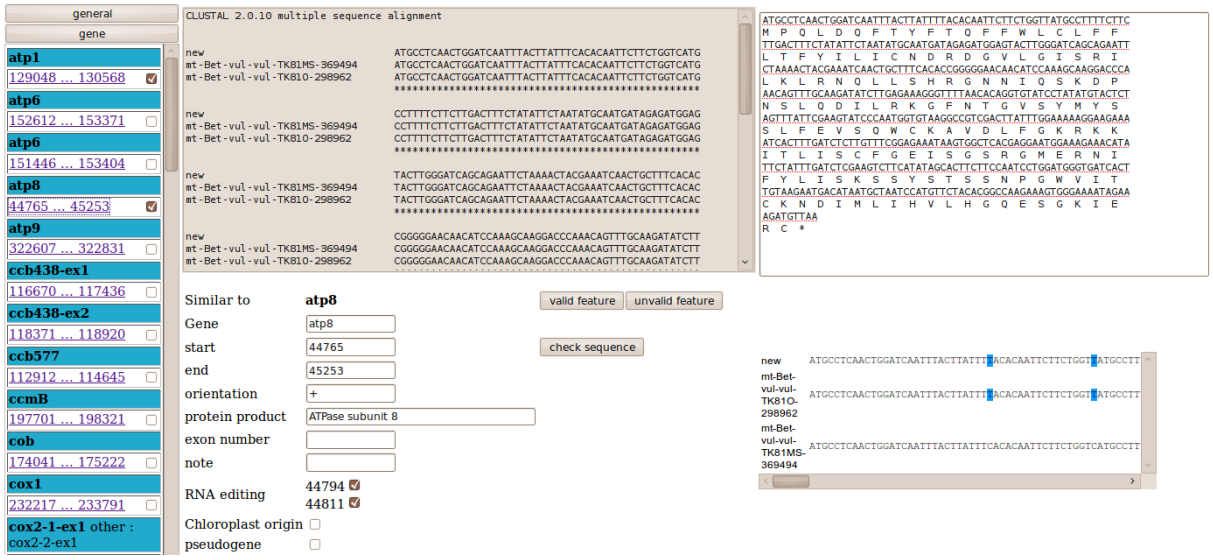
Une fois la soumission effectuée, l'utilisateur est redirigé vers une nouvelle page (Figure 3.7(a)) affichant les informations générales relatives au génome. Les champs *organelle*, *organism* et *taxon* sont pré-remplis suite aux informations données dans le formulaire précédent mais il est tout à fait possible de les modifier. L'utilisateur peut ensuite ouvrir et fermer les volets correspondants aux différents types d'éléments (gènes, ORF, ARNt, ARNr et duplications). Par exemple, lorsqu'il ouvre le volet correspondant aux gènes, il a accès à tous les gènes retrouvés dans le génome. Lorsqu'il clique sur un des gènes de la liste (*atp8* dans la Figure 3.7(b)) la page se décompose en quatre parties. En haut à gauche, l'alignement multiple de tous les gènes orthologues est présenté. En haut à droite, il est possible de voir la séquence correspondante du génome à annoter avec sa traduction en protéines. Ce cadre s'adapte (séquences nucléiques et protéiques) si l'utilisateur décide de modifier les bornes du gène (contenues dans le cadre en bas à gauche). Il s'adapte également lorsque l'on coche ou décoche les sites édités. Les champs du cadre en bas à gauche sont pré-remplis mais totalement modifiables par l'utilisateur. Le cadre en bas à droite permet une visualisation rapide des sites édités connus le long des séquences nucléiques (ils sont indiqués en bleu). Les autres volets sont présentés de la même façon, pour les ARNt, ARNr et duplications, il n'y a pas de traduction en protéine de la séquence, ni de champ *protein product*. Il n'y a pas non plus de sites édités. Lorsque l'utilisateur est satisfait, il peut soumettre ce formulaire et un fichier de type *embl* lui est alors renvoyé.

3.4 Conclusion

Nous avons donc procédé à la mise en place d'une base de données, PLAMIDB, dédiée à la comparaison des génomes mitochondriaux de plantes. Tous les génomes ne sont pas encore intégrés dans cette base, il reste notamment des génomes de plantes monocotylédones qui sont à vérifier avant leur intégration, c'est-à-dire qu'il faut vérifier que les champs des fichiers soient correctement remplis et les modifier si nécessaire. Nous avons également intégré des génomes chloroplastiques dans PLAMIDB qui s'avèrent être utiles lors de la recherche d'homologies d'ARNt. Nous avons de plus ajouté, à la fonction primaire de comparaison de la base, une fonction d'annotation des génomes. Comme les génomes mitochondriaux de plantes sont très peu mutés, il est possible de procéder à l'annotation d'un génome en utilisant uniquement d'autres génomes mitochondriaux de plantes, une fonction dont l'intérêt a été confirmé lorsque nous avons dû annoter les génomes mitochondriaux de betterave séquencés par le Génoscope.



(a) Exemple d'annotation, menu général.



(b) Exemple d'annotation, détail du gène *atp8*. Cette annotation est visible sur quatre cadres. En haut à gauche est présenté l'alignement de la séquence avec les gènes *atp8* contenus dans la base de données. En haut à droite, la traduction de la séquence en protéines. En bas à gauche, les champs d'annotation de ce gène (nom, positions start et stop, orientation, *etc.*). En bas à droite, les sites édités trouvés dans cette séquence par rapport aux annotations contenues dans la base de données.

FIG. 3.7 – Interface d'annotation de PLAMIDB, deuxième phase.

Chapitre 4

Analyse des réarrangements génomiques chez le maïs

Ce chapitre est dédié au premier travail d'analyse de la structure de génomes mitochondriaux que nous avons réalisé. L'étude a porté sur huit génomes : cinq génomes de maïs et trois de téosintes. Ce travail fait l'objet d'une publication dans *BMC Genomics*.

Les génomes de maïs ont été publiés récemment par Allen et ses collègues [Allen et al., 2007]. L'étude menée dans cette publication s'intéressait essentiellement à la description des génomes et de leur contenu avec une emphase particulière sur la description des nombreuses régions dupliquées dans chacun des génomes. Du point de vue des réarrangements, les auteurs décrivent les génomes comme étant fortement réarrangés sans proposer ni scénario ni histoire évolutive. Cet ensemble de génomes constituait un jeu de données parfait pour aborder le problème d'étude de l'architecture de génomes au niveau intraspécifique. A ces cinq génomes nous avons ajouté trois autres génomes, dont deux allaient servir de groupe externe, permettant à la fois d'orienter l'évolution mais aussi d'identifier des événements de duplication ancestraux.

Comme nous l'avons déjà décrit, il n'existe pas d'outil d'adapté à la comparaison de génomes comportant de marqueurs dupliqués, permettant de garder l'intégralité des marqueurs dupliqués dans ses analyses. Or dans ces génomes, de tels marqueurs se sont avérés être nombreux, composant 3,4% à 31,5% des génomes (Tableau 1), et donc nécessaires pour l'étude des réarrangements d'où notre volonté de mettre en place une stratégie permettant de les conserver. Remarquant que les duplications chez le maïs se trouvaient souvent en tandem (Figure 1) et que ce phénomène est décrit dans l'évolution des génomes mitochondriaux animaux, nous avons mis au point une méthode permettant de condenser et de distinguer les duplications, le but étant de n'avoir qu'une seule copie de chaque marqueur dupliqué afin d'utiliser les outils d'analyse d'événements de réarrangements.

Dans un premier temps, les séquences de marqueurs pour chaque génome ont été construites. Pour cela, nous avons procédé en deux temps. Tous d'abord nous avons extrait tous les marqueurs issus de l'annotation des génomes (gènes, ARN, ORF) communs à tous les génomes. Nous avons ensuite ajouté à ces séquence de marqueurs, des marqueurs situés dans des régions non codantes, détectés en utilisant le logiciel Mauve. Cet ajout s'est fait en deux étapes, étant donné que Mauve ne tient pas compte des région dupliquées. La première étape a consisté à extraire les marqueurs communs à tous les génomes en considérant les génomes sans duplicats. La deuxième étape fût de réintroduire les marqueurs dupliqués en utilisant l'outil YASS sur les séquences entières. Nous avons ensuite conservé tous les marqueurs qui ne chevauchaient pas de marqueurs issus

de l'annotation. Nous avons ainsi obtenu les séquences de marqueurs pour chacun des génomes, contenant les marqueurs communs à tous les génomes ainsi que les marqueurs dupliqués.

Étant donné que les outils d'analyse de réarrangement existants ne tiennent pas compte des marqueurs dupliqués, nous avons mis en place une stratégie pour exploiter les duplications. Nous avons considéré que toutes les duplications visibles dans les génomes provenaient de duplication ancestralement en tandem et la stratégie utilisée fût la suivante : condenser les duplications en tandem détectables dans les génomes puis identifier les paralogues et orthologues lorsque les duplications au sein des génomes ne se trouvaient plus en tandem. Nous avons donc condensé les duplications en tandem retrouvées uniques dans les génomes et donc considérées comme des événements propres à chacun des génomes. Nous obtenons donc, dans ce cas, pour chaque génome concerné, une séquence de marqueurs correspondant à une séquence ancestrale avant les événements de duplications en tandem et de pertes de certains marqueurs dans ces duplications (Figure 2). D'autres duplications ont été retrouvées : des duplications (non en tandem) communes à plusieurs génomes ou propres à un génome. Ces duplications furent considérées comme des duplications ancestrales en tandem mais remaniées au cours du temps. Dans ce cas, nous avons effectué une analyse de groupes de marqueurs afin de distinguer les marqueurs orthologues et paralogues. Cette méthode, présentée en Figure 3, nous a permis de condenser et distinguer la quasi-totalité des marqueurs dupliqués. Ils nous a ainsi été possible d'effectuer des phylogénies de réarrangements par Neighbor-Joining sur les distances d'inversion (Figure 5.B) et par parcimonie (Figure 6).

Nous avons mis en place un système de robustesse des nœuds de l'arbre de distance d'inversion, obtenu par Neighbor-Joining, en utilisant une méthode de Jackknife. La méthode consiste à ne garder qu'un pourcentage des éléments de la séquence (ici les marqueurs) et de construire un arbre phylogénétique avec ce nouveau jeu de données. Dans notre cas, nous avons généré 1000 séquences composées de 90% des marqueurs et ainsi construit 1000 arbres phylogénétiques. Le pourcentage, sur ces 1000 arbres de chaque nœud retrouvé identique à un des nœuds de l'arbre de distance d'inversion (composé de 100% des marqueurs) constitue alors la valeur de robustesse de ce nœud. Les nœuds de l'arbre de distances d'inversion sont relativement bien supportés, la plus faible valeur étant de 79,6% lorsque nous faisons un Jackknife à 90%.

De plus, les phylogénies de réarrangements obtenues, que ce soit par Neighbor-Joining sur les distances d'inversions ou par parcimonie, se sont révélées être congruentes avec une phylogénie de séquence que nous avions établie sur une large portion de séquences communes de ces génomes (Figure 5.A). Cette phylogénie de séquence a été réalisée sur la concaténation de l'ensemble des séquences communes aux génomes (codantes ou non) étant donné que le taux de substitution au niveau des séquences codantes pour des protéines est très faible. Nous avons donc cherché ici à maximiser l'information de séquence pour construire notre phylogénie. La phylogénie obtenue sur les séquences est elle aussi relativement bien supportée, que ce soit par les analyses de bootstrap effectuées avec BIONJ (Neighbor-Joining) ou TREE-PUZZLE (maximum de vraisemblance). La congruence obtenue entre phylogénies de séquences et de réarrangements montre que la méthode de tri des duplications, basée sur l'hypothèse de duplications en tandem remaniées au cours de l'évolution des génomes, est une bonne approche pour les génomes mitochondriaux de maïs.

La suite de ce chapitre reprend donc en détail ce que nous venons de décrire sous la forme de l'article que nous avons publié dans *BMC Genomics*. Les figures et tableaux supplémentaires sont données en Annexe Chapitre 4, pages 156 à 161.

Darracq *et al. BMC Genomics* 2010, **11**:233
http://www.biomedcentral.com/1471-2164/11/233

BMC Genomics

RESEARCH ARTICLE Open Access

A scenario of mitochondrial genome evolution in maize based on rearrangement events

Aude Darracq[1,2,3], Jean-Stéphane Varré[2,3], Pascal Touzet[1*]

Abstract

Background: Despite their monophyletic origin, animal and plant mitochondrial genomes have been described as exhibiting different modes of evolution. Indeed, plant mitochondrial genomes feature a larger size, a lower mutation rate and more rearrangements than their animal counterparts. Gene order variation in animal mitochondrial genomes is often described as being due to translocation and inversion events, but tandem duplication followed by loss has also been proposed as an alternative process. In plant mitochondrial genomes, at the species level, gene shuffling and duplicate occurrence are such that no clear phylogeny has ever been identified, when considering genome structure variation.

Results: In this study we analyzed the whole sequences of eight mitochondrial genomes from maize and teosintes in order to comprehend the events that led to their structural features, i.e. the order of genes, tRNAs, rRNAs, ORFs, pseudogenes and non-coding sequences shared by all mitogenomes and duplicate occurrences. We suggest a tandem duplication model similar to the one described in animals, except that some duplicates can remain. This model enabled us to develop a manual method to deal with duplicates, a recurrent problem in rearrangement analyses. The phylogenetic tree exclusively based on rearrangement and duplication events is congruent with the tree based on sequence polymorphism, validating our evolution model.

Conclusions: This study suggests more similarity than usually reported between plant and animal mitochondrial genomes in their mode of evolution. Further work will consist of developing new tools in order to automatically look for signatures of tandem duplication events in other plant mitogenomes and evaluate the occurrence of this process on a larger scale.

Background

All organelle genomes found in mitochondria of plant or animal cells are considered to have originated from an endosymbiotic form of α-Proteobacteria, and given rise to the emerging eukaryotic cell more than 10^9 years ago [1]. Despite their monophyletic origin, animal and plant mitochondrial genomes (mitogenomes) exhibit contrasted features, when considering size, compactness, mutation rate and gene-order variation [2]. Most animal mitogenomes are circular and compact, share the same gene content and have a size that does not exceed 20 kb. The high nucleotide mutation rate of their coding sequences has been commonly used in population genetic and phylogenetic studies [3]. However, in taxonomic studies, the introduction of gene-order variation to resolve specific nodes has proved to be a powerful tool [4]. In these animal rearranged mitogenomes, most gene rearrangements were due to inversions and translocations. But duplication events were also identified: in some cases, they were distant in the genomes, with or without loss of parts of the duplicated fragment [5]. In other cases, duplications occurred in tandem repeat and were followed either by non-random duplicate loss (cases of genes conserved side by side in the same orientation [6,7]) or by random loss (known as TDRL, Tandem Duplication with Random Loss) [8-10]. In most cases, when duplication involved a protein coding gene, only one functional copy remained.

In contrast, plant mitogenomes exhibit larger size (most are from 200 to 700 kb) and are less compact than their animal counterparts due to the occurrence of non-coding sequences and duplicated fragments. Moreover, plant mitogenomes are known to evolve rapidly in

* Correspondence: pascal.touzet@univ-lille1.fr
[1]Laboratoire de Genetique et Evolution des Populations Vegetales, UMR CNRS 8016, Universite Lille 1, 59655 Villeneuve d'Ascq Cedex, France

structure and slowly in sequence [11]. The occurrence of large repeated sequences has led to the idea of a complex genome, composed of alternative master chromosomes and sub-genomic molecules due to intragenomic recombination [12], even though whole sequenced genomes are usually represented as circular master circles [13,14]. At the intra-specific level, recombination through small repeat sequences is believed to be responsible for large gene-order shuffling and the emergence of new open reading frames, some of which have been involved in Cytoplasmic Male Sterility (CMS) [14,15]. In this context, the acquisition of whole sequence data for several mitogenomes found in a species opens new venues toward a better understanding of the evolutionary dynamics of this peculiar genome, especially when focusing on its high structural rearrangement rate and the origin of duplicated fragments.

The comparison of whole genomes using gene order has been an active field of research since the early 1990s. The first methods focused on the study of the minimal number of rearrangement events, mostly inversions, to go from one genome to another [16,17]. The resulting scenario could be seen as a putative evolutionary scenario. Phylogenetic reconstruction methods based on rearrangement events have also been proposed in order to compute scenarios and putative ancestors for a set of genomes [18,19]. Methods to study rearrangements that take duplicates into account have been investigated over the past decade. Since most of the mathematical models used to compute rearrangement distances and scenarios are based on the assumption that each gene or synteny block appears exactly once in each genome, methods designed for genomes without duplicates cannot be applied directly to plant mitochondrial genomes. One possible approach consists in keeping only one of the duplicates and removing the others from the genomes in order to obtain a dataset with one copy of each gene per genome [20,21]. The drawback of this solution is its high combinatorics if the number of duplicates is large. Moreover it does not provide any kind of explanation about duplication events. Other methods focus on the study of gene families, i.e. the evolutionary history of a gene and its duplicates [22]. The aim of these methods is to find the duplication events within a given phylogenetic tree. It follows that currently no method is able to reconstruct a rearrangement phylogenetic tree of genomes with duplicates. Therefore the 'manual approach' has to be used for resolving this type of evolutionary history [23].

Recently, Allen and colleagues [24] reported the whole sequencing of 5 mitogenomes in maize. As expected, the mitogenomes exhibited a large variation in size (from 535 to 740 kb) due mainly to large duplicated fragments, and gene shuffling was such that no clear evolutionary scenario could be pictured. However, on the basis of nucleotide divergence, groups of related mitogenomes could be defined and qualified as ancestral or derived though no phylogeny could be established. In the present study, we added three newly available whole mitogenome sequences of teosintes, species that are relatives of maize, to the five mitogenomes studied by Allen and colleagues [24] in order to conduct a phylogenetic analysis and ultimately comprehend the events that led to their structural features: sequence order and duplicates.

The analysis based on sequence polymorphism among the eight mitogenomes enabled us to build a robust reference tree for subsequent analyses solely based on genome structure information (sequence-order). We showed that mitogenome rearrangements could result from a mechanism similar to that found in animals, i.e. tandem duplication, but where some duplicates were partially lost. Using this evolution model, we developed a methodology to reconstruct a phylogeny based on rearrangement events that integrated most duplicates, and ended up with an evolutionary scenario of the mitochondrial genome in maize.

Results

Genome duplications

The analysis of maize and teosinte mitogenomes revealed the occurrence of duplications. Duplication length varied from 0.54 kbp to 120 kbp (Table 1). Duplicated fragments were an important part of the total genome length for the longest genomes, 23.4% for NA, 31.5% for CMS-C and 21.2% for *Zea mays* ssp. *parviglumis*, and more generally were the main cause of size differences among maize mitogenomes [24]. Six duplicated fragments were shared between maize [24] and *Zea mays ssp. parviglumis* mitogenomes : {NA, NB, CMS-C, CMS-S and *Zea mays* ssp. *parviglumis*} shared two duplications (11 and 17 kbp), {NA, NB, CMS-S and *Zea mays* ssp. *parviglumis*} a 0.7 kbp duplication, {NA, CMS-S, CMS-T and *Zea mays* ssp. *parviglumis*} a 5.3 kbp duplication, {NA, NB and *Zea mays* ssp. *parviglumis*} another 5.3 kbp duplication and {NA and *Zea mays* ssp. *parviglumis*} a 0.6 kbp duplication.

NA seemed to have a fragment duplicated in tandem, the two copies of its 120 kbp fragment were separated by only 9.3 kbp.

Backbone and genome structures

Backbone DNA sequences

Total backbone DNA sequence (including genes) represented a concatenation of all common fragments between all mitogenomes when considering only one copy of each duplicated sequence. Overall, in maize, *Zea mays ssp. parviglumis, Zea perennis* and *Zea luxurians*

Table 1 Length and percentage of duplicated fragments up to 500 bp

Genome	Genome length (kbp)	Duplication length (kbp)[a]	% of dupl. fragments in genome	Minimal dupl. length (kbp)[b]	Maximal dupl. length (kbp)[b]	Median dupl. length (kbp)	Number of dupl. fragments	Genome length without duplication (kbp)
NA	701.046	163.899	23.4%	0.60	120.0	5.316	8	537.147
NB	569.630	49.407	8.7%	0.54	17.0	8.183	6	520.223
CMS-C	739.719	232.947	31.5%	5.70	105.0	31.009	6	506.772
CMS-S	557.162	45.023	8.1%	0.72	17.0	4.589	8	512.139
CMS-T	535.825	25.884	5.3%	2.60	12.8	5.270	4	509.941
parvi	680.603	143.928	21.2%	0.60	55.0	8.207	8	536.675
lux	539.368	18.561	3.4%	1.70	10.1	6.737	3	517.175
per	570.354	53.719	9.3%	6.30	13.6	11.809	5	520.807

[a]All duplicated copies less one
[b]Length of one copy

mitogenomes, coding genes (including duplicated genes) represented 7.83 to 8.60% (median = 8.37%) of the total genome length whereas backbone DNA sequences represented 56.49 to 77.99% (median = 73.29%) of the total genome length (Table 2).

In all, there were 115 orthologous fragments over all mitogenomes (see Additional file 1). The smallest common fragment size was 94 bp and the largest was around 18,769 bp (median of 2,379 bp). Differences in size between orthologous fragments were due to indels (insertions and deletions). For each mitogenome, backbone sequence size was around 418 kbp, except for *Zea luxurians* with a size of 415 kbp. The multiple alignment length of the eight mitogenome backbones was 421,163 bp (counting gaps). Backbone repartition over *Zea* mitogenomes is given in Figure 1. We computed the gap sizes in the mitogenome sequences from the multiple alignment. Most of the gaps were 5 bp long as previously described by Allen and colleagues [24] and the insertions were small repetitions (data not shown). Compared to the other mitogenomes, *Zea luxurians* had more gaps longer than 5 bp. This mainly explains the backbone length difference between *Zea luxurians* and the other mitogenomes.

Genome structure sequence

The Genome Structure Sequence (GSS) is a block sequence characteristic of each mitogenome. It is built with block markers- which we hereafter call synteny anchors- that are common to all eight mitogenomes. Synteny anchors are composed of protein coding genes, tRNAs, rRNAs, ORFs, pseudogenes and non-coding sequences from the backbone DNA sequences (see Methods). Before paralog identification and synteny anchor collapsing ('bpisac'), GSSs contained 69 synteny anchors. They represented 69.99 to 74.21% of mitogenome lengths (median = 72.88%) (Table 2). Figure 1 provides a schematic view of GSS bpisac repartition over mitogenomes and shows that GSS bpisac uniformly covers all mitogenomes. It must be noted that in GSS, the numbers of synteny anchors correspond to one or more mitogenome markers: when they were systematically located together and in the same order in all eight mitogenomes, they were grouped into a single synteny anchor (see Additional file 2). Consequently there were 69 synteny anchors corresponding to 187 markers common to all mitogenomes. Synteny anchors contained from 1 (e.g. synteny anchor number 1) to 15 markers (e.g. synteny anchor number 59). As is generally the

Table 2 Backbone, GSS (Genome Structure Sequence) and protein coding gene proportions on the mitogenomes

Genome	Mitogenome length (kbp)	% of protein coding gene in mitogenome*	% of GSS length in mitogenome*	% of backbone length in mitogenome*
NA	707.046	8.22	70.44	59.60
NB	569.630	8.44	73.60	73.36
CMS-C	739.719	8.23	73.61	56.49
CMS-S	557.162	8.37	74.21	75.02
CMS-T	535.825	8.36	71.58	77.99
parvi	680.603	7.83	72.74	61.41
lux	539.368	8.60	69.99	76.84
per	570.354	8.51	73.01	73.21

*including duplicates

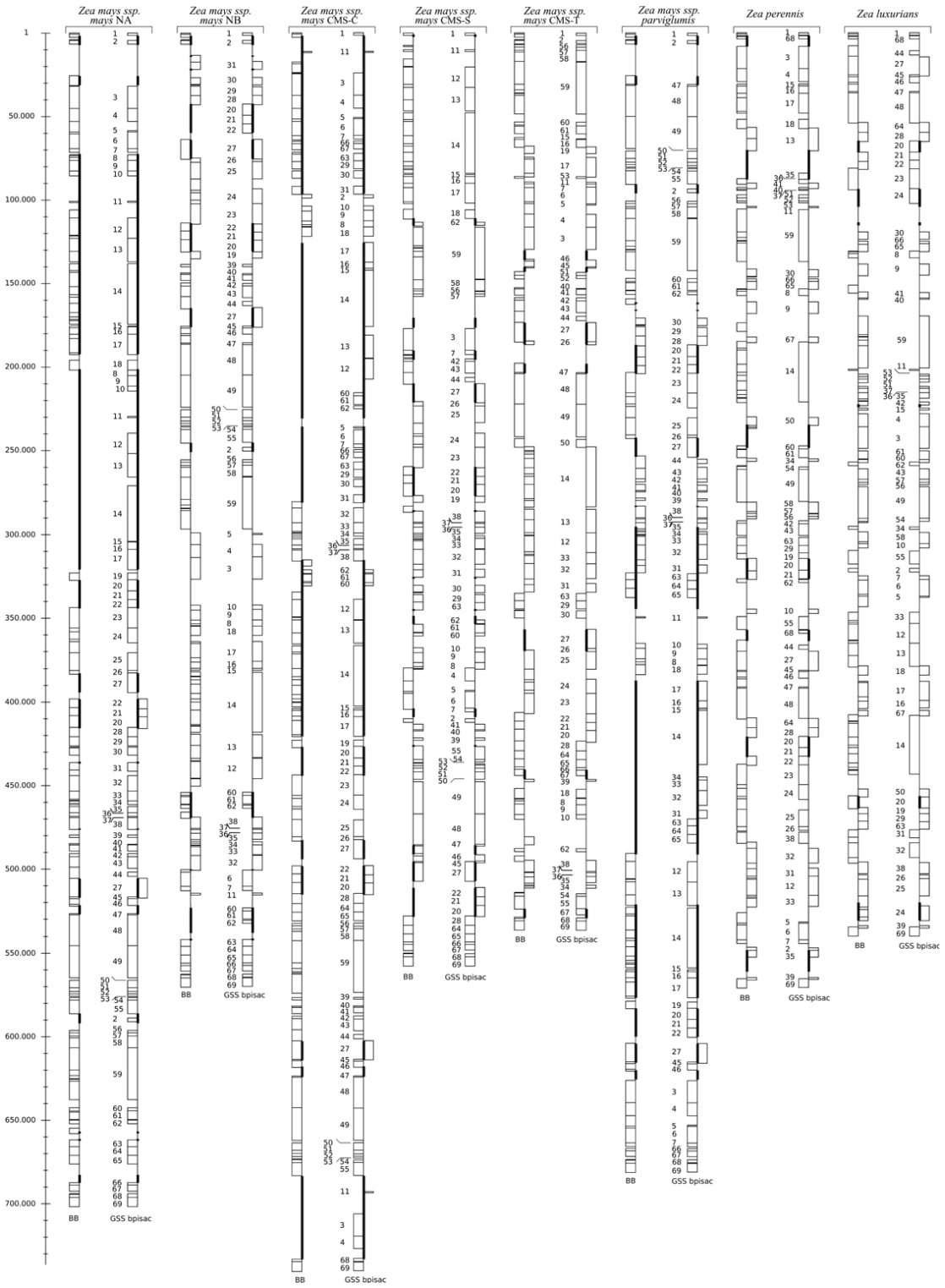


Figure 1 Repartition of Backbone DNA sequences and Genome Structure Sequences (GSSs) Sequences on each mitogenome. For each mitogenome, a pair of box sequences is represented : the backbone DNA sequence (BB) and the genome structure sequence before paralog identification and synteny anchor collapsing (GSS bpisac). Each box is either a Backbone DNA fragment for BB, or a synteny anchor for GSS bpisac. Boxes with the same number are homologous synteny anchors. The numbering of boxes was chosen according to *Zea mays* ssp. *mays* NA. Thus on BB and GSS bpisac, a box is drawn on the left side if it has the same orientation than its homolog in NA, on the right side otherwise. For each mitogenome, thick lines represent duplicated regions.

case in mitochondrial genomes, markers that were systematically grouped in our 8 mitogenomes were not composed of genes involved in the same metabolic pathway. Duplicated synteny anchors represented a large part of mitogenomes, particularly in NA, CMS-C and *Zea mays* ssp. *parviglumis*: 26.4% of the synteny anchors were duplicated in NA, 12.6% in NB, 36.8% in CMS-C, 12.6% in CMS-S, 2.3% in CMS-T, 20.7% in *Zea mays* ssp. *parviglumis*, and 9.2% in *Zea luxurians* and 10.34% in *Zea perennis*.

Using GSSs bpisac and assuming that tandem duplication was the underlying mechanism, we observed that most of the duplicated synteny anchors were indeed located in regions that could result from tandem duplication events. The fact that two regions did not share exactly the same synteny anchor content suggested deletion events of some duplicates after duplication. We called this mechanism Tandem Duplication with Partial Loss (TDPL). A hypothesis of TDPL in *Zea mays* ssp. *parviglumis* is shown in Figure 2.

Following the method described in Figure 3 and Methods (paralog identification and gene collapsing), GSS was obtained for each mitogenome, where duplicates were distinguished and/or collapsed. We identified 4 TDPLs specific to a mitogenome (one in NA, two in CMS-C and one in *Zea mays* ssp. *parviglumis*) where the two duplicates were still side by side, 2 TDPLs shared by some mitogenomes (one shared by all mitogenomes and the other by maize mitogenomes) where the two duplicates were physically separated and 4 tandem duplications specific to a mitogenome and where the copies were physically separated (CMS-S, CMS-T, *Zea luxurians* and *Zea perennis*). For these duplications, we hypothesized that the duplicates (originally in tandem) had been separated by rearrangement events after duplication. In the end, GSSs contained 72 blocks: the 69 original synteny anchors, minus 5 that were eliminated because orthologs and paralogs could not be distinguished, plus 8 additional blocks after paralog/ortholog identification. Transformation from GSS bpisac to GSS for CMS-C and *Zea perennis* is shown in Figure 4.

Sequence phylogeny

Bootstrap values (upper values in Figure 5.A.) indicated that the topology of the tree was relatively robust (from 94-99%) with some uncertainty regarding the separation between CMS-C and the remaining three *Zea mays* mitogenomes (74%). The Maximum Likelihood (ML) phylogenetic tree had the same topology as the NJ phylogenetic tree and bootstrap values (lower values in Figure 5.B.) were higher for all nodes. Molecular clock with ML was rejected ($p < 0.0001$).

We also constructed a phylogenetic tree with concatenated protein coding gene sequences which exhibited the same topology as the one from the backbone sequence but with shorter branch lengths (data not shown).

Rearrangement phylogeny

Phylogenetic analysis was based on rearrangement using GSSs. The phylogenetic tree was congruent with the one from backbone DNA sequence. Jackknife values were 96.1%, 99.5%, 79.6%, 100% and 100% for the five most terminal nodes (Figure 5.B). Tests were performed with different percentages of synteny anchors kept in

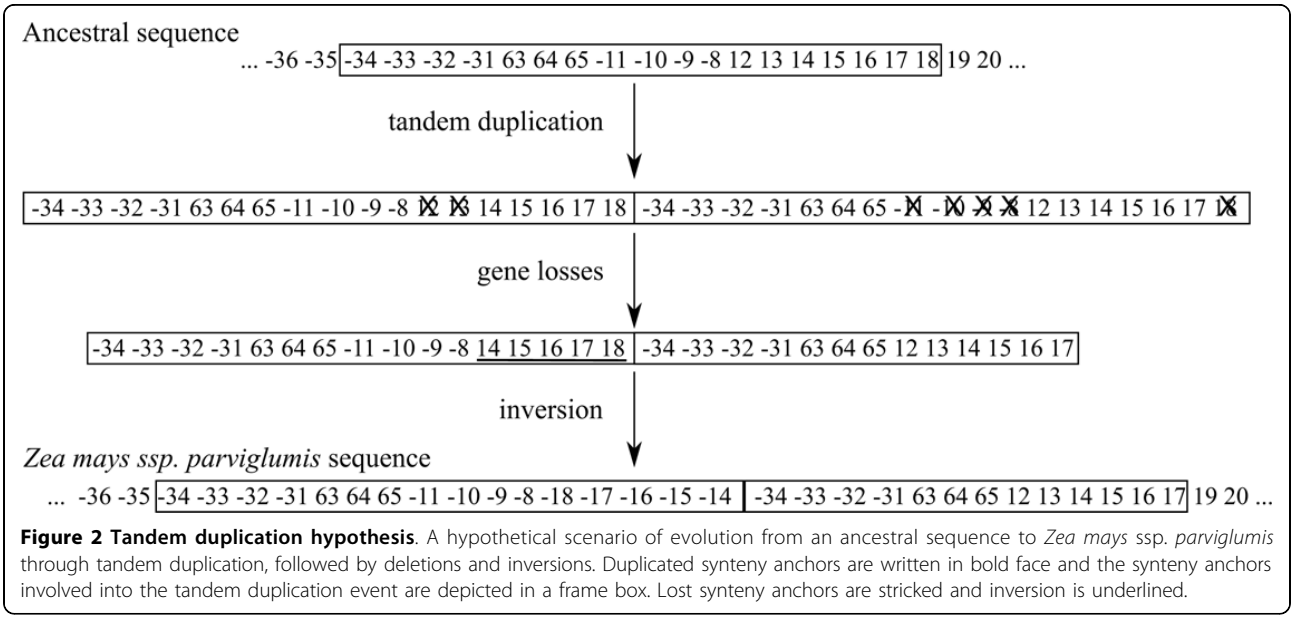


Figure 2 Tandem duplication hypothesis. A hypothetical scenario of evolution from an ancestral sequence to *Zea mays* ssp. *parviglumis* through tandem duplication, followed by deletions and inversions. Duplicated synteny anchors are written in bold face and the synteny anchors involved into the tandem duplication event are depicted in a frame box. Lost synteny anchors are stricked and inversion is underlined.

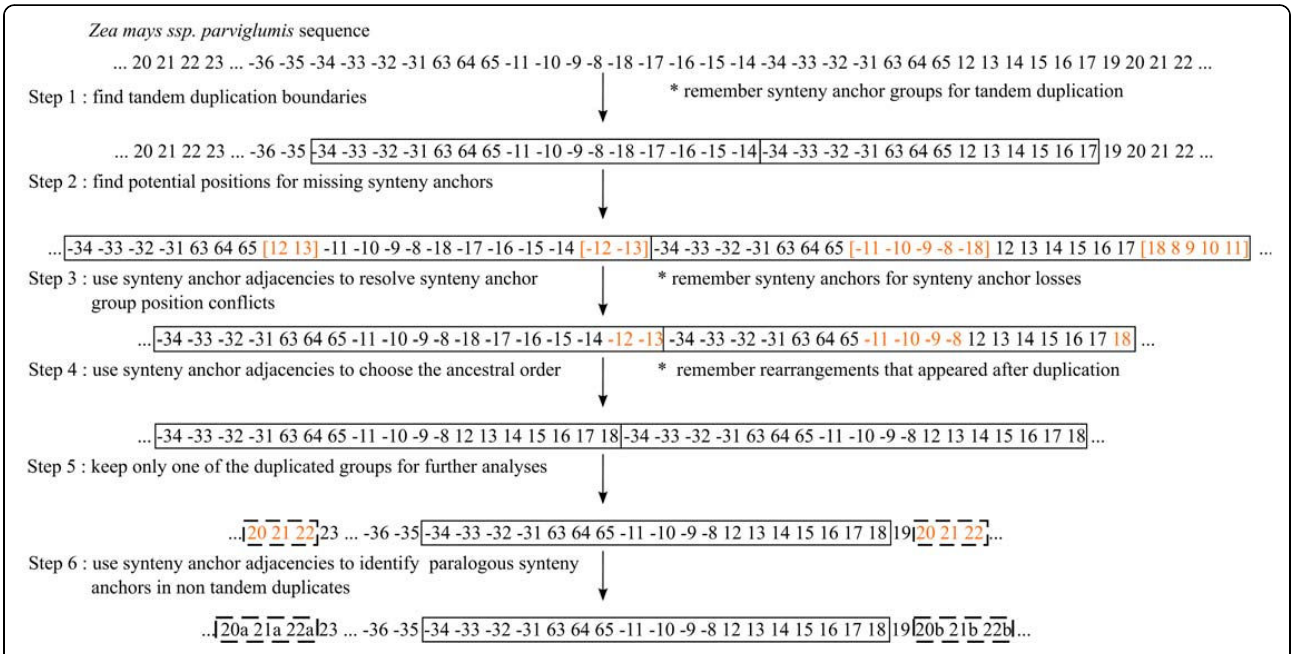


Figure 3 Method to deal with duplicates. Example of paralog identification and synteny anchor collapsing using *Zea mays* ssp. *parviglumis* GSS. Duplicated synteny anchors are in bold.

the jackknife computation (see Additional file 3). When 30 to 90% of the synteny anchors were kept, the main tree was congruent with the sequence tree.

We built a phylogenetic tree with a data set excluding blocks containing duplicated synteny anchors. It is noteworthy that the resulting tree was not congruent with the sequence tree. This highlights the importance of taking into account duplication events in the analysis. Moreover, when we deleted all copies of each duplicated synteny anchors, the data set went down from 72 to 28 synteny anchors.

Mitogenome rearrangement evolution

A parsimonious tree was constructed using MGR (Multiple Genome Rearrangements) with GSSs. This method has the advantage of providing a potential ancestral sequence at each node (A1 to A5) (Figure 6).

It was possible to reintroduce duplication events in the MGR tree. Indeed, duplication of synteny anchors {8 9 10 11 12 13 14 15 16 17 18} can be put on the NA branch, duplication of synteny anchors {5 6 7 66 67 63 29 30 31 32 33 34 35 36 37 38 -10 -9 -8 12 13 14 15 16 17 18}and {69 1 -11 3 4} on the CMS-C branch and duplication of synteny anchors {-34 -33 -32 -31 63 64 65 -11 -10 -9 -8 12 13 14 15 16 17 18} on the *Zea mays* ssp. *parviglumis* branch. These duplication events were followed by synteny anchor loss and inversions (as described in Figures 2 and 3). It was then possible to obtain a parsimonious evolutionary history of all eight mitogenomes. Likely events were positioned on each branch of the tree where (I) denotes an inversion, (TD) a tandem duplication, (TDPL) a tandem duplication with partial loss and (L) a loss. In Figure 7, an example of an evolutionary scenario is given from A5 to *Zea mays* ssp. *parviglumis* and NA.

It must be noted that some rearrangement events need not occur in an absolute order except for overlapping inversions, TDPLs and the last inversion in CMS-C and *Zea mays* ssp. *parviglumis*. It appears that overlapping inversions must be chronologically oriented in the evolution history: for example, from A5 to *Zea mays* ssp. *parviglumis*, inversion I{-31:-59} has to occur before inversion I{47:-3}. However, non-overlapping inversions can be permuted: for example, I{-20b:-20a} can occur either before or after I{-31:-59}. Two duplications have an ancestral position: TD{20:22} is ancestral to maize and teosinte mitogenomes and TD{27} is specific to maize mitogenomes (Figure 6). Over time, the duplicates were separated.

It is important to note that scenarios for all mitogenomes, computed by MGR, were consistent with rearrangement sites (i.e. breakpoint regions) observed at the sequence level by Allen and colleagues [24]. Indeed, many rearrangements predicted by MGR occurred between the second copies of synteny anchors 20 and 21 (*trn*N and orf99 in the region 140 kbp of NB); we also found rearrangement points near synteny anchors 4, 7, 9 and 18 (respectively *cob*, *nad*2 exon 1, *rbc*L and *cox*1 genes) whereas we did not find any rearrangements between synteny anchors 27 (first copy) and 21 (second

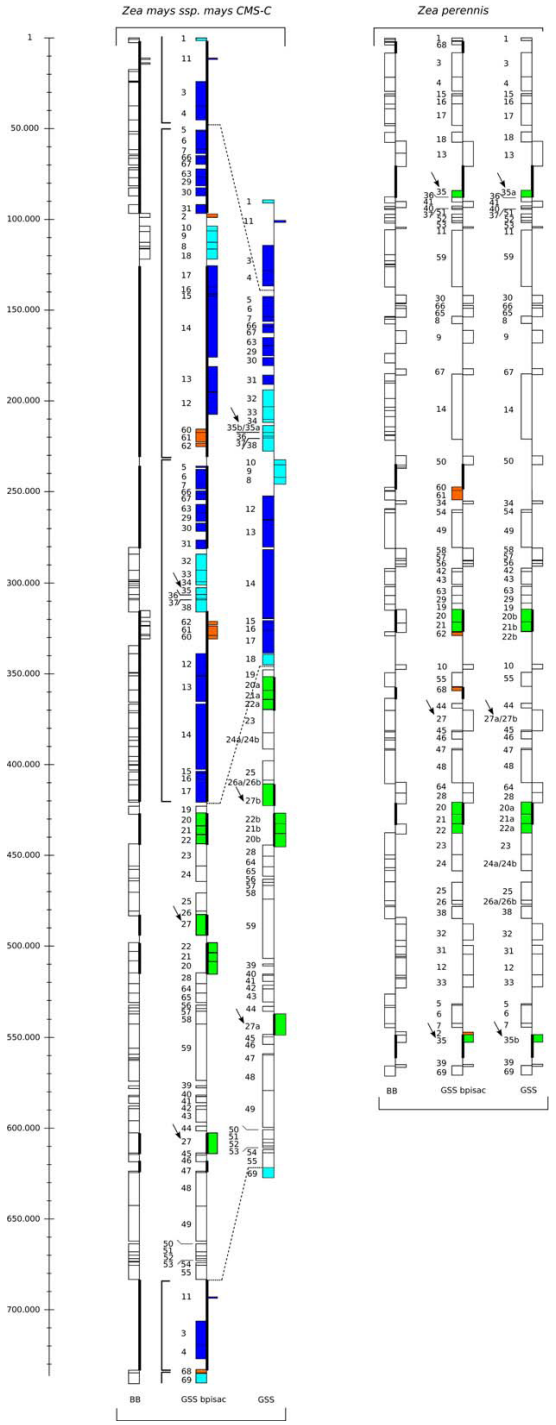


Figure 4 Backbone DNA sequence, GSS bpisac and GSS. Backbone DNA sequence (BB), GSS bpisac and GSS blocks repartition along CMS-C and *Zea perennis* mitogenomes. In CMS-C, dashed lines between GSS bpisac and GSS indicate the condensation of tandem duplicated synteny anchors (after the "collapsing" step). Vertical lines indicate each duplicated part. In CMS-C GSS, synteny anchors 35a and 35b are virtually added compared to GSS bpisac because synteny anchor 35 is duplicated in *Zea perennis* GSS bpisac and it is possible to distinguish them with their neighborhood. Conversely, 27 virtually duplicated (i.e. 27a and 27b are left side by side) in *Zea perennis* GSS is in two distinct copies in CMS-C GSS. For synteny anchor 40 duplicated in *Zea perennis* GSS bpisac, ortholog and paralog cannot be distinguished with their neighborhood. Consequently, synteny anchor 60, 61, 62 and 2 are deleted in GSS for all mitogenomes. Synteny anchors 27, 27a, 27b, 35, 35a and 35b are indicated by arrows. We applied the following color code: orange for deleted blocks, green for blocks for which paralogous from orthologous were distinguished, blue for tandemly duplicated blocks: dark blue when duplicates were conserved, light blue when one copy was lost.

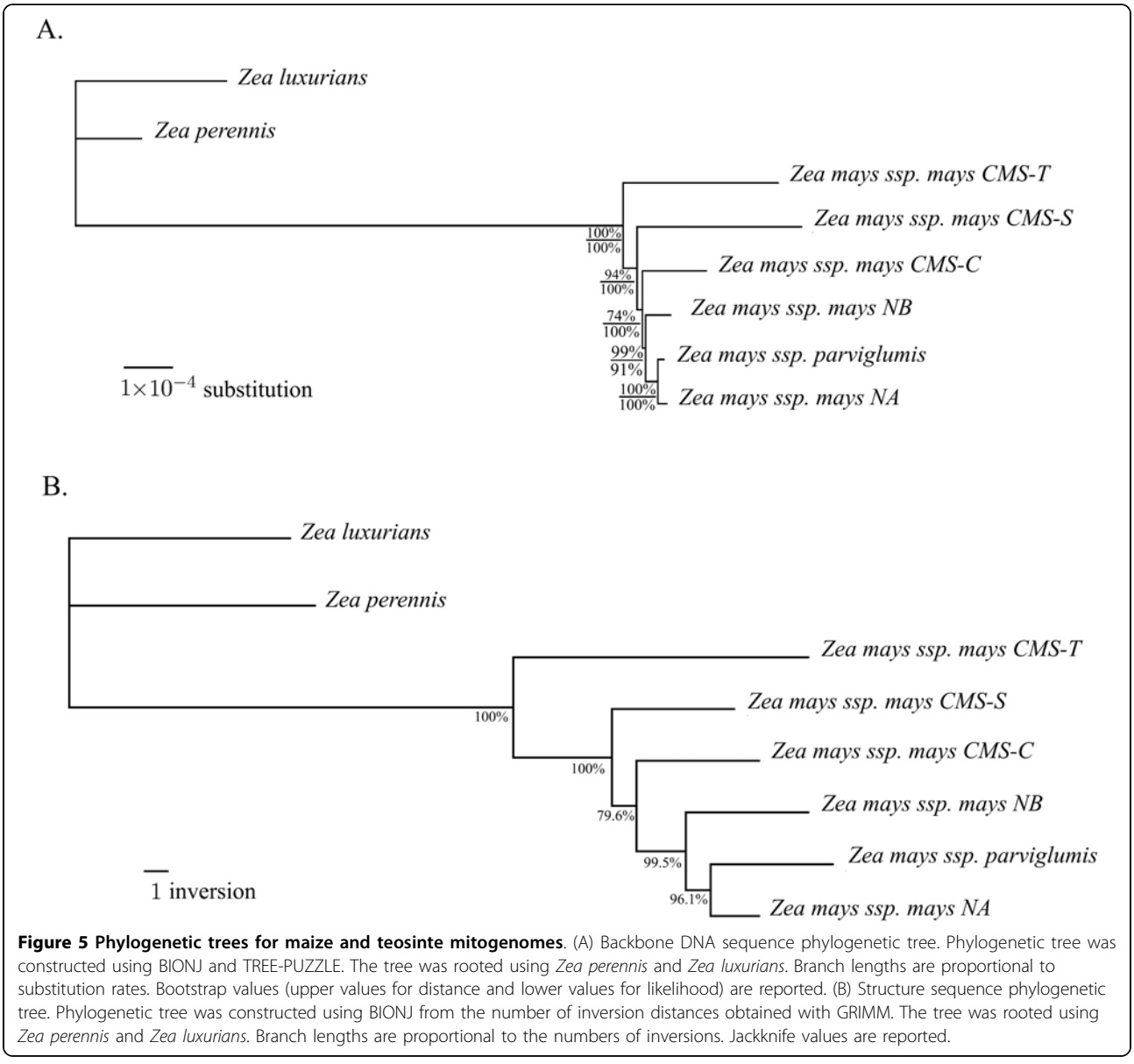


Figure 5 Phylogenetic trees for maize and teosinte mitogenomes. (A) Backbone DNA sequence phylogenetic tree. Phylogenetic tree was constructed using BIONJ and TREE-PUZZLE. The tree was rooted using *Zea perennis* and *Zea luxurians*. Branch lengths are proportional to substitution rates. Bootstrap values (upper values for distance and lower values for likelihood) are reported. (B) Structure sequence phylogenetic tree. Phylogenetic tree was constructed using BIONJ from the number of inversion distances obtained with GRIMM. The tree was rooted using *Zea perennis* and *Zea luxurians*. Branch lengths are proportional to the numbers of inversions. Jackknife values are reported.

copy) (*nad*1 exon 1 and *rps*3 exon 1 in the region 65 to 140 kbp of NB).

Discussion

We analyzed the evolution of mitochondrial genome structure within a plant species by concomitantly building a phylogenetic tree based on sequence polymorphism and a phylogenetic tree based on structural rearrangements among genomes. Both trees were congruent, suggesting that both sources of polymorphism are correlated, i.e. the more divergent a genome is, the more rearranged it is. Therefore it was possible to reconstruct an evolutionary scenario, suggest ancestral genome structures along the different nodes of the tree, and pinpoint tandem duplication as a possible mechanism in the important gene shuffling of plant mitochondrial genomes.

Methodology to deal with duplicates

From a methodological point of view, dealing with duplicates together with rearrangement events is a challenge. If one was able to distinguish between paralogous and orthologous synteny anchors, the problem would be reduced to the study of rearrangements with exactly one copy of each synteny anchor in each genome. Unfortunately, finding paralogous synteny anchors is usually a very difficult task (this is especially the case with the data analyzed here since mitochondrial synteny anchor duplicates are identical in most cases). Even if one was able to distinguish them, it

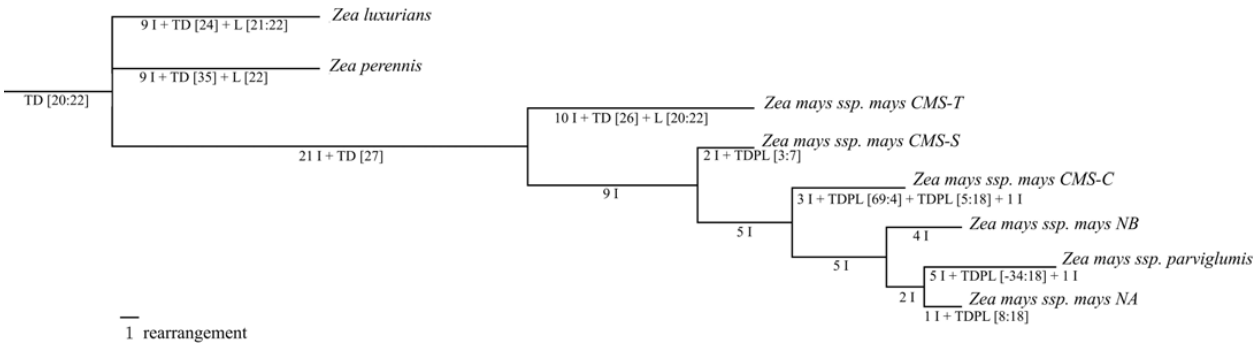


Figure 6 Maize and teosinte mitogenomes evolutionary tree. Parsimonious phylogenetic tree (built with MGR) using GSSs. Branch length is proportional to the number of rearrangement events (inversion, duplications, deletions). An Ancestral sequence is given for each node (A1 to A5). Rearrangement events are given on each branch: I = inversion, TD = tandem duplication, L = synteny anchor loss, TDPL = tandem duplication with partial loss. Numbers between vertical lines correspond to blocks affected by rearrangements.

remains that some duplicates are specific to a given genome or to a subset of genomes. Different methods have been proposed to deal with such datasets. In the exemplar model [20], only one copy of each duplicate is kept. In the maximum matching model [25], one keeps as many copies as the minimum number of copies of one duplicate found in a genome. The choice of which copy to keep is made according to an optimization function. For genome rearrangement purposes, this function consists in choosing the copies that minimize the evolutionary distance between two genomes. But such methods can be applied only if the number of duplicates remains small, otherwise the number of reduced genomes is too large. This is the case with our data. The exemplar genome approach would have led us to explore more than 16 million datasets from our eight mitogenomes. In the special case of tandem duplications, a method was previously described with random loss (TDRL) [26]. Unfortunately, in that model exactly one of each duplicate is immediately lost just after the duplication event, and the method proposed cannot be adapted because the underlying algorithms require that each marker synteny anchor be present only once in each genome.

Therefore we proposed a framework to analyze the rearrangement history of a set of genomes containing duplicates. In this framework we assumed that most of the duplicates came from tandem duplication events and that rearrangements occurring within a duplicated segment were independent from the other rearrangement events. Although this is not necessarily true in the general case, we found evidence of tandem duplication in parts of the genome. These hypotheses provided a means to deal with duplications and to allow us to propose both a scenario of rearrangements and a history of duplication events. We thus elaborated a four step method to account for duplicates. In short, we concealed duplicates to compute rearrangement scenarios and then we reintroduced them. The method was the following : i) identify TDPLs and collapse them in order to keep one copy of each synteny anchor, ii) distinguish between paralogs and orthologs for remaining duplicated synteny anchors, iii) apply the usual rearrangement algorithms (since no duplicate remains), iv) expand the previously collapsed TDPLs in step i) to recreate the TDPL event. The main difficulty of the first step is to correctly determine the boundary of the duplicated segment. We saw that using the information of the synteny anchor neighborhood shared by the genomes could help determine these boundaries (see Methods and Additional files 4). The second step proved to be more difficult since we had to deal with the problem of ortholog versus paralog identification. We supposed that the number of duplicated blocks involved in a TDPL but far apart from each other was rather limited and that the methods described above could thus be used. In this last case, though, the neighborhood could help distinguish between both duplicates (such as block 27 in the dataset). When the duplicates were not in tandem, we added the duplicated block in tandem with its counterpart in genomes in which it was missing because the block content had to be the same for the third step of the method. This did not change the distances among genomes nor did it modify scenarios. Indeed, adding the duplicated block next to its counterpart created an adjacency that was implicitly conserved when computing parsimonious scenarios. The last step consisted of replacing the collapsed TDPLs by their original block sequences. The duplication events were placed on the tree depending on whether TDPL was shared by several genomes or not. If the TDPL was specific to one genome, the duplication event necessarily occurred after the last speciation event. If a TDPL was shared by two or more genomes, the most parsimonious hypothesis

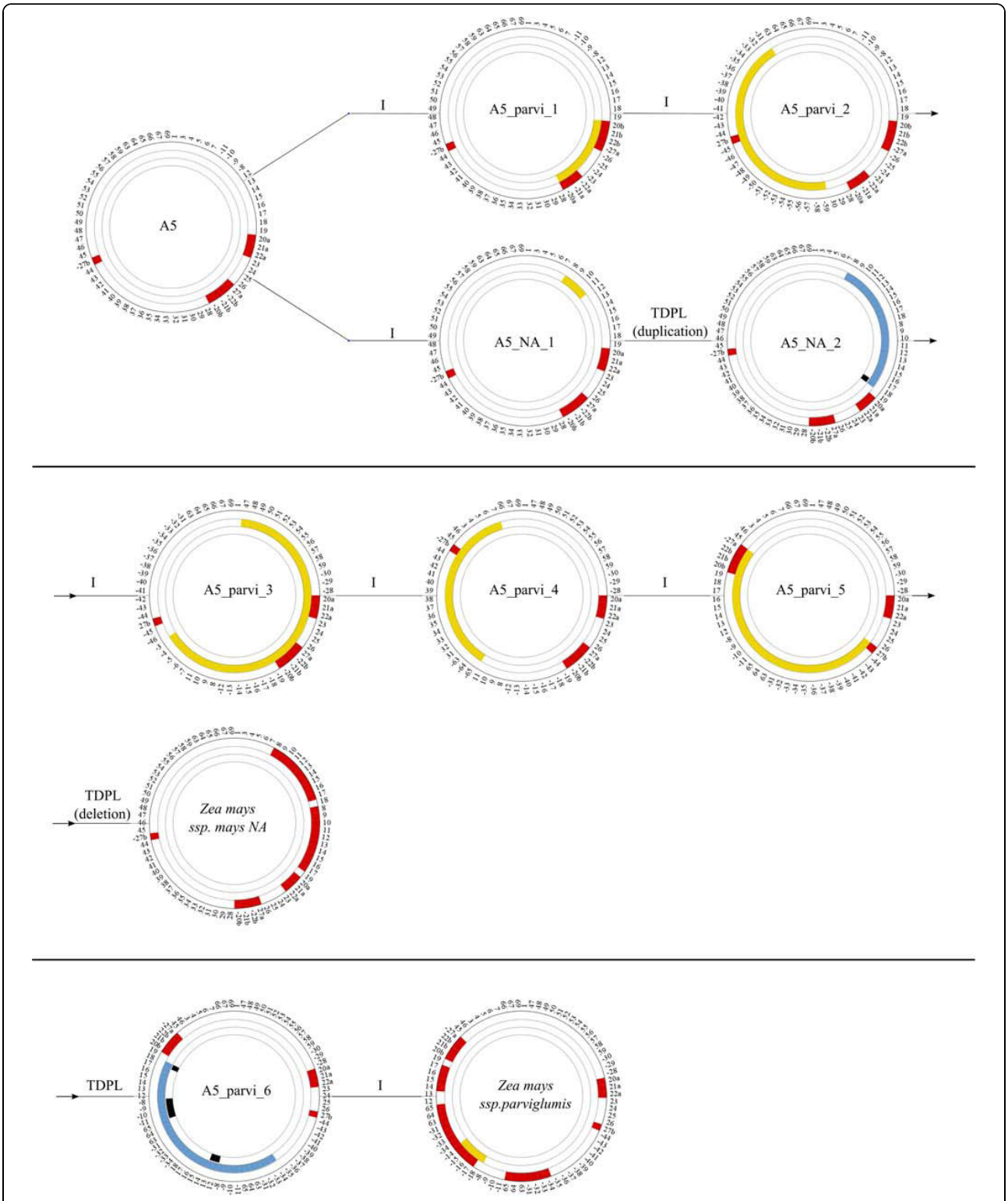


Figure 7 Predicted evolution from A5 to NA and *Zea mays* ssp. *Parviglumis*. Example of an evolutionary scenario from A5 to *Zea mays* ssp. *mays* NA and *Zea mays* ssp. *parviglumis*. A5 (the ancestral sequence computed by MGR) and intermediate mitogenomes after each rearrangement event are shown. Evolution of A5 mitogenome through inversions, tandem duplications (respectively yellow and blue in the middle circle) and synteny anchor loss (black in the internal circle) leads to NA (with 2 rearrangements) and *Zea mays* ssp. *parviglumis* (with 7 rearrangements). Duplicated synteny anchors are in red (outside circle). On each branch, the rearrangement event is indicated (I = Inversion, TDPL = Tandem Duplication with Partial Loss). For NA, the TDPL is indicated in two steps: a duplicated event followed by a deletion event.

was that the duplication event occurred just before the speciation event.

Phylogenetic relationships among Zea mitochondrial genomes

The phylogenetic relationships among maize mitogenomes concord with a former study by Allen and colleagues [24] where NA and NB were described as being the most-closely related mitogenomes, followed by CMS-C, CMS-S and CMS-T. On the basis of their nucleotide divergence, CMS-S and CMS-T were suggested to be the oldest cytoplasms. The introduction of two additional mitogenomes from the outgroup species of teosintes *Zea luxurians* and *Zea perennis* also suggested the ancestral position of CMS-S and T. Former studies on mitochondrial and chloroplastic diversity in *Zea* pointed out the fact that CMS-S was an old cytoplasm and most likely the result of introgression from teosinte *Zea mays* ssp. *mexicana*. But the phylogenetic location of CMS-T, due to a strong nucleotide divergence and a concomitant rearranged genome, is puzzling since CMS-T shares the same co-inherited chloroplastic genome with CMS-C and NB [27,28]. Consequently, the high divergence of CMS-T might not have occurred in a molecular clock tempo (as suggested by the rejection of the molecular clock hypothesis in the phylogenetic analysis). Chloroplastic sequence data could shed light on the relative ages of the cytoplasms studied. It is interesting to note that the same phenomenon was observed when considering the chloroplastic nucleotide diversity among several cytoplasms of wild beet: cytoplasm *Nv* and CMS *Owen* are closely related when considering chloroplastic nucleotide divergence [29] while mitochondrial genomes are highly rearranged and exhibit about 8% of specific sequences [30].

The phylogenetic location of *Zea mays* ssp. *parviglumis* included in the *Zea mays* clade concords with the scenario of a recent maize domestication from this teosinte subspecies [31]. Moreover, it highly suggests that the cytoplasms we studied differentiated before domestication.

Tandem duplication with partial loss as a plausible mechanism

Tandem duplication is a mechanism that has been demonstrated or at least suggested in mitochondrial genomes of several animal species, even though the underlying molecular mechanism is not always understood [32,33]. Tandem duplications have been mainly observed in Chordata, particularly in Vertebrates such as Lizards [33], Salamanders [9], Amphibians [34] or Gulper Eels [8]. Cases of tandem duplication are not restricted to Chordata, they have also been reported in Echinodermata [10], Insecta [35] and Lophotrochozoa (e.g. Mollusca) [36-38]. It must be noted that different types of tandem duplication have been observed in all these species: duplications of the whole genome, tandem duplications of genome parts, tandem duplications of non-coding regions or tandem duplications of one gene. In most cases, only one functional copy of the duplicates remains after duplication.

Mitochondrial genomes of maize and teosintes (*Zea mays* ssp. *parviglumis*, *Zea luxurians* and *Zea perennis*) could undergo the same mechanism of tandem duplication with loss as animal mitochondrial genomes. A possible mechanism could rely on the integration in the master chromosome of minicircles generated by homologous recombination between direct repeats from the original master circle, resulting in a duplication event [39]. But this would imply a preferential adjacent integration (see discussion by Fujita and colleagues [33] for animals). The low substitution rate in the maize mitogenome may explain why, in maize mitogenomes, one or more copies of duplicated synteny anchors remain, as opposed to animal mitogenomes where all gene copies but one are lost. More generally, a causal link has been suggested between mutation rate and genome compactness that could explain the large size and gene duplicate occurrence of plant mitochondrial genomes when compared with their animal counterparts [2]. The fact that the same mechanism could be involved in mitochondrial genomes of plants and animals falls in line with the monophyletic origin of animal and plant mitochondrial genomes [1]. For example, red algae [40], that form an independent lineage that radiated contemporarily with the other evolved eukaryotic lineages, demonstrates characteristics of both plant (gene with introns, ribosomal proteins) and animal mitochondria (modified genetic code, short mitochondrial sequence). Similar observations have been made for *Acanthamoeba castellanii* [41] or *Trichoplax adhaerens* [42].

Looking at the literature from the past decades, emphasis has been put on differences between animal and plant mitogenomes in their evolutionary dynamics and at the structure level [11,14]. While a compact circular genome is found in the majority of animal lineages, the plant mitogenome was described as a dynamic equilibrium of isoforms of a master circular chromosome and sub-molecules due to the occurrence of repeated sequences favoring intragenomic recombination. In this context, it is particularly interesting to notice that the evolutionary scenario based on rearrangement among master circles is congruent with the analysis based on sequence divergence among them. Therefore, it appears that master circles might reflect more than a virtual synthetic representation.

Conclusions

Despite important structural shuffling among genomes, even at the species level, we were able to build a phylogenetic tree using rearrangement events between plant mitochondrial genomes that was congruent with a sequence-based tree. To our knowledge this is the first evolutionary scenario of a plant mitogenome proposed solely on the basis of rearrangement events in complete DNA sequences. We showed that, under the hypothesis of structure evolution through inversions and tandem duplications with loss, an evolutionary path could be drawn for each genome. While such evolutionary events have been identified in animal mitogenomes, the hypothesis of a similar mechanism has never been discussed for plant mitogenomes. Further work will consist of developing new tools in order to automatically look for signatures of tandem duplication events in other plant mitogenomes and evaluate the occurrence of this process on a larger scale.

Methods

Data

Mitochondrial genomes used

The eight studied mitogenomes from *Zea* were downloaded from GenBank. Among the 5 recently sequenced mitogenomes from *Zea mays subsp. mays*, two of them are fertile cytotypes *NA* [GenBank:DQ490953] and *NB* [GenBank:AY506529], and three of them are cytoplasmic-male-sterile (CMS) cytotypes: *CMS-C* [GenBank: DQ645536], *CMS-S* [GenBank:DQ490951] and *CMS-T* [GenBank:DQ490953] [24]. We enriched the dataset with the mitogenomes of three teosinte species, *Zea mays subsp. parviglumis* [GenBank:DQ645539], *Zea luxurians* [GenBank:DQ645537] and *Zea perennis* [GenBank: DQ645538] (Allen *et al.*, unpublished results). The two last mitogenomes served as outgroups for phylogenetic analysis. Table 1 summarizes the genomes used.

We noted that all mitogenomes are in the master circle conformation and all our analyses were based on this conformation.

Synteny blocks

Synteny blocks, representing conserved sequence blocks between all mitogenomes, were computed using Mauve [43], a tool performing multiple genome alignments between sequences that can be rearranged. Mauve uses a set of genome DNA sequences as input. It locally computes co-linear blocks from anchors that are short unique similar DNA fragments. The anchors are then extended in order to produce longer common segments. Finally, the segments are clustered to locally produce co-linear blocks under the constraint that, for a given genome, segments have to be on the same strand. As the Mauve algorithm keeps short unique similar DNA fragments, duplicated DNA sequences are not taken into account. Mauve provides a backbone file containing synteny blocks and an alignment file containing the alignments of each synteny block.

Mauve parameters used are match weight seed = 9, minimum island = 15, maximum backbone gap size = 15, minimum backbone size = 50. Match weight seed parameter is essential in the multiple alignment and depends on the number of genomes to align and their lengths. Default weight seed is 11 for genomes of 1 MB length and increases with the genome size. As mitogenomes used in this study have a size comprised between 535 and 740 Kb, we set the weight seed at 9 (lower values were tested but a weight seed of 9 provided the best results). Minimum island is the minimum size for a fragment that is not common to all genomes. Maximum backbone gap size is the maximum size authorized for a gap in sequences common to all mitogenomes. If one mitogenome had a gap longer than to 15 bp in a sequence block, this block was split into two blocks at the gap. Minimum backbone size is the minimum size for a sequence block.

Backbone DNA sequence

In order to compare mitogenomes at the sequence level, for each genome we used the backbone and the alignment sequences provided by Mauve to build a sequence made of the concatenation of the synteny blocks, called *backbone DNA sequence*. As duplicated sequences are not taken into account in Mauve, we masked one copy of each duplicate (size >500 bp) for each mitogenome. A reference genome was chosen (here NA) in order to build the backbone DNA sequences. For each genome, the synteny blocks were concatenated, following the order of the synteny blocks on the reference genome. As we kept all common sequences between the eight genomes, the choice of one reference genome instead of another does not change the results. As the method used for computing synteny blocks allows insertions, deletions and substitutions, the length of a synteny block may vary depending on the genome and therefore the length of the backbone sequence may be different for each genome. The number of synteny blocks and the length of the backbone sequence for the eight genomes were summarized in Additional file 1. The repartition of synteny blocks for the mitogenomes was provided in Figure 1.

Genome structure sequence

In order to study genomic rearrangements we had to build a genome structure sequence (i.e. genome marker order) out of the genome DNA sequence. Such a genome structure sequence is an abstraction of the genome seen as a sequence of blocks that can be rearranged. The main difference when compared with the backbone sequence is that the DNA sequence within each block is no longer considered.

To build the genome structure sequence of each mitogenome, we applied the following strategy: i) first, we extracted annotated protein coding genes, tRNAs, rRNAs, ORFs (Open Reading Frame) and pseudogenes from the corresponding GenBank file, and then, ii) non-coding sequences from the backbones.

For coding sequences extracted from all eight mitogenomes we built a database. For each sequence, we used the YASS (Yet Another Similarity Searcher) software [44] against this database (excluding the sequence of interest). We conserved all reciprocal best hits in order to identify orthologous markers. As E-value depends on the sequence lengths compared, different E-values were used when sequences were shorter or longer than 100 bp. For the case of protein coding genes, rRNAs, ORFs and pseudogenes (with a length higher than 100 bp), we considered only RBHs with an E-value lower than $1e^{-170}$ and with an alignment length difference of less than 8%. For the case of tRNAs and some protein coding gene exons (with a length shorter than 100 bp) we chose an E-value of $1e^{-26}$ and an alignment length difference of less than 8%. When it was impossible to distinguish between two reciprocal best hits (same E-value and same sequence length), the copies were considered as homologous. If a marker was missing in a genome, we launched a search using YASS in order to check if it was a misannotation. If the marker was not found, the homologs (orthologs and paralogs) in other mitogenomes were excluded from the study.

For non-coding sequences, we used fragments from the backbone sequences that were larger than 100 bp. We did not consider those included in a coding region (because they would have been counted twice in the dataset). Using the YASS software, we only kept duplicates with an alignment length difference of less than 8%.

We thus obtained a set of 187 markers common to all genomes. If markers were found in the same order in all mitogenomes, we grouped them into marker groups, their boundaries corresponding to the flanking markers. Overall, the extraction procedure resulted in a total of 69 markers along mitogenomes that we call hereafter *synteny anchors*.

We obtained synteny anchor structure sequences by assigning a number to each synteny anchor. Using NA as the reference genome, each synteny anchor was assigned a number in ascending order from left to right. The numbering of the other genomes was based on NA (using another reference mitogenome does not change the results). A plus or minus was assigned to each synteny anchor depending on the strand where the synteny anchor occured in the NA genome. These structure sequences, where synteny anchor orthologs and paralogs had the same number, were called GSS bpisac (Genome Structure Sequence before paralog identification and synteny anchor collapsing). Additional file 2 provides the composition and numbering of synteny anchors used to build the GSS for each genome, Figure 1 depicts GSS bpisac blocks repartition along the eight genomes.

In order to test our hypothesis of tandem duplication in maize and teosinte mitogenomes, we needed to take into account duplicated synteny anchors. As paralogous synteny anchors have identical nucleotide sequences, we used the neighborhood graph (see below and Additional file 4) to distinguish them. Two different duplication types (of one or more synteny anchor groups) could be observed : unique to a mitogenome or shared by some or all mitogenomes.

If a duplication was specific to one genome and seemed to be tandem duplicated, we considered it as being a recent event. In order to integrate the duplicated synteny anchors in the dataset, we first looked for the bounds of the duplicated part, then we reintroduced all deleted synteny anchors yielding two juxtaposed identical parts, and finally collapsed the synteny anchors involved in the two parts by re-numbering them to obtain the part before tandem duplication.

If a duplication was shared between genomes (or specific to one genome and not tandem duplicated), we considered that there was a tandem duplication at the ancestral level. When synteny anchor copies were distant along the mitogenomes, we decided to distinguish the copies using their synteny anchor adjacencies in the eight genomes.

Through the neighborhood graph and the resulting hierarchical clustering (see Additional file 4) made on GSS before paralog identification and synteny anchor collapsing (bpisac), we determined the bounds of each duplicated part (duplicates are on a thick line on GSSs bpisac in Figure 1 and Figure 4). For example, for CMS-C, it was difficult to choose if synteny anchors {32 33 34 35 36 37 38} had to be clustered with {31} or with {60}. Thanks to the hierarchical clustering, {32 33 34 35 36 37 38} was put with {31} because {32 33 34 35 36 37 38} were clustered with {31}. After all obvious tandem duplications were collapsed, some duplications remained. Some of them were specific to a given mitogenome, while the others were shared by several mitogenomes. In the case of {20 21 22}, for which at least one copy was found in all mitogenomes, we made the hypothesis of an ancestral duplication of this group followed by loss of one copy of {21 22} in *Zea luxurians*, one copy of {22} in *Zea perennis* and all copies in CMS-T. Other mitogenomes had kept all copies. We renumbered one of the duplicates, depending on the neighborhood. For example, {20 21 22} was associated with {23 24 25 26}, that is why the first occurrence of {20 21 22} next to {23 24 25 26} was renumbered {20a 21a 22a} and the other occurrence was renumbered {20b 21b 22b}. We did the same for the group {27}, one copy (next to {44}) was renamed {27a} and the other was renamed {27b}. If a synteny anchor was duplicated (not in

tandem) in only one mitogenome, we also distinguished the two occurrences. Under the postulate of a tandem duplication event specific to this genome, we added the new number in tandem with the first occurrence in the other mitogenomes. This ensured that GRIMM kept synteny anchors together when computing evolving scenario between all other mitogenomes. It was the case for {24} in *Zea luxurians*, {26} and {67} in CMS-T, and {35} *Zea perennis* where paralogs were respectively renumbered {24b}, {26b},{67b} and {35b}. All duplicated synteny anchors were then distinguished except for {2} duplicated in NA and NB, {60, 61, 62} duplicated in NB, and {68} duplicated in *Zea perennis*. All copies of these five synteny anchors were thus deleted from the dataset.

It was thus possible to distinguish between paralogs and orthologs for 8 out of 13 duplicated synteny anchors (see Figure 3).

Then we were able to apply known rearrangement methods on this structure called GSS. The GSS was composed of 72 synteny anchors. Figure 4 provides a comparison of GSS bpisac and GSS for CMS-C and *Zea perennis* mitogenome.

Neighborhood graph and synteny anchor clusters

Neighborhood relationships between synteny anchors were modeled in a graph. Two synteny anchors were considered to be in the same neighborhood if they were separated by at most one synteny anchor. A weight function was defined between two synteny anchors as the number of times both synteny anchors were neighbor. For a given weight w, a cluster of synteny anchors was defined as a set of synteny anchors such that: i) for any synteny anchor s in the set there exists another synteny anchor s' such that s and s' are neighbor and the value of the weight function between them is greater than w, ii) for any synteny anchor s in the set and for any synteny anchor s' outside the set, s and s' are not neighbors or they are neighbors but the value of the weight function between them is lower than w. That is two synteny anchors were in the same cluster if they were separated by at most one synteny anchor at least w times. We used this definition of synteny anchor cluster because usual gene clusters such as common intervals [45] or gene teams [46] cannot be applied to our data: the definition is too restrictive and/or does not support duplicated genes.

Table 3 Ratio of pairwise genome substitution rate between backbone and protein coding sequences per 10 kb

	NB	CMS-C	CMS-S	CMS-T	parvi	lux	per
NA	4.002	1.296	1.428	0.994	1.770	1.453	1.299
NB	-	1.990	1.729	1.176	2.694	1.553	1.393
CMS-C	-	-	1.653	1.115	1.122	1.544	1.375
CMS-S	-	-	-	1.360	1.355	1.743	1.548
CMS-T	-	-	-	-	0.948	1.220	1.149
parvi	-	-	-	-	-	1.411	1.270
lux	-	-	-	-	-	-	1.055

Sequence analysis

Method for counting duplicated segments

Mitogenome statistics were performed with an in-house script using YASS in order to detect large duplicated segments (longer than 500 bp). YASS aligns pairwise sequences and finds conserved segments. As we were looking for highly conserved segments, we used a score of +1 for matches and a score of -3 for substitutions. Segments up to 500 bp (as in [24]) and with an E-value lower than $1e^{-300}$ were considered as paralogous.

Substitution rate

Sequence substitution rates were computed from the backbone DNA sequences and protein coding gene sequences for each mitogenome pairs. Protein coding gene sequence is the concatenation of one copy (since the copies are identical) of each protein coding gene, common to all mitogenomes. Substitution rate (for 10 kb) between two genomes was calculated as follows :

$$\frac{number\ of\ substitutions\ between\ genome1\ and\ genome2}{alignment\ length\ between\ genome1\ and\ genome2} \times 10000$$

Ratio of substitution rates between backbone DNA sequences and protein coding genes was also calculated (Table 3).

Structure sequence analysis

A simple way to measure a rearrangement distance between genomes is to count the number of breakpoints [47-49]. A breakpoint is a disruption of the genome sequence order, i.e. when adjacency between two genes in one genome disappears in another one. A breakpoint matrix distance among genomes provides a way to reconstruct a phylogenetic tree using distance methods [50]. But such a basic tool does not provide any information about the history of rearrangements.

To further pursue the analysis of genomic rearrangements, one might compute the rearrangement distance as the minimal number of rearrangement operations needed to transform a genome into another [51]. This distance can also be used to build a phylogenetic tree : the more similar two genomes are, the smaller the rearrangement distance between them. The computation of such a distance also provides the scenario of operations that rearranged a genome into another. This allows one to build parsimonious phylogenies and propose ancestral nodes [18]. We used the GRIMM software (Genome Rearrangements In Man and Mouse -this software is not specific to Human and mouse genomes) [52] to

compute inversion distances and scenarios. This software computes parsimonious inversion scenarios given a set of genomes as sequences of numbers without duplicates.

Phylogenetic analysis

At the DNA sequence level

Neighbor-Joining analyses were realized on the backbone DNA sequences using BIONJ [53]. Parameters used are bootstrap 1000× and Kimura-2 parameters distance for correction. Maximum likelihood and molecular clock were tested with TREE-PUZZLE [54] using the nucleotide model of Hasegawa-Kishino-Yano (HKY85) [55].

At the structure sequence level

Rearrangement analyses were performed using GRIMM onto the GSSs. We obtained a distance matrix and then used BIONJ on this matrix to obtain a phylogenetic tree. Unfortunately, no bootstrap method is available for rearrangement studies. In order to test the robustness of the reconstructed trees, we adapted a Jackknife test [56,57] on the GSSs as follows: we randomly kept ninety percent of the GSS blocks (65 blocks out of 72); on this subset we computed a GRIMM matrix and we built a phylogeny using BIONJ; 1000 tests were applied. We thus obtained 1000 trees. We reported the frequency of the nodes found in the original tree according to this set of trees. We performed tests for several percentages of kept GSS blocks (10%, 20%,...100%) using the same method (see Additional file 3). The MGR (Multiple Genome Rearrangements) software [18] answers the problem of computing a parsimonious phylogeny given a set of genomes represented as sequences of numbers without duplicates. Unfortunately this problem has been shown to be computationally hard (NP-hard). It follows that MGR provides an approximate solution which is often near optimal [18].

Additional file 1: Backbone DNA fragments. Each orthologous fragment between mitogenomes is represented by an arrow. Fragment with the smallest size is underlined in blue and fragment with the longest size in red.

Additional file 2: Synteny anchor numbers and compositions. Synteny anchors contained in GSSs. A synteny anchor often contains more than one genome marker (gene, tRNA, rRNA, ORF, pseudogene or non-coding sequence from backbone DNA sequence).

Additional file 3: Jackknife tests. Node values for percentage of conserved GSS blocks. For each percentage of conserved synteny anchors, 1000 GRIMM matrices were computed and 1000 trees were drawn from these matrices. Each node value obtained for the consensus of these 1000 trees was reported in the graph. For example, for 90% of conserved GSS synteny anchors, Jackknife value for the terminal node (separation between NB and the remaining two *Zea mays* mitogenomes) 96.1%.

Additional file 4: Hierarchical clustering. Hierarchical clustering obtained with the neighborhood graph using GSSs. Two synteny anchors closer to one another than the others were assigned to the same cluster.

Acknowledgements

The authors wish to thank V. Castric and F. Roux for their valuable comments on previous versions of the manuscript, A. Jacquemin for the software allowing to visualize rearrangement phylogeny trees (Figure 7) and two anonymous reviewers for valuable comments on a former version of the manuscript (particularly "reviewer #2"), Licia Huffman for copy-editing. This work was funded by a grant from the Agence Nationale de la Recherche (ANR-06-JCJC-0074) to PT, a grant from PPF Bioinformatique of University of Lille1 to PT and J-SV, and a PhD fellowship from French Research Ministry to AD.

Author details

[1]Laboratoire de Genetique et Evolution des Populations Vegetales, UMR CNRS 8016, Universite Lille 1, 59655 Villeneuve d'Ascq Cedex, France. [2]Laboratoire d'Informatique Fondamentale de Lille, UMR CNRS 8022, Universite Lille 1, 59655 Villeneuve d'Ascq Cedex, France. [3]INRIA Lille-Nord Europe, 59650 Villeneuve d'Ascq, France.

Authors' contributions

AD, JSV and PT designed the study. AD ran all the analyses and prepared all figures and tables. AD, JSV and PT interpreted the results and wrote the manuscript. All authors read and approved the final manuscript.

Received: 17 July 2009 Accepted: 9 April 2010 Published: 9 April 2010

References

1. Gray MW, Burger G, Lang BF: **Mitochondrial Evolution.** *Science* 1999, **283**:1476-1481.
2. Lynch M, Koskella B, Schaack S: **Mutation Pressure and the Evolution of Organelle Genomic Architecture.** *Science* 2006, **311**:1727-1730.
3. Boore JL: **Animal mitochondrial genomes.** *Nucleic Acids Res* 1999, **27(8)**:1767-1780.
4. Boore JL: **The use of genome-level characters for phylogenetic reconstruction.** *Trends Ecol Evol* 2006, **21(8)**:439-446.
5. Segawa RD, Aotsuka T: **The mitochondrial genome of the Japanese freshwater crab, Geothelphusa dehaani (Crustacea: Brachyura): Evidence for its evolution via gene duplication.** *Gene* 2005, **355**:28-39.
6. Lavrov DV, Boore JL, Brown WM: **Complete mtdna sequences of two millipedes suggest a new model for mitochondrial gene rearrangements : Duplication and nonrandom loss.** *Mol Biol Evol* 2002, **19**:163-169.
7. Wang X, Lavrov DV: **Mitochondrial Genome of the Homoscleromorph Oscarella carmela (Porifera, Demospongiae) Reveals Unexpected Complexity in the Common Ancestor of Sponges and Other Animals.** *Mol Biol Evol* 2007, **24(2)**:363-373.
8. Inoue JG, Miya M, Tsukamoto K, Nishida M: **Evolution of the Deep-Sea Gulper Eel Mitochondrial Genomes : Large-Scale Gene Rearrangements Originated Within the Eels.** *Mol Biol Evol* 2003, **20(11)**:1917-1924.
9. Lockridge Mueller R, Boore JL: **Molecular Mechanisms of Extensive Mitochondrial Gene Rearrangement in Plethodontid Salamanders.** *Mol Biol Evol* 2005, **22(10)**:2104-2112.
10. Perseke M, Fritzsch G, Ramsch K, Bernt M, Merkle D, Middendorf M, Bernhard D, Stadler PF, Schlegel M: **Evolution of mitochondrial gene orders in echinoderms.** *Mol Phylogenet Evol* 2008, **47(2)**:855-864.
11. Palmer JD, Herbon LA: **Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence.** *J Mol Evol* 1988, **28(1)**:87-97.
12. Lonsdale DM, Hodge TP, Fauron CMR: **The physical map and organisation of the mitochondrial genome from the fertile cytoplasm of maize.** *Nucleic Acids Res* 1984, **12(24)**:9249-9261.
13. Fauron CMR, Casper M: **A Second Type of Normal Maize Mitochondrial Genome: An Evolutionary Link.** *Genetics* 1994, **137**:875-882.
14. Kubo T, Newton KJ: **Angiosperm mitochondrial genomes and mutations.** *Mitochondrion* 2008, **8(1)**:5-14.
15. Knoop V: **The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective.** *Curr Genet* 2004, **46**:123-139.
16. Sankoff D, Leduc G, Antoine N, Paquin B, Lang B, Cedergren R: **Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome.** *Proceedings of the National Academy of Sciences* 1992, **89**:6575-6579.

17. Hannenhalli S, Pevzner P: **Transforming men into mice (polynomial algorithm for genomic distance problem).** *proceedings of the 36th Annual Symposium on Foundations of Computer Science (FOCS'95)* 1995, 581-592.
18. Bourque G, Pevzner PA: **Genome-scale evolution: Reconstructing gene orders in the ancestral species.** *Genome Res* 2002, **12(1)**:26-36.
19. Moret B, Tang J, Wang L, Warnow T: **Steps toward accurate reconstruction of phylogenies from gene-order data.** *J Comput Syst Sci* 2002, **65(3)**:508-525.
20. Sankoff D: **Genome rearrangement with gene families.** *Bioinformatics* 1999, **15(11)**:909-917.
21. Chen X, Zheng J, Fu Z, Nan P, Zhong Y, Lonardi S, Jiang T: **Assignment of orthologous genes via genome rearrangement.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2005, **2(4)**:302-315.
22. Chauve C, Doyon J, El-Mabrouk N: **Gene family evolution by duplication, speciation, and loss.** *Journal of Computational Biology* 2008, **15(8)**:1043-1062.
23. Gordon JL, Byrne KP, Wolfe KH: **Additions, Losses, and Rearrangements on the Evolutionary Route from a Reconstructed Ancestor to the Modern Saccharomyces cerevisiae Genome.** *PLoS Genetics* 2009, **5(5)**:e1000485.
24. Allen JO, Fauron CM, Minx P, (16 co-authors), *et al*: **Comparisons among two fertile and three male-sterile mitochondrial genomes of maize.** *Genetics* 2007, **177**:1173-1192.
25. Tang J, Moret BME: **Phylogenetic reconstruction from gene-rearrangement data with unequal gene content.** *Lecture Notes in Computer Science vol 2748 8th International Workshop on Algorithms and Data Structures (WABI 2003)* 2003, 37-46.
26. Bernt M, Merkle D, Ramsch K, Fritzsch G, Perseke M, Detlef B, Schlegel M, Stadler P, Middendorf M: **CREx: inferring genomic rearrangements based on common intervals.** *Bioinformatics* 2007, **23(21)**:2957-2958.
27. Pring DR, Levings III CS: **Heterogeneity of maize cytoplasmic genomes among male-sterile cytoplasms.** *Genetics* 1978, **89**:121-136.
28. Doebley J, Renfroe W, Blanton A: **Restriction Site Variation in the Zea Chloroplast Genome.** *Genetics* 1987, **117**:139-147.
29. Fénart S, Touzet P, Arnaud JF, Cuguen J: **Emergence of gynodioecy in wild beet (Beta vulgaris ssp. maritima L.): a genealogical approach using chloroplastic nucleotide sequences.** *Proc R Soc Lond B Biol Sci* 2006, **273**:1391-1398.
30. Satoh M, Kubo T, Nishizawa S, Estiati A, Itchoda N, Mikami T: **The cytoplasmic male-sterile type and normal type mitochondrial genomes of sugar beet share the same complement of genes of known function but differ in the content of expressed ORFs.** *Mol Genet Genomics* 2004, **272(3)**:247-256.
31. Doebley J: **The Genetics Of Maize Evolution.** *Annu Rev Gene* 2004, **38**:37-59.
32. Stanton DJ, Daehler LL, Moritz CC, Brown WM: **Sequences With the Potential to Form Stem-and-Loop Structures Are Associated With Coding-Region Duplications in Animal Mitochondrial DNA.** *Genetics* 1994, **137**:233-241.
33. Fujita MK, Boore JL, Moritz C: **Multiple Origins and Rapid Evolution of Duplicated Mitochondrial Genes in Parthenogenetic Geckos (Heteronotia binoei; Squamata, Gekkonidae).** *Mol Biol Evol* 2007, **24**:2775-2786.
34. San Mauro D, Gower DJ, Zardoya R, Wilkinson M: **A Hotspot of Gene Order Rearrangement by Tandem Duplication and Random Loss in the Vertebrate Mitochondrial Genome.** *Mol Biol Evol* 2006, **23(1)**:227-234.
35. Carapelli A, Vannini L, Nardi F, Boore JL, Beani L, Dallai R, Frati F: **The mitochondrial genome of the entomophagous endoparasite Xenos vesparum (Insecta: Strepsiptera).** *Gene* 2006, **376**:248-259.
36. Vallès Y, Boore JL: **Lophotrochozoan mitochondrial genomes.** *Integr Comp Biol* 2005, **46(4)**:544-557.
37. Grande C, Templado J, Zardoya R: **Evolution of gastropod mitochondrial genome arrangements.** *BMC Evolutionary Biology* 2008, **8**:61-75.
38. Yu Z, Wei Z, Kong X, Shi W: **Complete mitochondrial DNA sequence of oyster Crassostrea hongkongensis-a case of Tandem duplication-random loss for genome rearrangement in Crassostrea?** *BMC Genomics* 2008, **9**:477-489.
39. Small I, Suffolk R, Leaver CJ: **Evolution of plant mitochondrial genomes via substoichiometric intermediates.** *Cell* 1989, **58(1)**:69-76.
40. Leblanc C, Boyen C, Richard O, Bonnard G, Grienenberger JM, Kloareg B: **Complete Sequence of the Mitochondrial DNA of the Rhodophyte Chondrus crispus (Gigartinales). Gene Content and Genome Organization.** *J Mol Biol* 1995, **250**:484-495.
41. Burger G, Plante I, Lonergan KM, Gray MW: **The Mitochondrial DNA of the Amoeboid Protozoon, Acanthamoeba castellanii : Complete Sequence, Gene Content and Genome Organization.** *J Mol Evol* 1995, **245**:522-537.
42. Signorovitch AY, Buss LW, Dellaporta SL: **Comparative Genomics of Large Mitochondria in Placozoans.** *PLoS Genetics* 2007, **3(1)**:44-50.
43. Darling AC, Mau B, Blatter FR, Perna NT: **Mauve: multiple alignment of conserved genomic sequence with rearrangements.** *Genome Res* 2004, **14(7)**:1394-1403.
44. Noe L, Kucherov G: **YASS: enhancing the sensitivity of dna similarity search.** *Nucleic Acids Res* 2005, **33(2)**:W540-W543.
45. Uno T, Yagiura M: **Fast algorithms to enumerate all common intervals of two permutations.** *Algorithmica* 2000, **26(2)**:290-309.
46. Luc N, Risler JL, Bergeron A, Raffinot M: **Gene teams: a new formalization of gene clusters for comparative genomics.** *Comp Biol Chemistry* 2003, **27(1)**:59-67.
47. Watterson G, Ewens W, Hall T, Morgan A: **The chromosome inversion problem.** *J Theor Biol* 1982, **99**:1-7.
48. Blanchette M, Bourque G, Sankoff D: **Breakpoint phylogenies.** *Proceedings of the 8th Genome Informatics Workshop (GIW 1997)* University Academy Press. Tokyo 1997, 25-34.
49. Sankoff D, Bryant D, Deneault M, Lang F, Burger G: **Early Eukaryote Evolution Based on Mitochondrial Gene Order Breakpoints.** *J Comput Biol* 2000, **7(3-4)**:521-535.
50. Wang LS, Warnow T, Moret BME, Jansen RK, Raubeson LA: **Distance-Based Genome Rearrangement Phylogeny.** *J Mol Evol* 2006, **63(4)**:473-483.
51. Bader DA, Moret BME, Yan M: **A Linear-Time Algorithm for Computing Inversion Distance between Signed Permutations with an Experimental Study.** *J Comput Biol* 2001, **8(5)**:483-491.
52. Tesler G: **GRIMM: genome rearrangements web server.** *Bioinformatics* 2002, **18(3)**:492-493.
53. Gascuel O: **BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data.** *Mol Biol Evol* 1997, **14**:685-695.
54. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18**:502-504.
55. Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.** *Journal of Molecular Evolution* 1985, **22(2)**:160-174.
56. Quenouille MH: **Notes on bias in estimation.** *Biometrika* 1956, **43**:353-360.
57. Tukey JW: **Bias and confidence in not quite large samples (Abstract).** *Annals of Mathematical Statistics* 1958, **29**:614.

doi:10.1186/1471-2164-11-233
Cite this article as: Darracq *et al.*: A scenario of mitochondrial genome evolution in maize based on rearrangement events. *BMC Genomics* 2010 **11**:233.

Conclusion

Dans cette étude, nous avons donc établi une méthode basée sur des événements de duplication en tandem permettant de trier les marqueurs dupliqués dans un ensemble de génomes. Nous avons vu que ces marqueurs dupliqués, représentant une proportion importante de génomes, étaient essentiels pour reconstruire les événements de réarrangements des génomes. La méthode appliquée sur nos génomes nous a permis d'obtenir des séquences de marqueurs triés pouvant être utilisés dans les outils d'étude de réarrangements. Les phylogénies obtenues avec ces outils disponibles sont congruentes avec une phylogénie de séquences nucléotidiques réalisée sur une large portion de séquences communes entre les génomes. Cette congruence valide l'hypothèse d'évolution de ces génomes par le biais de duplications en tandem plus ou moins bien conservées dans les génomes, qui était l'hypothèse sur laquelle reposait notre méthode de tri des marqueurs dupliqués. Nous avons ainsi montré que, malgré les caractéristiques différentes établies entre les mitochondries animales et végétales (que ce soit au niveau de la taille, la structure, ou encore le taux de mutation), celles-ci pourraient évoluer par des mécanismes similaires, les différences observées pouvant s'expliquer par un taux de substitutions différent (hypothèse de pression de mutation [Lynch et al., 2006]). Cependant, la méthode appliquée sur ces génomes reste essentiellement manuelle, notamment au niveau de la détection des marqueurs dupliqués. Il devient nécessaire pour valider cette hypothèse sur un grand nombre de génomes de disposer d'une méthode automatique. La première étape consiste à identifier les duplications, que ce soit de duplications en tandem (avec pertes ou non de certains marqueurs) ou non et que ces duplications soient communes ou non a plusieurs génomes. Cette automatisation constitue la première étape dans la mise en place d'un processus complet de condensation et de distinction de marqueurs dupliqués au sein de plusieurs génomes.

Chapitre 5

Méthode de détection des duplications

Avec le séquençage complet de génomes et l'annotation qui en découle, la reconnaissance et l'analyse de gènes dupliqués prend de plus en plus d'importance. Par exemple, l'assignation de relations d'homologie entre gènes va aider à leur annotation fonctionnelle. Il existe des logiciels permettant de déterminer les relations d'orthologie (gènes homologues dérivant d'une spéciation) et de paralogie (gènes homologues dérivant d'une duplication) pour des gènes provenant de différents génomes. Ces logiciels peuvent s'appuyer sur différents critères (séquence génomique, synténie, duplication, réarrangement) afin de déterminer au mieux ces relations. Cependant, ils ne permettent pas de trouver des ensembles de gènes dupliqués provenant d'une duplication d'un fragment de génome.

Différentes méthodes permettant de retrouver des groupes de gènes orthologues ont été décrites. Afin de repérer ces groupes, les génomes sont représentés comme des permutations de leurs gènes. Les permutations sont des suites de gènes conservés entre plusieurs génomes. Par définition, on ne représente qu'une occurrence d'un gène dans une permutation. Le problème de cette formalisation est donc que les paralogues non distingués ne sont pas pris en compte. De nouvelles méthodes considérant les génomes comme des séquences et non des permutations permettent d'intégrer les paralogues. Nous verrons que ces nouvelles méthodes permettent de rechercher des ensembles de gènes communs entre plusieurs génomes (l'ordre et le sens ne sont pas pris en compte) tout en gardant les gènes dupliqués. Ces méthodes proposent aussi de retrouver des ensembles de gènes approchés entre plusieurs génomes, c'est-à-dire des ensembles de gènes comportant quelques gènes de différence. Nous partirons d'une méthode retrouvant les ensembles de gènes approchés (pouvant représenter des pertes de gènes entre deux ensembles) que nous adapterons à la recherche d'ensembles de gènes dupliqués entre un ou plusieurs génomes.

5.1 État de l'art

Dans cette partie, nous ferons un résumé des méthodes et logiciels permettant de retrouver des gènes ou ensemble de gènes communs, dupliqués ou non, parmi plusieurs génomes. Cette partie complète le Chapitre 2.

5.1.1 Méthodes de recherche d'orthologues et paralogues

Il existe différentes méthodes visant à retrouver des gènes orthologues et paralogues parmi deux ou plusieurs génomes. La plupart des logiciels proposés sont basés sur l'homologie de séquence afin d'effectuer les prédictions.

Beaucoup de logiciels de prédiction d'homologie s'appuient sur la séquence des gènes, que ce soit au niveau protéique ou au niveau nucléique. Parmi les plus connus, il existe InParanoid [Ostlund et al., 2010]. InParanoid est une base de données contenant des gènes orthologues identifiés chez les Eucaryotes. Le principe général de cet outil est de comparer, grâce à BLAST, les séquences protéiques de gènes entre paires de génomes afin d'en extraire les relations d'homologies. Les orthologues sont d'abord assignés (gènes les plus proches entre deux espèces) puis ce sont les paralogues (gènes les plus proches au sein d'une même espèce).

Les outils basés sur l'homologie de séquence ne peuvent pas être appliqués sur les génomes mitochondriaux de plantes puisque les séquences des paralogues sont identiques. Les séquences entre orthologues, à un niveau intra-spécifique, sont également identiques. Nous aurons le même problème avec tous les outils s'appuyant sur l'alignement de séquences pour prédire les relations d'homologie, que ce soit les outils basés sur des homologies de domaine protéique ou encore sur des phylogénies. Il faut, dans notre cas, absolument tenir compte d'autres critères, tels que la synténie ou les réarrangements, pour essayer de différentier les gènes dupliqués présents dans les génomes mitochondriaux de plantes.

D'autres méthodes, incluant des critères de synténie et de distance de réarrangement ont été développées afin d'améliorer ces prédictions. Parmi les plus récents outils de prédiction d'homologie intégrant les réarrangements de génome, il existe MSOAR2 [Shi et al., 2010]. MSOAR2 est une amélioration de MSOAR [Fu et al., 2007] qui tient compte des gènes dupliqués en tandem. Ce logiciel s'appuie sur l'homologie de séquence mais aussi sur la distance de réarrangements entre les gènes afin de prédire des groupes d'orthologues entre deux génomes. Trois problèmes font que cette méthode est inadaptée à notre cas. Le premier est que ce logiciel n'assigne des relations d'homologie qu'entre deux génomes. Le deuxième est que la première étape de l'algorithme consiste à repérer les orthologues en fonction de l'homologie de séquence. Le troisième problème est que les duplications en tandem autorisées sont des duplications présentant les deux copies côte à côte et donc ne tenant pas compte de la duplication de fragments de génome. Cet outil n'est donc pas utilisable pour les analyses que nous souhaitons faire.

Tous les logiciels permettant de trouver des relations d'orthologie et de paralogie entre les gènes vont s'appuyer sur des homologies de séquence. La plupart du temps, des phylogénies sont effectuées à partir de ces séquences afin d'améliorer les prédictions. Ce type de logiciel ne peut pas être appliqué sur les données de génomes mitochondriaux de plantes à cause du taux de mutation μ très faible dans ces génomes aboutissant à des séquences entre gènes homologues identiques et donc indifférentiables. De plus nous souhaitons retrouver des groupes de gènes homologues, les logiciels proposés retrouvent juste des relations entre les gènes homologues sans aucune information sur les ensembles dont ils font partie.

5.1.2 Méthodes de recherche de groupes de gènes

Différentes méthodes ont été développées afin de retrouver les groupes de gènes communs entre plusieurs génomes. La motivation de ces méthodes est de retrouver des groupes de gènes afin d'en prédire leurs fonctions ou leurs caractéristiques. En effet, chez les bactéries, les gènes évoluent en opérons. Il s'agit de groupes de gènes agissant dans une même voie métabolique étant, par exemple, sous l'influence d'un même promoteur. Ces groupes de gènes, ayant un avantage

biologique à rester ensemble, vont alors évoluer groupés entre différentes espèces. Les méthodes de prédiction de ces groupes de gènes conservés seront utiles pour déterminer d'éventuelles relations fonctionnelles. Les génomes mitochondriaux que nous avons analysés présentent des ensembles de gènes conservés entre génomes, cependant nous avons observé que les gènes composant ces groupes n'agissent pas dans les mêmes voies métaboliques. Bien que nous ne souhaitions pas prédire les fonctions des gènes, ces méthodes de recherche de groupes de gènes restent néanmoins applicables aux génomes mitochondriaux de plantes.

Il existe différents types de méthodes d'analyse de groupes de gènes en fonction de la définition donnée à un groupe de gènes : celles recherchant des *intervalles communs* [Heber and Stoye, 2001a], celles recherchant des *gene teams* [Bergeron et al., 2002] ou encore celles recherchant des *max-gap clusters* [Hoberman et al., 2005]. Ces méthodes utilisent la représentation des génomes sous forme de suites de marqueurs. Pour les analyses de fonction, les marqueurs sont des gènes mais nous pouvons également prendre en compte des pseudo-gènes, des ARN ou des fragments non-codants qui seraient communs à plusieurs génomes. Un marqueur représente donc, comme indiqué au Chapitre 2, la position d'une séquence (codante ou non) conservée entre les génomes comparés.

Intervalles Communs. La notion d'intervalles communs fut d'abord introduite par [Uno and Yagiura, 2000] puis appliquée à la recherche de groupes de gènes par [Heber and Stoye, 2001a]. Un intervalle commun est défini comme un intervalle entre deux génomes possédant les mêmes marqueurs. Les premiers algorithmes consistaient à considérer deux génomes comme des permutations (dont l'un des deux est la permutation identité) afin d'en extraire des groupes de gènes conservés. De nombreuses améliorations sur ces algorithmes ont été effectuées notamment au niveau de l'analyse à plus de deux génomes [Heber and Stoye, 2001b] mais aussi de l'optimisation de l'exécution [Heber et al., 2009].

Les intervalles communs recherchés à partir de permutations limitent les analyses. En effet, la non prise en compte des marqueurs dupliqués est un frein car les génomes sont composés de nombreux gènes dupliqués. Il est donc essentiel d'en tenir compte. Certaines méthodes ont proposé d'étendre la recherche d'intervalles communs contenant des marqueurs dupliqués en considérant les génomes non pas comme des permutations mais comme des séquences. Dans un premier temps cette amélioration a été appliquée à la recherche d'intervalles communs entre deux génomes [Didier, 2003] puis étendue à plusieurs génomes [Schmidt and Stoye, 2004]. Une autre amélioration fut apportée plus tard, toujours basé sur des séquences et en utilisant la notion de *character set* [Schmidt and Stoye, 2004]. Il s'agit de l'extension de la notion d'intervalles communs appliquée aux séquences de marqueurs, un *character set* représentera donc un ensemble de marqueurs conservés entre plusieurs génomes pouvant contenir des marqueurs dupliqués. Il sera alors possible de retrouver les intervalles communs avec erreurs, c'est-à-dire des intervalles communs dont le contenu en marqueurs n'est pas tout à fait identique. Dans une première approche, [Chauve et al., 2006] ont permis la recherche d'intervalles communs avec erreurs et marqueurs dupliqués à condition de donner le nombre d'erreurs autorisées entre les intervalles communs. Plus tard [Amir et al., 2007] ont proposé de retrouver tous les intervalles communs avec un nombre non défini d'erreurs en incluant les marqueurs dupliqués, grâce à la représentation des intervalles communs sous la forme d'un graphe.

Gene Teams. En parallèle de la notion d'intervalle commun est apparue la notion de *gene teams* [Bergeron et al., 2002, Luc et al., 2003]. Les *gene teams* sont des groupes de marqueurs conservés entre génomes. Leur différence majeure avec les intervalles communs est la prise en

compte de gènes non communs entre les différents génomes. Ainsi les génomes comparés peuvent avoir des jeux de marqueurs différents, ce qui se rapproche des données biologiques dans lesquelles ont peut avoir des pertes de gènes ou des gènes différents entre les génomes. La représentation des génomes pour les *gene teams* est différente de celle des intervalles communs, les marqueurs communs à tous les génomes seront représentés par des chiffres alors que les marqueurs uniques (homologues à aucun autre) seront représentés par un symbole spécial. Ce symbole servira à évaluer la distance (c'est-à-dire le nombre de trous) entre les marqueurs homologues composant les groupes de gènes. En effet, si on peut avoir plusieurs erreurs dans un groupe de marqueurs, la taille des trous (représentés par les marqueurs non homologues) est prise en compte. Les améliorations apportées aux *gene teams* ont mené à la notion de *max-gap cluster* [Hoberman et al., 2005, Ling et al., 2008]. Les *max-gap clusters* sont des régions contiguës du génome contenant un maximum de marqueurs homologues entre lesquels on va pouvoir retrouver un nombre défini, maximum, de marqueurs non homologues. Le problème de méthodes basées sur les *max-gap cluster* est que les comparaisons se font uniquement entre paires de génomes.

5.1.3 Méthodes retenues

D'un point de vue biologique, nous souhaitons trouver des traces de duplication entre plusieurs génomes. En général, suite à une duplication, certains des marqueurs dupliqués sont perdus dans l'une ou l'autre des copies. Nous voulons donc retrouver des groupes de marqueurs dupliqués avec erreur entre plusieurs génomes. Étant donné que les méthodes de recherche de groupes de marqueurs basées sur les *max-gap clusters* ne permettent de comparer que des paires de génomes, nous allons nous intéresser aux méthodes de recherche d'intervalles communs avec erreurs en autorisant l'inclusion de paralogues.

Intervalles connectés (CI). [Schmidt and Stoye, 2004] proposèrent une méthode permettant de retrouver les intervalles communs entre deux ou plusieurs génomes en intégrant les paralogues. Ceci est possible si l'on considère les génomes comme des séquences et non des permutations. Plusieurs définitions sont introduites. Nous considérons qu'une séquence de marqueurs est appelée *chaîne de caractères.*

Définition 1 (character set, [Schmidt and Stoye, 2004])**.** *Étant donné une chaîne de caractères* S*,* $S[k]$ $1 \leq k \leq n$ *désigne le* $k^{ème}$ *élément de* S*.* $S[i,j]$ *représente une sous-chaîne du* $i^{ème}$ *au* $j^{ème}$ *élément de* S*. Le* character set *d'une sous-chaîne* $S[i,j]$ *est* $\mathcal{CS}(S[i,j]) = \{S[k], i \leq k \leq j\}$*.*

Un character set représente le jeu de tous les marqueurs appartenant à un intervalle donné du génome, où l'ordre et le nombre d'occurrences des marqueurs paralogues n'a pas d'importance.

Définition 2 ($\mathcal{CS}$-position, $\mathcal{CS}$-position maximale, [Schmidt and Stoye, 2004])**.** *Étant donné une chaîne de caractères* S *sur un alphabet* Σ *et un sous-ensemble* $C \subseteq \Sigma$*, la paire* (i,j) *est une* $\mathcal{CS}$*-position de* C *dans* S *si et seulement si* $\mathcal{CS}(S[i,j]) = C$*. Une* $\mathcal{CS}$*-position* (i,j) *de* C *dans* S *est* gauche-maximale *si* $S[i-1] \notin \mathcal{CS}(S[i,j])$*, est* droite-maximale *si* $S[j+1] \notin \mathcal{CS}(S[i,j])$*, et est* maximale *si elle est à la fois gauche-maximale et droite-maximale.*

Une $\mathcal{CS}$-position d'un sous-ensemble C de Σ dans S est une région contiguë dans le génome qui contient exactement les marqueurs contenus dans C, autorisant les copies multiples. C a une $\mathcal{CS}$-position dans S si et seulement si C a une $\mathcal{CS}$-position maximale dans S. Par exemple, si l'on considère le génome S_1=(1 2 3 1 2 3 4 5), la $\mathcal{CS}$-position du $\mathcal{CS}$ $\{1,2,3\}$ ira de la position

1 dans S_1 à la position 6 étant donné qu'à la position 7, on trouve un marqueur qui ne fait pas partie du $\mathcal{CS}$ $\{1, 2, 3\}$.

Définition 3 ($\mathcal{CS}$-facteur commun à k chaînes de caractères, [Schmidt and Stoye, 2004])**.** *Étant donné un ensemble de k chaînes de caractères $\mathcal{S} = (S_1, S_2, ..., S_k)$ sur l'alphabet Σ, un sous-ensemble $C \subseteq \Sigma$ est un $\mathcal{CS}$-facteur commun de $\mathcal{S}$ si et seulement si C a une $\mathcal{CS}$-position maximale dans chaque $S_l, 1 \leq l \leq k$.*

Un $\mathcal{CS}$-facteur commun à k génomes représente un groupe de marqueurs qui apparaît dans chacun des k génomes. Le concept est similaire à celui d'intervalle commun à k permutations sauf qu'ici on autorise les marqueurs paralogues dans les génomes et en particulier dans les groupes de marqueurs.

Par exemple, si on considère les génomes S_1=(1 2 3 1 2 3 4 5) et S_2=(1 2 3 4 5), le $\mathcal{CS}$ $\{1, 2, 3\}$ est un $\mathcal{CS}$-facteur commun aux deux génomes puisqu'il a une $\mathcal{CS}$-position maximale chez S_1 (de 1 à 6) et chez S_2 (de 1 à 3).

Cette méthode va donc permettre de retrouver des ensembles de marqueurs conservés entre plusieurs génomes en incluant les marqueurs dupliqués.

Intervalles connectés approximatifs (ACI). En se basant sur la méthode précédente, la méthode ACI proposée par [Amir et al., 2007] permet de trouver des intervalles communs entre plusieurs génomes en incluant les marqueurs dupliqués et autorisant des erreurs entre les intervalles communs. Les définitions utilisées ici sont les mêmes que pour la méthode CI.

Le principe général de cette méthode consiste à extraire tous les $\mathcal{CS}$-positions maximales de tous les $\mathcal{CS}$ des génomes entrés. Chaque $\mathcal{CS}$ va être représenté par un vecteur binaire de taille $|\Sigma|$ où on aura à la $i^{ème}$ position du vecteur (représentant le $i^{ème}$ caractère de Σ) un 1 si ce caractère est présent dans le $\mathcal{CS}$ sinon un 0. Par exemple, pour le génome S_1 =(3 2 3 1 4 5), avec $\Sigma = \{1, 2, 3, 4, 5\}$. Le vecteur binaire correspondant au $\mathcal{CS}$ $\{2, 3\}$ est [0,1,1,0,0] car 2 et 3 sont respectivement le deuxième et troisième caractère de Σ.

Tous les $\mathcal{CS}$ de taille f, noté $\mathcal{CS}_f$ (qui contiennent exactement f caractères) seront stockés dans des listes (par exemple, $liste^1$ pour les $\mathcal{CS}$ de taille 1). Le nombre de listes est égal à $|\Sigma|$. Il faut ensuite construire un graphe où les $\mathcal{CS}$ seront des nœuds et les arêtes relieront les $\mathcal{CS}_f$ de taille f aux $\mathcal{CS}_{f+1}$ de taille $f + 1$ lorsque $\mathcal{CS}_f \subseteq \mathcal{CS}_{f+1}$. Pour retrouver les $\mathcal{CS}_f$ inclus dans les $\mathcal{CS}_{f+1}$ il suffira alors de comparer les vecteurs binaires des $\mathcal{CS}$ des $listes^f$ avec ceux des $listes^{f+1}$. Un $\mathcal{CS}_f$ inclus dans un $\mathcal{CS}_{f+1}$ est un $\mathcal{CS}$ qui a un seul marqueur de différence avec le $\mathcal{CS}_{f+1}$. Les arêtes, reliant les $\mathcal{CS}_f$ aux $\mathcal{CS}_{f+1}$ représenteront donc une différence de un marqueur entre les deux nœuds. Des données périphériques sont associées aux nœuds : les $\mathcal{CS}$-positions maximales. Les arêtes sont étiquetées avec le marqueur manquant entre les deux nœuds. Au final, on obtient un *graphe ACI* contenant tous les $\mathcal{CS}$ présents dans les génomes.

La recherche d'un $\mathcal{CS}$ C donné avec k erreurs peut alors s'effectuer en parcourant le graphe : les $\mathcal{CS}$ ayant k erreurs avec C correspondent aux nœuds distants de k arêtes du nœud représentant C. Une application de cette méthode est donnée dans l'exemple 1.

Exemple 1. On veut comparer les deux génomes S_1 = (1 3 2 1 2 3 4 1 2 3 1 4 5), et S_2 = (4 2 1 4 3 1 2 4 1 3 5). On obtient alors les listes suivantes :

- $liste^1 = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}$
- $liste^2 = \{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{3, 5\}, \{4, 5\}\}$
- $liste^3 = \{\{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{1, 3, 5\}, \{1, 4, 5\}, \{2, 3, 4\}\}$
- $liste^4 = \{\{1, 2, 3, 4\}, \{1, 3, 4, 5\}\}$

– $liste^5 = \{\{1,2,3,4,5\}\}$

Le graphe obtenu est présenté Figure 5.1. Si l'on souhaite trouver les positions des intervalles communs contenant les marqueurs 1,3,4 et 5 ayant une erreur, il suffit alors de chercher les nœuds connectés à {1,3,4,5} pour lesquels on a une différence de 1. On trouve alors les intervalles {1,3,4}, {1,3,5}, {1,4,5} et {1,2,3,4,5}. Les données périphériques associées à chaque nœud permettent d'obtenir les positions de ces intervalles dans les séquences.

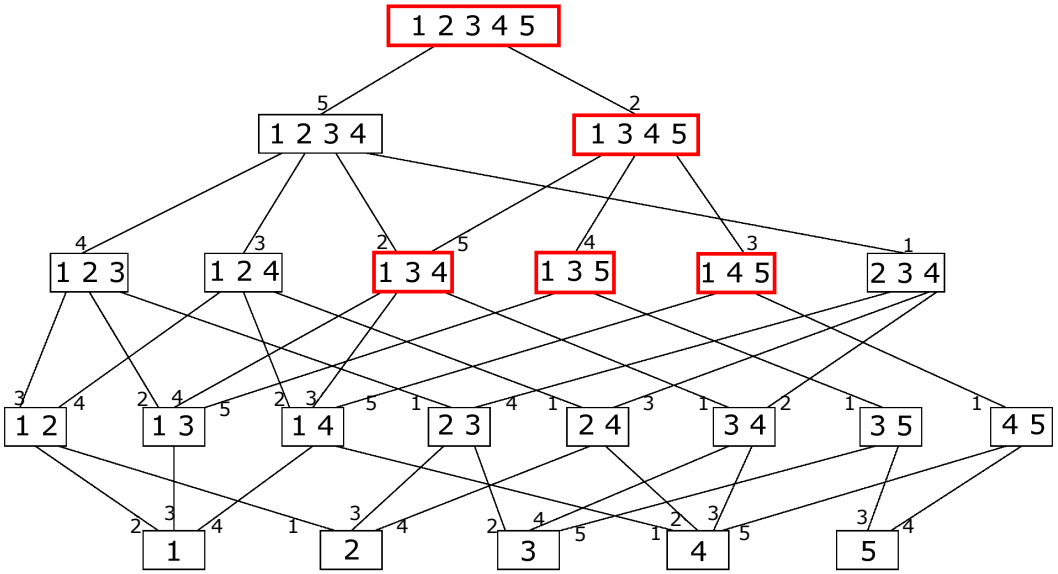


FIG. 5.1 – Graphe ACI [Amir et al., 2007] construit sur les génomes S_1 = (1 3 2 1 2 3 4 1 2 3 1 4 5), S_2 = (4 2 1 4 3 1 2 4 1 3 5) . Chaque rectangle représente un $\mathcal{CS}$. Les étiquettes des arêtes représentent les marqueurs perdus entre les $\mathcal{CS}$ reliés par ces arêtes. Les données périphériques des nœuds ne sont pas représentées.

5.2 Méthode proposée

La méthode ACI permet de retrouver des intervalles communs, approximatifs, entre plusieurs génomes en incluant les marqueurs dupliqués. Par contre elle n'est pas directement applicable pour la recherche d'intervalles communs dupliqués avec erreurs (correspondant à des duplications avec pertes). En effet, si l'on recherche uniquement des intervalles communs dupliqués, il faut alors regarder uniquement les nœuds du graphe pour lesquels, pour un génome donné, on trouve plusieurs positions. De plus, avoir des $\mathcal{CS}$-positions maximales va masquer les duplicats en tandem. Par exemple, dans le génome S_1=(1 2 3 1 2 3 4 5), le $\mathcal{CS}$ $\{1,2,3\}$ aura pour $\mathcal{CS}$-position maximale $(1,6)$. Dans ce cas, il n'aura qu'une seule position alors qu'il s'agit bien d'un groupe de marqueurs dupliqués.

Nous présentons maintenant l'adaptation de la méthode ACI à la recherche d'intervalles communs dupliqués afin d'obtenir des intervalles communs dupliqués avec erreurs. Du point de vue biologique, cela consiste à retrouver parmi un ensemble de génomes, des ensembles de

marqueurs dupliqués communs entre plusieurs génomes (ou unique à un génome) avec des pertes de marqueurs dans certaines copies des duplicats.

La seule différence avec la méthode ACI est que, pour chaque nœud du graphe ACI, les positions des $\mathcal{CS}$ seront établies à partir des $\mathcal{CS}$-positions minimales pour pouvoir repérer les $\mathcal{CS}$ dupliqués. Nous appelerons le graphe construit *graphe ADCI* (approximate duplicate common intervals). Nous avons également mis en place un système de filtre suite à la construction du graphe ADCI afin d'extraire les intervalles communs dupliqués les plus pertinents.

Définition 4 ($\mathcal{CS}$-position minimale)**.** *Soit σ un $\mathcal{CS}$, une $\mathcal{CS}$-position (i, j) de σ dans S est dite* minimale *si $S[i, j]$ est une permutation des éléments de σ.*

Une $\mathcal{CS}$-position minimale sera donc une chaîne de caractères ne contenant qu'une seule copie de chaque marqueur. Par exemple, dans le génome S_1=(1 2 3 1 2 3 4 5), pour le $\mathcal{CS}$ $\{1, 2, 3\}$, on aura plusieurs $\mathcal{CS}$-positions minimales : (1,3), (2,4), (3,5), (4,6).

Avoir un jeu de $\mathcal{CS}$-positions minimales ne va pas changer la topologie du graphe étant donné que le contenu en marqueurs d'une $\mathcal{CS}$-position minimale est le même que celui d'une $\mathcal{CS}$-position maximale. Seules les bornes sont différentes.

Définition 5 ($\mathcal{CS}$ dupliqué)**.** *Un $\mathcal{CS}$ σ de ℓ éléments est* dupliqué *dans une chaîne de caractères S s'il existe deux $\mathcal{CS}$-positions minimales non chevauchantes $C_1 = (i, i + \ell)$ et $C_2 = (j, j + \ell)$.*

Définition 6 ($\mathcal{CS}$ dupliqué en tandem)**.** *Un $\mathcal{CS}$ σ de ℓ éléments est* dupliqué en tandem *dans une chaîne de caractères S s'il existe deux $\mathcal{CS}$-positions minimales $C_1 = (i, i + \ell)$ et $C_2 = (i + \ell + 1, i + 2\ell + 1)$ dans S tels que $\forall n, 0 \leq n \leq \ell$, $S[i + n] = S[i + \ell + 1 + n]$.*

Définition 7 (bloc)**.** *On appellera* bloc *d'un $\mathcal{CS}$ dupliqué l'un ou l'autre des deux $\mathcal{CS}$-positions minimales sur la chaîne de caractères ou l'union des deux dans le cas d'un tandem.*

Par exemple, pour le $\mathcal{CS}$ $\{1, 2, 3\}$, dans le génome S_1=(1 2 3 1 2 3 4 5), il y aura un bloc et dans le génome S_2=(1 2 3 4 5 1 2 3) il y aura deux blocs.

Quand nous allons rechercher les $\mathcal{CS}$ dupliqués, il sera très important pour nous de ne considérer en tandem que les $\mathcal{CS}$ montrant de vrais motifs de duplication en tandem. Par exemple, pour le $\mathcal{CS}$ $\{1, 2, 3\}$ dans le génome S_1=(1 2 3 1 2 3 4 5), les $\mathcal{CS}$-positions minimales comportent bien les dupliqués dans le même ordre tandis que dans le génome S_2=(1 2 3 3 2 1 4 5), les $\mathcal{CS}$-positions minimales (par exemple aux positions (1,3) et (4,6)) montrent deux groupes de marqueurs dans des ordres différents, il s'agit donc d'une duplication en tandem remaniée que nous ne considèrerons pas comme une duplication en tandem.

Méthode. La méthode développée se décompose en 3 phases (Figure 5.2) :

- Dans la première (section 5.2.1), un graphe ADCI est construit avec un étiquetage pour chaque nœud du graphe. Un nœud représente un $\mathcal{CS}$. Pour chaque $\mathcal{CS}$ on aura pour chaque génome des informations permettant de retrouver les différents blocs.
- La deuxième phase (section 5.2.2) consistera à appliquer un filtre pour chaque $\mathcal{CS}$ d'un génome afin d'éliminer les $\mathcal{CS}$ les moins intéressants.
- Dans la troisième phase (section 5.2.3), les meilleurs $\mathcal{CS}$ seront sélectionnés en comparant les $\mathcal{CS}$ pour l'ensemble des génomes.

Un même exemple servira de guide afin de comprendre la méthode tout au long de son explication.

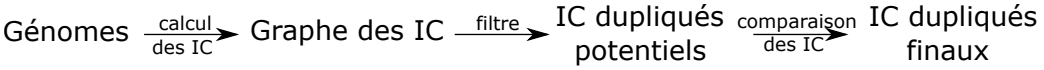


FIG. 5.2 – Procédure générale de la méthode de détection de fragments dupliqués en trois phases (IC = intervalles communs).

5.2.1 Phase 1 : construction du graphe ADCI

L'algorithme de [Amir et al., 2007] ne peut pas être directement utilisé car il est basé sur la recherche de $\mathcal{CS}$-positions maximales qui masquent les événements de duplications en tandem. Nous allons utiliser la notion de $\mathcal{CS}$-position minimale, ce qui ne va pas changer la méthode de construction du graphe. Par contre, on aura un plus grand nombre d'informations associées à chaque nœud du graphe. En effet, pour le génome S_1=(1 2 3 1 2 3 4 5), l'algorithme de [Amir et al., 2007] va associer la $\mathcal{CS}$-position maximale (1,6) au nœud {1,2,3} tandis que notre méthode associera les $\mathcal{CS}$-positions minimales (1,3),(2,4),(3,5) et (4,6).

Dans un premier temps, tous les $\mathcal{CS}$-positions minimales de tous les $\mathcal{CS}$ de tous les génomes sont extraites en commençant par les $\mathcal{CS}$ de taille 1 jusqu'aux $\mathcal{CS}$ de taille $|\Sigma|$. Les $\mathcal{CS}$, représentés sous forme de vecteurs binaires, sont placés dans des listes en fonction de leur taille. Un graphe est ensuite construit en comparant les $\mathcal{CS}_f$ aux $\mathcal{CS}_{f+1}$. Une arête existe entre deux $\mathcal{CS}$ de $\mathcal{CS}_f$ et $\mathcal{CS}_{f+1}$ si on observe une différence de 1 entre les vecteurs binaires correspondant à ces $\mathcal{CS}$. La procédure de construction du graphe, est décrite dans la Procédure 1 et l'Exemple 2 illustre l'application de cette procédure.

Procédure 1 Construction du graphe ADCI avec les $\mathcal{CS}$-positions minimales

Entrés: $\mathcal{S} = (S_1, S_2, ..., S_k)$
 # tous les génomes sous forme de chaîne de caractères
pour i=1 à $|\mathcal{S}|$ **faire**
  pour f=1 à $|\Sigma|$ **faire**
    extraire $\mathcal{CS}$-position minimale de taille f de S_i
    # cas de génomes circulaires, on cherche $\mathcal{CS}$-positions minimales de taille f dans S_i avec S_i=$S_i[1,|S_i|]$+$S_i[1,f-1]$
    créer vecteur binaire correspondant à chaque $\mathcal{CS}$-position minimale
    mémoriser bornes et génome de ce vecteur binaire
    ajouter vecteur binaire à $liste_f$
  fin pour
fin pour
pour f=1 à $|\Sigma|$ **faire**
  $liste_f$=tous les $\mathcal{CS}$ de taille f
  tri lexicographique de $liste_f$
  rassembler les $\mathcal{CS}_f$ identiques (en ajoutant les positions et génomes si f=1)
  # on ajoute uniquement les positions et génomes des $\mathcal{CS}_1$
  pour tout $\mathcal{CS}_f$ uniques dans $liste_f$ **faire**
    création d'un nœud correspondant au $\mathcal{CS}_f$ avec toutes les informations de bornes et génomes si f=1
  fin pour
fin pour
pour f=1 à $|\Sigma|-1$ **faire**
  pour tout $\mathcal{CS}_f$ dans $liste_f$ **faire**
    pour tout $\mathcal{CS}_{f+1}$ dans $liste_{f+1}$ **faire**
      comparaison des vecteurs binaires f et $f+1$
      si différence symétrique entre $\mathcal{CS}_f$ et $\mathcal{CS}_{f+1}$ = 1 **alors**
        création d'une arête entre $\mathcal{CS}_{f+1}$ et $liste_{f+1}$ avec la différence
        # les nœuds et arêtes sont stockés dans une matrice mat[$\mathcal{CS}_{f+1}$][$\mathcal{CS}_f$]=arête entre $\mathcal{CS}_f$ et $\mathcal{CS}_{f+1}$
      fin si
    fin pour
  fin pour
fin pour
Sortie: graphe ADCI

Exemple 2. On compare deux génomes S_1=(1 2 3 1 2 3 4 5) et S_2=(1 2 4 5 1 3). Les lignes 1 à 8 de la Procédure 1 vont créer les listes de $\mathcal{CS}$ en fonction de leur taille. Pour la liste de taille 1, les positions dans les génomes sont mémorisées. Étant donné que l'on ne considère que les $\mathcal{CS}$-positions minimales et que l'on compare plusieurs génomes, ces listes vont comporter des redondances (pour simplifier celles-ci n'ont pas été données).

- $liste_1 = \{\{1\},\{2\},\{3\},\{4\},\{5\}\}$:
  - $\{1\}$ aux positions (1,1) et (4,4) dans S_1 et (1,1) et (5,5) dans S_2
  - $\{2\}$ aux positions (2,2) et (5,5) dans S_1 et (2,2) dans S_2
  - $\{3\}$ aux positions (3,3) et (6,6) dans S_1 et (6,6) dans S_2
  - $\{4\}$ aux positions (7,7) dans S_1 et (3,3) dans S_2
  - $\{5\}$ aux positions (8,8) dans S_1 et (4,4) dans S_2
- $liste_2 = \{\{1,2\},\{1,3\},\{1,5\},\{2,3\},\{2,4\},\{3,4\},\{4,5\}\}$
- $liste_3 = \{\{1,2,3\},\{1,2,4\},\{1,3,5\},\{2,3,4\},\{2,4,5\},\{3,4,5\}\}$
- $liste_4 = \{\{1,2,3,4\},\{1,2,4,5\},\{1,3,4,5\},\{2,3,4,5\}\}$
- $liste_5 = \{\{1,2,3,4,5\}\}$

Les lignes 9 à 16 vont créer les nœuds correspondant à ces $\mathcal{CS}$ en rassemblant les $\mathcal{CS}$ redondants et les lignes 17 à 26 vont créer les arêtes entre chaque $\mathcal{CS}$ de taille f et $f+1$, auxquelles sera associée la différence. Le résultat obtenu est donné Figure 5.3. Pour des raisons de lisibilité seules les différences entre le $\mathcal{CS}$ $\{1\}$ et ses pères ($\mathcal{CS}$ de taille 2 ayant une différence de 1 avec le $\mathcal{CS}$ $\{1\}$) ont été représentées.

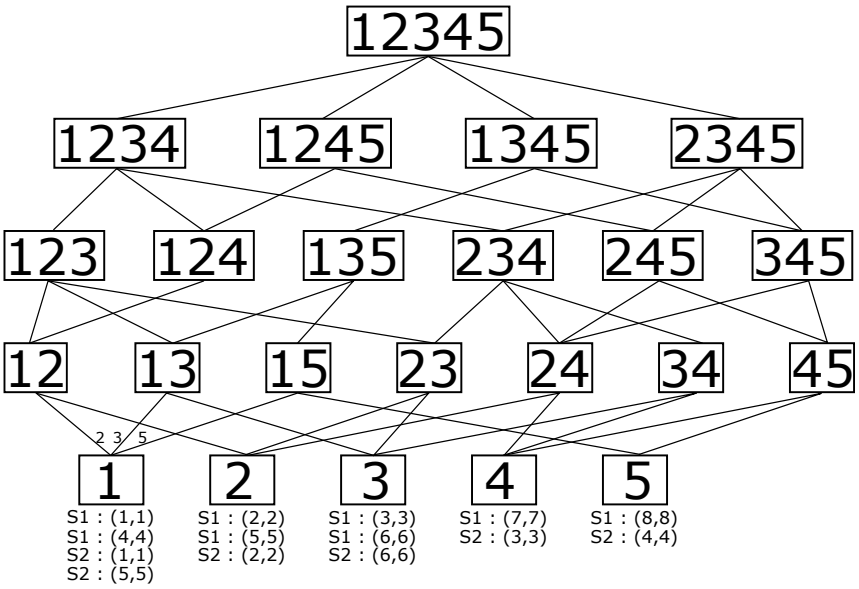


FIG. 5.3 – Graphe ADCI obtenu pour les génomes S_1=(1 2 3 1 2 3 4 5) et S_2=(1 2 4 5 1 3). Pour chaque génome, les $\mathcal{CS}$-positions minimales sont indiquées. Les valeurs sur les arêtes indiquent la différence entre un $\mathcal{CS}$ (en dessous) et son père (au dessus).

Une fois le graphe construit, il faut ajouter les bornes de chaque $\mathcal{CS}$ aux nœuds correspondants. Cet ajout va se faire en traitant le graphe par taille croissante des $\mathcal{CS}$ à partir des $\mathcal{CS}$ de taille 2, en considérant tous les $\mathcal{CS}$ fils d'un $\mathcal{CS}$. En effet, pour un $\mathcal{CS}$ on regarde tous les $\mathcal{CS}$ de taille juste inférieure qui lui sont attachés. Le fonctionnement général est décrit en Procédure 2.

Procédure 2 Ajout des bornes de chaque $\mathcal{CS}$

pour f de 2 à $|\Sigma|$ **faire**
  pour tout $\mathcal{CS}_f$ de chaque génome **faire**
    liste-fils ←liste des fils du $\mathcal{CS}_f$ regardé
    créer $liste - vu$ qui contiendra les fils de $\mathcal{CS}_f$ qui ont été vus
    pour i=0 à $|liste\text{-}fils| - 1$ **faire**
      $\mathcal{CS} - fils_i = liste\text{-}fils_i$
      pour j=$i + 1$ à $|liste - fils|$ **faire**
        $\mathcal{CS}$-$fils_j$=$liste$-$fils_j$
        comparer toutes les positions de $\mathcal{CS} - fils_i$ et $\mathcal{CS} - fils_j$
        si positions juxtaposées **alors**
          position-gauche est la position minimale entre $\mathcal{CS}$-$fils_i$ et $\mathcal{CS}$-$fils_j$
          position-droite est la position maximale entre $\mathcal{CS}$-$fils_i$ et $\mathcal{CS}$-$fils_j$
          si f=2 **alors**
            position-gauche($\mathcal{CS}_f$) ← *position-gauche*
            position-droite($\mathcal{CS}_f$) ← *position-droite*
            ajouter $\mathcal{CS} - fils_i$ et $\mathcal{CS} - fils_j$ à $liste - vu$
          fin si
          si f>2 **alors**
            $erreur$ ← une occurrence des erreurs contenues dans $\mathcal{CS}$-$fils_i$ et $\mathcal{CS}$-$fils_j$
            si (*position-droite*) − (*position-gauche*) + $|erreur| = |\mathcal{CS}_f|$ **alors**
              position-gauche($\mathcal{CS}_f$) ← *position-gauche*
              position-droite($\mathcal{CS}_f$) ← *position-droite*
              $erreurs(\mathcal{CS}_f)$ ← $erreurs(\mathcal{CS}\text{-}fils_i)$ qui ne sont pas compensées par les $erreurs(\mathcal{CS}\text{-}fils_j)$ et inversement
              ajouter $\mathcal{CS} - fils_i$ et $\mathcal{CS} - fils_j$ à $liste - vu$
            fin si
          fin si
        fin si
        si positions chevauchantes **alors**
          position-gauche est la position minimale entre $\mathcal{CS}$-$fils_i$ et $\mathcal{CS}$-$fils_j$
          position-droite est la position maximale entre $\mathcal{CS}$-$fils_i$ et $\mathcal{CS}$-$fils_j$
          position-gauche($\mathcal{CS}_f$) ← *position-gauche*
          position-droite($\mathcal{CS}_f$) ← *position-droite*
          $erreurs(\mathcal{CS}_f)$ ← $erreurs(\mathcal{CS}\text{-}fils_i)$ qui ne sont pas compensées par les $erreurs(\mathcal{CS}\text{-}fils_j)$ et inversement
          ajouter $\mathcal{CS} - fils_i$ et $\mathcal{CS} - fils_j$ à $liste - vu$
        fin si
      fin pour
      si une des positions de $\mathcal{CS} - fils_i$ n'est pas dans $liste - vu$ **alors**
        position-gauche($\mathcal{CS}_f$) ← *position-gauche*($\mathcal{CS}$-$fils_i$)
        position-droite($\mathcal{CS}_f$) ← *position-droite*($\mathcal{CS}$-$fils_i$)
        $erreurs(\mathcal{CS}_f)$ ← $erreurs(\mathcal{CS}\text{-}fils_i)$ et différence entre $\mathcal{CS}_f$ et $\mathcal{CS}$-$fils_i$
      fin si
    fin pour
    si une des positions de $\mathcal{CS} - fils_{|\Sigma|}$ n'est pas dans $liste - vu$ **alors**
      position-gauche($\mathcal{CS}_f$) ← *position-gauche*($\mathcal{CS}$-$fils_{|\Sigma|}$)
      position-droite($\mathcal{CS}_f$) ← *position-droite*($\mathcal{CS}$-$fils_{|\Sigma|}$)
      $erreurs(\mathcal{CS}_f)$ ← $erreurs(\mathcal{CS}\text{-}fils_{|\Sigma|})$ et différence entre $\mathcal{CS}_f$ et $\mathcal{CS}$-$fils_{|\Sigma|}$
    fin si
  fin pour
fin pour

Sortie: graphe ADCI avec les positions pour chaque nœud

L'Exemple 3 illustre son fonctionnement. Un point important est que pour prendre en compte des erreurs, ce ne sont pas nécessairement les $\mathcal{CS}$-positions minimales qui sont conservées dans le graphe mais une position (un intervalle sur la chaîne) et la liste des marqueurs manquants. Ainsi, si on ajoute virtuellement dans cet intervalle les marqueurs manquants, la position stockée correspondra à une $\mathcal{CS}$-position minimale. Cependant, pour chaque nœud il existe au moins un génome pour lequel c'est effectivement une $\mathcal{CS}$-position minimale qui est stockée puisqu'il faut au moins une occurrence d'un $\mathcal{CS}$ pour que le nœud correspondant apparaisse dans le graphe.

Exemple 3. Nous poursuivons la comparaison des génomes S_1=(1 2 3 1 2 3 4 5) et S_2=(1 2 4 5 1 3). La Figure 5.4 montre un sous-exemple du remplissage des bornes pour le graphe ADCI obtenu précédemment. On commence par remplir les bornes pour les $\mathcal{CS}$ de taille 2. Le premier $\mathcal{CS}$ de taille 2 est $\{1,2\}$, ses fils sont les $\mathcal{CS}$ $\{1\}$ et $\{2\}$.

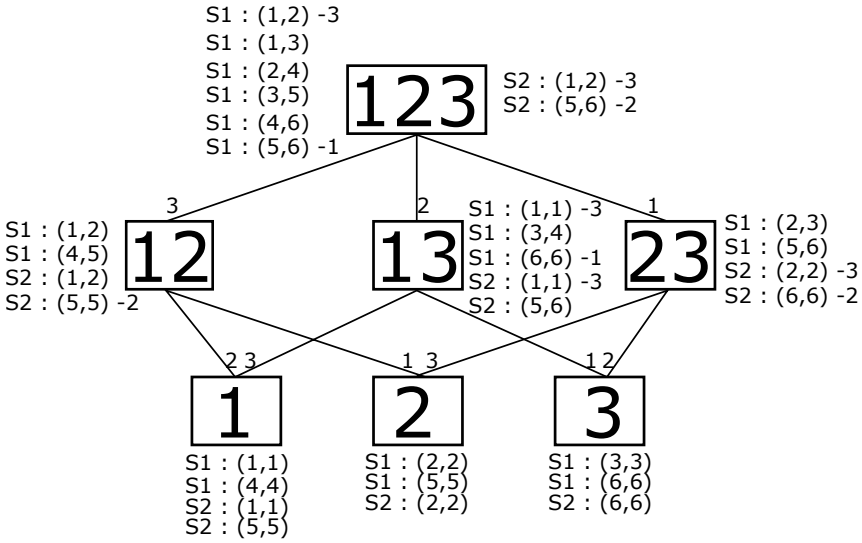


FIG. 5.4 – Remplissage des bornes des $\mathcal{CS}$ pour les génomes S_1=(1 2 3 1 2 3 4 5) et S_2=(1 2 4 5 1 3) sur un extrait du graphe présenté en Figure 5.3.

Pour le génome S_1, on a une occurrence du $\mathcal{CS}$ $\{1\}$ en position (1,1) et une occurrence du $\mathcal{CS}$ $\{2\}$ en position (2,2). Ces deux occurrences sont juxtaposées, la position pour le $\mathcal{CS}$ $\{1,2\}$ sera donc (1,2). Nous trouvons également une occurrence du $\mathcal{CS}$ $\{1\}$ en position (4,4) juxtaposée à une occurrence du $\mathcal{CS}$ $\{2\}$ en position (5,5), on aura donc une nouvelle position pour le $\mathcal{CS}$ $\{1,2\}$ qui sera (4,5).

Pour le génome S_2, l'occurrence du $\mathcal{CS}$ $\{1\}$ en position (1,1) est juxtaposée à celle du $\mathcal{CS}$ $\{2\}$ en position (2,2). La position du $\mathcal{CS}$ $\{1,2\}$ est donc (1,2). Nous avons également une occurrence du $\mathcal{CS}$ $\{1\}$ en position (1,1) qui n'est juxtaposée avec aucune autre occurrence du $\mathcal{CS}$ $\{2\}$. On aura dans ce cas, pour le $\mathcal{CS}$ $\{1,2\}$ une occurrence en position (1,1) avec la perte de 2 (perte notée sur l'arête reliant le $\mathcal{CS}$ $\{1\}$ au $\mathcal{CS}$ $\{1,2\}$).

Une fois toutes les positions des $\mathcal{CS}$ de taille 2 établies, on passe aux $\mathcal{CS}$ de taille 3. Ici regarde le $\mathcal{CS}$ $\{1,2,3\}$ qui a trois fils : $\{1,2\}$, $\{1,3\}$ et $\{2,3\}$.

Pour le génome S_1, si l'on compare les $\mathcal{CS}$ $\{1,2\}$ et $\{1,3\}$, on trouve une occurrence du $\mathcal{CS}$ $\{1,2\}$ en position (1,2) chevauchante à une occurrence du $\mathcal{CS}$ $\{1,3\}$ en position (1,1). La position du $\mathcal{CS}$ $\{1,2,3\}$ de cette occurrence sera (1,2) avec la perte du marqueur 3 provenant du $\mathcal{CS}$ $\{1,3\}$

qui n'est pas apporté par le $\mathcal{CS}$ $\{1,2\}$ (le $\mathcal{CS}$ $\{1,2\}$ ne contient pas le marqueur 3). On trouve aussi que le $\mathcal{CS}$ $\{1,2\}$ en position (1,2) est juxtaposé au $\mathcal{CS}$ $\{1,3\}$ en position (3,4). Ceci devrait nous conduire à une occurrence du $\mathcal{CS}$ $\{1,2,3\}$ en position (1,4). Mais la taille de cet intervalle dépasse la taille du $\mathcal{CS}$ $\{1,2,3\}$, il n'est donc pas pris en compte puisqu'on ne veut garder que des $\mathcal{CS}$-positions minimales, c'est-à-dire les $\mathcal{CS}$ ne contenant qu'une occurrence des dupliqués.

Pour le génome S_2, si l'on compare le $\mathcal{CS}$ $\{1,2\}$ au $\mathcal{CS}$ $\{2,3\}$ on observe une occurrence du $\mathcal{CS}$ $\{1,2\}$ en position (5,5) juxtaposée avec une occurrence du $\mathcal{CS}$ $\{2,3\}$ en position (6,6). Ces deux occurrences ont perdu le marqueur 2. L'occurrence du $\mathcal{CS}$ $\{1,2,3\}$ correspondante sera donc en position (5,6) avec la perte de 2. Dans ce cas l'intervalle (5,6) est de taille 2, avec la perte de 2 on arrive à une taille de 3 correspondant à la taille du $\mathcal{CS}$ $\{1,2,3\}$.

Lorsque nous avons toutes les positions de tous les $\mathcal{CS}$ du graphe, il faut fusionner les positions chevauchantes et repérer les éléments dupliqués. Une position strictement incluse dans une autre sera éliminée. Les positions chevauchantes seront rassemblées en considérant qu'il s'agit d'une duplication en tandem (si l'ordre d'apparition des dupliqués est conservé) avec perte des éléments chevauchants. Après cette phase, les informations périphériques des nœuds contiennent l'information sur les occurrences de blocs. Le fonctionnement est décrit en Procédure 3, l'Exemple 4 en montre le fonctionnement.

Exemple 4. Nous allons de nouveau nous appuyer sur la Figure 5.4 pour illustrer le fonctionnement de la Procédure 3. Dans les lignes 3 à 27 de la Procédure 3, nous rassemblons les positions chevauchantes des $\mathcal{CS}$. Lorsque deux positions d'un $\mathcal{CS}$ se chevauchent et que nous les rassemblons cela signifie que, pour la nouvelle position du $\mathcal{CS}$, nous aurons des marqueurs dupliqués. A chaque fois que nous rassemblons deux positions chevauchantes, il faut que le nombre de duplications d'un $\mathcal{CS}$ à la nouvelle position soit égal à la somme du nombre de duplications des positions chevauchantes. Dans notre cas, pour les $\mathcal{CS}$ de taille 1 et 2 il n'y a rien à faire. En effet, il n'y a pas de position chevauchante, ni de position juxtaposée.

Dans le cas du $\mathcal{CS}$ $\{1,2,3\}$, pour le génome S_1, la première position est (1,2) avec la perte de 3. Elle est strictement incluse dans (1,3) donc on élimine la position (1,2) et on passe à la position suivante (1,3). Celle-ci chevauche la position suivante (3,5) sur le marqueur en position 3, on aura donc une nouvelle position (1,5) avec perte du marqueur en position 3 (correspond au marqueur 1). Cette position (1,5) correspondra alors à la position d'une duplication en tandem du $\mathcal{CS}$ {1,2,3} avec perte du marqueur 1. La position suivante (4,6) chevauche (1,5) sur les marqueurs en position 4 et 5. On aura donc une nouvelle position (1,6) avec perte du marqueur 1 (provenant de la position (1,5)) mais aussi des marqueurs en position 4 et 5 qui sont chevauchants. La position (1,6) correspondra donc à une triplication (quand on ajoute une position, on incrémente le nombre de duplicats) avec perte des marqueurs 1, 2 et 3.

Dans le cas du génome S_2, le $\mathcal{CS}$ $\{1,2,3\}$ aura deux positions : une position en (1,2) avec perte du marqueur 3 (la position (2,2) est strictement incluse dans (1,2) donc on ne la prend pas en compte) et une position en (5,6) avec perte du marqueur 2.

Dans les lignes 28 à 32 de la Procédure 3, on diminue le nombre de duplicats prédits de un, tant que toutes les copies d'un marqueur sont perdues. Par exemple, pour le génome S_1, la position (1,6) du $\mathcal{CS}$ $\{1,2,3\}$ est décrite comme une triplication avec perte des marqueurs 1,2 et 3. On peut alors considérer qu'il s'agit d'une duplication en tandem du $\mathcal{CS}$ {1,2,3} sans aucune perte. La position (1,6) sera alors considérée comme une duplication en tandem et le nombre de copies à cette position sera deux. A ce stade, les positions d'un $\mathcal{CS}$ sont les positions des blocs constituant le $\mathcal{CS}$, le nombre de copies à une position est le nombre de copies d'un bloc et le nombre de positions du $\mathcal{CS}$ représente le nombre de blocs pour ce $\mathcal{CS}$. Par exemple pour le $\mathcal{CS}$

Procédure 3 Condensation des intervalles

pour tout $\mathcal{CS}$ de chaque génome **faire**
  ranger par ordre de *position-gauche* puis *position-droite* croissant toutes les positions du *cs*
  pour i=1 à nombre de positions du *cs* **faire**
    pos=*position*$_i$ du *cs*
    tant que pos_{i+1} chevauche *pos* et marqueurs dupliqués de pos_{i+1} sont dans le même ordre que ceux de *pos* **faire**
      si intervalle de pos_{i+1} n'est pas strictement inclus dans pos_i ou inversement **alors**
        incrémenter duplication de l'intervalle
        position-droite(*pos*)←position-droite(pos_{i+1})
        erreur(*pos*)←erreur(*pos*),erreur(pos_{i+1}),tous les éléments chevauchants entre *pos* et pos_{i+1}
        éliminer pos_{i+1}
        incrémenter i
      fin si
      si pos_{i+1} est strictement inclus dans *pos* **alors**
        incrémenter i
        pos_{i+1} est éliminée
      fin si
      si *pos* est strictement inclus dans pos_{i+1} **alors**
        $pos \leftarrow pos_{i+1}$
        # *b*ornes et erreurs sont uniquement celles de pos_{i+1}
        incrémenter i
      fin si
    fin tant que
    tant que toutes les occurrences des marqueurs du $\mathcal{CS}$, pour *pos*, sont perdues **faire**
      diminuer nombre de duplications de *pos* de 1
      enlever une occurrence de chaque marqueur dans erreur(*pos*)
    fin tant que
  fin pour
fin pour
*c*as de génomes circulaires
pour $\mathcal{CS}$ de chaque génome **faire**
  pos-min ← position avec la plus petite *position-gauche*
  pos-max ← position avec la plus grande *position-droite*
  si *pos-min* et *pos-max* ne sont pas juxtaposée **alors**
    si *position-gauche* de *pos-min*=1 et *position-droite* de *pos-max*=|*genome*| **alors**
      position-droite(*pos-max*)← *position-droite*(*pos-min*)
      erreur(*pos-max*)←erreur(*pos-max*)+erreur(*pos-min*)
      nombre de duplications de *pos-max* ←duplications de *pos-min*+duplications de *pos-max*
      éliminer *pos-min*
      tant que toutes les occurrences des marqueurs du $\mathcal{CS}$, pour *pos-max*, sont perdues **faire**
        diminuer nombre de duplications de *pos-max* de 1
        enlever une occurrence de chaque marqueur dans erreur(*pos-max*)
      fin tant que
    fin si
  fin si
fin pour
pour tout $\mathcal{CS}$ de chaque génome **faire**
  si le $\mathcal{CS}$ n'a qu'une position et que le nombre de duplicats de cette position<2 **alors**
    ce $\mathcal{CS}$ n'est pas dupliqué dans le génome
    la position est éliminée
  fin si
fin pour

$\{1, 2, 3\}$ on a un bloc (avec deux copies) dans le génome S_1 et deux blocs (tous les deux avec une copie) dans le génome S_2. Le résultat de cette procédure est donné Figure 5.5.

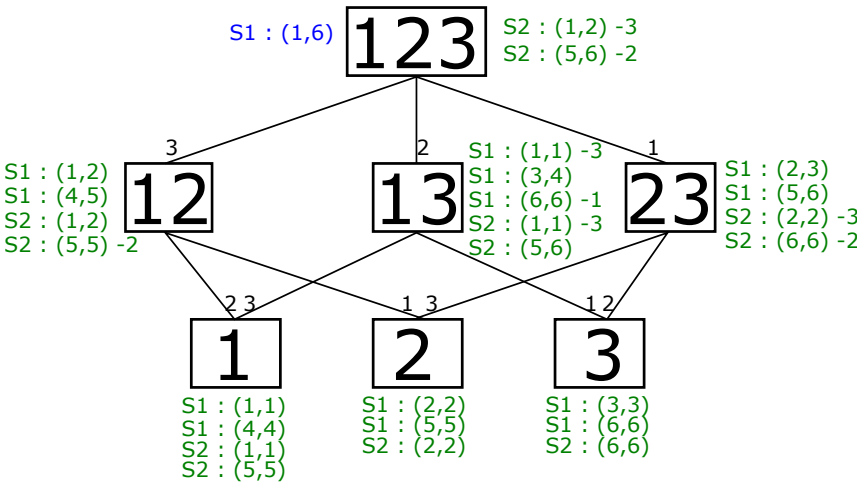


FIG. 5.5 – Condensation des positions des $\mathcal{CS}$ pour les génomes S_1=(1 2 3 1 2 3 4 5) et S_2=(1 2 4 5 1 3) sur un extrait du graphe présenté en Figure 5.3. Pour chaque $\mathcal{CS}$ les blocs ayant deux copies sont indiqués en bleu et ceux ayant une copie sont indiqués en vert.

Le graphe ADCI obtenu ici est donc un graphe représentant les différents intervalles communs à l'ensemble des génomes. Pour chacun de ces intervalles communs, on connaîtra grâce aux données périphériques le nombre de blocs, les positions des blocs, les pertes de marqueurs pour chacun des génomes.

La taille du graphe dépendra des intervalles communs trouvés parmi les génomes. En effet, il y aura autant de nœuds que d'intervalles communs différents entre tous les génomes. Donc plus les génomes seront proches (au niveau des suites de marqueurs) moins il y aura de nœuds. A l'opposé, plus les génomes seront réarrangés plus le graphe sera grand. Le graphe peut donc atteindre très facilement des centaines de nœuds. Étant donné que le nombre d'erreurs est indéterminé, le graphe ADCI construit aboutira à tous les ensembles possibles d'intervalles communs dupliqués avec erreurs.

Cependant, tous ces intervalles communs ne sont pas intéressants. Considérons par exemple le génome S_1=(1 2 3 4 5 6 7 8 9 1). Parmi les intervalles communs dupliqués on trouvera une duplication en tandem de {1,2,3,4,5,6,7,8,9} avec perte de 2,3,4,5,6,7,8 et 9, cet événement ne semble pas parcimonieux. On trouvera également des duplications comme {1} sans aucune perte ou encore {1,2} avec perte de 2. Par contre si on considère un second génome S_2=(1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9), un événement de duplication en tandem de {1,2,3,4,5,6,7,8,9} dans les deux génomes avec des pertes dans S_1 devient alors parcimonieux.

Afin de diminuer le nombres d'intervalles communs non pertinents et réduire la taille du graphe, nous allons mettre en place des filtres de sélection. On espère ainsi extraire les intervalles communs les plus intéressants que ce soit au niveau d'un génome et de l'ensemble des génomes.

5.2.2 Phase 2 : filtrage des intervalles communs

Ce filtre sert à éliminer, pour chaque génome, les intervalles communs les moins intéressants pour éviter de générer du bruit dans l'étape finale et ainsi réduire son temps d'exécution. Ce filtre prend en compte deux paramètres : le découpage et la conservation. Ces deux paramètres vont être couplés afin de donner un score à chaque intervalle commun de chaque génome. Ce score servira à déterminer les intervalles communs les plus intéressants. Comme nous cherchons ici les intervalles communs dupliqués entre les génomes, nous ne regarderons que les intervalles communs ayant plusieurs blocs. Si l'intervalle commun n'a qu'un bloc, il faut alors que le nombre de copies du bloc soit supérieur à un c'est-à-dire qu'il soit un intervalle commun dupliqué en tandem.

Découpage. Le découpage va permettre de garder, pour un génome, les intervalles communs dupliqués les moins séparés. Il sert donc à évaluer le nombre de réarrangements qui ont pu apparaître sur l'intervalle commun suite à sa duplication.

Pour un intervalle commun donné i, on introduit deux mesures. $d_{\text{obs}}(S, i)$ représente le nombre de blocs observés de l'intervalle commun i dans un génome donné S, c'est le nombre d'occurrences le plus élevé d'un des marqueurs de l'intervalle commun dans le génome S. $d_{\text{th}}(i)$ représente le nombre maximum de blocs observés de l'intervalle commun i dans l'un des génomes de l'ensemble des génomes $\mathcal{S}$ étudié, c'est $\max_{S \in \mathcal{S}} d_{\text{obs}}(S, i)$.

Définition 8 (découpage d'un intervalle commun)**.** *Soit un génome S, et i un intervalle commun de S. Le score de* découpage *est défini par :*

$$dec(S, i) = \frac{d_{th}(i)}{d_{obs}(S, i)}$$

On a la propriété que $dec(S, i) > 0$.

Un exemple du calcul du découpage est donné dans l'exemple 5.

Exemple 5. Considérons les génomes précédents S_1=(1 2 3 1 2 3 4 5) et S_2=(1 2 4 5 1 3). Dans le graphe ADCI (Figure 5.5), pour l'intervalle commun $\{1, 2, 3\}$ nous avons un bloc chez S_1, en position (1,6), et deux blocs chez S_2, en positions (1,2) avec perte du marqueur 3 et (5,6) avec perte du marqueur 2. On cherche $d_{\text{th}}(\{1, 2, 3\})$. Parmi tous les génomes, l'occurrence la plus élevée d'un des marqueurs de l'intervalle commun $\{1, 2, 3\}$ est deux et par conséquent $d_{\text{th}}(\{1, 2, 3\}) = 2$. Chez S_1 le nombre de blocs observés de l'intervalle commun $\{1, 2, 3\}$ est 1, donc $\text{dec}(S_1, \{1, 2, 3\}) = \frac{2}{1}$. Chez S_2 le nombre de blocs observés de l'intervalle commun $\{1, 2, 3\}$ est 2, donc $\text{dec}(S_2, (\{1, 2, 3\}) = \frac{2}{2}$.

Si nous avions eu un troisième génome S_3=(1 2 3 1 3 4 5 1 2) nous aurions eu $d_{\text{th}}(\{1, 2, 3\}) = 3$ puisqu'il y a trois occurrences du marqueur 1. L'ajout de ce génome entraîne donc une modification de d_{th} des intervalles communs contenant le marqueur 1. Les scores de découpage seront donc différents, prenant en compte une hypothétique triplication puisque l'on en trouve des traces chez S_3.

Conservation. Ce paramètre sert à évaluer le nombre de marqueurs perdus après la duplication d'un intervalle commun. Il va donc permettre d'éliminer les intervalles communs pour lesquels on a des traces trop faibles de duplication. Ce calcul tient compte du nombre de marqueurs impliqués dans les blocs de cet intervalle commun ainsi que le nombre de marqueurs qui devraient être présents si aucune copie n'avait disparu.

Préalablement, on définit $n(S, i)$ le nombre de marqueurs de l'ensemble des blocs de l'intervalle commun i dans S (n peut être vu comme la somme des tailles des blocs de l'intervalle commun).

Définition 9 (conservation d'un intervalle commun)**.** *Soit un génome S et i un intervalle commun donné de S. Le score de* conservation *est défini par :*

$$cons(S, i) = \frac{n(S, i)}{d_{th}(i) \times |i|}$$

On a la propriété que $0 \leq cons(i) \leq 1$.

Exemple 6. Considérons les génomes précédents S_1=(1 2 3 1 2 3 4 5) et S_2=(1 2 4 5 1 3). Dans le graphe ADCI (Figure 5.5), pour l'intervalle commun {1,2,3} qui est un intervalle de taille 3, on trouve un bloc de taille 6 chez S_1, $n(S_1, \{1, 2, 3\}) = 6$, et deux blocs de taille 2 chez S_2, $n(S_2, \{1, 2, 3\}) = 4$. Nous avons déterminé, dans l'Exemple 5, que $d_{\text{th}} = 2$. Donc d'après la Définition 9, $\text{cons}(S_1, (\{1, 2, 3\}) = \frac{6}{2\times3}$=1 et $\text{cons}(S_2, (\{1, 2, 3\}) = \frac{4}{2\times3}$=$\frac{2}{3}$.

On peut noter que $n(S, i)$ représente le nombre de marqueurs (incluant les dupliqués) composant un intervalle commun et $d_{\text{th}} \times |i|$ représente le nombre de marqueurs que l'on aurait dû avoir dans tous les blocs si aucune copie n'avait été perdue.

Filtrage. On définit maintenant le score utilisé pour le filtre. Celui-ci va prendre en compte à la fois le score de découpage et celui de conservation en multipliant simplement ces deux scores.

Définition 10 (score d'un intervalle commun)**.** *Soit S un génome et i un intervalle commun de S, on définit le* score *de i comme :*

$$f(S, i) = dec(S, i) \times cons(S, i)$$

Ce score va être comparé à un seuil. Si il est supérieur au seuil, l'intervalle commun pour ce génome est conservé sinon il est éliminé. Le seuil est fixé par l'utilisateur. C'est la seule variable de toute la méthode. La valeur de celui-ci sera discutée dans la Section 5.3.1.

5.2.3 Phase 3 : obtention des intervalles communs dupliqués

Les intervalles communs ayant passé le filtre vont ensuite être sélectionnés en prenant en compte leur répartition dans l'ensemble des génomes. La particularité de cette deuxième phase va être d'inclure plusieurs paramètres pour définir un poids afin de pouvoir comparer les intervalles communs partageant les mêmes marqueurs, qu'on qualifiera de chevauchants. On sélectionnera parmi les intervalles communs chevauchants le meilleur intervalle commun en considérant l'ensemble des génomes. En effet, sur des jeux de données comprenant plusieurs duplications, certains intervalles communs sélectionnés peuvent partager des marqueurs, il faudra alors choisir l'intervalle commun le plus intéressant pour une duplication (Exemple 7). Les intervalles communs chevauchant sont facilement repérables grâce aux vecteurs binaires qui leur sont associés.

Exemple 7. Considérons les génomes S_1=(1 2 3 1 2 3 4 5), S_2=(1 2 4 5 1 3) et S_3=(1 2 5 4 1 2 3) avec un seuil choisi de 0,5 ($f(S, i) \geqslant$0,5). Pour le génome S_3 on conservera un certain nombre d'intervalles dont {1,2} ($f_{\{1,2\}} = 1$) qui est dupliqué mais on aura aussi, par exemple, {1,2,3} qui est dupliqué avec perte de 3 ($f_{\{1,2,3\}} = \frac{5}{6}$). Ces deux intervalles se chevauchent puisqu'ils

impliquent les même dupliqués. Les chevauchements sont détectés grâce aux vecteurs binaires qui représentent les marqueurs composant l'intervalle commun. Le vecteur binaire associé à {1,2} est [1,1,0,0,0,0,0] et celui associé à {1,2,3} est [1,1,1,0,0,0,0]. Il ne faudra, au final, n'en conserver qu'un.

Quatre paramètres vont être intégrés pour la sélection des intervalles communs. Les deux premiers sont similaires aux paramètres utilisés pour le filtre présenté ci-dessus, mais ils sont appliqués à l'ensemble des génomes. On ajoute deux nouveaux paramètres : le nombre de génomes possédant l'intervalle commun dupliqué que nous évaluons et le nombre de marqueurs dupliqués dans cet intervalle commun.

Découpage total. Ce paramètre va servir à favoriser les intervalles communs les moins réarrangés. On calcule la moyenne des scores de découpage pour l'ensemble des génomes possédant l'intervalle commun.

Définition 11 (découpage total d'un intervalle commun)**.** *Soit un ensemble de génomes* $\mathcal{S} = \{S_1, S_2, ...S_n\}$*, et i un intervalle commun. Soit* $\mathcal{G}(i)$ *le sous-ensemble de* $\mathcal{S}$ *contenant les génomes possédant l'intervalle commun i dupliqué. Le score de* découpage total *de cet intervalle commun pour l'ensemble des génomes est défini par :*

$$dec_{total}(i) = \frac{\sum_{S \in \mathcal{G}(i)} dec(S, i)}{|\mathcal{G}(i)|}$$

On a la propriété que $dec_{total}(i) > 0$*.*

Conservation totale. Ce paramètre a pour objectif de favoriser les intervalles communs les mieux conservés (ceux qui ont perdu le moins de duplicats). Comme pour le découpage total, ce paramètre est similaire à celui du filtre (Définition 9) mais appliqué à l'ensemble des génomes.

Définition 12 (conservation totale d'un intervalle commun)**.** *Soit un ensemble de génomes* $\mathcal{S} = \{S_1, S_2, ...S_n\}$*, et i un intervalle commun. Soit* $\mathcal{G}(i)$ *le sous-ensemble de* $\mathcal{S}$ *contenant les génomes possédant l'intervalle commun i dupliqué. Le score de* conservation totale *de cet intervalle commun pour l'ensemble des génomes est défini par :*

$$cons_{total}(i) = \frac{\sum_{S \in \mathcal{G}(i)} cons(S, i)}{|\mathcal{G}(i)|}$$

On a la propriété que $0 < cons_{total}(i) \leq 1$*.*

Nombre de génomes. Ce paramètre favorisera les intervalles communs apparaissant dans le plus de génomes.

Définition 13 (nombre d'occurrences d'un intervalle commun dans les génomes)**.** *Soit un ensemble de génomes* $\mathcal{S} = \{S_1, S_2, ...S_n\}$*, et i un intervalle commun. Soit* $\mathcal{G}(i)$ *le sous-ensemble de* $\mathcal{S}$ *possédant l'intervalle commun i dupliqué. Le* nombre d'occurrences *de l'intervalle commun i dans les génomes est défini par :*

$$occ_{genome}(i) = \frac{|\mathcal{G}(i)|}{|\mathcal{S}|}$$

On a la propriété que $0 < occ_{genome}(i) \leq 1$*.*

Nombre de marqueurs dupliqués. Ce paramètre favorisera les intervalles communs ayant le plus de marqueurs dupliqués.

Définition 14 (nombre d'occurrences de marqueurs dupliqués dans un intervalle commun)**.** *Soit un ensemble de génomes $\mathcal{S} = \{S_1, S_2, ...S_n\}$, et i un intervalle commun. Soit $\mathcal{G}(i)$ le sous-ensemble de $\mathcal{S}$ possédant l'intervalle commun i dupliqué. Le* nombre d'occurrences *de marqueurs dupliqués de l'intervalle commun i pour l'ensemble des génomes est défini par :*

$$occ_{marqueur}(i) = \frac{n(S,i)}{|\Sigma|}$$

On a la propriété que $0 < occ_{marqueur}(i) \leq 1$.

Poids d'un intervalle commun. Les quatre paramètres précédents sont multipliés pour donner un poids à l'intervalle commun par rapport aux duplicats que l'on peut trouver dans l'ensemble des génomes.

Définition 15 (poids d'un intervalle commun)**.** *Soit un ensemble de génomes $\mathcal{S} = \{S_1, S_2, ...S_n\}$, et i un intervalle commun. Soit $\mathcal{G}(i)$ le sous-ensemble de $\mathcal{S}$ possédant l'intervalle commun i dupliqué. Le* poids *d'un intervalle commun est défini par :*

$$poids(i) = dec_{total}(i) \times cons_{total}(i) \times occ_{genome}(i) \times occ_{marqueur}(i)$$

Exemple 8. On considère toujours les génomes S_1=(1 2 3 1 2 3 4 5), S_2=(1 2 4 5 1 3) et S_3=(1 2 5 4 1 2 3). Si l'on reprend l'Exemple 7, et que l'on considère l'intervalle commun $\{1, 2\}$:

- $\text{occ}_{\text{genome}}(\{1,2\}) = \frac{3}{3}$ (les trois génomes possèdent l'intervalle commun dupliqué),
- $\text{occ}_{\text{marqueur}}(\{1,2\}) = \frac{2}{5}$ (sur les cinq marqueurs présents dans Σ, les deux marqueurs de l'intervalle commun sont dupliqués, tous génomes confondus),
- $\text{cons}_{\text{total}}(\{1,2\}) = \frac{\frac{4}{4}+\frac{3}{4}+\frac{4}{4}}{3}$,
- $\text{dec}_{\text{total}}(\{1,2\}) = \frac{\frac{2}{2}+\frac{2}{2}+\frac{2}{2}}{3}$,
- donc $\text{poids}(\{1,2\}) = 0,1$.

Considérons maintenant l'intervalle commun $\{1, 2, 3\}$:

- $\text{occ}_{\text{genome}}(\{1,2,3\}) = \frac{3}{3}$,
- $\text{occ}_{\text{marqueur}}(\{1,2,3\}) = \frac{3}{5}$,
- $\text{cons}_{\text{total}}(\{1,2,3\}) = \frac{\frac{6}{6}+\frac{4}{6}+\frac{5}{6}}{3}$,
- $\text{dec}_{\text{total}}(\{1,2,3\}) = \frac{\frac{2}{1}+\frac{2}{2}+\frac{2}{2}}{3}$,
- donc $\text{poids}(\{1,2,3\}) = 0,4$.

Entre l'intervalle commun {1,2} et l'intervalle commun {1,2,3}, l'intervalle commun {1,2,3} sera préféré ($\text{poids}(\{1,2,3\}) > \text{poids}(\{1,2\})$). Sur cet exemple, on obtiendra donc une duplication de {1,2,3} dans les trois génomes, en tandem sans perte dans S_1, avec perte de 3 dans une des copies et de 2 dans l'autre chez S_2, et perte de 3 dans une copie chez S_3 :

- S_1=(1 2 3 1 2 3 4 5),
- S_2=(1 2 4 5 1 3),
- $S3$=(1 2 5 4 1 2 3)).

On peut donc voir ici qu'on a bien retrouvé des duplications avec pertes qui seront communes à plusieurs génomes. En s'appuyant seulement sur le score d'un intervalle commun pour un génome considéré, on aurait plutôt gardé l'intervalle commun {1,2} pour S_3 au lieu de l'intervalle commun

{1,2,3} (pour S_3 on a $f_{\{1,2,3\}} \leq f_{\{1,2\}}$). Cependant quand on considère l'ensemble des génomes possédant un intervalle commun dupliqué, on peut alors repérer des duplications avec pertes communes à plusieurs d'entre-eux ce qui est plus parcimonieux que d'obtenir des duplications propres à chacun des génomes, correspondant à des intervalles communs similaires impliquant les mêmes marqueurs dupliqués.

5.3 Exemple d'application de la méthode

Nous avons appliqué notre méthode aux huit génomes mitochondriaux de *Zea* étudiés au chapitre précédent. Les génomes mitochondriaux de *Zea* sont composés de 69 marqueurs comprenant des marqueurs dupliqués. Tout d'abord, nous avons regardé les influences de la valeur de seuil du filtre sur les intervalles communs conservés par rapport à ces génomes. Nous avons ensuite choisi une valeur de seuil et appliqué la méthode. Nous discuterons des résultats obtenus.

5.3.1 Calibrage de la valeur du seuil du filtre

Le score d'un intervalle calculé pour le filtre dépend de la conservation d'un intervalle commun et de son découpage, pour un génome donné. Ce score est comparé à un seuil qui nous permet de décider si l'intervalle commun de ce génome est un intervalle commun à conserver pour la sélection des intervalles communs finaux ou non. La première étape de la méthode, consistant à construire le graphe ADCI et à n'en garder que les intervalles communs potentiellement dupliqués, nous retourne un graphe contenant 18054 nœuds représentant tous les intervalles communs différents pour tous les génomes. Pour chaque nœud, il peut y avoir un ou plusieurs génomes impliqués. Si l'on compte le nombre d'intervalles communs du graphe par rapport à chaque génome, nous obtenons 141057 intervalles communs.

Dans un premier temps nous avons regardé le nombre d'intervalles communs dupliqués conservés par rapport au seuil fixé. Nous avons fait varier le seuil de 0 à 1 par pas de 0,05. Les résultats sont présentés Figure 5.6. Quand le seuil est fixé à zéro, les 141057 intervalles communs passent le filtre. Plus le seuil est augmenté, plus le nombre d'intervalles communs éliminés est grand. Pour un seuil de 0,5 il ne reste que 8834 intervalles communs. Le score des intervalles communs dépendant de deux paramètres, il est difficile d'extraire avec cette figure les propriétés des intervalles communs éliminés.

Dans un deuxième temps, nous avons donc regardé comment évoluaient la conservation et le découpage des intervalles communs en fonction de la valeur du seuil. Pour cela nous avons tout d'abord regardé le nombre d'intervalles communs sélectionnés pour un seuil donné. Pour chaque seuil, les intervalles communs sélectionnés sont répartis en fonction du nombre de blocs les composant. La Figure 5.7 représente donc le nombre d'intervalles communs sélectionnés et leur répartition en nombre de blocs pour tous les génomes en fonction du seuil. Par exemple, pour un seuil de 0,35 (ce qui correspond aux intervalles communs dont le score est supérieur à $0,35$) on observe que la plupart des intervalles communs sont en deux, trois et quatre blocs, puis d'autres intervalles communs seront en 1 bloc. Enfin on aura une plus faible proportion d'intervalles communs en cinq, six et sept blocs. A partir d'un seuil de 0,6 on ne conserve plus que des intervalles communs décomposés au maximum en trois blocs.

Nous avons également regardé le pourcentage de conservation d'un intervalle commun par rapport au seuil. La Figure 5.8 représente le nombre d'intervalles communs sélectionnés pour chaque génome en fonction du seuil en considérant leur pourcentage de conservation. Pour une meilleure lisibilité, pour chaque seuil, nous avons représenté le logarithme du nombre d'intervalles

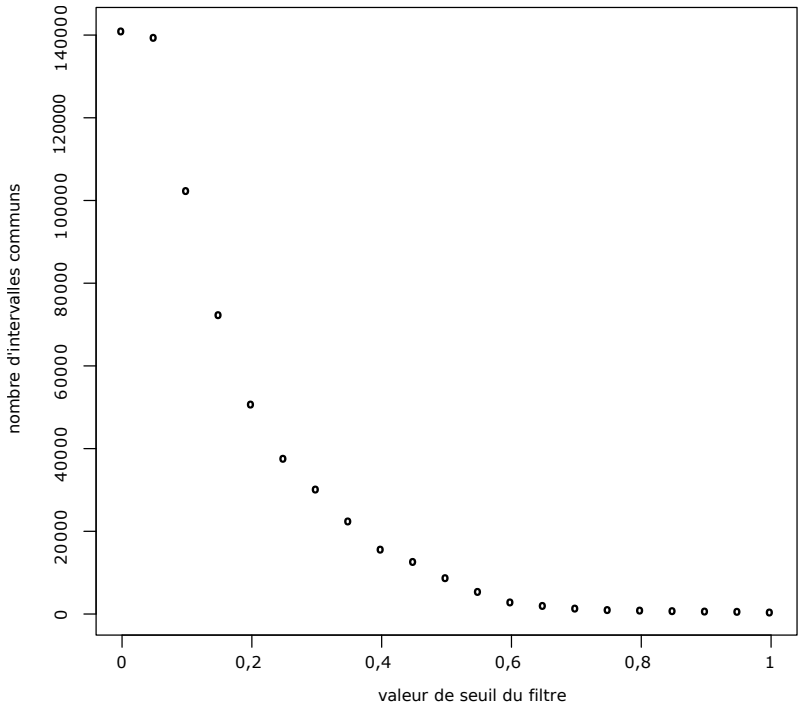


FIG. 5.6 – Nombre d'intervalles communs sélectionnés en fonction de la valeur de seuil du filtre.

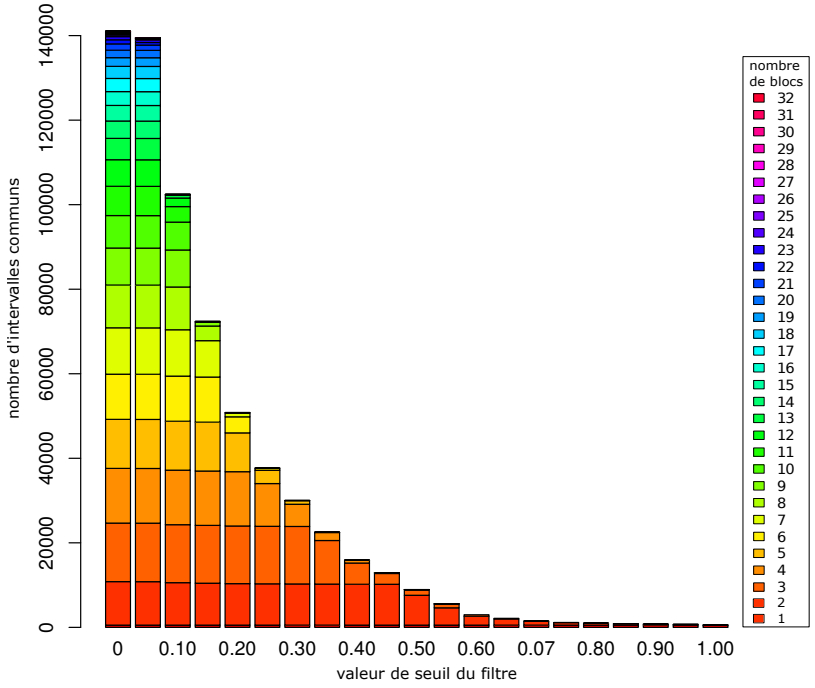


FIG. 5.7 – Nombre d'intervalles communs sélectionnés en fonction de la valeur de seuil du filtre. Ces intervalles communs sont répartis en fonction du nombre de blocs qui les composent. Par exemple, pour une valeur de seuil du filtre de 30% on obtient environ 10000 IC composés de 2 blocs, 15000 IC composés de 3 blocs et 5000 IC composés de 4 et 5 blocs.

pour une conservation donnée. Par exemple, pour un seuil de 0,50, les intervalles communs ayant un score f supérieur à ce seuil sont des intervalles communs dont la conservation varie de 0,5 à 1. Nous pouvons constater qu'à partir d'un seuil de 0,65, on conserve le même nombre d'intervalles communs pour les pourcentages de conservation faibles. En effet, pour des valeurs de seuil supérieures à 0,65, on a le même nombre d'intervalles communs ayant une conservation comprise entre 0,5 et 0,65 non inclus. D'autre part, entre un seuil de 0,65 et un seuil de 0,70 on perd les intervalles communs ayant une conservation comprise entre 0,65 et 0,70. Une valeur de seuil fixée à 0,65 semble donc un bon compromis : on stabilise les intervalles communs dont la conservation est faible et en même temps on ne perd pas trop d'intervalles communs mieux conservés. Nous avons donc choisi une valeur de seuil de 0,65 appliquée à ce filtre afin de ne conserver que les intervalles communs les plus intéressants pour la suite de la méthode.

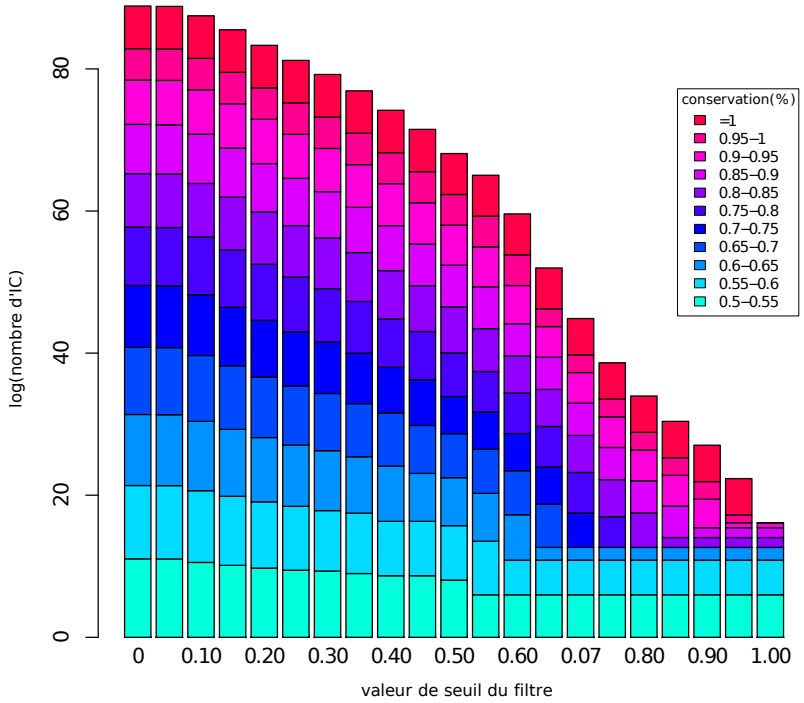


FIG. 5.8 – Nombre d'intervalles communs sélectionnés en fonction de la valeur de seuil du filtre. Ces intervalles communs sont répartis en fonction de leur pourcentage de conservation. Par exemple, pour une valeur de seuil du filtre de 50% nous obtenons des intervalles communs dont la conservation varie de 50% à 100%.

5.3.2 Résultats obtenus

Pour rappel, lors de l'étude effectuée sur *Zea*, l'inspection manuelle des génomes nous avait permis d'identifier des duplications communes à plusieurs génomes ainsi que des duplications propres à un seul génome (pouvant être conservées en tandem ou non). Nous avons comparé ces résultats pour différentes valeurs du seuil. Les résultats obtenus sont présentés de manière synthétique dans le Tableau 5.1 et sont discutés plus en détails dans les paragraphes suivants.

Groupe de marqueurs dupliqués	Génomes possédant la duplication	Duplication détectée en tandem	Comparaison avec les observations manuelles
Duplications manuellement détectées			
{20,21,22}	tous les *Zea*	non	
{27}	tous les *Zea mays* et parviglumis	non	
{8,9,10,11,12,13,14,15,16,17,18}	NA	oui	
{5,6,7,66,67,63,29,30,31,32,33,34,60,61,62 35,36,37,38,2,10,9,8,12,13,14,15,16,17,18}	CMS-C	oui	
{68,69,1,11,3,4}	CMS-C	oui	
{34,33,32,31,63,64,65,11,10,9,8, 12,13,14,15,16,17,18}	parviglumis	oui	
{2}	NA,NB et parviglumis	non	
{60,61,62}	NB	non	
{3,4,5,6,7}	CMS-S	non	
{62}	CMS-S	non	
{26}	CMS-T	non	
{67}	CMS-T	non	
{68}	*Zea perennis*	non	
{35}	*Zea perennis*	non	
{24}	*Zea luxurians*	non	
Seuil 0,65			
{20,21,22}	tous les *Zea* sauf CMS-T	non	•
{26,27}	tous les *Zea mays* et parviglumis	non	R
{8,9,10,11,12,13,14,15,16,17,18}	NA et parviglumis	oui chez NA	•et S
{68,69,1,11,3,4}	CMS-C	oui	•
{5,6,7,66,67,63,29,30,31}	CMS-C	non	S
{12,13,14,15,16,17}	CMS-C	non	S
{60,61,62}	CMS-C et CMS-S	non	S et E
{12,13,14,15,16,17,60,61,62}	NB	non	E
{34,33,32,31,63,64,65}	parviglumis	non	S
{68,3,4}	*Zea perennis*	non	E
{2}	NA et parviglumis	non	•
{2,30,31}	NB	non	E
{5,6,7}	CMS-S	non	Ra
{66,67}	CMS-T	non	E
{35}	*Zea perennis*	non	•
{24,25,26}	*Zea luxurians*	non	E
Seuil 0,55			
{21,22,23,24,25,26}	tous les *Zea*	non	R et S
{20}	tous les *Zea* sauf CMS-T	non	S
{27}	tous les *Zea mays* et parviglumis	non	S
{8,9,10,11,12,13,14,15,16,17,18}	NA et parviglumis	oui chez NA	ffl et S
{3,4,5,6,7}	CMS-C et CMS-S	non	S et •
duplication total du génome	NB	oui	
{66,67}	CMS-C et CMS-T	non	E
{63,29,30,31,32,33,34}	CMS-C et parviglumis	non	S et E
{12,13,14,15,16,17,60,61,62}	CMS-C	non	S
{11,12,13,14,15,16,17,18,60,61,62}	CMS-S	non	E
{1,2,3}	NA	non	E
{2}	parviglumis	non	•
{64,65}	parviglumis	non	S
{68,3,4}	*Zea perennis*	non	E
{35}	*Zea perennis*	non	•
Seuil 0,70			
{20,21,22}	tous les *Zea* sauf CMS-T et *Zea luxurians*	non	•
{20,21}	*Zea luxurians*	non	Ra
{26,27}	tous les *Zea mays* et parviglumis	non	E
{8,9,10,11,12,13,14,15,16,17,18}	NA et parviglumis	oui	•et S
{68,69,1,11,3,4}	CMS-C	oui	•
{5,6,7,66,67,63,29,30,31}	CMS-C	non	S
{12,13,14,15,16,17}	CMS-C	non	
{60,61,62}	CMS-C	non	S
{60,61}	CMS-S	non	E
{12,13,14,15,60,61,62}	NB	non	E
{34,33,32,31,53,54,55}	parviglumis	non	S
{68}	*Zea perennis*	non	•
{2}	NA et parviglumis	non	•
{2,31}	NB	non	E
{6,7}	CMS-S	non	Ra
{66,67}	CMS-T	non	E
{35}	*Zea perennis*	non	•
{24}	*Zea luxurians*	non	•

TAB. 5.1 – Comparaison des duplications détectées dans les génomes mitochondriaux de *Zea* pour différentes valeur de seuil du filtre avec celles observées suite à l'inspection manuelle des génomes. • : identique, E : étendu, R : regroupé, Ra : raccourci, S : séparé.

Seuil à 0,65. Les groupes de marqueurs dupliqués prédits sur les génomes mitochondriaux de *Zea* à l'aide de notre méthode sont présentés Figure 5.9. Les résultats montrent des intervalles communs dupliqués partagés entre plusieurs génomes mais aussi des intervalles communs dupliqués propres à certains génomes.

On retrouve l'intervalle commun {20,21,22} dupliqué chez tous les *Zea* sauf chez CMS-T. En effet, chez CMS-T il ne reste aucune copie de cette duplication, elle n'a donc pas été détectée.

On retrouve également la duplication de {27} mais associée à {26}. En d'autres termes, nous avions vu manuellement une duplication de {27} chez tous les maïs et parviglumis, et une duplication de {26} propre à CMS-T. La méthode prédit une duplication de {26,27} dans tous ces génomes avec la perte d'une copie de {26} dans tous sauf CMS-T.

La duplication en tandem de l'intervalle commun {8,9,10,11,12,13,14,15,16,17,18} est retrouvée chez NA. On retrouve également la duplication en tandem de {68,69,1,11,3,4} chez CMS-C. Par contre les autres duplications en tandem ne sont pas retrouvées chez CMS-C et parviglumis. On retrouve bien les gènes impliqués dans ces duplications mais séparés en plusieurs blocs. Par exemple chez parviglumis, au niveau de ce qui avait été défini comme une duplication en tandem de {34,33,32,31,63,64,65,11,10,9,8,12,13,14,15,16,17,18}, on trouve une duplication de {34,33,32,31,63,64,65} et une duplication de {11,10,9,8,12,13,14,15,16,17,18}. Il y a deux raisons pour lesquelles ces motifs en tandem ne sont pas retrouvés. D'une part, ces duplications en tandem sont des duplications en tandem remaniées. Or la méthode ne détermine une duplication en tandem que si les marqueurs dupliqués apparaissent dans le même ordre. D'autre part, les duplications en tandem de CMS-C et parviglumis ont été découpées car les sous-intervalles les composant sont des intervalles communs dupliqués retrouvés dans plusieurs génomes (contrairement à la duplication en tandem qui est propre aux deux génomes). Ce qui est logique puisqu'un sous-intervalle commun à plusieurs de génomes et bien conservé aura plus de poids qu'un intervalle commun plus grand mais propre à un génome. Chez parviglumis, on retrouve les mêmes bornes que celles de la duplication en tandem estimée, mis à part qu'elle est ici trouvée en deux intervalles communs dupliqués. Un intervalle commun est identique à l'intervalle commun dupliqué en tandem chez NA, l'autre est unique à parviglumis. Chez CMS-C la grande duplication en tandem manuellement établie est retrouvée ici comme trois intervalles communs dupliqués. Le premier {5,6,7,66,61,29,30,31} est propre à CMS-C. Le deuxième {12,13,14,15,16,17} est également unique à CMS-C mais est identique à une sous-partie de la duplication en tandem retrouvée chez NA. Le troisième {60,61,62} est aussi retrouvé chez CMS-S. Chez NB on retrouve un intervalle commun dupliqué présentant exactement les mêmes gènes que dans les deux derniers intervalles communs de CMS-C, soit {12,13,14,15,16,17,60,61,62}. En fait, chez NB seuls les marqueurs 60,61 et 62 sont réellement dupliqués. L'intervalle commun {12,13,14,15,16,17,60,61,62} est retenu chez NB car son poids est supérieur à celui de {60,61,62}. En effet, il contient plus de gènes dupliqués et il est également présent chez CMS-C. Par contre, il n'a pas été retenu chez CMS-C car le sous-groupe {12,13,14,15,16,17} avait plus de poids (il s'agissait d'un intervalle commun dont plus de génomes étaient impliqués par rapport à l'intervalle commun {12,13,14,15,16,17,60,61,62}). Un intervalle commun supplémentaire est prédit, {2} pour NA et parviglumis. Par contre, pour NB qui présente le marqueur {2} dupliqué , l'intervalle commun associé sera {2,30,31} car ce motif est retrouvé chez CMS-C.

Les autres intervalles communs dupliqués trouvés dans les autres génomes sont propres à chaque génome.

La méthode proposée est donc bien capable d'identifier des groupes de marqueurs dupliqués entre plusieurs génomes, ces groupes pouvant avoir subi des pertes de marqueurs au cours de l'évolution. Nous avons ensuite fait deux tests, en variant le seuil au dessus et en dessous de 0,65

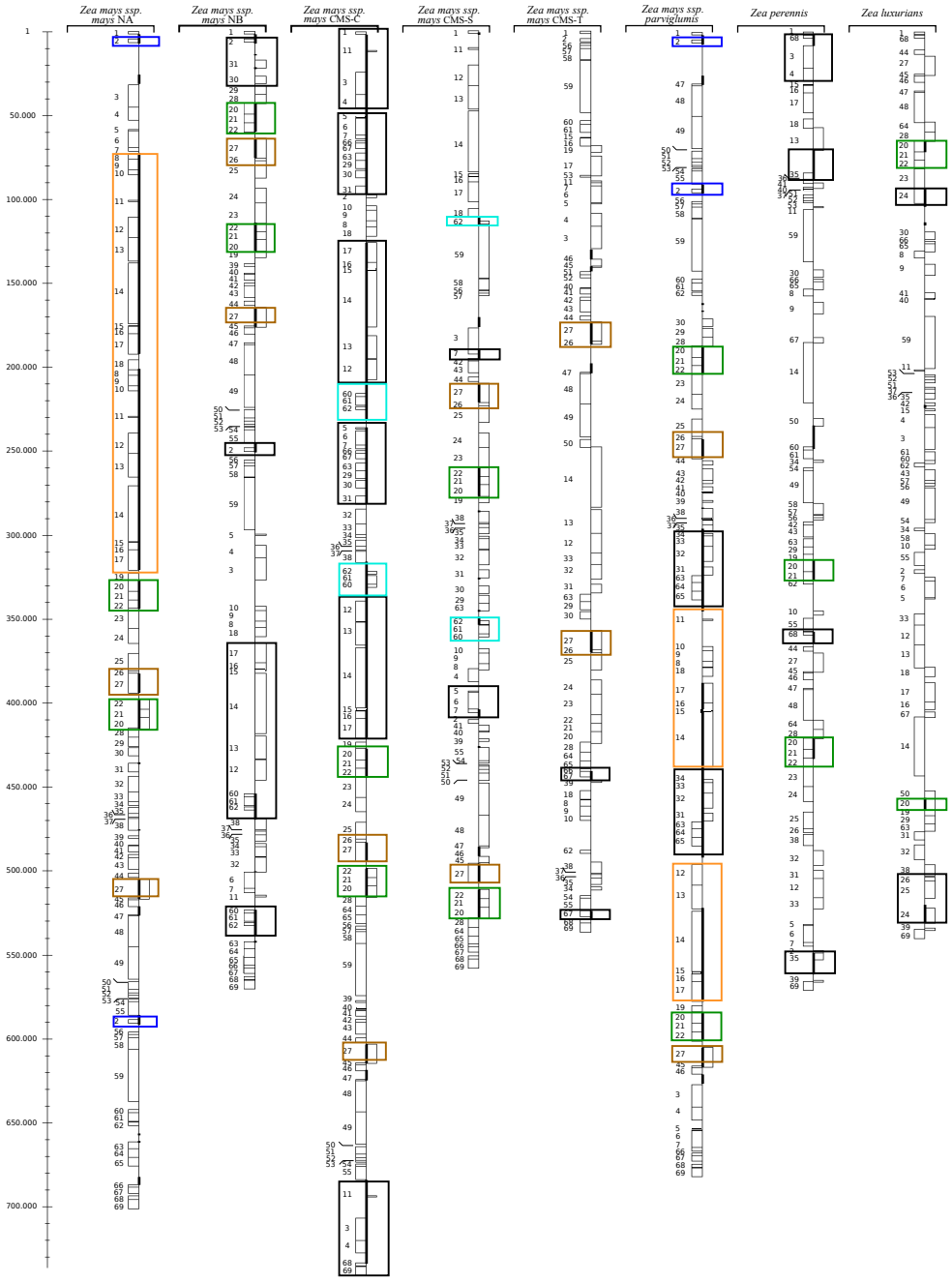


FIG. 5.9 – Duplications détectées avec un seuil fixé à 0,65. Les duplications en noir sont propres à un génome. Les duplications en couleur sont communes à plusieurs génomes.

afin d'en évaluer l'impact sur les intervalles communs prédits.

Seuil à 0,55. Pour un seuil fixé à 0,55, les résultats changent légèrement (Table 5.1). En effet, par rapport au seuil de 0,65, celui-ci laissera passer au filtre des intervalles communs moins conservés et qui peuvent avoir plus de blocs. Il est évident que certains intervalles communs vont venir bruiter les résultats. En effet, les grandes duplications chez CMS-C et parviglumis se retrouvent fragmentées en blocs plus petits. Cela se produit car des intervalles communs de faible poids dans d'autre génomes vont passer le filtre et vont alors influencer le poids de l'intervalle commun final. De plus, chez NB il est prédit une duplication totale de génome alors que celui-ci ne présente que peu de marqueurs dupliqués. Il semblerait que l'on soit ici trop permissif.

Seuil à 0,70. Pour un seuil de 0,70 on devient plus restrictif et des intervalles communs avec beaucoup de délétions seront éliminés au niveau du filtre (Table 5.1). Ainsi, l'intervalle commun {20,21,22} est prédit pour tous les génomes sauf *Zea perennis* qui a eu trop de pertes dans cet intervalle. Chez *Zea perennis* il sera prédit une duplication de {20,21}. De manière générale, les intervalles communs prédits tendent à avoir le moins d'erreurs possible. On devient alors trop restrictif.

On voit que vraisemblablement, le seuil de 0,65 donne les meilleures prédictions. Cependant, le paramètre de découpage peut avoir une forte influence. Par exemple, lorsqu'un intervalle commun conservé à 100% est décomposé en quatre blocs, il ne sera pas pris en compte. Prenons le génome S_1=(1 2 5 6 3 4 7 8 1 2 9 10 3 4), l'intervalle commun {1,2,3,4} est bien présent mais remanié en quatre blocs : (1,2), (3,4), (1,2) et (3,4). La valeur du paramètre de découpage pour ce génome est alors de 0,5 et il ne passera pas le filtre. En revanche, les sous-intervalles {1,2} et {3,4} passeront le filtre puisque, par rapport à ce génome, ils ne sont pas découpés.

Il est possible de replacer les intervalles communs dupliqués prédits sur l'arbre phylogénétique de *Zea* obtenu précédemment (Figure 5.10 et Table 5.1). Pour les duplications communes à plusieurs génomes, nous obtenons donc une duplication ancestrale aux *Zea* de {20,21,22} avec perte de {22} et {20,21} respectivement chez *Zea luxurians* et *Zea perennis*. Nous avons également une duplication de {26,27} ancestrale aux maïs suivie de la perte de {27} au niveau de la séparation de CMS-T et des autres maïs. CMS-S, CMS-T et NB partagent une duplication de {60,61,62}. Cette duplication chez NB fait partie du groupe {12,13,14,15,16,17,60,61,62}. NB est le seul à posséder cet intervalle commun dupliqué, de plus les copies de {12,13,14,15,16,17} ont été perdues. Cependant, on retrouve {12,13,14,15,16,17} dupliqué chez CMS-C. D'un point de vue phylogénétique soit on a une duplication ancestrale de {60,61,62} après la divergence avec CMS-T dont les copies {60,61} ont été perdues chez CMS-S et {60,61,62} ont été perdues après la divergence avec NB (cette duplication n'est pas retrouvée chez NA et parviglumis). Soit on a une duplication ancestrale de {12,13,14,15,16,17,60,61,62} après la divergence avec CMS-T dont les copies {12,13,14,15,16,17,60,61} ont été perdues chez CMS-S, {12,13,14,15,16,17} ont été perdues chez NB et {60,61,62} ont été perdues après la divergence avec NB. Le plus parcimonieux semble être une duplication ancestrale de {60,61,62} en considérant que la duplication de {12,13,14,15,16,17,60,61,62} puisse être un artefact dû à la conservation de la synténie de ces marqueurs entre CMS-C et NB. Il en est de même pour {2} qui est dupliqué chez NA et parviglumis. On le retrouve également chez NB mais dans {2,30,31} avec la perte de {30,31}. Là aussi, l'intervalle commun {2,30,31} est retrouvé chez CMS-C, il est donc possible que ce soit également un artefact et que {2} soit dupliqué au niveau de la séparation de CMS-C avec les autres espèces. On retrouve la duplication en tandem chez NA qui serait une duplication

ancestrale à NA et parviglumis avec perte de {8,9,10,11,12,13,18} chez parviglumis et perte de {18} chez NA. Les duplications en tandem chez parviglumis et CMS-C ne sont pas retrouvées telles quelles sauf celle de {68,69,11,1,3,4} chez CMS-C. Elles sont découpées en fragments communs aux autres génomes. Nous l'avons vu, il semble que {12,13,14,15,16,17} apparaisse dupliqué plusieurs fois dans la phylogénie et c'est en majorité ce groupe qui fait que les duplications en tandem chez CMS-C et parviglumis ne sont pas retrouvées. Il peut s'agir soit d'un groupe dupliqué ancestralement, qui se retrouverait dans une duplication en tandem quasiment conservée chez NA, soit d'un phénomène d'homoplasie, supposant que certaines régions du génome seraient plus soumises aux duplications que d'autres.

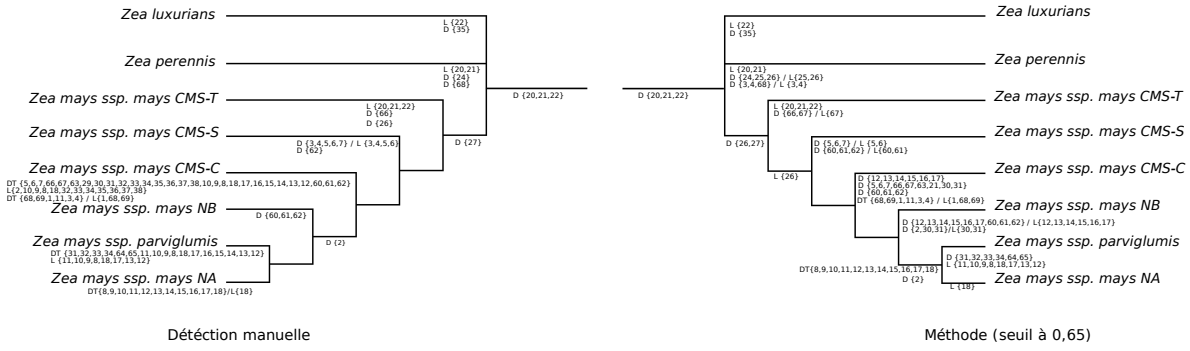


FIG. 5.10 – Arbres phylogénétiques des génomes mitochondriaux de *Zea* obtenus à partir de séquences nucléotidiques sur lesquels ont été replacés les événements de duplications trouvés manuellement et avec la méthode.

5.4 Conclusion

Le méthode proposée est donc bien capable de retrouver des groupes de marqueurs dupliqués, dans un ou plusieurs génomes. Par rapport aux analyses manuelles, certaines des duplications en tandem vues manuellement sont retrouvées mais découpées en plusieurs blocs lorsque des motifs dupliqués sont retrouvés entre plusieurs génomes. Pour retrouver des duplications en tandem propres à un génome, sans qu'elles soient découpées à cause de l'influence des autres génomes, il suffit à l'utilisateur d'analyser un seul génome à la fois. D'un autre côté, étant donné les génomes que nous avons analysés, nous ne pouvons pas savoir si les prédictions manuelles sont meilleures que les prédictions de notre méthode. En effet, nous ne connaissons pas les événements de duplication réels qui sont apparus sur ces génomes au cours de l'évolution. Il faudrait dans un premier temps, tester cette méthode sur un ensemble de génomes virtuels, dont l'évolution (en terme de spéciation, duplication et rearrangement) aurait été simulée afin de pouvoir comparer les résultats obtenus aux événements de duplication induits.

Une limitation de cette méthode reste le temps de calcul, en effet, la phase de construction du graphe et de rassemblement des blocs peut prendre plusieurs heures. Cela dépend du nombre de génomes, du nombre de marqueurs ainsi que du nombre d'intervalles communs existants entre ces génomes.

La méthode développée permet de traiter des génomes ayant connu des triplications (ou plus).

Par contre, pour l'instant, les motifs tripliqués sont recherchés comme des motifs identiques c'est-à-dire qu'on ne tient pas compte de la possible duplication d'un groupe de marqueurs suivie de la duplication d'un sous intervalle de ce groupe.

Une amélioration possible de cette méthode serait d'intégrer une étape suivant la récupération des intervalles communs finaux, visant à affiner les prédictions. Par exemple dans le cas de la duplication de l'intervalle commun {2} chez NA et parviglumis et {2,30,31} chez NB, on pourrait alors prédire une duplication de {2} dans ces trois génomes sachant que {2,30,31} n'est, au final, retrouvé que chez NB qui a en plus perdu les copies de 30 et 31. Dans cette étape, nous pourrions également regrouper les blocs des éventuelles duplications en tandem remaniées.

Nous avons donc réussi à développer un outil d'aide à l'analyse de régions dupliquées dans les génomes. L'utilisateur peut faire varier le seuil du filtre lui permettant de retrouver des motifs plus ou moins conservés. De plus, nous voulions mettre en place une méthode dans laquelle on se dispense de l'information de phylogénie entre les génomes. Par rapport aux génomes testés, les duplications communes à plusieurs génomes sont relativement bien placées sur les branches de la phylogénie. La méthode fournit donc des résultats satisfaisants sans avoir besoin d'une information phylogénétique supplémentaire qui peut parfois être un frein aux analyses.

Chapitre 6

Analyse des génomes de betterave

Une des parties majeures de cette thèse fut le séquençage des génomes mitochondriaux de *Beta vulgaris ssp. maritima* (betterave sauvage) afin d'en analyser et d'en comparer les contenus ainsi que les structures.
Le laboratoire GEPV possède une collection de graines de betteraves sauvages sur différents cytoplasmes (on connaît une vingtaine de cytoplasmes différents grâce à des marqueurs RFLP mitochondriaux). *Beta vulgaris* est une espèce dite gynodoïque : en populations naturelles, on observe des individus hermaphrodites et des individus femelles (ou mâles stériles). Les individus femelles sont en fait des individus hermaphrodites ayant perdu la fonction mâle et nous savons que les facteurs de stérilité mâle des plantes cultivées sont codés par le génome mitochondrial. Ces génomes mâles stériles sont appelés CMS (*Cytoplasmic Male Sterile*). D'autre part, existe des facteurs nucléaires, capables de restaurer la fertilité mâle.

Nous avons sélectionné quatre génomes de *Beta vulgaris ssp. maritima* appelées A, B, CMS-E et CMS-G. A et B sont des cytoplasmes mâles fertiles tandis que CMS-E et CMS-G sont des CMS. La Figure 6.1 montre la répartition de ces cytoplasmes, basée sur les séquences chloroplastiques, réalisée lors d'une première étude [Fénart et al., 2006]. En plus de ces quatre génomes, nous avons séquencé un génome mitochondrial de *Beta macrocarpa* qui servira de groupe externe lors des analyses phylogénétiques. Deux autres génomes de *Beta vulgaris* ont été séquencés par le passé : TK81-O [Kubo et al., 2000] et TK81-MS [Satoh et al., 2004]. Ces génomes sont considérés comme respectivement très proches de Nv et Sv (Figure 6.1).

Ce chapitre sera composé de quatre parties. Dans la première partie, nous traiterons du séquençage et de l'analyse de fragments chloroplastiques. Ces fragments constitueront une donnée supplémentaire nous permettant de retracer une histoire phylogénétique indépendante de celles que nous obtiendrons à partir des données mitochondriales des cytoplasmes séquencés. En effet, une coévolution entre les cytoplasmes mitochondriaux et chloroplastiques est attendue. Cette donnée nous permettra donc de valider les phylogénies basées sur les réarrangements et séquences mitochondriales. Dans la deuxième partie, nous discuterons d'une méthode que nous avons mise en place afin de reconstruire les génomes mitochondriaux composés de plusieurs fragments (contigs) issus du séquençage. Dans la troisième partie, nous ferons l'analyse de la composition et des taux de substitution des génomes mitochondriaux séquencés. Enfin dans la quatrième partie, nous ferons l'analyse des structures de ces génomes mitochondriaux en utilisant les méthodes développées sur le maïs ainsi que la méthode de détection de duplications décrite dans le chapitre précédent.

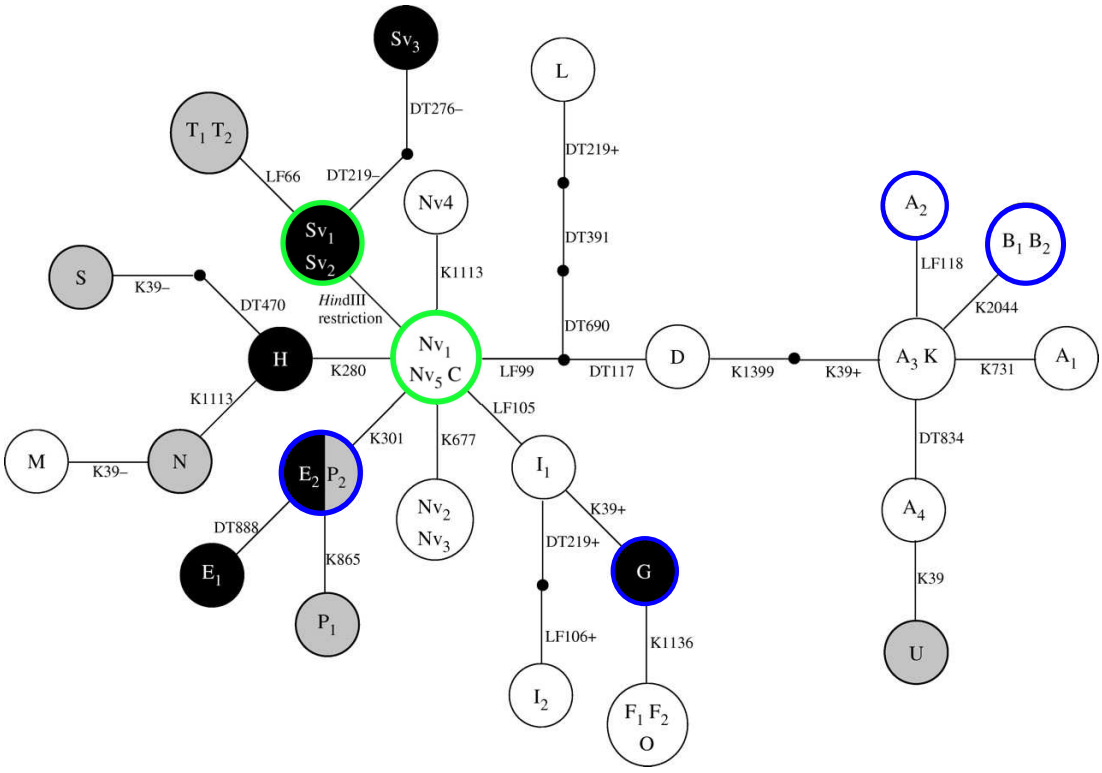


FIG. 6.1 – Réseau d'haplotypes chloroplastiques correspondant aux différents génomes mitochondriaux. Extrait de [Fénart et al., 2006]. En bleu les génomes que nous avons séquencés, en vert les génomes séquencés lors de précédentes études. Les cercles noirs représentent les cytoplasmes stériles, les cercles blancs les cytoplasmes fertiles et les cercles gris les cytoplasmes dont on ne connait pas l'effet sur le sexe (cytoplasmes rares). Certains cytoplasmes, non différenciés au niveau chloroplastique se retrouvent donc dans un même cercle.

6.1 Analyse chloroplastique

Afin de valider l'analyse phylogénétique des réarrangements des génomes mitochondriaux de betterave, nous avons réalisé le séquençage de fragments chloroplastiques pour chacun des génomes étudiés (A, B, CMS-E, CMS-G, TK81-O, TK81-MS et macrocarpa) et nous avons également séquencé des fragments des génomes D, Nv et Sv afin de comparer la phylogénie obtenue sur ces fragments avec le réseau d'haplotypes que nous avons vu plus haut (Figure 6.1).

6.1.1 Méthodes

Afin de réaliser le séquençage de fragments chloroplastiques, nous avons défini des amorces sur le génome chloroplastique du tabac en essayant d'éviter les régions codantes qui sont très peu polymorphes. Les fragments choisis pour le séquençage sont représentés sur la Figure 6.2 et les amorces résumées dans le Tableau 6.1. Ils représentent une taille totale de 34392 pb (22% du génome). Le séquençage obtenu, en comptant tous les fragments amplifiés a donné une taille totale de 24619 pb sur les génomes de *Beta vulgaris* et *Beta macrocarpa*.

En tenant compte seulement des fragments amplifiés contenant au moins un site polymorphe, la taille totale de ces fragments est de 17740 pb.

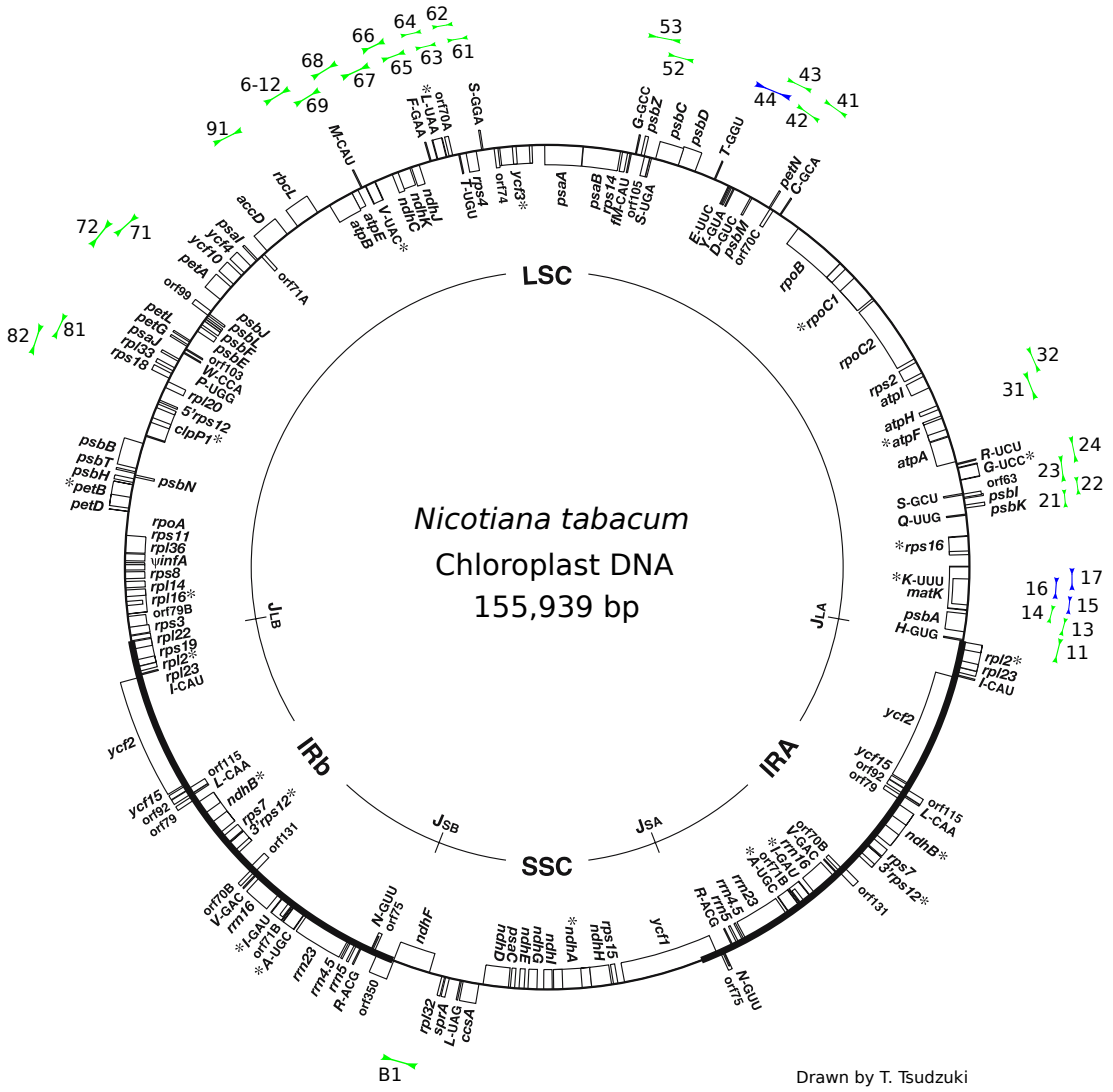


FIG. 6.2 – Fragments chloroplastiques séquencés sur *Beta vulgaris* par rapport à *Nicotiana tabacum*. En vert, le fragments séquencés, en bleu les fragments dont nous possédions déjà les amorces.

num	Nom	Amorce forward	Séquence	Amorce Reverse	Séquence	Position dans le tabac	Taille	Tm	Résultat
11	HpsbA	*trnH*	ACGGGAATTGAACCCGCGCA	*psbA*	GCTGCTTGGCCTGTAGTAGG	14-759	745	53	P
13	psbAK1	*psbA*	AGCTGCAACAGGAGCTGAAT	*trnK1*	CGACTAGTTCCGGGTTCGA**	1128-1833	708	53	P
14	psbAmatK	*psbA*	CGCTTTCGCGTCTCTCTAAA	*matK*	AAGATCGGGCTCGGAATTAT	1554-2277	724	48/65	NA
15	K1matK1	*trnK1*	GTTGCCCCGGGATTCGAA*	*matK*	TTAGGGCATCCCATTAGTA*	1816-2501	686	50	P
16	matK2matK6b	*matK*	CTAGCACAAGAAAGTCGAAG**	*matK*	GGATTTCTAACCATCTTGTT**	2335-3872	1357	54	NP
17	matK6bK2	*matK*	ATTCTGTTGATACATTCGAG*	*trnK2*	GAGTACTCGGCTTTTAAGTG*	3451-4390	940	54	NP
21	QpsbI	*trnQ*	AACCCGTTGCCTTACCACTT	*psbI*	GGATTACGTCCGGGATCATT	7461-8497	1037	53	NP
22	psbKS	*psbK*	TGTTTGGCAAGCTGCTGTAA	*trnS*	GGAGAGATGGCTGAGTGGAC	7995-8724	730	57	P
23	SG2	*trnS*	TCGAACCCTCGGTACGATTA	*trnG2*	CGTTAGCTTGGAAGGCTAGG	8652-10239	1588	57	NP
24	G1atpA	*trnG1*	GCGGGTATAGTTTAGTGGTAAAA	*atpA*	TTTACCGAGGAAGCAGAAGC	9504-10711	1208	55	P
31	atpFatpH	*atpF*	CGGTATTAAACCCGAAACTCC	*atpH*	ATTTCTGCCGCTTCCGTTAT	13386-14094	708	53	NP
32	atpHatpI	*atpH*	CTCGCGATACCCTCTACAGC	*atpI*	CCGCGGCTTATATAGGTGAA	13984-15303	1320	53	P
41	Cycf6	*trnC*	CCAGTTCAAATCCGGGTGTC	*ycf6*	GAGTCCACTTCTTCCCCACA	28850-29610	781	53	NP
42	ycf6psbM	*ycf6*	TAAGTCTTGCTTGGGCTGCT	*psbM*	TCTTGCATTTATTGCTACTGCAC	23554-24856	1303	53	P
43	psbMY	*psbM*	CGTTTTGACTGACTGTTTTTACG	*trnY*	CGTTGGCAATATGTCTACGC	24771-26065	1395	48/65	NA
44	DT	*trnD*	ACCAATTGAACTACAATCCC***	*trnT*	CTACCACTGAGTTAAAAGGG***	25995-26207	1213	56,5	P
52	psbCS	*psbC*	GCTTCTCAAGCTCAAGCATTT	*trnS*	GATGGCTGAGCGGTTGATAG	36408-37231	824	53	NP
53	SfM	*trnS*	GGGATTCGAACCCTCGATAG	*trnfM*	ACCTTGAGGTCACGGGTTC	37155-38396	1242	53	NP
61	Srps4	*trnS*	CGAGGGTTCGAATCCCTCTC	*rps4*	AAGCCTAGGAACGGAAGTGA	47187-48076	890	57	NP
62	rps4T	*rps4*	GAAGCCATACCCAATCGAAA	*trnT*	GAGGTTAGAGCATCGCATTTG	47844-48577	734	53	NP
63	TL1	*trnT*	CGGAATCGAACCGATGAC	*trnL1*	TCCGTAGCGTCTACCGATTT	48529-49333	805	53	P
64	L1F	*trnL1*	GGATATGGCGAAATCGGTAG	*trnF*	GCCAGGAACCAGATTTGAAC	49305-50319	1014	53	P
65	FndhJ	*trnF*	CCGGGATAGCTCAGTTGGTA	*ndhJ*	CGCTCTCAATGTGCCTATGA	50247-51318	1072	53	P
66	ndhJndhC	*ndhJ*	GGTTGATCCACACCATCCTC	*ndhC*	GTGTTAGCCCCGATTAGCAA	51230-52586	1357	53	P
67	ndhCV1	*ndhC*	TCGCCCATTGGTTCTATACC	*trnV1*	GAAGGTCTACGGTTCGAGTCC	52516-53791	1276	53	P
68	V1M	*trnV1*	GGGCTATACGGACTCGAACC	*trnM*	CCGCCGTATGAAAGCAGTA	53762-54636	875	53	P
69	V2atpB	*trnV2*	CGAGTTGCTCTACCAACTGAGC	*atpB*	CTACCGCGAAGGCTATGAAC	54373-55335	963	53	P
6-12	atpBrbcL	*atpB*	GCTGTACCTCACAAGCCACA	*rbcL*	TCGGTCCATACAGTTGTCCA	56584-57814	1231	53	NP
71	petApsbJ	*petA*	GCATCTGTTATTTTGGCACA	*psbJ*	TGGCCGATACTACTGGAAGG	65219-66489	1271	53	P
72	psbJpsbE	*psbJ*	GAACCCAATCCGGAATATGA	*psbE*	CGGGTTGGTTATTTGTCAGC	66378-67037	660	53	NP
81	rps18rpl20	*rps18*	GAAACAATTTGAAAGAACCGAGT	*rpl20*	CGGGGATATATAGCTCGGAGA	70765-71399	635	53	NP
82	rpl20clpP	*rpl20*	ATCCCGATGAGCCGAAACTA	*clpP1*	GCCCAAGCTTATGGAATTGTT	71268-72517	1250	55	P
91	rbcLaccD	*rbcL*	GAAATTATTCGCGAGGCTTG	*accD*	TTAAACCACCATCTTTCCAT	58926-59868	943	48/65	NA
B1	ndhFrpl32	*ndhF*	ACTGGAAGTGGAATGAAAGG	*rpl32*	TTCCTTTTCCAAATATTTTTACG	114242-115130	889	53	P

TAB. 6.1 – Amorces choisies pour le séquençage chloroplastique. : *[Fénart et al., 2006], **[Grivet and Petit, 2002], ***[Grivet and Petit, 2003]
P : fragment polymorphe, NA : fragment non amplifié, NP : fragment non polymorphe.

6.1.2 Phylogénie chloroplastique

Les fragments amplifiés contenant au moins un site polymorphe ont été concaténés afin de produire un arbre phylogénétique chloroplastique. Nous avons réalisé deux arbres, l'un en utilisant une méthode de *Neighbor Joining* (NJ) et l'autre en utilisant une méthode de *maximum de vraisemblance* (ML). L'arbre NJ a été réalisé en utilisant BioNJ [Gascuel, 1997] (bootstrap × 1000, Kimura 2 paramètres) et l'arbre ML en utilisant TreePuzzle [Strimmer and von Haeseler, 1996] (bootstrap × 1000, paramètre HKY85). Les résultats sont présentés en Figure 6.3. Les deux méthodes produisent des arbres ayant la même topologie. Le séquençage de fragments chloroplastiques ne nous a pas permis d'améliorer le réseau haplotypique vu en Figure 6.1. En effet, nous retrouvons bien la séparation du groupe Nv, D, A et B des autres espèces, nous avons également A et B qui sont plus proches. Cependant, nous avons un râteau ne permettant pas de situer exactement CMS-E, CMS-G, TK81-MS et Sv. Nous sommes à un niveau intra-spécifique où les espèces sont très proches. Même au niveau chloroplastique, et où il est donc assez difficile d'établir une phylogénie. Cette phylogénie chloroplastique nous confirme que TK81-MS est très proche de Sv et TK81-O de Nv.

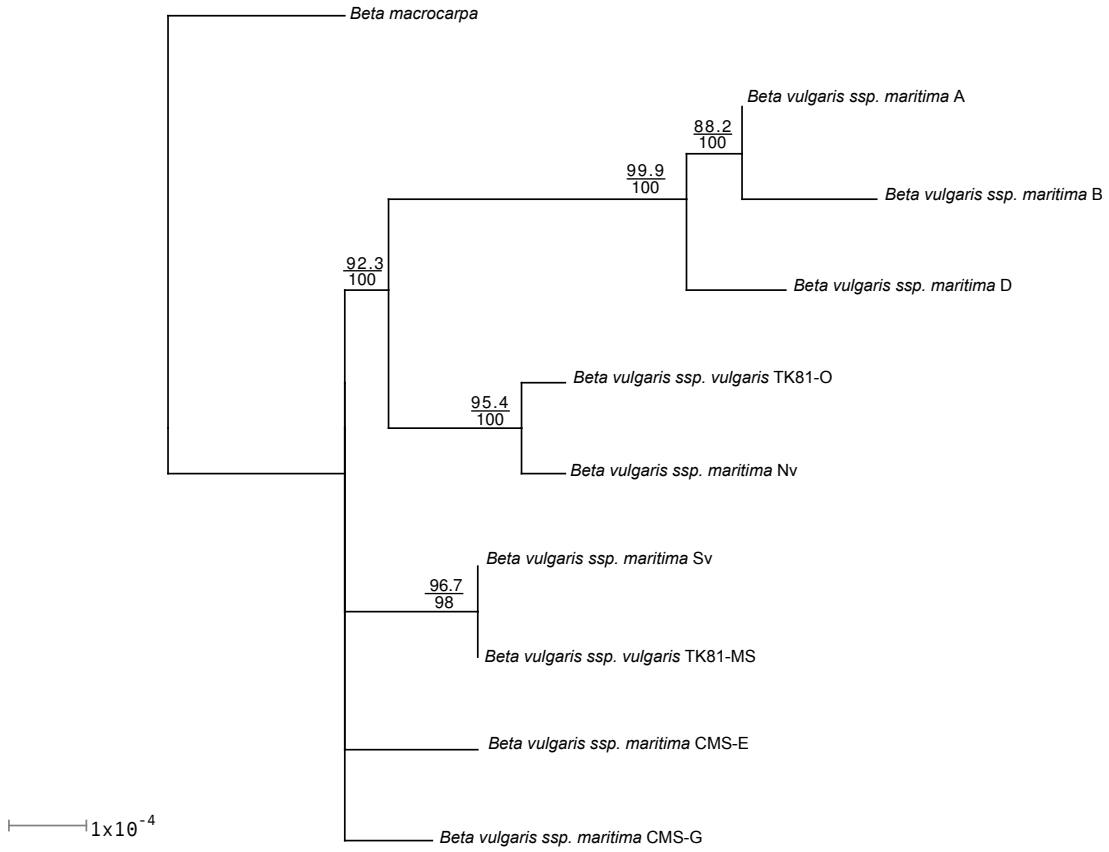


FIG. 6.3 – Phylogénie chloroplastique réalisée avec BioNJ. Valeurs de bootstrap : haut BioNJ, bas : Tree-Puzzle.

6.2 Méthode de reconstruction des génomes

Les génomes mitochondriaux des cytoplasmes choisis ont été extraits selon la procédure décrite par Scotti et collaborateurs [Scotti et al., 2001]. Cette extraction a été réalisée à l'Institut de Biologie Moléculaire des Plantes (IBMP, Strasbourg) avec l'aide de Laurence Maréchal-Drouard. Les génomes ont ensuite été séquencés au Génoscope. L'extraction mitochondriale a été réalisée sur les racines, limitant ainsi les contaminations chloroplastiques, pour les génomes A, B, CMS-E et CMS-G et sur les feuilles pour *Beta macrocarpa* qui ne présentait pas suffisamment de matériel racinaire puisqu'il s'agit d'une plante annuelle.

A la fin du séquençage des génomes mitochondriaux de *Beta vulgaris* A,B CMS-E, CMS-G et *Beta macrocarpa*, trois des génomes sont composés de deux contigs. Pour chacun de ces génomes (CMS-E, CMS-G et macrocarpa) nous avons un contig circulaire et un contig linéaire sauf pour macrocarpa où les deux contigs sont linéaires. Lorsque nous avons récupéré les génomes séquencés, le génome A était aussi dans cette configuration (contig 434 linéaire et contig 462 circulaire). Étant donné que les deux génomes séquencés de *Beta vulgaris* (TK81-O et TK81-MS) sont sous forme d'un seul cercle et qu'aucune molécule linéaire n'a été détectée dans ces espèces [Backert et al., 1997], nous avons choisi d'essayer de rassembler ces contigs en un seul cercle.

Circularisation de A

La stratégie adoptée pour A fut la suivante : dans un premier temps nous avons effectué un dotplot entre A et la séquence de TK81-O (voir Figure 6.4) afin de déterminer les permutations correspondant à ces deux génomes. Nous avons alors obtenu les permutations suivantes :

```
TK81-O : 1 2 3 4 5 6 7 -5 -6 8 9 10 11 -5 -4 -3
A contig 462 (circulaire) : 11 -5 -4 -3 1 10
A contig 434 (linéaire)   : -3 -2 -9 -5 8 7 -6 -5
```

En analysant les données, nous avons pu constater que le problème d'assemblage de ces deux contigs venait très certainement de la présence de régions dupliquées. En effet, lorsqu'une région dupliquée dépasse la taille des fragments séquencés (5 kbp dans notre cas), cela peut engendrer des erreurs lors de l'assemblage et les régions flanquantes de ces régions dupliquées peuvent être inversées. D'après nos données, TK81-O possède trois copies de 5, deux copies de 3 et deux copies de 4 alors que chez A nous avons bien les trois copies de 5 et les deux copies de 3 mais une seule copie de 4. Étant donné que chez TK81-O le fragment 4 est toujours à côté de 5, on peut supposer que le contig 434 de A ne se circularise pas par absence de 4 à la fin de ce contig. Nous avons alors déterminé des régions pour trouver des amorces PCR afin de tester nos hypothèses. Les contigs et les hypothèses de PCR sont schématisés en Figure 6.5.

Pour circulariser le contig 434 et tester la présence de -4 entre -5 et -3, deux PCR suffisent. Il faut alors choisir des fragments uniques afin d'être sûr de la position des amorces. Nous avons donc choisi les blocs -6 et -2 pour tester la présence éventuelle du bloc 4 dans le contig 434. Les PCR préconisées ont été effectuées au Génoscope. Ces deux PCR ont fonctionné, confirmant la présence du bloc 4 et la circularisation du contig 434. Le bloc 4, faisant environ 10 kbp et étant toujours entouré de 5 et 3, avait été replacé dans le contig 462 et non dans le contig 434. L'analyse de la profondeur (c'est-à-dire le nombre de séquences) de cette région confirme cette duplication.

Nous avons donc obtenu deux contigs circulaires. Il existe alors deux possibilités : soit ces deux contigs sont tels quels dans les mitochondries, soit ils forment un cercle maître en intégrant

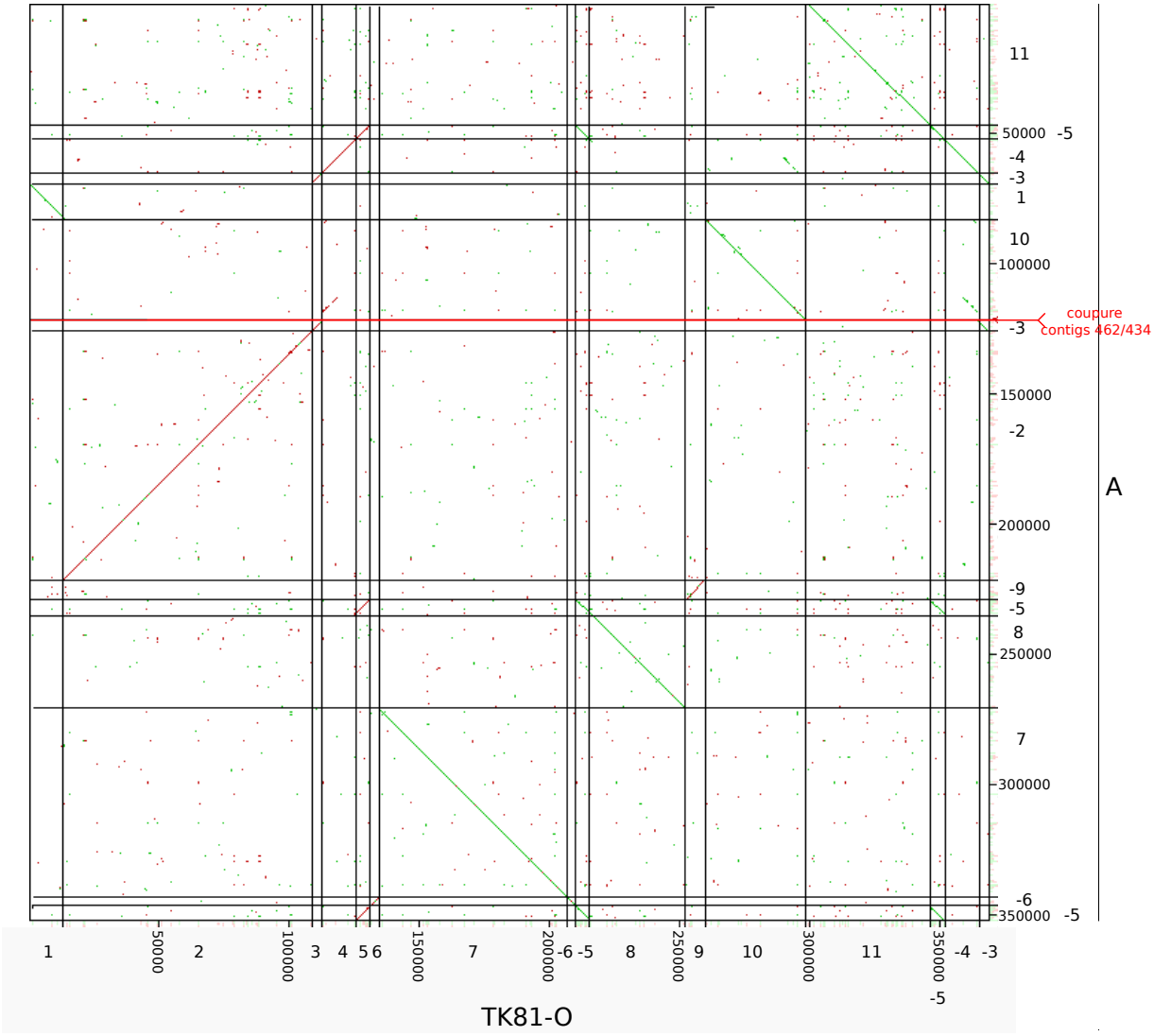


FIG. 6.4 – Dotplot obtenu grâce à YASS entre les contigs de A et la séquence de TK81-O. Les numéros indiqués, variant de 1 à 11, correspondent aux numéros donnés aux blocs homologues. Lorsque deux régions s'alignent entre les génomes, ce qui signifie qu'elles sont homologues, une droite est représentée entres les régions correspondantes. Les droites en vert correspondent à des fragments homologues ayant la même orientation sur les génomes, celles en rouge correspondent à des fragments ayant des orientations différentes.

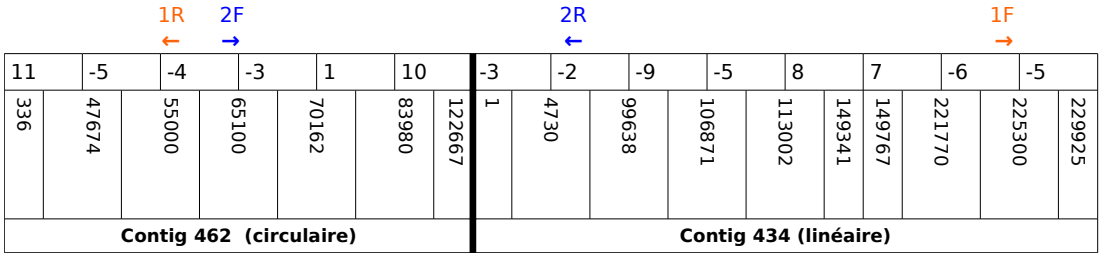


FIG. 6.5 – Contigs présents chez *Beta vulgaris ssp. maritima A* avant la reconstruction du cercle maître.

le contig 434 dans le contig 462. Étant donné que TK81-O et TK81-MS ainsi que la majorité des génomes mitochondriaux sont sous forme de cercle maître, nous avons choisi d'intégrer les deux contigs (représenté en Figure 6.6).

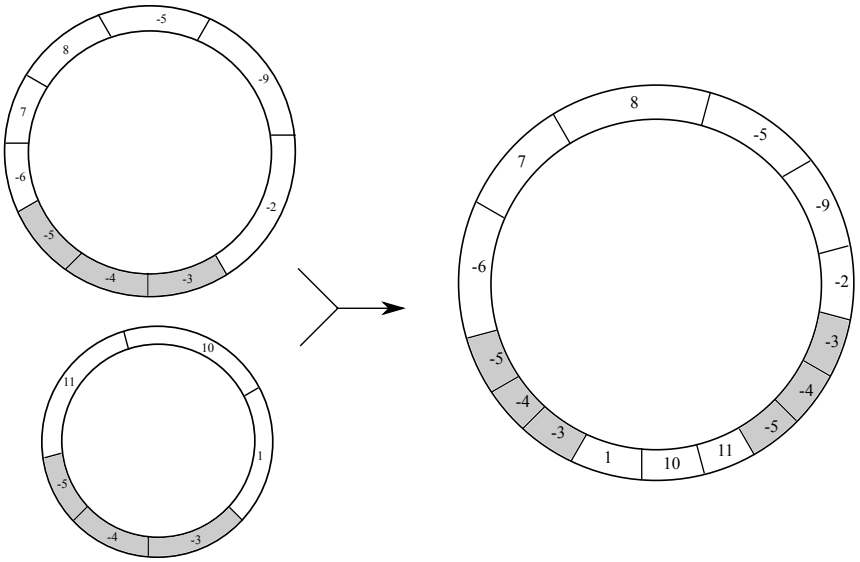


FIG. 6.6 – Intégration des contigs circulaires chez *Beta vulgaris ssp. maritima A* (à gauche) pour la reconstruction du cercle maître (à droite). L'intégration est faite au niveau de la duplication {-5,-4,-3}.

Les génomes CMS-E, CMS-G et macrocarpa sont également en deux contigs. De premières analyses ont été réalisées mais n'ont pas pu être terminées. Nous n'avons donc pas pu circulariser les contigs linéaires afin des les intégrer dans les circulaires.

Circularisation de CMS-G

Dans le cas du génome G, nous obtenons les permutations suivantes :

```
TK81-O : 1 2 3 4 5 6 7 8 9 -6 -5 10 11 12 13 14 15 16 17 -5 -4
CMS-G contig 249 (circulaire) : 6 -9 1 -12 4 5 6 11 -3 16 2 17 15 -13 -10 5
CMS-G contig 244 (linéaire)   : 6 -9 -8 -7 14
```

Ici, nous pourrions penser tout simplement à l'absence du bloc 5, sans lequel l'espace entre 6 et 14 dans le contig 244 est trop grand pour voir la circularisation. Celui-ci pourrait alors s'intégrer à la suite du contig 244 et donnerait un cercle maître de la forme 6 -9 1 -12 4 5 6 11 -3 16 2 17 15 -13 -10 5 6 -9 -8 -7 14 5. Cependant nous avons testé une PCR entre le bloc -10 et 1 qui donne la taille trouvée dans le premier contig, prouvant l'existence de la structure -10 5 6 -9 1 dans ce génome, ainsi qu'une PCR entre -10 et -8 qui n'a pas fonctionné, éliminant la présence de -10 en amont du contig 244. L'intégration du contig 244 ne peut donc pas être à la fin du contig 249. Nous avons ensuite pensé que le contig 244 pouvait s'intégrer au niveau de -12 s'il existe les blocs 4 et 5 en amont du contig 244. Nous avons donc tenté une PCR entre les blocs 4 et -9 (ce motif n'existe pas dans le contig linéaire). Cette PCR a fonctionné, par contre elle n'a pas été séquencée. Cela dit, une PCR entre les blocs 14 et 4 n'a pas fonctionné ce qui signifie qu'il y a certainement plus de 20 kpb de séquence manquante entre 4 et 14. Le contig 244 n'est donc pas circularisé. Nous avons constaté que la taille du génome CMS-G était beaucoup plus petite que celle des autres génomes et qu'il manquait environ 30 kbp pour arriver à une taille similaire. A partir des contigs non assemblés, nous avons constaté certains contigs présents chez TK81-O mais non assemblés chez CMS-G. Des PCR entre ces contigs et le bloc 4 ont confirmé leur présence (en amont du contig 244). Cependant, elles n'ont pas non plus été séquencées et les investigations sur ce génome ont été arrêtées par le Génoscope.

Circularisation de CMS-E

Dans le cas du génome CMS-E, nous obtenons les permutations suivantes :

```
TK81-O : 1 2 3 4 5 6 7 8 9 10 11 12 -11 -10 -9 13 14 5 15 -10 -9 -8 -7
CMS-E contig 281 (circulaire) : -2 3 4 5 15 -11 -10 -12 -13 9 10 11 -14
CMS-E contig 270 (linéaire)   : -8 -7 1 4 5 6 8
```

Dans la cas de ce génome, nous avons confirmé la présence du bloc -9 en amont du contig 270 (PCR entre -9 et 7) ainsi que du bloc 9 à la fin de ce même contig (PCR entre 6 et 9). Deux PCR, entre 3 et 15 et entre 1 et 5, confirment bien que ces deux blocs sont reliés (et non pas 1 avec 15 et 3 avec 5 qui n'ont pas fonctionné). Les analyses PCR se sont arrêtées ici pour ce génome et les PCR localisant les blocs 9 aux extrémités du contig 270 n'ont pas été séquencées. Cependant, étant donné qu'il existe trois copies des blocs 9 et 10 chez TK81-O, nous pouvons penser qu'il y en a également trois chez CMS-E. Il manquerait donc une copie de 10. Le contig 270 pourrait alors être circulaire, de la forme -10 -9 -8 -7 1 4 5 6 7 9, et pourrait s'intégrer entre -10 et -12 pour donner le cercle maître -2 3 4 5 15 -11 -10 -9 -8 -7 1 4 5 6 7 9 -10 -12 -13 9 10 11 -14.

Circularisation de macrocarpa

Dans le cas de macrocarpa, nous obtenons les permutations suivantes :

```
TK81-O : 1 2 3 4 5 6 7 8 9 10 -7 -6 11 12 13 14 -6 -5 -4
macro contig 1112 (linéaire) : -6 11 8 -2 -12 10 -7 -6 -5 -4 -3 9 14 -6
macro contig 1167 (linéaire)   : -4 -1 13
```

Pour ce génome, les PCR sont en cours au moment de l'écriture de ce document et la présence du bloc -5 en amont du contig 1167 est déjà confirmée. Nous cherchons maintenant à savoir si

celui-ci se circularise et s'il s'intègre dans le contig 1112. Deux positions sont possibles, soit au niveau des blocs -6 -5 soit à la fin du contig 1112.

La méthode que nous avons appliquée pour reconstruire les génomes donne de bons résultats. Nous avons réussi à reconstituer le cercle maître de A et agrandi les contigs de CMS-E, CMS-G et macrocarpa avec des hypothèses de duplication. Malheureusement, le manque de temps ne nous a pas permis d'aller jusqu'au bout de nos hypothèses et ces trois génomes sont encore sous forme de deux contigs.

6.3 Analyse du contenu

Nous présenterons dans cette partie l'analyse des six génomes mitochondriaux de *Beta vulgaris* et de celui de *Beta macrocarpa* décrits précédemment. Cette analyse consiste en une comparaison de l'ensemble de ces génomes, que ce soit au niveau de leur contenu en gènes ou de leur taux d'évolution. Nous noterons, malgré la présence de trois génomes non finalisés, que leur contenu en gènes est identique. Cependant, ces génomes contiennent des ORF spécifiques et certains de leurs gènes ont des structures différentes. Par exemple, dans le génome CMS-G, tous les gènes appartenant au complexe IV sont modifiés par mutations (substitutions et insertion/délétions) aboutissant, pour certains des gènes, à une protéine tronquée. Dans cette étude, nous réaliserons également une phylogénie basée sur les séquences communes entre les génomes ainsi que des analyses de leurs taux de substitutions synonymes et non synonymes. Ces analyses montreront que les taux de substitutions entre les génomes CMS et non-CMS sont différents, laissant penser qu'un taux de substitutions plus élevé chez les CMS serait à l'origine de l'émergence de la stérilité mâle. Enfin nous comparerons l'évolution des génomes mitochondriaux de *Beta* et de *Zea* et constaterons que les forces d'évolution entre ces deux espèces ne sont pas tout à fait identiques.

Cette partie est également rédigée en Annexe Chapitre 6 (Page 164) sous la forme d'un article qui est en préparation pour une soumission dans *Nucleic Acid Research*.

6.3.1 Méthodes

Annotation des génomes

L'annotation des génomes a été réalisée en utilisant Plamidb. Les séquences de gènes codant pour des protéines, ARN et ORF ont été comparées aux annotations des génomes mitochondriaux des deux génomes de betterave TK81-O et TK81-MS déjà séquencés ainsi qu'à ceux d'*Arabidopsis thaliana* et *Nicotiana tabacum*. Les sites édités ont été déterminés par rapport aux annotations établies dans le génome de TK81-O qui avait bénéficié d'une analyse expérimentale [Mower and Palmer, 2006]. Comme il est décrit dans le Chapitre 3, l'annotation des ARN est vérifiée avec une analyse par tRNAScan-SE [Lowe and Eddy, 1997] et d'éventuelles ORF supplémentaires sont recherchées (ORF d'au moins 300 pb).

De plus, nous avons vérifié les ORF trouvées en réalisant un Blast sur la base de données non-redondante. Les ORF chimériques, c'est-à-dire composées de fragments de gènes connus, ont été recherchées par deux méthodes. La première consistait à utiliser YASS [Noe and Kucherov, 2005] pour chaque ORF contre les séquences de gènes et ARN annotés avec une E-value de 0,1 et en ne conservant que les fragments homologues à 100%. La deuxième méthode consistait à réaliser un Blast de ces ORF sur GenBank avec une E-value de 0,1 et une taille de mot de 16.

Complexité des génomes

La complexité des génomes (*genome complexity*, en anglais) correspond à la taille des génomes lorsque l'on ne considère qu'une copie de chaque fragment dupliqué, de taille supérieure ou égale à 500 pb. Nous avons détecté ces régions dupliquées à l'aide de YASS en conservant les séquences de taille supérieure à 500 pb avec une E-value supérieure à $1e^{-30}$ et un score de +1 pour les match et -3 pour les substitutions.

Analyse des séquences d'origine chloroplastiques

Les séquences insérées, d'origine chloroplastique, ont été détectées en utilisant YASS avec une E-value supérieure à $1e^{-30}$ et un score de +1 pour les match et -3 pour les substitutions. Nous avons conservé les régions d'au moins 30 pb contenant moins de 10% de mutations (substitutions et insertions/délétions) par rapport aux séquences des génomes chloroplastiques.

Analyses phylogénétiques

Afin d'obtenir une phylogénie des génomes au niveau des séquences nucléotidiques, nous avons réalisé une concaténation des séquences communes à tous les génomes en utilisant Mauve sur les séquences des complexités des génomes. Les paramètres que nous avons utilisés avec Mauve [Darling et al., 2004] sont les mêmes que ceux décrits dans le Chapitre 4. Les séquences ont été concaténées par rapport à leur ordre dans le génome A (l'ordre n'ayant pas d'importance ici).

Les analyses de Neighbor-Joining ont été réalisées en utilisant BIONJ (bootstrap $\times$1000, Kimura 2 paramètres) [Gascuel, 1997]. Les analyses de maximum de vraisemblance et les tests d'horloge moléculaire ont été effectués avec TREE-PUZZLE [Schmidt et al., 2002] en utilisant le modèle de substitution Hasegawa-Kishino-Yano (HKY85) [Hasegawa et al., 1985].

Analyse des taux de substitutions

Les analyses de diversité et divergence nucléotidiques ont été réalisées avec DnaSP [Rozas et al., 2003] sur les 29 gènes concaténés de *Beta vulgaris*. Nous avons regardé le nombre moyen de substitutions synonymes par site synonyme entre les génomes CMS ou entre non-CMS (π_s), le nombre de substitutions non synonymes par site non synonyme entre les génomes CMS ou entre non-CMS (π_a) ainsi que les divergences synonymes (K_s) et non synonymes (K_a) de ces génomes par rapport à *Beta macrocarpa*.
La taille estimée de la protéine *cox1* de CMS-G a été établie avec le programme pI/MW d'Expasy (http ://scansite.mit.edu/cgi-bin/calcpi) et les analyses statistiques ont été faites avec le logiciel Minitab.

6.3.2 Résultats

Taille des génomes et composition

Rappelons que nous avons séquencé cinq génomes mitochondriaux (quatre de *Beta vulgaris* et un de *Beta macrocarpa*). Les deux génomes fertiles ont été complètement circularisés (A et B) tandis que les trois autres sont encore en deux contigs (un contig circulaire et un linéaire pour les deux génomes stériles CMS-E et CMS-G et deux contigs linéaires pour *Beta macrocarpa*). Les génomes et leur contenu sont représentés en Figure 6.7. Les tailles obtenues varient entre 341,257 kpb pour CMS-G et 378,457 kpb pour CMS-E. Par rapport aux génomes précedemment

séquencés, ces tailles sont plus proches de la taille du génome mitochondrial fertile (TK81-O 368,801 kbp) que de celle de la CMS (TK81-MS 501,020 kbp). Les données sur le contenu des cinq génomes séquencés ainsi que des deux génomes existants sont résumées dans le Tableau 6.2. La médiane des pourcentages de GC est de 43,89% ce qui correspond aux valeurs généralement trouvées chez les plantes [Allen et al., 2007]. Nous trouvons un ratio de 1,47 entre les tailles du plus petit et du plus grand génome. La variation entres les tailles des génomes est principalement due aux duplications qui représentent 4,55% de la taille du génome pour CMS-G à 29,98% pour TK81-MS (coefficient de corrélation de Pearson r= 0,968 ; p=0,000). La complexité des génomes, représentant le contenu génétique sans redondance de chaque génome, est quant à elle moins variable entre les génomes, allant de 325,716 kpb pour CMS-G à 357,125 kpb pour CMS-E avec une médiane de 335,262 kpb (ratio de 1,10 entre le génome le plus grand et le génome le moins grand). Nous pouvons noter que la complexité des génomes n'est pas corrélée avec la taille (coefficient de corrélation de Pearson r=0,542 ; p=0,209).

Séquences chloroplastiques intégrées

La taille totale des séquences chloroplastiques intégrées dans les génomes mitochondriaux varie de 4202 pb pour CMS-G à 8123 pb pour CMS-E. Elle représente 1,27% à 2,33% des complexités de génomes (Tableau 6.2). La taille des séquences intégrées varie entre 22 pb pour tous les génomes et 3368 pb dans les génomes B, E et macrocarpa (Tableau supplémentaire S1, page 195) avec une médiane par génome variant de 34 pb pour CMS-G à 41 pb pour TK81-O et macrocarpa.

Gènes conservés

Tous les génomes analysés possèdent le même contenu en gènes (Tableau 6.3). Ils sont composés de 29 gènes codant pour des protéines : 18 sont impliqués dans la chaîne de transport des électrons : 9 dans le complexe I (*nad* 1, 2, 3, 4, 4L, 5, 6, 7, 9), 1 dans le complexe III (*cob*), 3 dans le complexe IV (*cox* 1, 2, 3) et 5 dans le complexe V (*atp* 1, 6, 8, 9, orf25) ; 3 gènes sont impliqués dans la biogenèse du cytochrome c (*ccmB*, *ccmFC*, *ccmFN*) ; 6 gènes codent pour des protéines ribosomiques (rpl5, *rsp* 3, 4, 7, 12, 13) ; on trouve également un gène impliqué dans le système de translocation (*tatC*) et une maturase (*mat-r*).
En utilisant TK81-O comme référence, nous avons déterminé 33 sites édités sur les gènes codant pour des protéines. Nous pouvons noter trois cas d'édition du codon start (ACG en ATG dans les gènes *atp6*, *nad1* et *nad4L*) et deux cas d'édition du codon stop (CAA en TAA pour *atp6* et CGA en TGA pour *atp9*). Le gène *tatC* ne possède pas de codon start AUG, le codon start prédit chez TK81-O, AUA, ne semble pas être édité.
L'utilisation des codons stop est UAA pour 16 gènes (en incluant *atp6* édité) : *atp* 1, 6, 8, *cox* 1, 2, *nad* 1, 2, 3, 4L, 5, 6, 9 (sauf une des copies de CMS-G), *rpl5*, *rps* 3, 4, 7 ; UGA pour 10 gènes : *atp9*, *ccmB*, *ccmC*, *ccmFN*, *cob*, *cox2* (chez CMS-G et une des copies de TK81-MS), *cox3*, *nad4*, *rps* 12, 13 et UAG pour 4 gènes : *mat-r*, *nad7*, *nad9* (TK81-MS et une des copies de CMS-G), *tatC*.
Nous avons trouvé 20 introns pour 7 protéines, comme cela avait été décrit précédemment [Kubo et al., 2000]. Six sont épissés en trans pour trois gènes (*nad* 1, 2, 5) et quatorze sont épissés en cis pour six gènes (*nad* 1, 2, 4, 5, 7 et *ccmFC*). Nous pouvons noter que, contrairement aux autres plantes, le gène *rps3* contenu dans les génomes mitochondriaux de *Beta* ne possède pas d'intron. Cependant, une séquence supplémentaire, homologue à l'intron 1, est retrouvée dans les génomes mitochondriaux de *Beta* (nous l'avons alors compté comme un pseudogène).

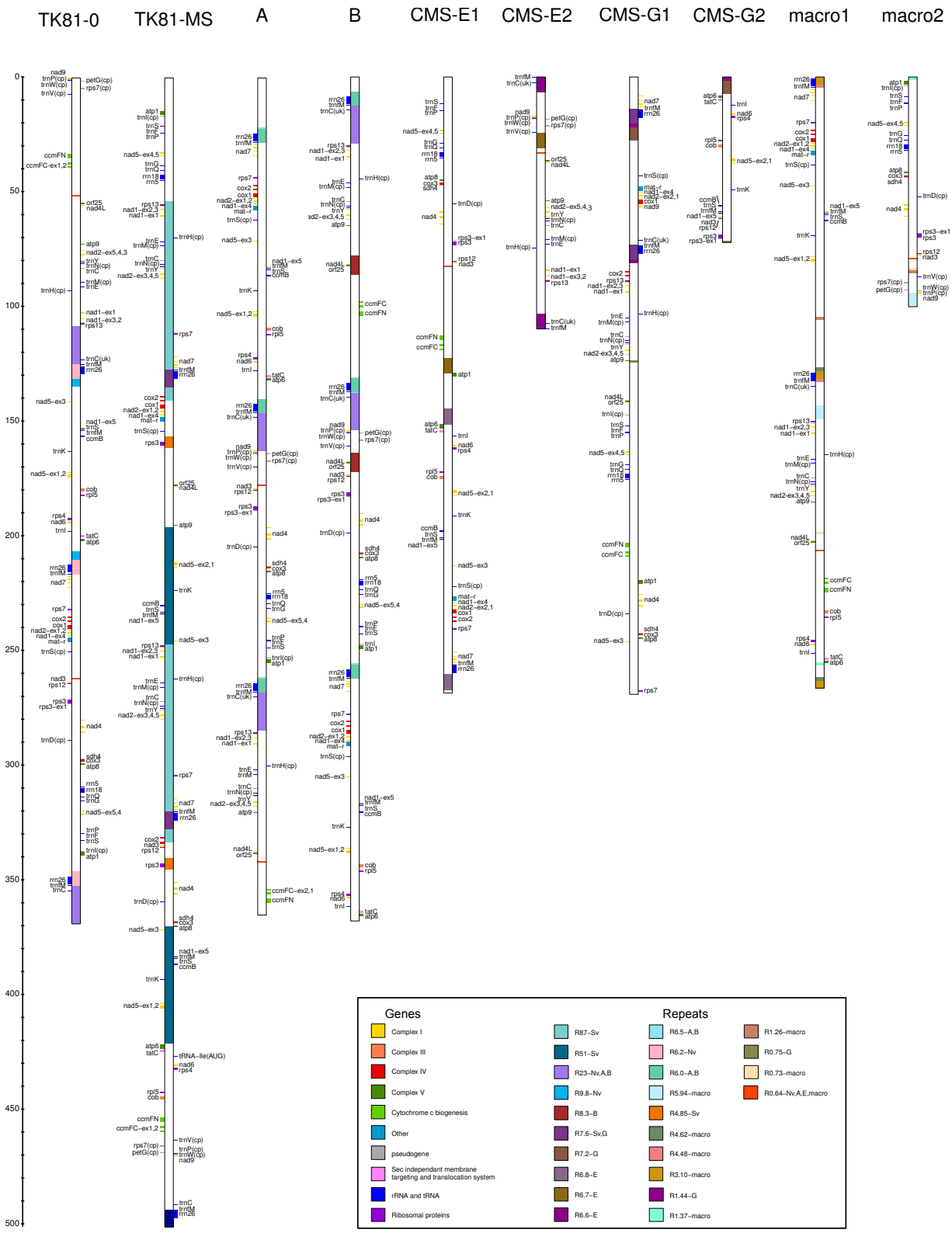


FIG. 6.7 – Représentation schématique des génomes mitochondriaux de *Beta*. Tous les génomes ont été représentés sous forme linéaire. Pour CMS-E, CMS-G et macrocarpa, les deux contigs non assemblés sont représentés. Pour un génome donné, les rectangles externes correspondent aux gènes, ORF et ARN et les rectangles internes correspondent aux duplications.

	Génomes						
Contenu	TK81-O	TK81-MS	A	B	CMS-E	CMS-G	macro
Génomes							
Taille des génomes	368801	501020	364950	367943	378457 circ : 268616 lin : 109841	341257 circ : 269136 lin : 72121	366580 lin : 266432 lin : 100148
%GC	43,86	43,89	43,91	43,89	43,88	43,92	43,96
Séquences répétées	34313	150214	29688	37461	21332	15541	19613
Séquences répétées, % du genome	9,30	29,98	8,13	10,18	5,64	4,55	5,35
Complexités des génomes	334488	350806	335262	330482	357125	325716	346967
Gènes							
Gènes protéiques	27693	37485	27693	28593	28845	28118	27693
1 copie des gènes protéiques	27693	28839/29142	27693	27693	28845	28118	27693
Introns cis	18727	30641	18749	18749	18748	18746	18749
1 copie des introns cis	18727	18733	18749	18749	18748	18746	18749
ARNr	12065	12065	12065	12065	5389	8727	8727
1 copie des ARNr	5389	5389	5389	5389	5389	5389	5389
ARNt	1746	2282	1746	1746	1746	1453	1599
1 copie des ARNt	1449	1449	1449	1449	1449	1303	1449
Pseudogènes et pseudo-exons	1180	1177	1180	1180	1180	524	1180
1 copie des pseudogènes et pseudo-exons	1180	1106	1180	1180	1180	524	1180
Total du codant							
Genes connus	41457	51760	41432	42332	35908	38298	37947
1 copie des gènes	34482	35602/35905	34459	34459	35611	34810	34459
Genes, % du genome	11,24	10,33	11,35	11,51	9,49	11,22	10,35
1 copie des genes, % de la complexité du génome	10,31	10,15/10,24	10,28	10,43	9,97	10,69	9,93
Total des ORF							
ORF	69801	96427	72985	69432	72317	66536	68279
1 copie des ORF	63147	72944	65218	62973	70850	66536	66704
ORF, % du genome	18,93	19,25	20,00	18,87	19,11	19,50	18,63
1 copie des ORF, % de la complexité du génome	18,88	20,79	19,45	19,05	19,84	20,43	19,22
Séquences chloroplastiques intégrées							
Séquences insérées	7862	7344	7740	7743	8123	4202	7865
1 copies des séquences insérées	7812	6671	7698	7668	8098	4132	7815
% du genome	2,13	1,47	2,12	2,10	2,15	1,23	2,15
% de la complexité des génomes	2,33	1,90	2,30	2,32	2,27	1,27	2,27

TAB. 6.2 – Contenu des génomes mitochondriaux de *Beta*

Groupes de gènes	Gènes	TK81-O	TK81-MS	A	B	CMS-E	CMS-G	macro
Complexe I								
	nad1 (5 exons)	+	+[a]	+	+	+	+	+
	nad2 (5 exons)	+	+[b]	+	+	+	+	+
	nad3	+	+	+	+	+	+	+
	nad4 (3 exons)	+	+	+	+	+	+	+
	nad4L	+	+	+	2+	+	+	+
	nad5	+	+[c]	+	+	+	+	+
	nad6	+	+	+	+	+	+	+
	nad7 (5 exons)	+	2+	+	+	+	+	+
	nad9	+	+	+	+	+	+ (2*)	+
Complexe III								
	cob	+	+	+	+	+	+	+
Complexe IV								
	cox1	+	+	+	+	+	+	+
	cox2 (2 exons)	+	2+	+	+	+	+	+
	cox3	+	+	+	+	+	+	+
Complexe V								
	atp1	+	+	+	+	+	+	+
	atp6	+	+	+	+	+	+	+
	atp8	+	+	+	+	+	+	+
	atp9	+	+	+	+	+	+	+
	orf25 (*atp4*)	+	+	+	2	+	+	+
Cytochrome-c-biogenesis								
	ccmB	+	2+	+	+	+	+	+
	ccmFC (2 exons)	+	+	+	+	+	+	+
	ccmFN	+	+	+	+	+	+	+
Protéines ribosomiques								
	rpl5	+	+	+	+	+	+	+
	rps3	+	2+	+	+	+	+	+
	rps4	+	+	+	+	+	+	+
	rps7	+	2+	+	+	+	+	+
	rps12	+	+	+	+	+	+	+
	rps13	+	2+	+	+	+	+	+
Autres protéines								
	mat-r	+	+	+	+	+	+	+
Pseudogènes								
	sdh4	+	+	+	+	+	+	+
	petG cp-like	+	+	+	+	+	(*)	+
	rps7 cp-like	+	+	+	+	+	(*)	+
	rps3 exon 1	+	-	+	+	+	+	+
Sec-independant membrane targeting and translocation system								
	tatC	+	+	+	+	+	+	+
ARN ribosomiques								
	rrn5S	+	+	+	+	+	+	+
	rrn18S	+	+	+	+	+	+	+
	rrn26S	3+	3+	3+	3+	+	2+	2+
ARN de transfert								
	Natifs							
	trnC1-GCA	ψ	2ψ	ψ	ψ	ψ	ψ	ψ
	trnE-UUC	+	2+	+	+	+	+	+
	trnF-GAA	+	+	+	+	+	+	+
	trnG-GCC	+	+	+	+	+	+	+
	trnI-CAU	+	+	+	+	+	+	+
	tnrP-UGG	+	+	+	+	+	+	+
	trnQ-UUG	+	+	+	+	+	+	+
	trnS-GCU	+	+	+	+	+	+	+
	trnS-UGA	+	2+	+	+	+	+	+
	trnY-GUA	+	2+	+	+	+	+	+
	trnK-UUU	+	2+	+	+	+	+	+
	trnfM-CAU	4+	5+	4+	4+	4+	3+	3+
	Origine chloroplastique							
	trnD-GUC	+	+	+	+	+	+	+
	trnH-GUG	+	2+	+	+	+	+	+
	trnI-CAU	ψ	ψ	ψ	ψ	-	ψ	ψ
	trnN-GUU	+	2+	+	+	+	+	+
	trnM-CAU	+	2+	+	+	+	+	+
	trnP-UGG	ψ	ψ	ψ	ψ	ψ	-	ψ
	trnS-GGA	+	+	+	+	+	+	+
	trnV-GAC	+	+	+	+	+	-	+
	trnW-CCA	+	+	+	+	+	(*)	+
	Origine inconnue							
	trnC2-GCA	2+	+	2+	2+	2+	+	+

TAB. 6.3 – Gènes, ARNr et ARNt contenus dans les génomes mitochondriaux des *Beta*.
+ : présent, - : absent, ψ : pseudogène, le nombre de copies est donné.
[a] : 2 copies des exons 1,2,3 et 5,
[b] : 2 copies des exons 3,4 et 5,
[c] : 2 copies des exons 1,2 et 3.
(*) : est retrouvé dans un contig non assemblé.

Les génomes de *Beta* possèdent trois gènes ARNr (*rrn* 5S, 8S et 26S). Le gène rrn26S est retrouvé en trois copies chez TK81-O, TK81-MS, A et B, en deux copies chez CMS-G et macrocarpa et un seule copie chez CMS-E. Cette absence des trois copies dans les trois derniers génomes, vient probablement du fait que ces génomes ne sont pas complets et pour lesquels il est possible qu'il manque des fragments dupliqués.

Dix-huit ARNt sont retrouvés ainsi que cinq pseudo-ARNt. Sur les dix-huit ARNt, onze sont d'origine mitochondriale, six d'origine chloroplastique et un d'origine inconnue (*trnC2* [Kubo et al., 2000]). L'ARNt d'origine chloroplastique *trnW* est retrouvé chez G dans un contig de 5 kbp qui n'a pas été assemblé aux contigs circulaire et linéaire. Sur les cinq pseudo-ARNt décrits par Kubo et ses collaborateurs [Kubo et al., 2000], *trnP* et *trnV* ne sont pas retrouvés chez CMS-G et macrocarpa et *trnI* n'est pas retrouvé chez CMS-E.

Polymorphisme des gènes codant pour des protéines

Sur les sept génomes mitochondriaux analysés, nous avons trouvé dix-neuf gènes polymorphes codant pour des protéines (Tableau 6.4). Nous avons détecté 53 substitutions et un gène contenant une partie 5' modifiée chez CMS-E et TK81-MS (*atp6*). Sur les 53 substitutions, 23 sont spécifiques à CMS-G, 14 à TK81-MS, 6 à TK81-O, 3 à CMS-E et 1 à macrocarpa. Quatre substitutions sont partagées par CMS-G et TK81-MS (l'allèle alternatif est partagé par CMS-E, TK81-O, A, B et macrocarpa), une est partagée par CMS-E, CMS-G et TK81-MS (l'allèle alternatif est partagé par TK81-O, A, B et macrocarpa). Sur un site, on retrouve un allèle unique pour TK81-MS, un partagé par A et B, les autres génomes partagent un troisième allèle. Sur ces 53 substitutions, 14 sont synonymes et 39 sont non synonymes. Sur les 39 substitutions non synonymes, une substitution modifie le codon start du gène *cox1* chez CMS-G entraînant un gène *cox1* soit plus court de 87 pb soit plus long de 408 pb. Comme cela avait déjà été décrit par notre équipe [Ducos et al., 2001], nous avons observé qu'une substitution dans l'exon2 de *cox2* entraîne l'apparition d'un codon stop prématuré et qu'une substitution au niveau du codon stop de *nad9* entraîne séquence codante plus longue de 42 pb. Trois des polymorphismes spécifiques de TK81-O (sur *nad2* exon4, *rsp3* et *rps12*) sont situés au niveau des sites édités sur les séquences non modifiées.

Nous avons également retrouvé, comme cela avait déjà été décrit [Satoh et al., 2004], deux copies de l'exon2 de *cox2* chez TK81-MS avec une copie identique aux exon2 de *cox2* dans les six autres génomes mitochondriaux et l'autre copie possèdant les 158 premiers nucléotides identiques aux autres génomes suivis de 506 pb qui lui sont spécifiques.

Enfin, comme il a été décrit [Yamamoto et al., 2005], le gène *atp6* possède une séquence supplémentaire de 1172 pb en amont du gène chez TK81-MS. CMS-E présente également une séquence plus longue en 5' du gène *atp6* ayant une homologie de 88% avec la séquence de TK81-MS et 1% d'indels. Cette séquence spécifique chez CMS-E est identique à l'*atp6* décrite pour I-12CMS(3) trouvée au Pakistan, suggérant que CMS-E et I-12CMS(3) seraient identiques [Onodera et al., 1999].

ORF

Nous avons trouvé 235 ORF (qui sont de potentielles régions codantes d'au moins 300 pb) parmi les sept génomes mitochondriaux, dont 34 se superposent à des gènes, 20 à des régions chloroplastiques et 13 sont chimériques. Ces séquences chimériques contiennent au moins 16 pb d'un gène mitochondrial ou chloroplastique (Tableau supplémentaire page 196 et Tableau 6.5 pour les ORF chimériques). Si l'on ne considère que les ORF spécifiques aux CMS, qui sont des

Genes	Position dans TK81-O	Genomes						
		TK81-O	TK81-MS	A	B	CMS-E	CMS-G	macro
atp1	1386	gaT→ D	gaT→ D	gaT→ D	gaT→ D	gaT→ D	gaG→E	gaT→ D
atp6			First 1171 bp Fspecific to Sv			First 1177 bp specific to CMS-E		
	264	gtT→V	gtT→V	gtT→V	gtT→V	gtG→V	gtT→V	gtT→V
	489	ccT→P	ccC→P	ccT→P	ccT→P	ccT→P	ccT→P	ccT→P
ccb438 exon 1	306	ttA→L	ttA→L	ttA→L	ttA→L	ttA→L	ttC→F	ttA→L
ccb438 exon 2	288	cCa→P	cCa→P	cCa→P	cCa→P	cCa→P	cAa→Q	cCa→P
cox1	1-3	AtG→M	AtG→M	AtG→M	AtG→M	AtG→M	TtT→F[a]	AtG→M
	14	gTt→V	gTt→V	gTt→V	gTt→V	gTt→V	gAt→D	gTt→V
	131	cGa→R	cGa→R	cGa→R	cGa→R	cGa→R	cAa→Q	cGa→R
	153	ggT→G	ggC→G	ggT→G	ggT→G	ggT→G	ggT→G	ggT→G
	966	atC→I	atA→I	atC→I	atC→I	atC→I	atC→I	atC→I
	1179	gcA→A	gcG→A	gcA→A	gcA→A	gcA→A	gcA→A	gcA→A
	1206	atC→I	atA→I	atC→I	atC→I	atC→I	atC→I	atC→I
	1207	Ttt→F	Ttt→F	Ttt→F	Ttt→F	Ttt→F	Gtt→V	Ttt→F
cox2 exon 2	376	tTa→L	tTa→L[b]	tTa→L	tTa→L	tTa→L	tGa→*	tTa→L
cox3	151	Att→I	Att→I	Att→I	Att→I	Att→I	Ctt→L	Att→I
mat-r	750	aaT→N	aaT→N	aaT→N	aaT→N	aaG→K	aaT→N	aaT→N
	1086	gtC→V	gtA→V	gtC→V	gtC→V	gtC→V	gtC→V	gtC→V
	1215	aaT→N	aaG→K	aaT→N	aaT→N	aaT→N	aaT→N	aaT→N
nad1 exon 1	7	Ata→T	Ata→T	Ata→T	Ata→T	Ata→T	Gta→V	Ata→T
nad1 exon 3	160-161	CGt→R	GCt→A	GCt→A	GCt→A	GCt→A	GCt→A	GCt→A
nad2 exon 4	14	ccC→P	ccT→P	ccT→P	ccT→P	ccT→P	ccT→P	ccT→P
	74	atA→I	atA→I	atA→I	atA→I	atA→I	atC→I	atA→I
nad4L	17	tAt→Y	tAt→Y	tAt→Y	tAt→Y	tAt→Y	tTt→F	tAt→Y
	19	Ttt→F	Ttt→F	Ttt→F	Ttt→F	Ttt→F	Gtt→V	Ttt→F
nad5 exon 1	13	Atc→I	Atc→I	Atc→I	Atc→I	Ctc→L	Atc→I	Atc→I
nad5 exon 4	3	Aat→N	Cat→H	Aat→N	Aat→N	Aat→N	Cat→H	Aat→N
nad7 exon 1	15	atC→I	atC→I	atC→I	atC→I	atC→I	atG→M	atC→I
nad9	59	aAa→K	aAa→K	aAa→K	aAa→K	aAa→K	aCa→T	aAa→K
	66	atA→I	atC→I	atA→I	atA→I	atA→I	atA→I	atA→I
	74	tCa→S	tCa→S	tCa→S	tCa→S	tCa→S	tTa→L	tCa→S
	118	Caa→Q	Aaa→K	Caa→Q	Caa→Q	Caa→Q	Caa→Q	Caa→Q
	261	cgG→R	cgC→R	cgC→R	cgC→R	cgC→R	cgC→R	cgC→R
	262	Cta→L	Gta→V	Gta→V	Gta→V	Gta→V	Gta→V	Gta→V
	318	ccA→P	ccA→P	ccA→P	ccA→P	ccA→P	ccG→P	ccA→P
	525	ttT→F	ttT→F	ttT→F	ttT→F	ttT→F	ttG→L	ttT→F
	559	Cgt→R	Cgt→R	Cgt→R	Cgt→R	Cgt→R	Ggt→G	Cgt→R
	557	Taa→*	Taa→*	Taa→*	Taa→*	Taa→*	Gaa→E[c]	Taa→*
orf25	463	Cac→H	Cac→H	Cac→H	Cac→H	Cac→H	Aac→N	Cac→H
rps3	85	Agt→S	Ggt→G	Agt→S	Agt→S	Agt→S	Ggt→G	Agt→S
	106	Ctc→L	Atc→I	Atc→I	Atc→I	Atc→I	Atc→I	Atc→I
	755	tCc→S	tTc→F	tTc→F	tTc→F	tTc→F	tTc→F	tTc→F
	1106	aTa→I	aTa→I	aTa→I	aTa→I	aTa→I	aGa→R	aTa→I
	1232	aTa→I	aGa→R	aTa→I	aTa→I	aTa→I	aGa→R	aTa→I
	1240	Gct→A	Cct→P	Gct→A	Gct→A	Gct→A	Cct→P	Gct→A
rps4	103	Aag→K	Gag→E	Aag→K	Aag→K	Aag→K	Aag→K	Aag→K
	527	cTg→L	cGg→R	cTg→L	cTg→L	cTg→L	cTg→L	cTg→L
	573	cgC→R	cgC→R	cgC→R	cgC→R	cgC→R	cgA→R	cgC→R
	745-746	TAt→Y	GAt→D	TCt→S	TCt→S	TAt→Y	TAt→Y	TAt→Y
rps7	198	gtC→V	gtA→V	gtC→V	gtC→V	gtA→V	gtA→V	gtC→V
rps12	269	tCg→S	tTg→L	tCg→S	tCg→S	tCg→S	tCg→S	tCg→S
	326	gAt→D	gGt→G	gAt→D	gAt→D	gAt→D	gAt→D	gAt→D
tatC	323	aGa→R	aGa→R	aGa→R	aGa→R	aGa→R	aGa→R	aTa→I
	567	tcT→S	tcC→S	tcT→S	tcT→S	tcT→S	tcT→S	tcT→S

Tab. 6.4 – Substitutions observées dans les gènes codant pour des protéines au sein des génomes mitochondriaux de *Beta*.

[a] : codon start non défini (commence 408 pb avant ou 87 pb après),

[b] : 2 copies, l'une est identique à TK81-O, l'autre possède les 198 premières pb identiques à TK81-O et 506 pb uniques,

[c] : entraine 42 pb supplémentaires,

* : codon stop.

facteurs potentiels de stérilité mâle, nous avons trouvé 28 ORF spécifiques à TK81-MS, 21 à CMS-G et 11 à CMS-E. Quatre ORF sont partagées par CMS-G et TK81-MS, trois par CMS-E et TK81-MS, deux par CMS-E et CMS-G et trois sont communes aux trois CMS.
En référence à [Satoh et al., 2004] où TK81-O et TK81-MS avaient été comparés, l'*orf317* avait été trouvé chez TK81-O uniquement. Nous retrouvons cette ORF dans les cinq autres génomes. L'*orf518* unique à TK81-O a été retrouvée chez A et macrocarpa mais est absente dans les autres génomes, par contre l'ORF correspondante *orf496*, unique à TK81-MS, est retrouvé chez B, CMS-E et CMS-G. L'*orf129b* décrite comme unique à TK81-O n'est effectivement pas retrouvée dans les autres génomes, par contre, son homologue chez TK81-MS (*orf122b*) est retrouvée dans les cinq autres génomes. Les ORF *orf324* et *orf119c*, décrites comme uniques à TK81-MS, ne sont effectivement pas retrouvées dans les autres génomes.
Parmi les ORF spécifiques à CMS-E, nous retrouvons l'*orf129* décrite comme étant candidate dans la stérilité mâle de I-12CMS(3) [Yamamoto et al., 2005], ce qui ajoute un argument en faveur de la proximité, si ce n'est l'identité de ces deux cytoplasmes.
Au niveau des ORF spécifiques aux non-CMS ou absentes dans certaines CMS, il faut noter que, étant donné que les génomes des CMS-E et CMS-G ne sont pas complets, leur spécificité n'est pas garantie. Nous avons donc trouvé 27 ORF spécifiques à TK81-O, 10 dans tous les génomes sauf TK81-MS, 5 dans tous les génomes sauf CMS-G et TK81-MS, 3 spécifiques à TK81-O, A et macrocarpa, 9 dans tous les génomes sauf CMS-G, 1 spécifique à macrocarpa, 1 commune à A et macrocarpa et 1 spécifique à A (Tableau supplémentaire page 196).
Les ORF chimériques, trouvées parmi les sept génomes, sont résumées dans le Tableau 6.5. Seuls TK81-O et TK81-MS possèdent deux ORF chimériques qui leurs sont spécifiques.

ORF et fragments chimériques	TK81-O	TK81-MS	A	B	CMS-E	CMS-G	macro
orf100e : 30pb *atp8*	262102-262404						
orf105a : 59pb *matK*	19539-19856		8023-8340	117299-117616	99299-99616[a]		90547-90864[a]
orf105d : 27pb *rrn18*		57459-57776 249975-250292	287788-288105	32176-32493	86329-86646[b]	91046-91363[a]	152611-152928[a]
orf105e : 34pb *atp9*					140735-141052[a]	29317-29634[a]	
orf107a : 27pb,25pb *atp8*	72786-73109		319702-320025	64090-64413	54410-54733[b]	122961-123284[a]	184529-184852[a]
orf119a : 20pb *ycf2* plastidique	117425-117784 359394-359753	485856-486215	153118-153477 275014-275373	19402-19761 144555-144914	7506-7865[b]	65812-66171[a]	139831-140190[a]
orf162 : 20pb *ycf2* plastidique	157959-158448	228055-228544 388149-388638	87868-88357	322031-322520	195890-196379[a]	52895-53384[b]	64246-64735[a]
orf221[t] : 274pb cox2 exon1		213770-214435 402259-402924					
orf246[t] : 74bp *rsp3* exon 1 ;24bp,30bp *rps3* exon 2	272854-273594						
orf273 : 74bp *rps3* exon 1			188701-189522	183003-183824	70563-71384[a]	70539-71360[b]	67251-68072[b]
orf281 : 30pb *atp8*			177940-178785	172242-173087	81300-82145[a]	77988-78833[b]	
orf282 : 828pb *nad9*			284458-285306	28846-29694			
orf317[t] : 274bp *cox2* exon 1	170862-171815		100788-101741	334951-335904	182507-183460[a]	37990-38943[b]	77166-78119[a]

TAB. 6.5 – ORF chimériques observées dans les génomes mitochondriaux de *Beta*. Les valeurs des positions sont indiquées. [a] : se trouve dans le premier contig, [b] : se trouve dans le deuxième contig, [t] : ORF décrite comme transcrite [Satoh et al., 2004].

Séquences répétées

Les régions dupliquées dans chacun des génomes sont représentées dans la Figure 6.7. Au total, nous avons identifié 24 répétitions de taille supérieure à 500 pb (Tableau 6.6). Les deux plus grandes duplications (R87 et R51 de, respectivement, 87 kpb et 51 kpb) sont trouvées dans le plus grand génome, c'est-à-dire TK81-MS. TK81-O, A et B partagent une duplication de 23 kpb (R23) qui est un sous-fragment de R87. Toutes les autres duplications sont inférieures à 10 kpb et sont quasiment toutes uniques à un génome donné. Certaines duplications dans un génome peuvent avoir un fragment commun dans un autre génome. Par exemple, la triplication R6.0 spécifique de A et B est partiellement incluse dans la triplication R6.2 de TK81-O. TK81-MS possède également une triplication R7.6 dont R6.2 et R6.0 sont partiellement incluses. Ceci pourrait par exemple être la signature d'une triplication ancestrale réarrangée au cours de l'évolution dans les génomes mitochondriaux de *Beta*.

Analyse de l'évolution des génomes mitochondriaux de *Beta*

Afin d'étudier les relations phylogénétiques entre ces génomes, nous avons construit une concaténation des séquences communes entre les sept génomes (incluant les régions non codantes) à partir des séquences de complexités de génomes (c'est-à-dire sans les duplicats). L'alignement multiple de ces concaténations aboutit à une séquence consensus de 267,160 kpb (en incluant les indels) et, pour chaque génome, les tailles varient entre 266,554 kpb (TK81-O) et 266,763 kpb (CMS-G) (moyenne de 266,702 kpb et médiane de 266,694 kpb). A partir de cette concaténation nous avons pu déterminer le taux de substitutions entre paires de génomes pour 10 kpb (Tableau 6.7(a)) ainsi que le taux d'indels (insertion/délétion) pour des indels de moins de 16 pb (Tableau 6.7(b)) et le taux de mutations (Tableau 6.7(c)). Ces taux de substitution, indel et mutation correspondent respectivement au nombre de substitutions, indels, et substitutions et indels comptabilisé sur l'alignement de deux génomes et pour 10kpb. Le taux de substitutions varie de 1,087 substitution pour 10 kpb (entre A et B) à 34,075 substitutions pour 10 kpb (entre CMS-G et TK81-MS) et la valeur médiane est de 26,435. Au niveau du taux d'indels, les valeurs varient de 0,224 indels pour 10 kpb (entre A et B) à 25,660 indels pour 10 kbp (entre TK81-O et TK81-MS), la valeur médiane est de 15,785. Lorsque l'on compare les substitutions et les indels, le taux de mutations varie de 1,311 mutations pour 10 kpb (entre A et B) à 56,454 mutations pour 10 kpb (entre CMS-G et TK81-MS), la valeur médiane est de 44,620.
En utilisant la même méthode, nous avons calculé les matrices de substitutions, indels et mutations pour les huit génomes publiés de *Zea* [Allen et al., 2007, Darracq et al., 2010] (Tableaux 6.8(a), 6.8(b), 6.8(c)). Les taux de substitutions varient de 0,550 à 18,292 substitutions pour 10 kpb avec une valeur médiane de 5,807, cette valeur est cinq fois plus petite que la médiane du taux de substitutions dans les génomes mitochondriaux de *Beta*. Inversement, le taux d'indels varie de 2,514 à 164,876 indels pour 10 kpb avec une médiane de 28,460 qui est huit fois plus élevée que la médiane du taux d'indels chez *Beta*. Au niveau du taux de mutations, celui-ci varie chez les *Zea* de 3,064 à 182,203 mutations pour 10 kpb avec une médiane de 34,912 qui est du même ordre de grandeur que la médiane observée chez *Beta*.
Si l'on regarde ces valeurs à un niveau intraspécifique (six génomes de *Beta vulgaris* et cinq génomes de *Zea mays*), les taux de substitutions et d'indels sont fortement corrélés chez *Beta vulgaris* (r=0,989 et p=0,000) et le sont dans une moindre mesure chez *Zea mays* (r=0,774 et p=0,009). Nous pouvons noter que le taux moyen de substitutions est significativement plus élevé chez *Beta vulgaris* (18,9) que chez *Zea mays* (4,80) (T=4,18, p=0,001 et df=14) alors que les taux moyens d'indels (13,66 pour *Bv* et 18,78 pour *Zm*) et de mutations (23,57 pour *Bv* et 32,7

Répétitions	Genomes						
	Nv	Sv	A	B	CMS-E	CMS-G	macro
R87	-, R23, R*9.8*, *R6.2*	+, R7.6	-, R23, *R6.5*, *R6.0*	-, R23, *R8.3*, *R6.5*, *R6.6*, *R6.0*	-, R6.8	-, *R7.6*, *R7.2*, R1.44	-, R5.94, R4.62, R4.48, R3.10, *R1.37*
R51	-	+	-	-	-	-	-
R23	+, R9.8, R6.2	-, *R7.6*, *R4.85*	+, R6.5, R6.0	+, R6.5, R6.0	-, R6.8, R6.6, R0.60	-, *R7.6*	-, R5.94, R4.62, R4.48, R3.10
R9.8	+, R6.2	-, *R7.6*	-, *R6.5*, *R6.0*	-, *R6.5*, *R6.0*	-, *R6.8*, *R6.6*	-, *R7.6*, *R7.2*, R1.44	-, R4.62, R4.48, R3.10
R8.3	-, R0.64, R0.58	-	-, R0.64	+	-, R0.64	-	-, R0.73, R0.64
R7.6	-, *R6.2*	+, 3×	-, *R6.5*, *R6.0*	-, *R6.5*, *R6.0*	-, *R6.8*, *R6.6*	+, *R7.2*, R1.44	-, R4.62, R4.48, R3.10
R7.2	-	-	-	-	-, *R6.8*, *R6.6*	+, R1.44	-
R6.8	-, *R6.2*	-	-, *R6.5*, *R6.0*	-, *R6.5*, R6.0 *R6.0*	+	-, R1.44	-, *R4.62*
R6.7	-	-	-	-	+	-	-
R6.6	-, *R6.2*	-	-, *R6.5*, *R6.0*	-, *R6.5*, *R6.0*	+	-	-, *R4.48*
R6.5	-, *R6.2*	-	+, R6.0	+, R6.0	-	-	-, R4.62, R4.48, R3.10
R6.2	+, 3×	-	-, *R6.0*	-, *R6.0*	-	-	-, R4.62, R4.48, R3.10
R6.0	-	-	+, 3×	+, 3×	-	-	-, *R4.62*, *R4.48*, *R3.10*
R5.94	-	-	-	-	-, R0.60	-	+
R4.85	-	+	-	-	-	-	-
R4.62	-	-	-	-	-	-	+, R4.48, R3.10
R4.48	-	-	-	-	-	-	+, R3.10
R3.10	-	-	-	-	-	-	+, 3×
R1.44	-	-	-	-	-	+, 3×	-
R1.37	-	-	-	-	-	-	+
R1.26	-	-	-	-	-	-	+
R0.75	-	-	-	-	-	+	-
R0.73	-, R0.58	-	-	-	-	-	+
R0.64	+	-	+	-	+	-	+

TAB. 6.6 – Répétitions de taille supérieure à 500 pb. Le nom des répétitions correspond à leur taille, par exemple, R87 fait 87 kpb. Pour chaque répétition de chaque génome, les répétitions sont indiquées présentes (+) ou non dans les génomes (-) et si d'autres répétitions sont incluses, entièrement ou partiellement (en italique), dans la répétition concernée. 3× indique que la répétition est présente en trois copies.

Génomes	TK81-MS	A	B	CMS-E	CMS-G	macro
TK81-O	28,812	4,724	4,837	5,662	31,301	4,911
TK81-MS	-	26,734	26,547	26,435	34,075	26,471
A	-	-	1,087	3,337	28,759	2,587
B	-	-	-	3,224	28,834	2,474
CMS-E	-	-	-	-	29,248	3,374
CMS-G	-	-	-	-	-	28,827

(a) Substitutions pour 10 kpb

Génomes	TK81-MS	A	B	CMS-E	CMS-G	macro
TK81-O	25,660	7,874	7,724	8,249	22,904	7,949
TK81-MS	-	19,948	19,797	19,873	22,379	20,022
A	-	-	0,224	1,199	15,860	0,824
B	-	-	-	1,049	15,785	0,674
CMS-E	-	-	-	-	16,311	1,199
CMS-G	-	-	-	-	-	16,006

(b) Indels pour 10 kpb

Génomes	TK81-MS	A	B	CMS-E	CMS-G	macro
TK81-O	54,473	12,598	12,561	13,911	54,205	12,860
TK81-MS	-	46,683	46,345	46,309	56,454	46,494
A	-	-	1,311	4,537	44,620	3,412
B	-	-	-	4,274	44,620	3,1496
CMS-E	-	-	-	-	45,559	4,574
CMS-G	-	-	-	-	-	44,833

(c) Mutations pour 10 kpb

TAB. 6.7 – Substitutions, indels et mutations pour 10 kpb entre paires de génomes chez *Beta*.

pour *Zm*) ne sont pas significativement différents entre ces espèces (respectivement T=-1,53, p=0,142, df=20 et T=-1,49, p=0,153, df=20).

A partir des séquences concaténées, nous avons construit un arbre phylogénétique avec macrocarpa comme groupe externe (Figure 6.8). Les valeurs obtenues par Neighbor-Joining (valeurs supérieures) et par maximum de vraisemblance (valeurs inférieures) supportent bien cet arbre. L'arbre obtenu est composé de deux clades, l'un formé par TK81-O, A et B constituant le clade des non-CMS, l'autre est constitué de CMS-E, CMS-G et TK81-MS constituant le clade des CMS. Nous pouvons noter que les branches de CMS-G et TK81-MS sont beaucoup plus longues que celles des autres génomes. Nous avons alors regardé s'il existe une dynamique évolutive différente entre les CMS et les non CMS chez *Beta vulgaris*. Nous avons pour cela évalué la moyenne des diversités nucléotidiques synonymes et non synonymes (π_s et π_a) au sein des CMS (CMS-E, CMS-G et TK81-MS) ou des non-CMS (A, B et TK81-O), ainsi que leurs divergences nucléotidiques synonymes et non synonymes (K_s et K_a) par rapport à *Beta macrocarpa* (Tableau 6.9). Ces divergences et diversités ont été analysées par rapport à la concaténation des 29 gènes codant pour des protéines (taille totale de 29344 pb).

Lorsque l'on considère la moyenne des K_s des CMS et non-CMS avec macrocarpa, nous trouvons une moyenne 6,6 fois supérieure dans les CMS par rapport aux non-CMS. Le même phénomène est observé au niveau de la diversité synonyme (π_s) : les CMS ont une diversité cinq fois plus élevée par rapport aux non-CMS, suggérant un taux de substitution synonyme plus élevés dans la lignée CMS.

La divergence non synonyme des gènes mitochondriaux avec macrocarpa est en moyenne six fois plus élevée dans les CMS par rapport aux non-CMS, cependant, le ratio K_a/K_s est similaire. Egalement, les diversités nucléotidiques non synonymes sont environ cinq fois supérieures dans les CMS par rapport aux non-CMS, aboutissant à un ration π_a/π_s similaire entre CMS et non-CMS. Par rapport aux CMS, CMS-G a un K_a élevé conduisant à un K_a/K_s supérieur à 1 suggérant

Génomes	NB	CMS-C	CMS-S	CMS-T	parvi	lux	per
NA	2,490	2,481	4,884	4,621	0,550	15,539	14,055
NB	-	3,712	5,915	5,699	2,514	16,764	15,112
CMS-C	-	-	5,652	5,341	2,443	16,526	15,017
CMS-S	-	-	-	7,206	4,980	18,292	16,568
CMS-T	-	-	-	-	4,645	17,002	15,517
parvi	-	-	-	-	-	15,564	14,056
lux	-	-	-	-	-	-	4,888

(a) Substitutions pour 10 kpb

Génomes	NB	CMS-C	CMS-S	CMS-T	parvi	lux	per
NA	5,866	12,211	13,169	24,638	2,514	160,760	107,125
NB	-	14,058	14,776	26,440	6,370	162,017	108,131
CMS-C	-	-	19,736	26,418	12,382	159,949	110,943
CMS-S	-	-	-	30,479	14,102	163,911	109,753
CMS-T	-	-	-	-	24,761	164,876	114,108
parvi	-	-	-	-	-	161,228	107,158
lux	-	-	-	-	-	-	114,122

(b) Indels pour 10 kpb

Génomes	NB	CMS-C	CMS-S	CMS-T	parvi	lux	per
NA	8,356	14,629	18,053	29,259	3,064	176,300	121,180
NB	-	17,770	20,692	32,140	8,885	178,782	123,243
CMS-C	-	-	25,388	31,759	14,826	176,475	125,961
CMS-S	-	-	-	37,685	19,082	182,203	126,321
CMS-T	-	-	-	-	29,407	181,879	129,626
parvi	-	-	-	-	-	176,793	121,214
lux	-	-	-	-	-	-	119,011

(c) Mutations pour 10 kpb

TAB. 6.8 – Substitutions, indels et mutations pour 10 kpb entre paires de génomes chez *Zea*.

Gènes (29344 pb)	π_s	π_a	π_a/π_s	K_s	K_a	K_a/K_s
CMS	0,00124	0,00112	0,916	0,00073	0,00069	0,944
E				0,00063	0,00035	0,551
G				0,00067	0,00095	1,416
TK81-MS				0,00174	0,0007	0,401
non CMS	0,00021	0,0002	0,945	0,00011	0,00012	1,102
A				0	0,00005	-
B				0	0,00005	-
TK81-O				0,00032	0,00025	0,787

TAB. 6.9 – Diversités et divergences nucléotidiques synonymes et non synonymes des CMS et non-CMS par rapport à *Beta macrocarpa*. Analyses effectuées sur les alignements de la concaténation des 29 gènes mitochondriaux communs aux *Beta*.

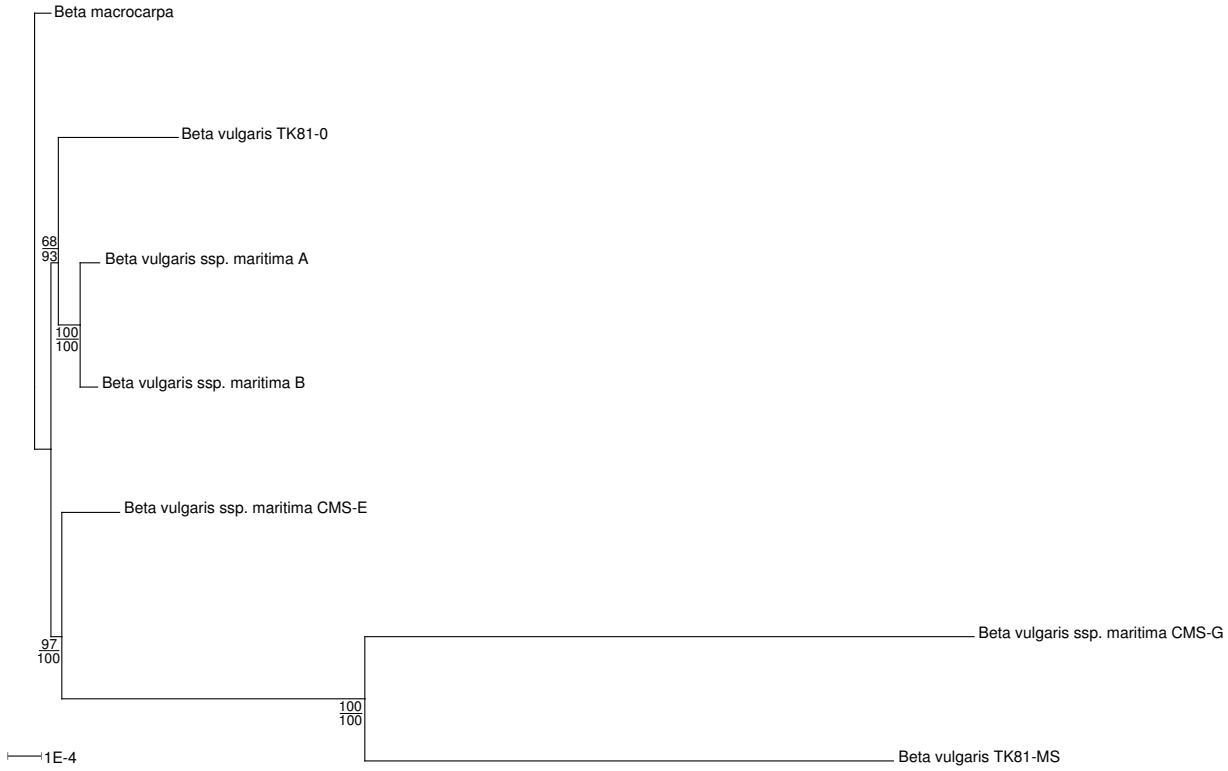


FIG. 6.8 – Arbre de Neighbor-Joining (réalisé avec BIONJ) sur les séquences communes au génomes mitochondriaux de *Beta* (267160 pb), enraciné avec *Beta macrocarpa*. Valeurs de bootstrap : valeurs du haut : BIONJ, bas : valeurs de maximum de vraisemblance obtenues avec TREE-PUZZLE. L'hypothèse d'horloge moléculaire est rejetée (TREE-PUZZLE).

que ce génome est soumis à une sélection positive.

6.3.3 Discussion et conclusion

Nous avons donc présenté ici les résultats issus du séquençage, quasiment complet, de cinq génomes mitochondriaux du genre *Beta*, quatre provenant de *Beta vulgaris* et un d'une espèce sœur, *Beta macrocarpa*. En ajoutant à ces génomes deux génomes déjà séquencés, nous avons obtenu un jeu de données sur lequel il nous a été possible de comparer la diversité à un niveau intraspécifique. La stérilité mâle est un facteur important dans le système de reproduction de *Beta vulgaris* (appelé gynodioécie) et est suspecté d'affecter le contenu en gènes et ORF mitochondriaux à travers l'émergence et la sélection de gènes et ORF spécifiques des CMS [Charlesworth, 2002, Touzet and Delph, 2009]. Nous avons donc comparé les deux groupes de génomes mitochondriaux séquencés (CMS et non-CMS) par rapport à notre groupe externe (*Beta macrocarpa*) et montré qu'ils possèdent une différence au niveau des diversités et divergences nucléotidiques. Nous avons également comparé les génomes mitochondriaux des dicotylédones *Beta vulgaris* et des monocotylédones *Zea mays* et observé des différences d'évolution de ces génomes mitochondriaux à un niveau intraspécifique.

CMS chez *Beta vulgaris*

Un aspect de cette étude a été de pouvoir identifier les CMS que nous avons séquencées par rapport à celles caractérisées dans différentes études. En effet, il n'existe pas de nomenclature générale pour les génomes mitochondriaux de *Beta vulgaris*. De ce fait, il semblerait que la CMS que nous appelons CMS-E dans cette étude, en suivant la nomenclature de [Cuguen et al., 1994], corresponde à la CMS appelée I-12CMS(3) par l'équipe d'Hokkaido [Onodera et al., 1999, Yamamoto et al., 2008]. En effet, CMS-E possède une séquence au début du gène *atp6* qui lui est spécifique ainsi que l'*orf129*, qui sont deux caractéristiques propres à I-12CMS(3). La CMS-E est la source la plus fréquente de CMS dans les populations sauvages de betterave le long des côtes Européennes [Dufaÿ et al., 2009] et semble avoir une échelle géographique plus importante puisque I-12CMS(3) (vraisemblablement CMS-E) est trouvée dans des populations sauvages au Pakistan. On peut alors se demander combien de CMS existent dans les populations sauvages de betterave, et d'après ce que nous savons, il en existerait quatre (CMS-E/I-12CMS(3), CMS-G, Owen/Sv/TK81-MS et CMS-H) parmi 20 génomes mitochondriaux recensés [Desplanque et al., 2000], les autres étant des génomes fertiles tels que A, B et TK81-O.

Les différentes études sur les CMS ont tenté de trouver des candidats à la stérilité mâle.
Tout d'abord, les études sur les CMS Owen/Sv/TK81-MS (utilisées dans la culture de betteraves sucrières) ont montré que sur les quatre ORF spécifiques à TK81-MS, lors de sa comparaison avec TK81-O, seule une ORF contenant *apt6* et sa pré-séquence spécifique (*preSatp6*) est exprimée [Yamamoto et al., 2005]. Cependant, les effets de la restauration nucléaire n'ont pas été montrés et aucune transformation expérimentale n'a validé les effets stérilisants de *preSatp6* [Yamamoto et al., 2008].
En ce qui concerne les CMS-E/I-12CMS(3), il a été démontré que l'*orf129* qui leur est spécifique est bien transcrite et code pour un polypeptide spécifique de 12 kDa qui serait accumulé dans les mitochondries de fleurs, racines et feuilles [Yamamoto et al., 2008]. Une expression transgénique de l'*orf129* dans le tabac a mené à l'obtention de plants stériles, démontrant ainsi l'effet stérilisant de l'*orf129*. L'action des restaurateurs n'est pas encore connue et aucune accumulation de l'*orf129*

n'a été détectée dans les plants restaurés.
Au niveau de CMS-G, une première étude a montré que la séquence du gène *cox2* était modifiée, aboutissant à une protéine tronquée au niveau C-terminal [Ducos et al., 2001]. De plus, il a été montré que l'activité *in vivo* de la cytochrome *c* oxydase est diminuée de 50% dans les feuilles, laissant supposer un effet de cette mutation sur l'activité du complexe. De plus, nous n'avons pas retrouvé le complexe IV par *blue native PAGE* chez CMS-G, soulevant la question de la stabilité et des propriétés chimiques de ce complexe dans les CMS-G. Dans cette étude, nous avons bien retrouvé le gène *cox2* modifié dans le génome mitochondrial de CMS-G et nous avons également détecté des mutations au niveau des autres gènes impliqués dans le complexe IV, c'est-à-dire *cox1* et *cox3*. Au niveau de *cox1*, nous avons trouvé une mutation du codon start pouvant entraîner la traduction d'une protéine plus longue du côté N-terminal (660 aa par rapport aux 524 dans les autres génomes ; poids estimé de 73,5 kDa par rapport au poids de 57,5 kDa dans les autres génomes) ou plus courte (495 aa ; poids estimé de 54,2 kDa). Nous pouvons noter qu'une étude récente sur ce même génome confirme qu'il n'existe qu'une copie du gène *cox1* chez CMS-G [Kawanishi et al., 2010]. Nous n'avons pas détecté de variants de *cox1* plus long par SDS-PAGE suggérant que la forme de *cox1* dans CMS-G est la forme la plus courte dont la variation de taille avec la forme normale est insuffisante pour pouvoir être détectée avec cette technique. Cependant, dans la forme tronquée en N-terminal, certains acides aminés de cette région tronquée sont impliqués dans les liaisons entre les sous-unités I/III (S_{10}), I/VIIc (A_{25}) et dans le transfert de protons (L_{19}). De plus, dans la partie non tronquée de la séquence, nous avons trouvé deux mutations non synonymes : R_{180}/Q et F_{403}/V, cependant ces codons ne sont pas impliqués dans une fonction connue. Au niveau de *cox3*, nous avons détecté une mutation non synonyme I_{51}/L qui n'est pas non plus associée à une fonction connue. Les polymorphismes observés sur les gènes du complexe IV peuvent être le résultat d'un relâchement de la sélection, permettant alors l'accumulation de mutations non-synonymes suite à la perturbation de l'activité du complexe IV. Dans ce cas, on pourrait alors avoir des mutations compensatoires au niveau des gènes nucléaires impliqués dans le complexe IV, posant la question des interactions noyau/mitochondries et d'une éventuelle coévolution entre les gènes impliqués dans un même complexe de la voie respiratoire.

Phylogénie des génomes mitochondriaux

Une première étude phylogénétique basée sur des séquences chloroplastiques effectuée au GEPV [Fénart et al., 2006] laissait supposer que les cytoplasmes mâles stériles avaient émergé indépendamment à partir d'un cytoplasme fertile. Cependant, la résolution de cette phylogénie était très faible, due au manque de polymorphisme entre les séquences. De plus, nous n'avons pas réussi à mieux résoudre cette phylogénie malgré un effort important de séquençage (Section 6.1). Dans cette étude, nous avons utilisé une grande portion de séquences mitochondriales (séquence consensus de 267,160 kpb) afin de réaliser un arbre phylogénétique en prenant *Beta macrocarpa* comme groupe externe. L'arbre obtenu va dans le sens de la phylogénie chloroplastique mais est mieux résolu. Nous avons ainsi déterminé deux clades sur cet arbre, l'un composé des trois CMS et l'autre composé des trois non-CMS. Nous avons alors regardé s'il existait des différences entre ces deux clades en termes de diversité nucléotidiques. Les CMS semblent avoir de plus fortes diversité et divergence nucléotidiques par rapport à *Beta macrocarpa*, suggérant un taux de mutation μ plus fort dans ce clade.

Comparaison de *Beta* et *Zea* : hypothèse de pression de mutation

Lorsque l'on compare les deux jeux de données intraspécifiques, l'un chez une espèce monocotylédone (*Zea mays*), l'autre chez une espèce dicotylédone (*Beta vulgaris*), différentes caractéristiques peuvent être observées. Dans les deux espèces, la principale source de variabilité de la taille des génomes est due aux séquences répétées. De plus, dans ces deux espèces, comme cela est souvent observé, l'ordre des gènes (ou de marqueurs) entre les individus d'une même espèce est différent. Malgré ces similitudes, nous pouvons noter que, bien que le contenu en gènes entre ces espèces soit quasiment identique, il existe une forte variabilité entre les tailles des génomes. En effet, la médiane de la taille des génomes chez les *Zea* est de 569,992 kpb alors que celle chez les *Beta* est de 341,115 kpb. On peut alors se demander quelles forces évolutives ont conduit à cette forte variabilité de taille entre ces deux espèces. L'hypothèse de pression de mutation consiste à considérer que le faible taux de mutation μ trouvé dans les génomes mitochondriaux des végétaux facilite l'accumulation de séquences non codantes et, de ce fait, participe à une augmentation de la taille des génomes mitochondriaux lorsqu'ils sont comparés aux génomes mitochondriaux des animaux [Lynch et al., 2006]. Il est intéressant de noter que le taux de substitutions moyen au sein de *Beta vulgaris* est 4 fois plus élevé que chez *Zea mays*. La différence de taille entre les génomes mitochondriaux de *Beta* (plus petits) et de *Zea* (plus grands) pourrait donc être la signature d'un taux de mutation μ différent entre ces deux espèces. Dans ce cas, ces observations pourraient être congruentes avec l'hypothèse de pression de mutation, prédisant une corrélation négative entre le taux de mutation μ et la taille des génomes, sous l'hypothèse d'une taille efficace des population égale.

Nous avons donc, dans cette partie, effectué l'analyse du contenu en gènes des génomes mitochondriaux séquencés. Nous avons également pu réaliser un arbre phylogénétique sur les séquences mitochondriales obtenues et observé que les branches des CMS étaient plus longues. Nous avons, de ce fait commencé une analyse des diversité et divergences des CMS et non-CMS par rapport à *Beta macrocarpa*, ces analyses ont montré des différences entre ces deux groupes. Nous avons également commencé à explorer la piste des causes de la stérilité mâle chez CMS-G et avons trouvé comme candidats les gènes impliqués dans le complexe IV. Enfin, nous avons observé des différences de forces évolutives entre les génomes *Beta* et *Zea* qui seraient en faveur de l'hypothèse de pression de mutation.

Dans cette étude, nous avons mis l'emphase sur les génomes au niveau de leurs séquences, il devient alors intéressant de poursuivre avec leur analyse au niveau des structures et réarrangements.

6.4 Analyse des réarrangements

Dans cette partie, nous allons essayer d'établir une histoire évolutive basée sur les réarrangements, comme nous l'avions fait pour le maïs. Ici deux difficultés supplémentaires rendent l'analyse plus complexe : la présence de marqueurs tripliqués, et l'incertitude sur l'ordre des marqueurs pour CMS-G, CMS-E et macrocarpa. De plus, ce jeu de données de génomes de *Beta* va nous permettre de tester la méthode de détection de duplications que nous avons développé précédemment.

6.4.1 Méthodes

L'étude des réarrangements passe par la construction des suites de blocs correspondant aux génomes. Comme pour le cas du maïs, nous nous sommes intéressés aux blocs communs (peu importe le nombre de blocs paralogues) entre tous les génomes. Contrairement aux analyses sur le maïs, nous avons ici simplifié la méthode de détection des blocs communs. Au lieu de commencer par considérer les gènes puis les fragments non codants, qui ne chevauchaient pas de gènes, en comparant l'ensemble des génomes, nous avons directement détecté les blocs communs sur les séquences entières de génomes. L'alignement des génomes a été effectué avec Mauve. Le problème de ce logiciel est qu'il ne prend pas en compte les régions dupliquées. Nous avons donc masqué, pour chaque génome, toutes les copies des dupliqués sauf une, afin de pouvoir utiliser Mauve qui nous renvoie l'ensemble des fragments de séquence communs entre les génomes. Les alignements des fragments ont ensuite été vérifiés afin de détecter et éliminer d'éventuels faux positifs. Chacun des fragments est ensuite recherché grâce à YASS sur les génomes complets (avec toutes les copies des duplicats) afin d'en trouver les homologues. Nous avons ainsi constitué le jeu de fragments communs pour chaque génome (aussi appelé backbone). Il est alors possible de transformer le backbone de chaque génome en permutations afin d'en analyser les réarrangements. Chaque élément des permutations est alors appelé marqueur. Un marqueur correspond à un fragment de la séquence nucléotidique. La méthode décrite est résumée en Figure 6.9.

La génération des permutations est ici plus complexe que pour le maïs. En effet, trois des génomes mitochondriaux (CMS-E, CMS-G et macrocarpa) sont encore en deux contigs qui n'ont pas pu être rattachés. Nous sommes donc partis de l'hypothèse la plus simple, c'est-à-dire réintégrer les contigs linéaires là où nous pensons qu'ils devraient être, comme nous l'avons décrit dans la Section 6.2 en attendant la finalisation de ces génomes.

Une fois les permutations construites, nous avons utilisé la méthode développée précédemment (voir Chapitre 5) afin de trouver les ensembles de blocs dupliqués communs ou non entre les différents génomes. Nous avons également utilisé la méthode de détection de groupes de marqueurs par rapport à leur voisinage afin de retrouver des associations entre les différents marqueurs pour identifier les orthologues et paralogues. Comme pour le maïs, les paralogues sont identifiés et condensés si nécessaire.

Lorsque nous avons déterminé les permutations des génomes avec les copies des duplicats identifiés, nous pouvons alors utiliser GRIMM pour obtenir une matrice de distances. Cette matrice de distances est ensuite analysée avec BioNJ afin d'obtenir un arbre de NJ. Comme pour le maïs, nous avons utilisé un Jackknife (de 90 à 10% par pas de 10%, 1000 fois) afin d'ajouter des valeurs de robustesse aux nœuds de notre arbre. Nous avons également utilisé MGR pour obtenir un arbre de parcimonie.

6.4.2 Résultats

Construction des backbones

Mauve a permis d'identifier 99 fragments communs entre les sept génomes. La taille de la séquence consensus de l'alignement des séquences des sept backbones est de 271577 pb. Le plus long fragment commun est de 36634 pb et le plus petit de 52 bp (moyenne 2733 pb, médiane 207 bp). Suite à l'étape de nettoyage, 36 fragments ont été éliminés et la séquence consensus est alors longue de 267160 pb. Les fragments font en moyenne 4169 pb (médiane=902 pb). Les fragments éliminés faisaient partie des plus petits fragments avec des tailles de 52 à 315 pb (moyenne=206 pb, médiane=89 pb). Les backbones nettoyés utilisés ici sont les mêmes que ceux utilisés lors

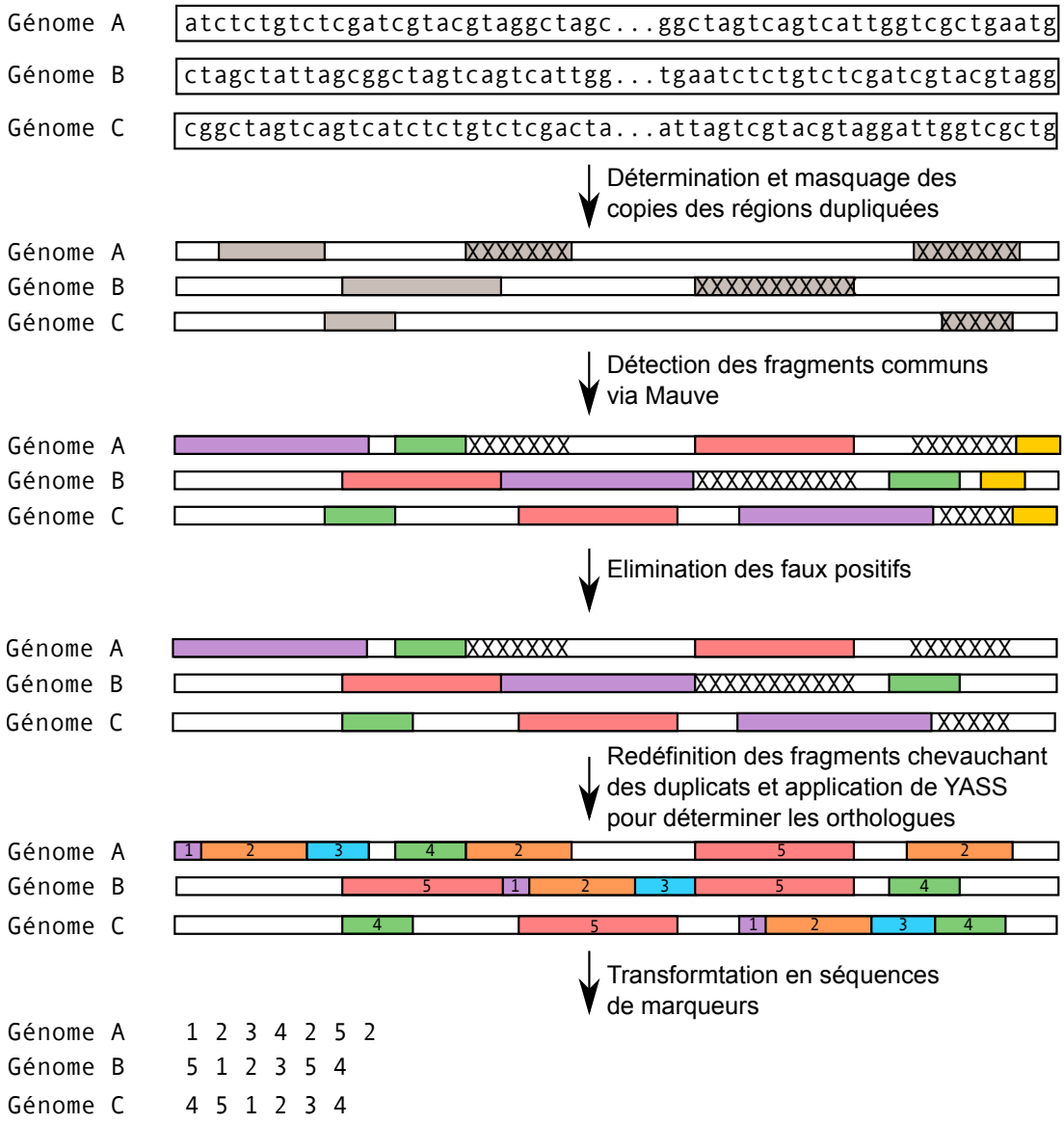


FIG. 6.9 – Méthode pour transformer les génomes en suites de marqueurs. Cinq étapes permettent d'obtenir une suite de marqueurs communs entre plusieurs génomes à partir de leurs séquences nucléotidiques. Les rectangles gris correspondent aux régions dupliquées dans chacun des génomes et les '×' symbolisent les régions masquées lors de l'analyse avec Mauve. Les autres rectangles colorés représentent les fragments orthologues et paralogues dans les génomes (une même couleur est attribuée aux fragments homologues).

Groupe de marqueurs dupliqués	Génomes possédant la duplication	Triplications dans certains génomes
{15,16,27,28,29,30,31}	tous les génomes	oui
{11,12,13}	tous les génomes	oui
{32,33,34,35}	tous les génomes sauf TK81-MS et CMS-G	non
{41,42,43}	tous les génomes sauf TK81-MS et CMS-G	non
{38,39,40}	CMS-G	non
{36,24,25,-26,-5,-4}	TK81-MS	non
{2,10,37}	TK81-MS	non
{-21}	TK81-MS	oui
{7}	TK81-MS	non
{19}	TK81-MS	non
{24}	macrocarpa	non
{-45}	CMS-E	non

Tab. 6.10 – Récapitulatif des duplications détectées dans les génomes mitochondriaux de *Beta*.

de la phylogénie basée sur les séquences mitochondriales (Figure 6.8). Parmi les 63 marqueurs restant dans chacun des backbones, nous avons condensé ceux qui étaient toujours en synténie entre tous les génomes. Nous arrivons donc à un jeu final de 46 marqueurs communs à tous les génomes. Une représentation des fragments le long des génomes est présentée en Figure 6.10.

Intégration des contigs

Comme nous l'avons précisé auparavant, les génomes CMS-E, CMS-G et macrocarpa sont formés de deux contigs. Afin de conserver un maximum de génomes, nous avons décidé de former les cercles maîtres avec les hypothèses de duplication provoquant la non intégration des contigs. Ainsi, nous avons intégré le deuxième contig (E2 dans la Figure 6.10) de CMS-E entre les blocs 28 et -14 du premier. Pour CMS-G, nous avons intégré le deuxième contig (G2 dans la Figure 6.10) entre les *marqueurs* -16 et -15 (deuxième copie) du premier contig. Pour macrocarpa, nous avons inséré le deuxième contig (marco2 dans la Figure 6.10) en sens reverse juste à la fin du premier contig (après le bloc 29).

Détection des duplicats

Nous avons utilisé la méthode développée au chapitre précédent afin de détecter les duplicats dans les génomes, communs ou non entre eux. Les duplicats trouvés sont reportés dans le Tableau 6.10.

Nous trouvons douze duplications pouvant contenir un ou plusieurs marqueurs. Quatre duplications sont communes à plusieurs génomes : deux communes à tous les génomes et deux communes à tous les génomes sauf TK81-MS et CMS-G. Ces quatre duplications comprennent plusieurs marqueurs. Cinq duplications sont propres à TK81-MS (contenant un à six marqueurs). En observant la position de ces duplicats, il est tout à fait possible qu'il s'agisse en fait d'une duplication en tandem remaniée. Nous trouvons également une duplication spécifique à CMS-G, une spécifique à CMS-E et une spécifique à macrocarpa. Aucune duplication n'a été détectée comme étant en tandem.

6.4.3 Identification des paralogues et condensation

Les résultats, obtenus sur les duplicats existant entre les génomes, vont nous permettre d'identifier et de condenser les paralogues. La méthode utilisée ici est la même que pour le maïs.

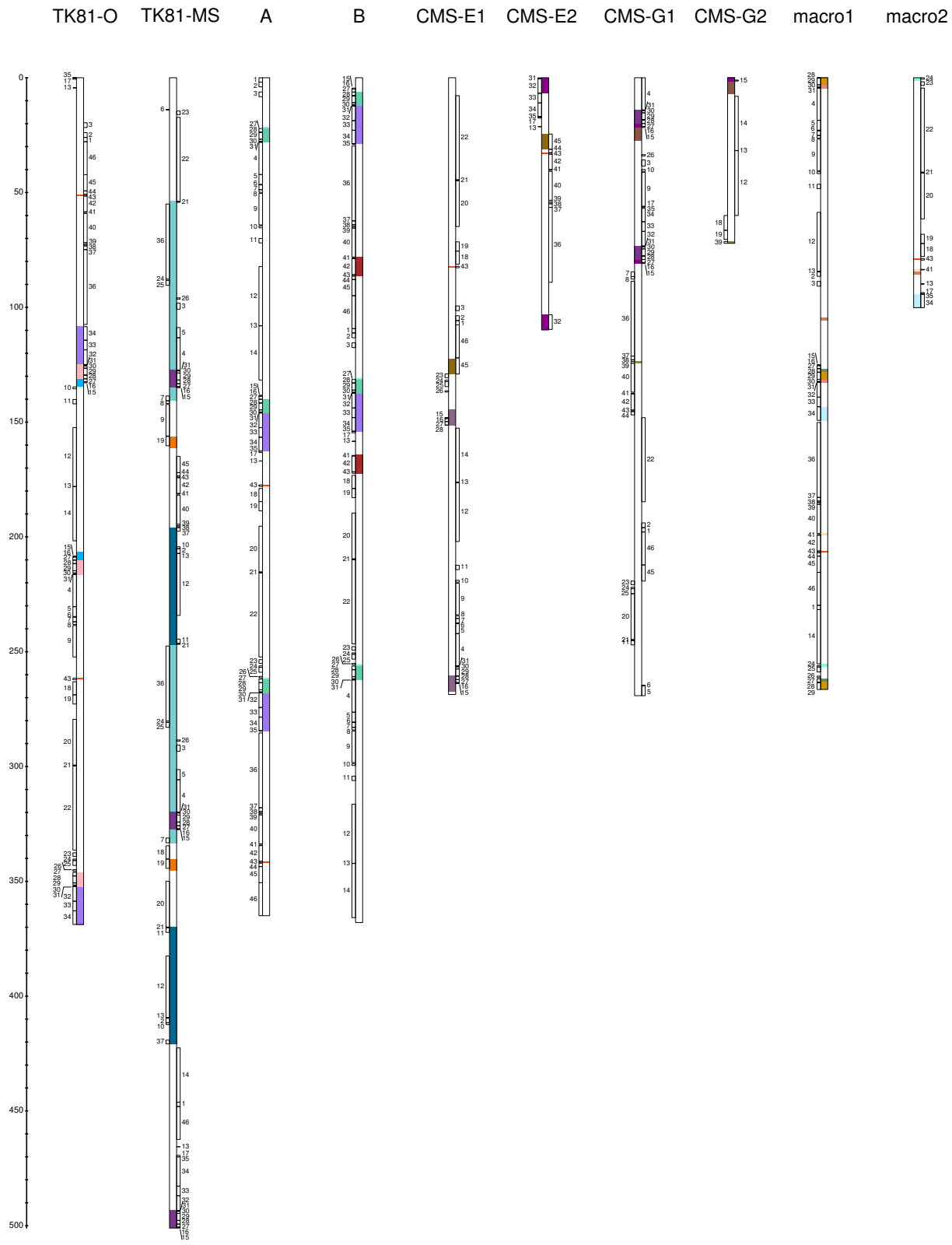


FIG. 6.10 – Répartition des fragments nettoyés, trouvés par Mauve, le long des génomes (les paralogues sont inclus). Les rectangles colorés représentent les duplications. Les régions dupliquées homologues, dans ou entre les génomes, ont une même couleur. Les marqueurs présents sur les génomes sont représentés par des rectangles, à gauche ou à droite des génomes en fonction de leur orientation. Les marqueurs homologues portent un même numéro.

Dans un premier temps, nous recherchons d'éventuelles duplications en tandem propres à chacun des génomes afin de les condenser et dans un deuxième temps, nous identifions les paralogues entre chaque génome.

Chez TK81-MS, six des dupliqués peuvent en fait s'expliquer par une grande duplication en tandem remaniée. Une seule inversion semble nécessaire pour retrouver cette duplication en tandem. En effet, une inversion entre les marqueurs dupliqués -21 et 21 est suffisante pour orienter et rassembler les deux parties de la duplication en tandem. Il devient alors simple de la condenser (Figure 6.11 A.). Cette duplication semble être la seule duplication en tandem existante dans ces génomes. Une fois cette duplication condensée, il reste, parmi tous les génomes, des duplications partagées et des duplications propres à un génome.

Duplications propres

Commençons par les duplications propres à chacun des génomes. Nous considérons ces duplications comme étant apparues après spéciation, ce qui nous permet d'ajouter un dupliqué virtuel, dans les autres génomes, à côté du marqueur existant. Chez TK81-MS, il reste deux copies du marqueur 21 (la troisième a été condensée). Une copie est à côté du marqueur 22, l'autre à côté du marqueur 2. Dans tous les autres génomes, 21 est entre 20 et 22. La copie de 21 proche de 22 sera alors appelée 21a et celle proche de 22 appelée 21b.

Chez macrocarpa, le marqueur 24 est dupliqué avec une copie à côté du marqueur 23, l'autre à côté du marqueur 25. Dans tous les autres génomes, 24 est toujours à côté de 25. La copie de 24 à côté de 25 chez macrocarpa sera nommée 24b et celle à côté de 23, 24a. Dans les autres génomes, on remplace 24 par 24a 24b avec 24b à côté de 25.

Chez CMS-E, nous avons une duplication du marqueur 45, avec une copie à côté du marqueur 44, l'autre à côté du marqueur 46. Comme dans les cas précédents, dans tous les autres génomes, 45 est proche de 44 ou de 46. Il suffit alors de remplacer 45 par 45a lorsqu'il est à côté de 44 et 45 par 45b lorsqu'il est à côté de 46. Dans les autres génomes, on renumérote 45 par 45a 45b.

Chez CMS-G, nous voyons une duplication du marqueur 39, avec une copie entre les marqueurs 38 et 40, l'autre entre les marqueurs 19 et 15. Dans tous les autres génomes, 39 est toujours entre 38 et 40. Dans ce cadre, il nous est impossible de resituer la deuxième copie de 39 dans les autres génomes. Nous avons donc éliminé le marqueur 39 pour la suite des analyses.

Parmi les quatre duplications propres à un génome, nous avons réussi à en identifier trois.

Duplications partagées

Au niveau des duplications communes à plusieurs génomes, nous avions détecté {11,12,13} dupliquée dans tous les génomes, {15,16,27,28,29,30,31} tripliquée dans tous les génomes ainsi que {32,33,34,35} et {41,42,43} dupliquées dans tous les génomes sauf TK81-MS et CMS-G. Concernant {11,12,13}, avant condensation on avait trois copies chez TK81-MS. Après condensation, il semblerait que la duplication de {11,12,13} ne soit en fait qu'une duplication de {13} puisque nous n'avons aucune duplication de 11 et 12 parmi tous les génomes. Chez CMS-G, la duplication de {11,12,13} a bien été détectée, cependant, aucun de ces marqueurs n'est dupliqué. Sachant que dans les autres génomes, une copie de 13 est toujours retrouvée à côté de 14 et l'autre à côté de 17, ce qui a été détecté chez CMS-G ne semble pas être une duplication de {11,12,13} mais juste son remaniement. Il est fort probable que chez CMS-G, la copie de 13 ait été perdue. Sachant que dans tous les génomes 13 est soit à côté de 14, soit à côté de 17 et que chez CMS-G seule la copie à côté de 14 est présente, nous avons virtuellement replacé la copie manquante à côté de 17. Ainsi, parmi tous les génomes, la copie à côté de 17 est renommée 13b et l'autre 13a.

La triplication de {15,16,27,28,29,30,31} est retrouvée dans tous les génomes. En fait, chez TK81-MS, après condensation il ne reste que deux copies. Si l'on observe tous les génomes, deux copies de {15,16,27,28,29,30,31} sont toujours à côté de {32,33,34,35}. Chez TK81-MS il ne reste qu'un exemplaire de ces deux groupes. Sachant que les copies sont toutes retrouvées dans les autres génomes, on peut penser à la perte d'une copie de ces deux groupes chez TK81-MS. Chez CMS-G, il semble manquer une copie de {32,33,34,35}. Cependant, les PCR (qui ont fonctionné mais non séquencées) semblent confirmer la présence de la deuxième copie en amont du marqueur {-15} dans G2. Pour les trois copies de {15,16,27,28,29,30,31}, celle qui se retrouvera à côté de 4 sera renommée D{15-31}a, celle à côté de 17 sera renommée D{15-31}b et celle à côté de 36 sera renommée D{15-31}c. Les marqueurs dupliqués {32,33,34,35} sont toujours retrouvés entre D{15-31}b et 17 et entre D{15-31}c et 36. Ainsi, ceux à côté de 17 seront renommés D{32-35}a et ceux à côté de 36 {32-35}b. Pour TK81-MS, les copies manquantes de {15,16,27,28,29,30,31} et {32,33,34,35} sont celles à côté de 36, nous avons donc ajouté D{15-31}c et D{32-35}b en amont de 36.

La duplication de {41,42,43} est retrouvée dans tous les génomes sauf TK81-MS et CMS-G. Dans tous les autres génomes, nous trouvons une copie entre 40 et 44 et l'autre à côté de 18. Chez CMS-G et TK81-MS nous n'avons que la copie entre 40 et 44. Il pourrait alors s'agir d'une perte de l'autre copie dans ces deux génomes (la duplication existant chez macrocarpa). Nous avons donc replacé la copie manquante dans ces deux génomes à côté de 18. Dans tous les génomes, la copie de {41,42,43} entre 40 et 44 est renommée D{41-43}a et celle à côté de 18 D{41-43}b.

Au final, parmi tous les marqueurs dupliqués, un seul n'a pas pu être identifié et a donc été éliminé. Les permutations sont maintenant constituées d'un jeu de 42 marqueurs où tous les paralogues ont été résolus. Ces jeux de permutations vont pouvoir être utilisés dans les outils classiques d'analyse de réarrangements tels que GRIMM et MGR.

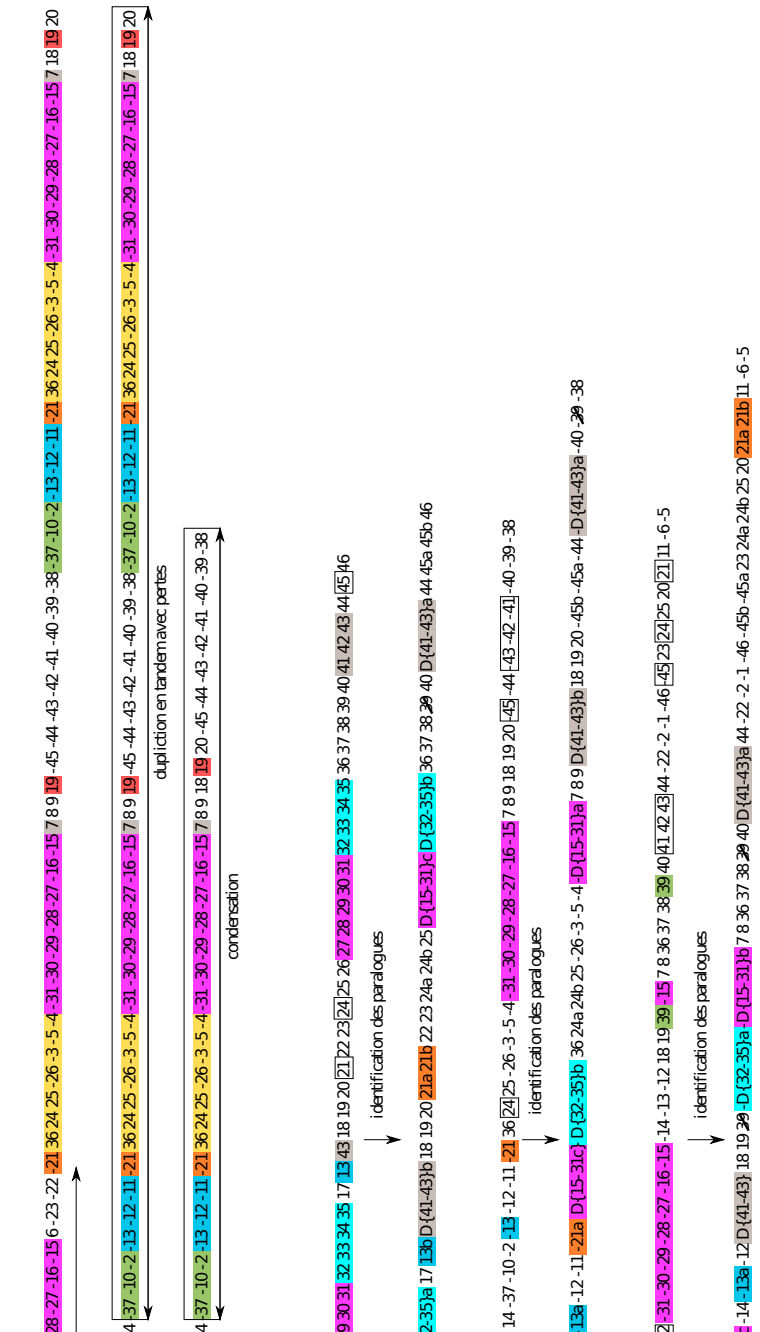


FIG. 6.11 – Identification et condensation des paralogues. A) Repérage et condensation de la duplication en tandem chez TK81-MS. B) Exemples d'identification de paralogues chez TK81-MS, A et CMS-G. Pour chaque génome, une même couleur représente les groupes de marqueurs homologues. Dans la partie B, les marqueurs encadrés sont des marqueurs non dupliqués dans le génome mais qui ont été virtuellement dupliqués car ils existaient en plusieurs copies dans un autre génome. Les marqueurs barrés sont les marqueurs dont les copies n'ont pas pu être distinguées et qui ont été éliminés pour le reste des analyses.

6.4.4 Analyse des réarrangements

Les arbres obtenus avec GRIMM et MGR sont montrés Figure 6.12. Les deux arbres obtenus ne présentent pas les mêmes topologies. En effet, si l'arbre obtenu avec GRIMM présente une topologie proche des arbres phylogénétiques réalisés sur les séquences mitochondriales (Figure 6.8) et chloroplastiques (Figure 6.3), l'arbre MGR est assez différent. Dans l'arbre MGR, TK81-O a une position basale, nous trouvons ensuite un râteau où A et CMS-E sont ensemble, TK81-MS et CMS-G sont ensemble et B est seul. Même si TK81-MS et CMS-G restent groupés, ils n'ont plus la position basale déterminée avec les séquences mitochondriales et chloroplastiques.

L'arbre obtenu avec GRIMM se rapproche un peu plus de ce que nous avions obtenu avec les séquences mitochondriales et chloroplastiques. Ici, seule la position de CMS-E diverge. Au lieu de retrouver CMS-E au niveau du groupe TK81-MS et CMS-G, on le retrouve avec B. Les nœuds semblent assez robustes. En effet, pour un Jackknife à 90%, nous trouvons une valeur de 100% au niveau de la séparation de TK81-MS et CMS-G des autres espèces, 94% au niveau de la séparation de TK81-O des autres espèces, 91% au niveau de la séparation de A des autres espèces et 78% au niveau de B et CMS-E. Cependant, les valeurs de Jackknife chutent assez vite lorsqu'on diminue le pourcentage de marqueurs conservés. Cette topologie d'arbre est supportée jusqu'à 60% (où les valeurs de robustesse deviennent très faibles).

Les phylogénies réalisées avec deux méthodes distinctes montrent donc des résultats différents. L'arbre obtenu par MGR (parcimonie) présente une topologie très différente de celle obtenue avec les séquences mitochondriales et chloroplastiques. Au contraire, l'arbre obtenu avec GRIMM (neighbor joining) présente des résultats plus proches. Seul le génome CMS-E a une position différente de ce que nous avions pu voir avant. Les valeurs de Jackknife obtenues supportent bien la topologie de l'arbre avec une valeur plus faible au niveau de B et CMS-E.

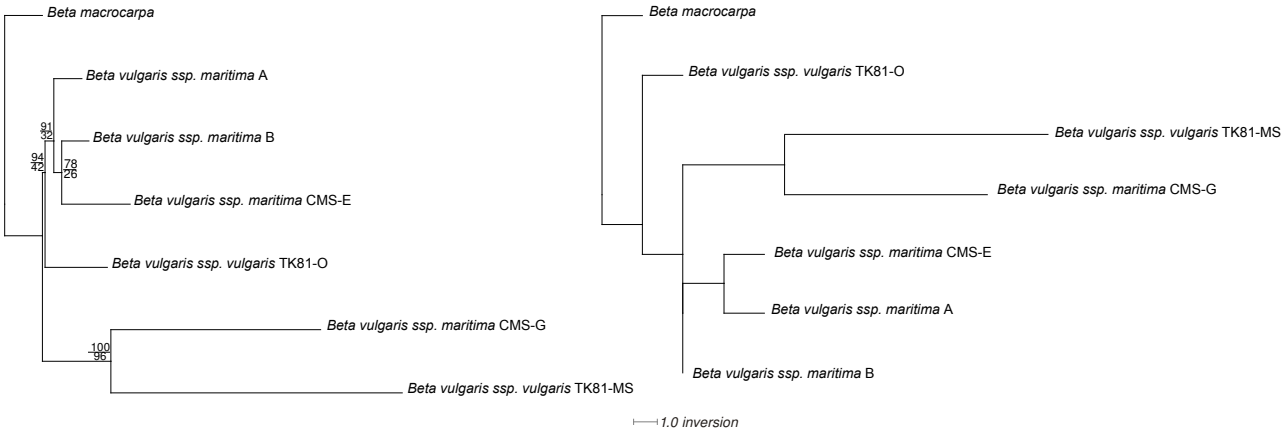


FIG. 6.12 – Phylogénie de réarrangements basée sur les distances d'inversion. A gauche arbre obtenu avec GRIMM, à droite arbre obtenu avec MGR. Les valeurs de nœuds indiquées correspondent aux Jackknife 90% (en haut) et 60% (en bas).

6.4.5 Évolution des génomes

Comme nous venons de le voir, l'arbre obtenu par MGR n'est pas concordant avec les autres arbres phylogénétiques. Nous ne pourrons donc ni utiliser cet arbre, ni les potentielles séquences

ancestrales qu'il retourne. Cependant, grâce aux duplicats obtenus avec notre méthode, nous pouvons toujours retracer ces événements le long d'un arbre phylogénétique. Étant donné que l'arbre GRIMM est basé sur la structure des génomes et que trois des génomes ont été "manuellement" reconstitués, nous allons nous appuyer ici sur l'arbre obtenu sur les séquences mitochondriales (lequel est mieux résolu que l'arbre chloroplastique).

Lors de l'analyse des duplicats, nous avons détécté des duplications communes à tous les génomes. Le groupe de marqueurs {15,16,27,28,29,30,31} est tripliqué dans tous les génomes. Comme nous l'avons expliqué, il y a vraisemblablement une perte d'une des copies chez TK81-MS. Chez G, pour une des copies, il ne reste que le marqueur {15} : les PCR obtenues laissent penser à la présence du groupe complet mais nous n'avons pas la séquence correspondante. Chez macrocarpa, il manque {15,16} dans une des copies indiquant une perte de ces marqueurs. Il manque également {15,16,27} dans une autre copie. Cette dernière est en bord de contig. Etant donné que ce contig n'est pas circularisé, nous ne pouvons pas affirmer qu'il s'agisse vraiment d'une perte. Chez CMS-E aussi nous pouvons constater des pertes dans les copies. Ce génome n'étant pas fini, nous ne pouvons pas affirmer qu'il s'agisse de pertes au cours de l'évolution. Chez TK81-O, il manque les marqueurs {15,16} dans une des copies, perdus au cours de l'évolution. Chez A et B {15,16} ont été perdus dans deux copies. Les marqueurs {15,16,27,28,29,30,31} auraient donc été tripliqués ancestralement avec des pertes au cours de l'évolution.
Le marqueur {13} est retrouvé dupliqué dans tous les génomes sauf chez CMS-G (faux positif dû à un remaniement) et TK81-MS (faux positif dû à une duplication en tandem plus récente). Le marqueur {13} aurait donc été dupliqué ancestralement et perdu au niveau de CMS-G et TK81-MS.
Le groupe de marqueurs {32,33,34,35} est également retrouvé dupliqué dans tous les génomes sauf CMS-G et TK81-MS. Chez CMS-E il ne reste que {32} dans une des copies. Une nouvelle fois, ce marqueur est en bord de contig, nous ne pouvons donc pas dire s'il s'agit d'une perte de {33,34,35} ou d'un manque dû au séquençage. Chez macrocarpa et TK81-O, une copie de {35} aurait été perdue. Les marqueurs {32,33,34,35} ont donc été ancestralement dupliqués avec perte de toutes les copies chez TK81-MS et CMS-G et perte de {35} chez macrocarpa et TK81-0.
Le groupe de marqueurs {41,42,43} est également retrouvé dupliqué dans tous les génomes sauf CMS-G et TK81-MS. On peut noter la perte d'une des copies {42} chez macrocarpa et d'une des copies de {41,42} chez A, CMS-E et TK81-O. B conserve toutes les copies. Ici aussi le groupe de marqueurs {41,42,43} semble avoir été dupliqué ancestralement.

Au niveau des dupliqués propres aux génomes, on a noté la duplication en tandem remaniée chez TK81-MS avec pertes (Figure 6.11 A.). Nous avons également noté la duplication de {45} chez E et {24} chez macrocarpa (rappelons que {39} chez CMS-G avait été éliminé).

L'évolution de ces duplications a été replacée sur la phylogénie des séquences mitochondriales (Figure 6.13).

6.5 Conclusion

Dans ce chapitre, nous avons procédé à l'analyse complète des génomes mitochondriaux de betterave. Nous sommes partis du séquençage des génomes mitochondriaux en passant par

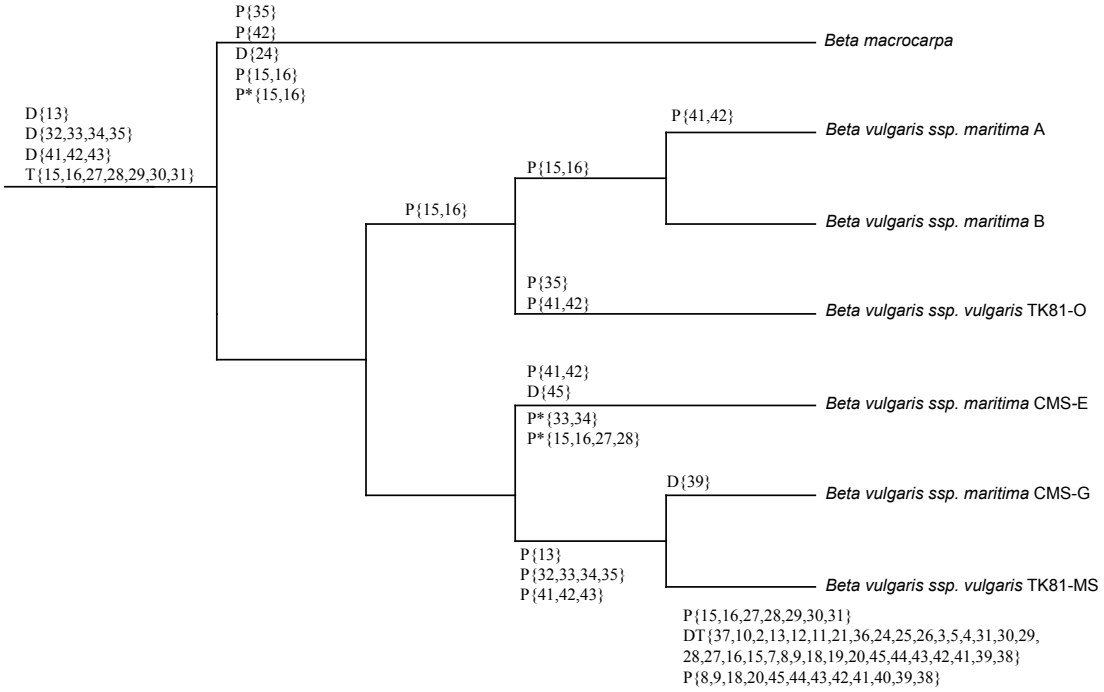


FIG. 6.13 – Evolution des génomes mitochondriaux de betterave, aperçu des événements de duplication. D=duplication, P=perte, DT=duplication en tandem, P*=perte potentielle non prouvée.

leur annotation et leur analyse pour enfin pouvoir étudier leur évolution. Afin de comparer les phylogénies obtenues sur les séquences mitochondriales et les réarrangements, nous avons séquencé des fragments chloroplastiques. Malheureusement, le polymorphisme chloroplastique chez cette espèce est très faible, de ce fait, les données chloroplastiques ne nous ont pas beaucoup aidés. Elle nous ont juste permis de confimer l'existence de la séparation de TK81-O, A, B des autres espèces, et que TK81-O est proche de Nv et TK81-MS de Sv. La phylogénie mitochondriale a donné un arbre mieux résolu qui reste en accord avec l'arbre chloroplastique. Cette meilleure résolution vient certainement du fait que nous avions plus de séquences pour faire l'analyse mitochondriale (environ 272 kpb et environ 18 kpb pour le chloroplaste).

L'analyse des réarrangements, basée sur les distances d'inversion, a donné une phylogénie assez concordante avec celles obtenues sur les séquences. Cependant, le génome CMS-E n'était pas placé au même endroit. Nous nous sommes retrouvés ici confrontés à un réel problème qui est celui des génomes non complets. Si le manque d'information n'a pas influencé l'analyse des contenus (tous les gènes étaient présents), cela a été plus gênant au niveau de l'analyse des réarrangements. En effet, pour trois des génomes, nous avons rassemblé les contigs en les intégrant à l'endroit qui semblait le plus parcimonieux. Si pour CMS-G et macrocarpa cette intégration semblait évidente, elle ne l'a pas autant été pour CMS-E. De plus, l'intégration a été faite en comparant les génomes à TK81-O, ce qui nous a peut être induit à rapprocher artificiellement, au niveau de leur structure, ces génomes de TK81-O. Le fait que trois des génomes ne soient pas complets pose aussi un problème au niveau de l'étude des duplications. En effet, si les génomes n'ont pas été circularisés, il est fort probable que ce soit à cause de régions dupliquées. De ce fait, nous ne pouvons pas affirmer que les pertes de marqueurs observées dans ces génomes

soient réellement des pertes. Cependant, la phylogénie obtenue est plutôt bonne et l'analyse des duplications (au niveau duplications et pertes) va dans le sens de la phylogénie des séquences mitochondriales obtenue.

Nous pouvons également constater que la méthode de détection de duplications développée fonctionne bien avec ces génomes, avec la capacité de détecter des événements de triplication. Cependant nous avons toujours le problème, sur les marqueurs simples dupliqués, de la détection de groupes de marqueurs (détection de plus de marqueurs qu'il n'en est). Ce problème vient vraisemblablement d'un phénomène d'attraction. En effet, lorsque nous comparons plusieurs génomes, un poids est mis sur les groupes de marqueurs. Chez TK81-MS, la duplication en tandem contenait des marqueurs dupliqués dans d'autres génomes, le poids de groupes de marqueurs devient alors plus fort que le poids d'un marqueur seul.

Conclusions et perspectives

Dans cette thèse, nous nous sommes intéressés à l'évolution au niveau de la structure, de génomes décrits comme très réarrangés pour lesquels aucune étude n'avait été réalisée du point de vue structural. Rappelons que dans les génomes des mitochondries végétales, le taux de substitution est faible et que la source de polymorphisme de ces génomes se fait essentiellement par le biais de réarrangements. L'étude, au niveau intraspécifique, de huit génomes mitochondriaux de *Zea* nous a permis d'observer des signes de duplication en tandem de larges régions dans ces génomes. La difficulté rencontrée ici concerne les outils d'analyse d'évolution de structure de génomes, qui ne tiennent pas compte des événements de duplication lorsque toutes les copies ne sont pas éliminées. Or, les génomes mitochondriaux de maïs sont en grande partie constitués de marqueurs dupliqués. En se basant sur cette observation et sur la description de duplications en tandem existantes mais très peu conservées chez les animaux, nous avons émis l'hypothèse que les duplications dans les génomes mitochondriaux de plantes proviennent d'événements de duplication en tandem. Contrairement à ce qui est observé au niveau animal, les duplications se sont pas intégralement éliminées, sans doute à cause du taux de substitution très faible dans les mitochondries végétales. En effet, selon l'hypothèse de la pression de mutation [Lynch et al., 2006] quand l'on compare des espèces dont la taille efficace de population est égale, il existe une corrélation négative entre le taux de substitution et la taille des génomes mitochondriaux : un génome ayant un faible taux de substitution accumulera plus facilement des séquences dupliquées ou d'origine extra-mitochondriale. A partir de l'hypothèse que les génomes mitochondriaux de plantes connaissent des événements de duplication en tandem au cours de leur évolution, nous avons établi un protocole permettant de distinguer la majeure partie des marqueurs dupliqués observés, nous permettant alors de les intégrer pour des analyses d'évolution en terme de réarrangements. Une histoire évolutive sur les réarrangements, incluant des marqueurs dupliqués, a pu être établie sur les génomes mitochondriaux de *Zea* (ainsi que la détermination de la structure des génomes ancestraux potentiels). Notre hypothèse ainsi que le protocole qui en découle, nous ont permis d'obtenir, en terme de réarrangements, une phylogénie identique à celle observée sur les séquences nucléotidiques. Sur cette hypothèse, nous avons également développé une méthode de détection d'éléments dupliqués, en tandem ou non, communs ou non à plusieurs génomes. Cette méthode a produit des résultats encourageants sur les analyses des génomes mitochondriaux de *Zea*. En effet, même si toutes les duplications détectées par la méthode n'étaient pas strictement identiques aux duplications que nous avions observées, nous avons pu détecter les marqueurs communs dupliqués entre deux ou plusieurs génomes et également des duplications en tandem propres à un génome. Nous sommes ainsi parvenus à mettre en place des outils permettant de réaliser, sur des génomes contenant des marqueurs dupliqués, des analyses évolutives en terme de réarrangements. Ces outils peuvent voir leur utilisation s'étendre au delà de l'analyse des génomes mitochondriaux de plantes, puisqu'ils permettent de traiter les problèmes d'événements de duplications (lors de l'analyse de l'évolution des structures de génomes) qui concernent les

génomes d'autres organismes et d'autres organites. Ces outils sont pour le moment au stade expérimental mais constituent une bonne base pour les approches d'études d'arrangements de génomes comportant des marqueurs dupliqués. Le but est maintenant de créer un processus complet, permettant de produire une phylogénie contenant des événements de duplication, à partir de séquences génomiques. Cependant, l'automatisation complète pose des questions une fois les duplications retrouvées grâce à la méthode proposée. Les duplications en tandem sont condensées mais il serait intéressant de définir ce qu'est la meilleure condensation possible et comment la calculer. Le point délicat reste encore la détection des relations de paralogie et d'orthologie pour les duplications qui ne sont plus en tandem, même si les travaux sur la recherche de matchings pourront être utilisés pour résoudre les cas difficiles. Cependant on peut imaginer la mise au point de méthodes visant non pas à identifier paralogues et orthologues mais à trouver un scénario qui place ces marqueurs côte-à-côte, afin de faire apparaître des duplications en tandem. On aboutit dans tous les cas à des génomes dépourvus de duplicats. A partir de ce jeu de données, il est alors possible d'utiliser un logiciel d'analyse d'événements de réarrangements pour établir une histoire évolutive de ces génomes et replacer les événements de duplication au niveau des branches. La procédure décrite ici nécessite donc essentiellement un travail au niveau de l'identification et de la condensation des marqueurs dupliqués à l'issue de la méthode de détection.

Cette procédure, pour le moment non chaînée, a également été appliquée sur sept génomes mitochondriaux de *Beta*. Au cours de cette thèse, nous avons, avec l'aide du Génoscope, réalisé le séquençage de quatre génomes de *Beta vulgaris* dont deux sont des génomes dits mâles fertiles et deux sont mâles stériles (appelés CMS). Nous avons également séquencé le génome de *Beta macrocarpa* dans le but d'avoir un groupe externe lors de l'analyse de l'évolution des génomes dans la section *Beta*. A ces génomes séquencés, nous avons ajouté deux génomes de *Beta vulgaris* qui avaient déjà été séquencés. En terme de reconnaissance de motifs dupliqués communs, la méthode que nous avons développée s'est avérée fonctionnelle. Effectivement, nous avons réussi à déterminer les différents événements de duplication au cours de l'évolution de ces génomes. La tâche ne fut pas aussi évidente au niveau de l'étude des réarrangements des génomes puisque trois d'entre eux n'ont pas été complètement reconstitués, rendant l'ordre des contigs incertain. Cependant, l'expertise manuelle que nous avons apportée au niveau de séquençage a permis la reconstitution totale d'un des génomes et de progresser sur celle des trois génomes fragmentés. Cette technique manuelle mérite d'être explorée afin d'en tirer un processus automatique d'aide à la reconstitution se basant sur l'hypothèse que, si les contigs ne peuvent pas être assemblés, il y a peut être un problème de duplication et donc de reconstitution des fragments. Ce logiciel pourrait alors aider à déterminer les régions critiques et proposer d'éventuelles PCR pour aider à la reconstitution des génomes qui peut s'avérer être problématique dans le cas de génomes contenant de grandes duplications. Cependant, les nouvelles technologies de séquençage visent à obtenir des fragments plus longs. Avec de longs fragments, l'assemblage de génomes contenant des fragments dupliqués peut alors devenir beaucoup plus simple, si les fragments dupliqués ont une taille inférieure à celle des fragments séquencés [Metzker, 2010, Nagarajan et al., 2010]. Nous avons également développé PLAMIDB, une base de donnée dédiée aux génomes mitochondriaux de plantes, qui s'est révélé être un très bon outil d'aide à la comparaison des génomes et d'annotation pour les génomes de *Beta*.

L'étude des génomes mitochondriaux de betterave nous a permis d'observer certaines spécificités de génomes mâle stériles. Rappelons que les facteurs de stérilité mâle mitochondriaux chez les plantes sont différents en fonction des organismes. Par exemple, chez le maïs,

la stérilité des trois organismes séquencés ne dépend pas des mêmes facteurs. Par rapport à notre étude sur les betteraves, pour la CMS-G, nous avons pu établir qu'en plus du gène *cox2* (composant du complexe IV), tronqué au niveau de l'extrémité 3', les autres gènes de ce complexe sont également modifiés (*cox1* étant tronqué et *cox3* ayant une substitution non synonyme). Il reste à savoir si la réduction de l'activité du complexe IV, observée dans une étude précédente, est due à ces mutations et si elle joue un rôle dans la stérilité de ce génome. En effet, ce complexe, intervenant dans la production d'ATP, peut avoir une influence étant donné que la production des gamètes nécessite de l'énergie. Si la plante produit moins d'énergie, il peut alors y avoir un effet négatif sur la production de pollen. Nous avons également trouvé des ORF spécifiques des CMS-E et CMS-G. Chez certaines plantes, il a été montré que des ORF interviennent dans la stérilité mâle [Chase, 2007]. Une des ORF spécifique de la CMS-E a été décrite récemment comme potentiellement intervenant dans la stérilité mâle de I12-CMS(3). Ces deux CMS correspondraient au même génome mais porteraient des noms différents. En effet, l'*atp6* dans ces deux génomes est identique (et différente de celle des autres génomes) et l'ORF intervenant dans la stérilité mâle de I12-CMS(3) est une ORF spécifique de cette CMS [Onodera et al., 1999, Yamamoto et al., 2008]. Plus généralement, il serait intéressant de poursuivre l'analyse de ces ORF candidates, déterminées comme spécifiques de chacune des CMS E et G, par des analyses d'expression afin d'établir lesquelles sont vraiment exprimées dans les mitochondries et pourraient donc être responsables de la stérilité mâle. De plus, les analyses sur le taux de substitutions ont montré, au niveau mitochondrial, un taux beaucoup plus élevé chez les CMS par rapport au non-CMS, alors que ce taux au niveau chloroplastique n'est pas différent. Il semble ici que les histoires évolutives entre les CMS et non-CMS soient différentes au niveau mitochondrial et qu'un fort taux de substitution permettrait l'émergence de la stérilité mâle. On peut également noter que la distance d'évolution des réarrangements semble aller dans le sens du taux de mutation μ. Même si les génomes de CMS ne sont pas complets, on peut observer que les distances d'inversion sont beaucoup plus élevées chez les CMS que les non-CMS. On peut alors penser que le taux d'évolution plus fort chez les CMS pourrait expliquer la stérilité mâle, en raison de mutations sur les gènes ou de réarrangements chromosomiques pouvant intervenir dans l'apparition de nouvelles ORF.

Lorsque nous comparons les évolutions des génomes de *Zea* et de *Beta*, nous pouvons constater que les forces évolutives semblent différentes. En effet, chez *Zea*, le taux de substitutions est beaucoup plus faible et le taux d'insertion/délétion (indel) beaucoup plus élevé par rapport à *Beta*. Bien que nous n'ayons pas l'intégralité des génomes de *Beta*, il semblerait que l'on trouve plus de motifs de duplication en tandem et de duplications spécifiques chez *Zea*, et plus de signes de duplications partagées chez *Beta*. On peut alors se demander si les duplications en tandem que nous avons détectées chez le maïs sont un phénomène évolutif propre à cette espèce ou s'il est retrouvé chez d'autres monocotylédones (par exemple le blé ou le riz). Il existe maintenant plusieurs espèces séquencées pour leurs génomes mitochondriaux. Ces génomes peuvent former deux jeux de données intéressants, constitués d'un côté de monocotylédones et de l'autre de dicotylédones. L'intérêt serait de comparer les évolutions des génomes mitochondriaux à un niveau interspécifique, pour chacun de ces groupes, afin d'établir s'il existe des différences évolutives entre monocotyledones et dicotylédones (en termes de duplication en tandem, taux de substitutions ou encore de taille des génomes). La question de l'évolution des génomes mitochondriaux de plantes commence à prendre de plus en plus d'intérêt. Nous avons pu le constater, les taux de substitution et d'indels ainsi que les événements de duplication sont différents entre le maïs et la betterave. De plus, une étude récente sur les Curcubitacés montre qu'il existe une différence de

taille et de taux de mutation entre *Citrullus lanatus* et *Curcubita pepo* [Alverson et al., 2010]. Les mécanismes d'évolution de ces génomes ne sont pas encore connus et comme nous l'avons vu, il est difficile d'en avoir une idée générale en se concentrant uniquement sur une espèce. Cependant, une analyse à un niveau interspécifique en multipliant les jeux de données pourrait aider à avoir une image plus générale des différents mécanismes mis en jeu en fonction des espèces. Nous avons montré qu'à un niveau intraspécifique, il existe une forte variabilité et une hétérogénéité dans l'évolution des génomes. Par conséquent les informations sur un seul génome d'une espèce ne sont pas forcément représentatives de cette espèce. De ce fait, pour construire une histoire évolutive à un niveau interspécifique, il sera impératif de disposer de plusieurs génomes pour une même espèce.

Bibliographie

[Aguilera and Gómez-González, 2008] Aguilera, A. and Gómez-González, B. (2008). Genome instability : a mechanistic view of its causes and consequences. *Nature Reviews Genetics*, 9 :204–217.

[Albert et al., 1996] Albert, B., Godelle, B., Atlan, A., Paepe, R. D., and Gouyon, P. H. (1996). Dynamics of Plant Mitochondrial Genome : Model of a Three-Level Selection Process. *Genetics*, 144 :369–382.

[Allen et al., 2007] Allen, J. O., Fauron, C. M., Minx, P., Roark, L., Oddiraju, S., Lin, G. N., Meyer, L., Sun, H., Kim, K., Wang, C., Du, F., Xu, D., Gibson, M., Cifrese, J., Clifton, S. W., and Newton, K. J. (2007). Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. *Genetics*, 177 :1173–1192.

[Alverson et al., 2010] Alverson, A., Wei, X., Rice, D., Stern, D., Barry, K., and Palmer, J. (2010). Insights into the Evolution of Mitochondrial Genome Size from Complete Sequences of Citrullus lanatus and Cucurbita pepo (Cucurbitaceae). *Molecular Biology and Evolution*, 27(6) :1436–1448.

[Amir et al., 2007] Amir, A., Gasieniec, L., and Shalom, R. (2007). Improved approximate common interval. *Information Processing Letters*, 103(4) :142–149.

[Backert et al., 1997] Backert, S., Nietsen, B. L., and BSmer, T. (1997). The mystery of the rings : structure and replication of mitochondrial genomes from higher plants. *Trends in Plant Science*, 2(12) :477–483.

[Bader et al., 2001] Bader, D., Moret, B., , and Yan, M. (2001). A linear-time algorithm for computing inversion distances between signed permutations with an experimental study. *Journal of Computational Biology*, 8(5) :483–491.

[Bader, 2009] Bader, M. (2009). Sorting by reversals, block interchanges, tandem duplications, and deletions. *BMC Bioinformatics*, 10(Suppl 1) :S9.

[Bader, 2010] Bader, M. (2010). Genome rearrangements with duplications. *BMC Bioinformatics*, 11(Suppl 1) :S27.

[Bafna and Pevzner, 1993] Bafna, V. and Pevzner, P. A. (1993). Genome rearrangements and sorting by reversals. In *Proceedings of the 34th Annual Symposium on Foundations of Computer Science, FOCS'93*, pages 148–157. IEEE.

[Bailey and Eichler, 2006] Bailey, J. and Eichler, E. (2006). Primate segmental duplications : crucibles of evolution, diversity and disease. *Nature Reviews Genetics*, 7 :552–564.

[Ballard and Dean, 2001] Ballard, J. W. O. and Dean, M. D. (2001). The mitochondrial genome : mutation, selection and recombination. *Current Opinion in Genetics and Development*, 11 :667–672.

[Ballard and Rand, 2005] Ballard, J. W. O. and Rand, D. M. (2005). The population biology of mitochondrial DNA and its phylogenetic implications. *Annual Review of Ecology, Evolution, and Systematics*, 36 :621–642.

[Barr et al., 2005] Barr, C. M., Neiman, M., and Taylor, D. R. (2005). Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. *New Phytologist*, 168(1) :39–50.

[Bendich, 1993] Bendich, A. J. (1993). Reaching for the ring : the study of mitochondrial genome structure. *Current Genetics*, 24 :279–290.

[Bergeron, 2005] Bergeron, A. (2005). A very elementary presentation of the hannenhalli-pevzner theory. *Discrete Applied Mathematics*, 146(2) :134–145.

[Bergeron et al., 2002] Bergeron, A., Corteel, S., and Raffinot, M. (2002). The algorithmic of gene teams. In Guigó, R. and Gusfield, D., editors, *Proceedings of the 2nd International Workshop on Algorithms in Bioinformatics, WABI'02*, volume 2452 of *Lecture Notes in Computer Science*, pages 464–476. Springer.

[Bergeron et al., 2006] Bergeron, A., Mixtacki, J., and Stoye, J. (2006). On sorting by translocations. *Journal of Computational Biology*, 13(2) :567–578.

[Berman et al., 2001] Berman, P., Hannenhalli, S., and Karpinski, M. (2001). 1.375-approximation algorithm for sorting by reversals. *Electronic Colloquium on Computational Complexity (ECCC)*, 8(47).

[Birky, 2001] Birky, C. J. (2001). The inheritance of genes in mitochondria and chloroplasts : Laws, Mechanisms, and Models. *Annual Review of Genetics*, 35 :125–148.

[Blanchard and Lynch, 2000] Blanchard, J.-L. and Lynch, M. (2000). Organellar genes why do they end up in the nucleus ? *Trends in Genetics*, 16(7) :315–320.

[Blanchette et al., 1996] Blanchette, M., Kunisawa, T., and Sankoff, D. (1996). Parametric genome rearrangement. *Gene*, 172(1) :GC11–GC17.

[Blanchette et al., 1999] Blanchette, M., Kunisawa, T., and Sankoff, D. (1999). Gene order breakpoint evidence in animal mitochondrial phylogeny. *Journal of Molecular Evolution*, 19(2) :193–203.

[Boore and Brown, 1998] Boore, J. and Brown, W. (1998). Big trees from little genomes : mitochondrial gene order as a phylogenetic tool. *Current Opinion in Genetics and Development*, 8(6) :668–674.

[Bourque and Pevzner, 2002] Bourque, G. and Pevzner, P. A. (2002). Genome-scale evolution : Reconstructing gene orders in the ancestral species. *Genome Research*, 12(1) :26–36.

[Bowe and dePamphilis, 1996] Bowe, L. M. and dePamphilis, C. W. (1996). Effects of RNA Editing and Gene Processing on Phylogenetic Reconstruction. *Molecular Biology and Evolution*, 13(9) :1159–1166.

[Brudno et al., 2003a] Brudno, M., Do, C., Cooper, G., Kim, M., E, D., Program, N. C. S., Green, E., Sidow, A., and Batzoglou, S. (2003a). LAGAN and Multi-LAGAN : efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research*, 13(4) :721–731.

[Brudno et al., 2003b] Brudno, M., Malde, S., Poliakov, A., Do, C., Couronne, O., Dubchak, I., and Batzoglou, S. (2003b). Glocal alignment : finding rearrangements during alignment . *Bioinformatics*, 19 :i54–i62.

[Burger et al., 1995] Burger, G., Plante, I., Lonergan, K. M., and Gray, M. W. (1995). The Mitochondrial DNA of the Amoeboid Protozoon, Acanthamoeba castellanii : Complete Sequence, Gene Content and Genome Organization. *Journal of Molecular Biology*, 245 :522–537.

[Burzynski et al., 2003] Burzynski, A., Zbawicka, M., Skibinski, D. O. F., and Wenne, R. (2003). Evidence for Recombination of mtDNA in the Marine Mussel Mytilus trossulus from the Baltic. *Molecular Biology and Evolution*, 20(3) :388–392.

[Campbell and Barker, 1999] Campbell, N. J. H. and Barker, S. C. (1999). The Novel Mitochondrial Gene Arrangement of the Cattle Tick, Boophilus microplus : Fivefold Tandem Repetition of a Coding Region. *Molecular Biology and Evolution*, 16(6) :732–740.

[Caprara, 1997] Caprara, A. (1997). Sorting by reversals is difficult. In *Proceedings of the First Annual International Conference on Research in Computational Molecular Biology, RECOMB'97*, pages 75–83. ACM Press.

[Caprara, 1999] Caprara, A. (1999). Formulations and hardness of multiple sorting by reversals. In *Proceedings of the 3rd annual international conference on Computational molecular biology, RECOMB'99*, pages 84–93. ACM Press.

[Carapelli et al., 2006] Carapelli, A., Vannini, L., Nardi, F., Boore, J. L., Beani, L., Dallai, R., and Frati, F. (2006). The mitochondrial genome of the entomophagous endoparasite xenos vesparum (insecta : Strepsiptera). *Gene*, 376(2) :248–259.

[Charlesworth, 2002] Charlesworth, D. (2002). What maintains male-sterility factors in plant populations. *Heredity*, 89 :408–409.

[Chase, 2007] Chase, C. D. (2007). Cytoplasmic male sterility : a window to the world of plant mitochondrial-nuclear interactions. *Trends in Genetics*, 23 :81–90.

[Chauve et al., 2006] Chauve, C., Diekmann, Y., Heber, S., Mixtacki, J., Rahmann, S., and Stoye, J. (2006). On common intervals with errors. Technical report, University of Bielefeld.

[Cho et al., 2009] Cho, H., Eda, M., Nishida, S., Yasukochi, Y., Chong, J., and Koike, H. (2009). Tandem duplication of mitochondrial DNA in the black-faced spoonbill, Platalea minor. *Genes and Genetic Systems*, 84(4) :297–305.

[Christie, 1996] Christie, D. A. (1996). Sorting permutations by block-interchanges. *Information Processing Letters*, 60(4) :165–169.

[Coghlan et al., 2005] Coghlan, A., Eichler, E., Oliver, G., and Paterson, AH Stein, L. (2005). Chromosome evolution in eukaryotes : a multi-kingdom perspective. *Trends in Genetics*, 21(12) :673–682.

[Cuguen et al., 1994] Cuguen, J., Wattier, R., Saumitou-Laprade, P., Forcioli, D., Mörchen, M., Van Dijk, H., and Vernet, P. (1994). Gynodioecy and mitochondrial DNA polymorphism in natural populations of Beta vulgaris ssp maritima. *Gen. Sel. Evol.*, 26 :87–101.

[Dalevi et al., 2002] Dalevi, D., Eriksen, N., K., E., and S.G.E., A. (2002). Measuring Genome Divergence in Bacteria : A Case Study Using Chlamydian Data . *Journal of Molecular Evolution*, 55(1) :24–36.

[Darling et al., 2004] Darling, A. C., Mau, B., Blatter, F. R., and Perna, N. T. (2004). Mauve : multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7) :1394–1403.

[Darracq et al., 2010] Darracq, A., Varré, J.-S., and P., T. (2010). A scenario of mitochondrial genome evolution in maize based on rearrangement events. *BMC Genomics*, 11 :233–249.

[de Haas et al., 1991] de Haas, J. M., Hille, J., Kors, F., van der Meet, B., Kool, A. J., Folkerts, O., and Nijkamp, H. J. J. (1991). Two potential Petunia hybrida mitochondrial DNA replication origins show structural and in vitro functional homology with the animal mitochondrial DNA heavy and light strand replication origins. *Current Genetics*, 20 :503–513.

[Delcher et al., 1999] Delcher, A., Kasif, S., Fleischmann, R.D. Peterson, J., White, O., and Salzberg, S. (1999). Alignment of Whole Genomes. *Nucleic Acids Research*, 27(11) :2369–2376.

[Desplanque et al., 2000] Desplanque, B., Viard, F., Bernard, J., Forcioli, D., Saumitou-Laprade, P., Cuguen, J., and Van Dijk, H. (2000). The linkage disequilibrium between chloroplast DNA and mitochondrial DNA haplotypes in Beta vulgaris ssp. maritima (L.) : the usefulness of both genomes for population genetic studies. *Mol. Ecol.*, 9 :141–154.

[Dias and Meidanis, 2001] Dias, Z. and Meidanis, J. (2001). Genome rearrangements distance by fusion, fission, and transposition is easy. In *Proceedings of the 8th Symposium on String Processing and Information Retrieval, SPIRE'01*, pages 250–253.

[Didier, 2003] Didier, G. (2003). Common intervals of two sequences. In *Algorithms in Bioinformatics*, volume 2812 of *Lecture Notes in Computer Science*, pages 17–24. Springer.

[Dobzhansky and Sturtevant, 1938] Dobzhansky, T. and Sturtevant, A. H. (1938). Inversions in the chromosomes of drosophila pseudoobscura. *Genetics*, 23(1) :28–64.

[Dowton and Campbell, 2001] Dowton, M. and Campbell, N. (2001). Intramitochondrial recombination – is it why some mitochondrial genes sleep around ? *Trends in Ecology and Evolution*, 16(6) :269–271.

[Ducos et al., 2001] Ducos, E., Touzet, P., and Boutry, M. (2001). The male sterile G cytoplasm of wild beet displays modified mitochondrial respiratory complexes. *Plant J.*, 26 :71–180.

[Dufaÿ et al., 2009] Dufaÿ, M., Cuguen, J., Arnaud, J.-F., and Touzet, P. (2009). Sex ratio variation among gynodioecious populations of wild beet : can it be explained by negative frequency-dependent selection ? *Evolution*, 63 :1483–1497.

[El-Mabrouk, 2001] El-Mabrouk, N. (2001). Sorting signed permutations by reversals and insertions/deletions of contiguous segments. *Journal of Discrete Algorithms*, 1(1) :105–122.

[Elias and Hartman, 2006] Elias, I. and Hartman, T. (2006). Approximation algorithm for sorting by transpositions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(4) :369–379.

[Fénart et al., 2006] Fénart, S., Touzet, P., Arnaud, J.-F., and Cuguen, J. (2006). Emergence of gynodioecy in wild beet (beta vulgaris ssp. maritima l.) : a genealogical approach using chloroplastic nucleotide sequences. *Proceedings of the Royal Society B : Biological Sciences*, 273 :1391–1398.

[Fu et al., 2007] Fu, Z., Chen, X., Vacic, V., Nan, P., Zhong, Y., and Jiang, T. (2007). Msoar : a high-throughput ortholog assignment system based on genome rearrangement. *Journal of Computational Biology*, 14(9) :1160–75.

[Fujita et al., 2007] Fujita, M. K., Boore, J. L., and Moritz, C. (2007). Multiple origins and rapid evolution of duplicated mitochondrial genes in parthenogenetic geckos (heteronotia binoei ; squamata, gekkonidae). *Molecular Biology and Evolution*, 24(12) :2775–2786.

[Gascuel, 1997] Gascuel, O. (1997). Bionj : an improved version of the nj algorithm based on a simple model of sequence data. *Molecular Biology Evolution*, 14 :685–695.

[Grande et al., 2008] Grande, C., Templado, J., and Zardoya, R. (2008). Evolution of gastropod mitochondrial genome arrangements. *BMC Evolutionary Biology*, 8 :61.

[Gray et al., 1999] Gray, M. W., Burger, G., and Lang, B. F. (1999). Mitochondrial Evolution. *Science*, 283 :1476–1481.

[Gray et al., 1992] Gray, M. W., Hanic-Joyce, P. J., and Covello, P. S. (1992). Transcription, processing and editing in plant mitochondria. *Annual Review of Plant Physiology and Plant Molecular Biology*, 43 :145–175.

[Grivet and Petit, 2002] Grivet, D. and Petit, J. (2002). Phylogeography of the common ivy (hedera sp.) in europe : genetic differenciation through space and time. *Molecular Ecology*, 11 :1351–1362.

[Grivet and Petit, 2003] Grivet, D. and Petit, J. (2003). Chloroplast dna phylogeography of the hornbeam in europe : evidence for a bottleneck at the outset of postglacial colonization. *Conservation Genetics*, 4 :47–56.

[Hannenhalli and Pevzner, 1995] Hannenhalli, S. and Pevzner, P. (1995). Transforming cabbage into turnip : polynomial algorithm for sorting signed permutations by reversals. In *Proceedings of the 27th annual ACM symposium on Theory of computing*, pages 178–189. ACM Press.

[Hasegawa et al., 1985] Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2) :160–174.

[Heber et al., 2009] Heber, S., Mayr, R., and Stoye, J. (2009). Common intervals of multiple permutations. *Algorithmica*. doi:10.1007/s00453-009-9332-1.

[Heber and Stoye, 2001a] Heber, S. and Stoye, J. (2001a). Algorithms for finding gene clusters. In Heidelberg, S. B. ., editor, *Proceedings of the 1st International Workshop on Algorithms in Bioinformatics, WABI'01*, volume 2149 of *Lecture Notes in Computer Science*, pages 252–263.

[Heber and Stoye, 2001b] Heber, S. and Stoye, J. (2001b). Finding all common intervals of k permutations. In *Proceedings of the 12th Annual Symposium on Combinatorial Pattern Matching, CPM'01*, volume 2089 of *Lecture Notes in Computer Science*, pages 207–218. Springer.

[Hoberman et al., 2005] Hoberman, R., Sankoff, D., and Durand, D. (2005). The statistical analysis of spatially clustered genes under the maximum gap criterion. *Journal of Computational Biology*, 12(8) :1083–102.

[Ingman et al., 2000] Ingman, M., Kaessmann, H., Paabo, S., and Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. *Nature*, 408 :708–713.

[Inoue et al., 2003] Inoue, J. G., Miya, M., Tsukamoto, K., and Nishida, M. (2003). Evolution of the Deep-Sea Gulper Eel Mitochondrial Genomes : Large-Scale Gene Rearrangements Originated Within the Eels. *Molecular Biology and Evolution*, 20(11) :1917–1924.

[Kanehisa and Goto, 2000] Kanehisa, M. and Goto, S. (2000). Kegg : Kyoto encyclopedia of genes and genomes. *Nucleic Acid Research*, 28(1) :27–30.

[Kawanishi et al., 2010] Kawanishi, Y., Shinada, H., Matsunaga, M., Masaki, Y., Mikami, T., and Kubo, T. (2010). A new source of cytoplasmic male sterility found in wild beet and its relationship to other CMS types. *Genome*, 53 :251–256.

[Kececioglu and Sankoff, 1993] Kececioglu, J. and Sankoff, D. (1993). Exact and approximation algorithms for the inversion distance between two chromosomes. In Apostolico, A., Crochemore, M., Galil, Z., and Manber, U., editors, *Proceedings of the 4th Annual Symposium on Combinatorial Pattern Matching, CPM'93*, volume 684 of *Lecture Notes in Computer Science*, pages 87–105. Springer.

[Kececioglu and Sankoff, 1995] Kececioglu, J. and Sankoff, D. (1995). Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica*, 13 :180–210.

[Kubo et al., 2000] Kubo, T., Nishizawa, S., Sugawara, A., Itchoda, N., Estiati, A., and Mikami, T. (2000). The complete nucleotide sequence of the mitochondrial genome of sugar beet (beta vulgaris l.) reveals a novel gene for trnacys(gca). *Nucleic Acids Research*, 28(13) :2571–2576.

[Kvist et al., 2003] Kvist, L., Martens, J., Nazarenko, A., and Orell, M. (2003). Paternal Leakage of Mitochondrial DNA in the Great Tit (Parus major). *Molecular Biology and Evolution*, 20(2) :243–247.

[Ladoukakis and Zouros, 2001] Ladoukakis, E. D. and Zouros, E. (2001). Direct Evidence for Homologous Recombination in Mussel (Mytilus galloprovincialis) Mitochondrial DNA . *Molecular Biology and Evolution*, 18(7) :1168–1175.

[Lang et al., 1997] Lang, B. F., Burger, G., O'Kelly, C. J., Cedergren, R., Golding, G. B., Lemieux, C., Sankoff, D., Turmel, M., and Gray, M. W. (1997). An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature*, 387 :493 – 497.

[Lavrov et al., 2002] Lavrov, D. V., Boore, J. L., and Brown, W. M. (2002). Complete mtdna sequences of two millipedes suggest a new model for mitochondrial gene rearrangements : Duplication and nonrandom loss. *Molecular Biology and Evolution*, 19(2) :163–169.

[Leblanc et al., 1995] Leblanc, C., Boyen, C., Bonnard, O. R. G., Grienenberger, J.-M., and Kloareg, B. (1995). Complete Sequence of the Mitochondrial DNA of the Rhodophyte Chondrus crispus (Gigartinales). Gene Content and Genome Organization. *Journal of Molecular Biology*, 250(4) :484–495.

[Lefebvre et al., 2003] Lefebvre, J., El-Mabrouk, N., Tillier, E., and Sankoff, D. (2003). Detection and validation of single gene inversions. *Bioinformatics*, 19(Suppl 1) :i190–i196.

[Lin et al., 2006] Lin, Y. C., Lu, C. L., Liu, Y.-C., and Tang, C. Y. (2006). Spring : a tool for the analysis of genome rearrangement using reversals and block-interchanges. *Nucleic Acids Research*, 34(Web-Server-Issue) :696–699.

[Ling et al., 2008] Ling, X., He, X., Xin, D., and Han, J. (2008). Efficiently identifying max-gap clusters in pairwise genome comparison. *Journal of Computational Biology*, 15(6) :593–609.

[Lowe and Eddy, 1997] Lowe, T. and Eddy, S. (1997). tRNAscan-SE : a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5) :955–964.

[Lu et al., 2006] Lu, C., Huang, Y., Wang, T., and Chiu, H. (2006). Analysis of circular genome rearrangement by fusions, fissions and block-interchanges. *BMC Bioinformatics*, 7 :295.

[Luc et al., 2003] Luc, N., Risler, J.-L., Bergeron, A., and Raffinot, M. (2003). Gene teams : a new formalization of gene clusters for comparative genomics. *Computational Biology and Chemistry*, 27(1) :59–67.

[Lunt and Hyman, 1997] Lunt, D. H. and Hyman, B. C. (1997). Animal mitochondrial DMA recombination. *Nature*, 387(6630) :247.

[Lynch, 2007] Lynch, M. (2007). *The origins of genome architecture*. Sinauer Associates, Sunderland.

[Lynch et al., 2006] Lynch, M., Koskella, B., and Schaack, S. (2006). Mutation pressure and the evolution of organelle genomic architecture. *Science*, 311(5768) :1727–1730.

[Ma et al., 2006] Ma, J., Zhang, L., Suh, B. B., Raney, B. J., Burhans, R. C., Kent, W. J., Blanchette, M., Haussler, D., and Miller, W. (2006). Reconstructing contiguous regions of an ancestral genome. *Genome Research*, 16(12) :1557–1565.

[Mackenzie and McIntosh, 1999] Mackenzie, S. and McIntosh, L. (1999). Higher plant mitochondria. *The Plant Cell*, 11 :571–585.

[Margulis, 1970] Margulis, L. (1970). *Origin of Eukaryotic Cells : Evidence and Research Implications for A Theory of The Origin and Evolution of Microbial, Plant, and Animal Cells on The Precambrian Earth.* Yale Univ. Press,New Haven.

[Marienfeld et al., 1997] Marienfeld, J. R., Unseld, M., Brandt, P., and Brennicke, A. (1997). Mosaic open reading frames in the Arabidopsis thaliana mitochondrial genome. *Biological chemistry*, 378(8) :859–862.

[Martin and Müller, 1998] Martin, W. and Müller, M. (1998). The hydrogen hypothesis for the first eukaryote. *Nature*, 392 :37–41.

[Mauro et al., 2006] Mauro, D. S., Gower, D. J., Zardoya, R., and Wilkinson, M. (2006). A Hotspot of Gene Order Rearrangement by Tandem Duplication and Random Loss in the Vertebrate Mitochondrial Genome. *Molecular Biology and Evolution*, 23(1) :227–234.

[McLysaght et al., 2002] McLysaght, A., Seoighe, C., and Wolfe, K. (2002). *Comparative Genomics*, chapter High frequency of inversions during eukaryote gene order evolution. Kluwer Academic Press, NY.

[Metzker, 2010] Metzker, M. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11 :31–46.

[Moret et al., 2002] Moret, B., Tang, J., Wang, L., and Warnow, T. (2002). Steps toward accurate reconstruction of phylogenies from gene-order data. *Journal of Computer and System Sciences*, 65(3) :508–525.

[Moret et al., 2001] Moret, B. M. E., Wang, L., Warnow, T., and Wyman, S. (2001). New approaches for reconstructing phylogenies from gene order data. *Bioinformatics*, 17(90001) :S165–S173.

[Mower and Palmer, 2006] Mower, J. and Palmer, J. (2006). Patterns of partial RNA editing in mitochondrial genes of Beta vulgaris. *Molecular Genetics and Genomics*, 276 :285–293.

[Mueller and Boore, 2005] Mueller, R. L. and Boore, J. L. (2005). Molecular Mechanisms of Extensive Mitochondrial Gene Rearrangement in Plethodontid Salamanders. *Molecular Biology and Evolution*, 22(10) :2104–2112.

[Nagarajan et al., 2010] Nagarajan, N., Cook, C., Di Bonaventura, M., Ge, H., Richards, A., Bishop-Lilly, K., DeSalle, R., Read, T., and Pop, M. (2010). Finishing genomes with limited resources : lessons from an ensemble of microbial genomes. *BMC Genomics*, 11 :242.

[Needleman and Wunsch, 1970] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3) :443–453.

[Noe and Kucherov, 2005] Noe, L. and Kucherov, G. (2005). YASS : enhancing the sensitivity of dna similarity search. *Nucleic Acids Research*, 33(2) :W540–W543.

[Notsu et al., 2002] Notsu, Y., Masood, S., Nishikawa, T., Kubo, N., Akiduki, G., Nakazono, M., Hirai, A., and Kadowaki, K. (2002). The complete sequence of the rice (oryza sativa l.) mitochondrial genome : frequent dna sequence acquisition and loss during the evolution of flowering plants. *Molecular Genetics and Genomics*, 268 :434–445.

[Onodera et al., 1999] Onodera, Y., Yamamoto, M., Kubo, T., and Mikami, T. (1999). Heterogeneity of the atp6 presequences in normal and different sources of male-sterile cytoplasms of sugar beet. *Journal of Plant Physiology*, 155 :656–660.

[Osawa et al., 1992] Osawa, S., Jukes, T., Watanabe, K., and Muto, A. (1992). Recent evidence for evolution of the genetic code. *Microbiology and Molecular Biology Reviews*, 56(1) :229–264.

[Ostlund et al., 2010] Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D. N., Roopra, S., Frings, O., and Sonnhammer, E. L. L. (2010). Inparanoid 7 : new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research*, 38(Database issue) :D196–203.

[Perseke et al., 2008] Perseke, M., Fritzsch, G., Ramsch, K., Bernt, M., Merkle, D., Middendorf, M., Bernhard, D., Stadler, P. F., and Schlegel, M. (2008). Evolution of mitochondrial gene orders in echinoderms. *Molecular Phylogenetics and Evolution*, 47(2) :855–864.

[Pevzner and Tesler, 2003a] Pevzner, P. and Tesler, G. (2003a). Genome rearrangements in mammalian evolution : lessons from human and mouse genomes. *Genome Research*, 13(1) :37–45.

[Pevzner and Tesler, 2003b] Pevzner, P. and Tesler, G. (2003b). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. USA*, 100(13) :7672–7677.

[Pfeiffer et al., 2000] Pfeiffer, P., Goedecke, W., and Obe, G. (2000). Mechanisms of DNA double-strand break repair and their potential to induce chromosomal aberrations. *Mutagenesis*, 15(4) :289–302.

[Rand, 2001] Rand, D. (2001). The Units Of Selection On Mitochondrial DNA. *Annual Review of Ecology and Systematics*, 32 :415–448.

[Rokas et al., 2003] Rokas, A., Ladoukakis, E., and Zouros, E. (2003). Animal mitochondrial DNA recombination revisited. *TRENDS in Ecology and Evolution*, 18(8) :411–417.

[Rozas et al., 2003] Rozas, J., Sanshez-Delbarrio, J. C., Messeguer, X., and Rozas, R. (2003). Bioinformatics. *DnaSP, DNA polymorphism analyses by the coalescent and other methods*, 19 :2496–2497.

[Sankoff, 1999] Sankoff, D. (1999). Genome rearrangement with gene families. *Bioinformatics*, 15(11) :909–917.

[Sankoff and Blanchette, 1998] Sankoff, D. and Blanchette, M. (1998). Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology*, 5(3) :555–570.

[Sankoff et al., 1997] Sankoff, D., Ferretti, V., and Nadeau, J. (1997). Conserved Segment Identification. *J. Comput. Biol.*, 4(4) :559–565.

[Sankoff et al., 2000] Sankoff, D., Parent, M., and Bryant, D. (2000). *Comparative Genomics*, chapter Accuracy and robustness of analyses based on numbers of genes in observed segments, pages 299–336. Kluwer Academic Press, NY.

[Sankoff et al., 1996] Sankoff, D., Sundaram, G., and Kececioglu, J. D. (1996). Steiner points in the space of genome rearrangements. *Int. J. Found. Comput. Sci.*, 7(1) :1–9.

[Satoh et al., 2004] Satoh, M., Kubo, T., Nishizawa, S., Estiati, A., Itchoda, N., and Mikami, T. (2004). The cytoplasmic male-sterile type and normal type mitochondrial genomes of sugar beet share the same complement of genes of known function but differ in the content of expressed orfs. *Molecular Genetics and Genomics*, 272(3) :247–256.

[Schmidt et al., 2002] Schmidt, H. A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). TREE-PUZZLE : maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18 :502–504.

[Schmidt and Stoye, 2004] Schmidt, T. and Stoye, J. (2004). Quadratic time algorithms for finding common intervals in two and more sequences. In Sahinalp, S. C., Muthukrishnan, S., and Dogrusoz, U., editors, *Proceedings of the 15th Annual Symposium on Combinatorail Pattern Matching, CPM'04*, volume 3109 of *Lecture Notes in Computer Science*, pages 347–358. Springer.

[Schuster and Brennicke, 1994] Schuster, W. and Brennicke, A. (1994). The plant mitochondrial genome : Physical Structure, Information Content, RNA Editing, and Gene Migration to the Nucleus. *Annual Review of Plant Physiology and Plant Molecular Biology*, 45 :61–78.

[Scotti et al., 2001] Scotti, N., Cardi, T., and Maréchal-Drouard, L. (2001). Mitochondrial DNA and RNA isolation from small amounts of potato tissue. *Plant. Mol. Biol. Rep.*, 19 :1–8.

[Setubal and Meidanis, 1997] Setubal, C. and Meidanis, J. (1997). *Introduction to Computational Molecular Biology*. PWS Publishing.

[Shadel and Clayton, 1997] Shadel, G. S. and Clayton, D. A. (1997). Mitochondrial DNA Maintenance in Vertebrates. *Annual Review of Biochemistry*, 66 :409–435.

[Shao and Barker, 2003] Shao, R. and Barker, S. C. (2003). The Highly Rearranged Mitochondrial Genome of the Plague Thrips, Thrips imaginis (Insecta : Thysanoptera) : Convergence of Two Novel Gene Boundaries and an Extraordinary Arrangement of rRNA Genes. *Molecular Biology and Evolution*, 20(3) :362–370.

[Shi et al., 2010] Shi, G., Zhang, L., and Jiang, T. (2010). Msoar 2.0 : Incorporating tandem duplications into ortholog assignment based on genome rearrangement. *BMC Bioinformatics*, 11 :10.

[Signorovitch et al., 2007] Signorovitch, A. Y., Buss, L. W., and Dellaporta, S. L. (2007). Comparative genomics of large mitochondria in placozoans. *PLoS Genetics*, 3(1) :44–50.

[Smith et al., 2010] Smith, D., Lee, R., Cushman, J., Magnuson, J., Tran, D., and Polle, J. (2010). The Dunaliella salina organelle genomes : large sequences, inflated with intronic and intergenic DNA. *BMC Plant Biology*, 10 :83.

[Snel et al., 1999] Snel, B., Bork, P., and Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nature*, 21 :108–110.

[Strimmer and von Haeseler, 1996] Strimmer, K. and von Haeseler, A. (1996). Quartet puzzling : a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, 13(7) :964–969.

[Sugiyama et al., 2005] Sugiyama, Y., Watase, Y., Nagase, M., Makita, N., Yagura, S., Hirai, A., and Sugiura, M. (2005). The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome : comparative analysis of mitochondrial genomes in higher plants. *Molecular Genetics and Genomics*, 272 :603–615.

[Sun et al., 2005] Sun, H., Zhou, K., and Song, D. (2005). Mitochondrial genome of the Chinese mitten crab Eriocheir japonica sinenesis (Brachyura : Thoracotremata : Grapsoidea) reveals a novel gene order and two target regions of gene rearrangements. *Gene*, 249 :207–217.

[Suyama and Bork, 2001] Suyama, M. and Bork, P. (2001). Evolution of prokaryotic gene order : genome rearrangements in closely related species. *Trends in Genetics*, 17 :10–13.

[Swidan et al., 2006] Swidan, F., Rocha, E., Shmoish, M., and Pinter, R. (2006). An Integrative Method for Accurate Comparative Genome Mapping. *PLoS Computational Biology*, 2(8) :e75.

[Swidan and Shamir, 2009] Swidan, F. and Shamir, R. (2009). Assessing the Quality of Whole Genome Alignments in Bacteria. *Advances in Bioinformatics*, 2009 :1–8.

[Taanman, 1999] Taanman, J.-W. (1999). The mitochondrial genome : structure, transcription, translation and replication. *Biochimica et Biophysica Acta*, 1410 :103–123.

[Tang and Moret, 2003] Tang, J. and Moret, B. M. E. (2003). Phylogenetic reconstruction from gene-rearrangement data with unequal gene content. In *Proceedings of the 8th International Workshop on Algorithms and Data Structures, WABI'03*, volume 2748 of *Lecture Notes in Computer Science*, pages 37–46. Springer.

[Tannier and Sagot, 2004] Tannier, E. and Sagot, M.-F. (2004). Sorting by Reversals in Subquadratic Time. In *Proceedings of the 15th Annual Symposium on Combinatorail Pattern Matching, CPM'04*, volume 3109 of *Lecture Notes in Computer Science*, pages 1–13. Springer.

[Tesler, 2002] Tesler, G. (2002). GRIMM : genome rearrangements web server. *Bioinformatics*, 18(3) :492–493.

[Touzet and Delph, 2009] Touzet, P. and Delph, L. (2009). The effect of breeding system on polymorphism in mitochondrial genes of Silene. *Genetics*, 181 :631–644.

[Uno and Yagiura, 2000] Uno, T. and Yagiura, M. (2000). Fast algorithms to enumerate all common intervals of two permutations. *Algorithmica*, 26(2) :290–309.

[Vallès and Boore, 2005] Vallès, Y. and Boore, J. L. (2005). Lophotrochozoan mitochondrial genomes. *Integrative and Comparative Biology*, 46(4) :544–557.

[Wang et al., 2002] Wang, L.-S., Jansen, R. K., Moret, B. M. E., Raubeson, L. A., and Warnow, T. (2002). Fast phylogenetic methods for the analysis of genome rearrangement data : An empirical study. In *Proceedings of the 6th Pacific Symposium on Biocomputing, PSB'02*, pages 524–535.

[Welch et al., 2006] Welch, M. E., Darnell1, M. Z., and McCauley, D. E. (2006). Variable Populations Within Variable Populations : Quantifying Mitochondrial Heteroplasmy in Natural Populations of the Gynodioecious Plant Silene vulgaris. *Genetics*, 174 :829–837.

[Whittle and Johnston, 2002] Whittle, C.-A. and Johnston, M. O. (2002). Male-Driven Evolution of Mitochondrial and Chloroplastidial DNA Sequences in Plants. *Molecular Biology and Evolution*, 19(6) :938–949.

[Wolfe et al., 1987] Wolfe, K. H., Li, W.-H., and Sharp, P. M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. USA*, 84 :9054–9058.

[Woo et al., 2007] Woo, H., Lee, Y., Park, S., Lim, J., Jang, K., Choi, E., Choi, Y., and Hwang, U. (2007). Complete mitochondrial genome of a troglobite millipede Antrokoreana gracilipes (Diplopoda, Juliformia, Julida), and juliformian phylogeny. *Molecules and Cells*, 23(2) :182–191.

[Yamamoto et al., 2005] Yamamoto, M. P., Kubo, T., and Mikami, T. (2005). The 5'-leader sequence of sugar beet mitochondrial atp6 encodes a novel polypeptide that is characteristic of Owen cytoplasmic male sterility. *Mol. Gen. Genom.*, 273 :342–349.

[Yamamoto et al., 2008] Yamamoto, M. P., Shinada, H., Onodera, Y., Komaki, C., Mikami, T., and T, K. (2008). A male-sterility-associated mitochondrial protein in wild beets causes pollen disruption in transgenic plants. *The Plant Journal*, 54(6) :1027–1036.

[Yancopoulos et al., 2005] Yancopoulos, S., Attie, O., and Friedberg, R. (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16) :3340–3346.

[Yu et al., 2008] Yu, Z., Wei, Z., Kong, X., and Shi, W. (2008). Complete mitochondrial DNA sequence of oyster Crassostrea hongkongensis-a case of Tandem duplication-random loss for genome rearrangement in Crassostrea ? *BMC Genomics*, 9(1) :477.

[Zillig et al., 1989] Zillig, W., Palm, P., and Klenk, H.-P. (1989). Did eukaryotes originate by a fusion event ? *Endocytobiosis and Cell Research*, 6 :1–25.

Annexes Chapitre 4

Table S1 – Length and position of backbone fragments used for backbone DNA sequences. All fragments orientation depended on NA (if fragment positions begin by -, fragment is in reverse orientation compared to NA)
Blue line indicates the smallest fragment and red line largest fragment

Genomes	NA			NB			CMS-C			CMS-S			CMS-T			*Zea mays ssp. parviglum*			*Zea luxurians*			*Zea perennis*		
Fragment number	Start	Stop	Length	Start	Stop	Length	Start	Stop	Length	Start	Stop	Length	Start	Stop	Length	Start	Stop	Length	Start	Stop	Length	Start	Stop	Length
1	1	927	927	1	927	927	1	927	927	1	927	927	1	937	937	1	927	927	1	927	927	1	927	927
2	931	1873	943	931	1873	943	931	1873	943	192850	193788	939	941	1888	948	931	1873	943	931	1872	942	931	1873	943
3	4420	4515	96	4420	4515	96	-98252	-98347	96	409940	410035	96	4421	4516	96	4420	4515	96	320446	320541	96	-548419	-548514	96
4	4530	5947	1418	4520	5937	1418	-96825	-98242	1418	410050	411462	1413	4521	5953	1433	4530	5947	1418	320556	321968	1413	-546982	-548394	1413
5	25609	30792	5184	180222	185397	5176	17937	23112	5176	-485571	-490760	5190	197783	202951	5169	25584	30755	5172	30019	35163	5145	385692	390870	5179
6	31984	32077	94	-326084	-326207	124	24300	24393	94	176893	176986	94	-128956	-129049	94	626397	626490	94	-248047	-248140	94	8260	8353	94
7	32119	45067	12949	-313147	-326081	12935	24435	37388	12954	177028	189974	12947	-115972	-128919	12948	626532	639480	12949	-235163	-248010	12848	8390	21272	12883
8	45285	52442	7158	-305777	-312929	7153	37606	44768	7163	379613	386765	7153	-108605	-115752	7148	639698	646855	7158	-227936	-234964	7029	21486	28608	7123
9	53051	59103	6053	-299116	-305168	6053	45377	51429	6053	387374	393432	6059	-101928	-107991	6064	647464	653521	6058	-336674	-342669	5996	526019	532050	6032
10	59331	60590	1260	500411	501670	1260	51657	52916	1260	393660	394919	1260	-100441	-101700	1260	653749	655008	1260	-335200	-336452	1253	532278	533542	1265
11	60614	69205	8592	501682	510265	8584	52940	61556	8617	394949	403561	8613	-91812	-100423	8612	655020	663603	8584	-326661	-335200	8540	533614	542204	8591
12	69217	70947	1731	510277	512007	1731	61568	63298	1731	189974	191704	1731	-90070	-91800	1731	663615	665345	1731	-324934	-326649	1716	542216	543959	1744
13	72952	73260	309	-353758	-354066	309	-115213	-115521	309	-379294	-379602	309	457601	457909	309	-376903	-377211	309	130736	131044	309	153282	153595	314
14	73260	74458	1199	-352560	-353758	1199	-114015	-115213	1199	-378096	-379294	1199	457909	459107	1199	-375705	-376903	1199	131044	132218	1175	153611	154799	1189
15	74458	76458	2001	-350560	-352560	2001	-112015	-114015	2001	-376101	-378096	1996	459112	461112	2001	-373705	-375705	2001	132223	134200	1978	154824	156815	1992
16	76465	82610	6146	-344408	-350553	6146	-105868	-112008	6141	-369954	-376094	6141	461119	467269	6151	-367553	-373698	6146	-138490	-144625	6136	-161153	-167325	6173
17	82610	84768	2159	-342254	-344408	2155	-103723	-105868	2146	-367792	-369954	2163	467269	469423	2155	-365399	-367553	2155	-305399	-307529	2131	-344727	-346869	2143
18	97832	97976	145	-517153	-517297	145	-13894	-14038	145	7507	7651	145	513820	513941	122	-352191	-352335	145	-292193	-292312	120	259025	259144	120
19	100651	100763	113	-514366	-514478	113	-11107	-11219	113	10326	10438	113	-89409	-89521	113	-349404	-349516	113	201386	201498	113	-104133	-104245	113
20	106111	110514	4404	-445230	-449633	4404	334451	338834	4384	15786	20184	4399	-480494	-484883	4390	490730	495125	4396	150960	155311	4352	173767	178157	4391
21	110664	110830	167	-444914	-445080	167	338984	339150	167	20334	20500	167	-310435	-310601	167	495275	495441	167	352968	353134	167	-515473	-515639	167
22	110830	121026	10197	-434718	-444914	10197	339150	349356	10207	20500	30702	10203	-300235	-310435	10201	495441	505637	10197	353134	363248	10115	-505266	-515443	10178
23	121041	121373	333	-434365	-434697	333	349371	349703	333	30711	31048	338	-299870	-300202	333	505652	505984	333	363248	363594	347	-504912	-505257	346
24	121379	122313	935	-433419	-434353	935	349721	350655	935	31048	31982	935	-298930	-299864	935	505984	506918	935	363606	364541	936	-503889	-504828	940
25	122969	130565	7597	-425167	-432763	7597	351339	358926	7588	32644	40250	7607	-290669	-298275	7607	507574	515183	7610	365188	372696	7509	-62689	-70264	7576
26	130576	136380	5805	-419352	-425156	5805	358938	364745	5808	40283	46092	5810	-284849	-290658	5810	515219	521011	5793	372696	378438	5743	-56906	-62689	5784
27	138269	142080	3812	-413648	-417463	3816	366634	370445	3812	47981	51801	3821	-279142	-282965	3824	522900	526711	3812	407869	411653	3785	184920	188752	3833
28	142117	142233	117	-413495	-413611	117	370482	370598	117	51838	51954	117	-278989	-279105	117	526748	526864	117	411653	411781	129	188789	188917	129
29	142233	143279	1047	-412449	-413495	1047	370598	371644	1047	51954	53000	1047	-277943	-278989	1047	526864	527910	1047	411801	412847	1047	188937	189983	1047
30	143328	151730	8403	-403998	-412400	8403	371693	380106	8414	53049	61462	8414	-269472	-277894	8423	527959	536361	8403	412847	421151	8305	190032	198423	8392
31	151746	156607	4862	-399121	-403982	4862	380122	384985	4864	61478	66344	4867	-264585	-269456	4872	536377	541238	4862	421151	425973	4823	198423	203282	4860
32	156607	158195	1589	-397533	-399121	1589	384989	386569	1581	66349	67937	1589	-263012	-264585	1574	541238	542826	1589	425973	427537	1565	203307	204894	1588
33	158207	161728	3522	-394000	-397521	3522	386587	390108	3522	67943	71464	3522	-259485	-263006	3522	542838	546359	3522	427537	431058	3522	204906	208443	3538
34	161737	167344	5608	-388379	-393991	5613	390129	395733	5605	71491	77088	5598	-253876	-259470	5595	546368	551975	5608	431058	436579	5522	208476	214093	5618
35	167354	169851	2498	-385862	-388359	2498	395743	398240	2498	77088	79581	2494	-251386	-253876	2491	551985	554482	2498	436579	439064	2486	214119	216622	2504
36	169867	171471	1605	-384242	-385846	1605	398256	399860	1605	79597	81201	1605	-249771	-251370	1600	554498	556102	1605	439064	440610	1547	216630	218199	1570
37	171476	172053	578	-383665	-384242	578	399870	400447	578	81206	81791	586	-249174	-249751	578	556112	556689	578	440615	441189	575	218289	218874	586
38	172110	173707	1598	-382016	-383608	1593	400504	402089	1586	81848	83440	1593	-247525	-249117	1593	556746	558338	1593	441214	442762	1549	218931	220495	1565
39	175664	175838	175	-379880	-380054	175	404038	404212	175	85392	85566	175	63384	63558	175	560295	560469	175	-224754	-224928	175	31643	31817	175
40	175838	180127	4290	-375586	-379880	4295	404212	408498	4287	85566	89863	4298	63558	67849	4292	560469	564763	4295	-398912	-403178	4267	31817	36127	4311
41	180297	182841	2545	-372880	-375416	2537	408668	411204	2537	90029	92565	2537	-83259	-85793	2535	564933	567477	2545	-396224	-398749	2526	36292	38852	2561
42	182841	191254	8414	-364456	-372875	8420	411229	419632	8404	92580	100993	8414	-74845	-83259	8415	567477	575890	8414	-387942	-396214	8273	38922	47308	8387
43	191271	191606	336	-364104	-364439	336	419649	419984	336	101010	101345	336	-74493	-74828	336	575907	576242	336	-387632	-387942	311	47308	47626	319
44	195969	200892	4924	-354818	-359741	4924	-116268	-121191	4924	105713	110641	4929	451940	456854	4915	-377963	-382886	4924	-378438	-383327	4890	51983	56906	4924
45	323192	326779	3588	-130718	-134295	3578	422878	426465	3588	-276669	-280251	3583	-68017	-71599	3583	579132	582719	3588	-463308	-466865	3558	310715	314292	3578
46	343852	356039	12188	-101472	-113649	12178	443545	455722	12178	-247407	-259583	12177	-394248	-406420	12173	203327	215504	12178	81333	93462	12130	437456	449636	12181
47	356039	358104	2066	-99407	-101472	2066	455722	457779	2058	-245342	-247407	2066	-392196	-394248	2053	215504	217569	2066	-528078	-530112	2035	449636	451696	2061
48	358104	362728	4625	-94794	-99407	4614	457779	462392	4614	-240711	-245324	4614	-387579	-392196	4618	217569	222193	4625	-523441	-528078	4638	451696	456333	4638
49	362740	364211	1472	-93323	-94794	1472	462416	463887	1472	-239208	-240675	1468	-386090	-387561	1472	222199	223670	1472	-521968	-523441	1474	456453	457934	1482
50	364221	370723	6503	-86811	-93313	6503	463897	470394	6498	-232686	-239188	6503	-379578	-386085	6508	223685	230187	6503	-515471	-521958	6488	457934	464432	6499
51	370733	380958	10226	-76581	-86801	10221	470409	480635	10227	-222442	-232676	10235	-369354	-379578	10225	230202	240427	10226	-505229	-515451	10223	464477	474716	10240
52	380958	382669	1712	-74870	-76581	1712	480635	482346	1712	-220731	-222442	1712	-184248	-185959	1712	240427	242138	1712	-503505	-505229	1725	474716	476432	1717
53	398167	403220	5054	113649	118698	5050	497833	502886	5054	259583	264632	5050	406420	411469	5050	-198274	-203327	5054	-76300	-81333	5034	-432388	-437456	5069
54	403220	408140	4921	118698	123618	4921	502886	507813	4928	264632	269564	4933	411469	416394	4926	-193347	-198274	4928	-71386	-76300	4915	-321364	-326294	4931
55	408169	415041	6873	123647	130519	6873	507842	514714	6873	269593	276470	6878	416423	423284	6862	-186446	-193318	6873	456399	463116	6718	-314484	-321335	6852
56	415041	420716	5676	-36827	-42502	5676	514714	520389	5676	527829	533499	5671	423284	428968	5685	-180771	-186446	5676	-58990	-64644	5655	-414912	-420578	5667
57	420790	425445	4656	-32098	-36753	4656	76280	80924	4645	-335858	-340505	4648	339205	343840	4636	-176042	-180697	4656	-466881	-471469	4589	306046	310699	4654
58	427013	427293	281	-30250	-30530	281	82492	82772	281	-334001	-334281	281	345408	345688	281	-174194	-174474	281	-122836	-123106	271	-145209	-145479	271
59	427322	430907	3586	-26636	-30221	3586	82801	86376	3576	-330387	-333972	3586	345717	349297	3581	-170580	-174165	3586	-119293	-122836	3544	-141604	-145180	3577
60	436235	440620	4386	-16928	-21313	4386	91729	96108	4380	-320661	-325057	4397	-329404	-333794	4391	-317992	-322377	4386	476501	480862	4362	-499264	-503657	4394
61	440620	444133	3514	-13415	-16928	3514	280678	284191	3514	-317140	-320661	3522	-325891	-329404	3514	-314479	-317992	3514	480862	484328	3467	-495760	-499264	3505
62	444138	452485	8348	-491510	-499857	8348	284196	292547	8352	-308793	-317135	8343	-317538	-325886	8349	-306132	-314479	8348	484333	492607	8275	-487434	-495755	8322
63	453094	458083	4990	-485904	-490893	4990	293156	298155	5000	-303182	-308176	4995	-311938	-316929	4992	-300534	-305523	4990	346696	351656	4961	-516976	-521970	4995
64	458112	458939	828	-485048	-485875	828	298184	299011	828	-302326	-303153	828	-311082	-311909	828	-299678	-300505	828	351656	352477	822	-516155	-516976	822
65	458944	459361	418	-484626	-485043	418	299021	299438	418	-301904	-302321	418	-310650	-311067	418	-299261	-299678	418	352497	352919	423	-515688	-516105	418
66	459361	460489	1129	-483498	-484626	1129	299438	300566	1129	-300776	-301904	1129	-509531	-510659	1129	-298133	-299261	1129	295453	296563	1111	-254752	-255870	1119
67	460517	461359	843	-482628	-483470	843	300594	301436	843	-299906	-300748	843	-508661	-509503	843	-297263	-298105	843	-219462	-220300	839	-235464	-236315	852
68	461364	462547	1184	-481440	-482623	1184	301441	302624	1184	-298723	-299906	1184	-507463	-508646	1184	-296075	-297258	1184	-218264	-219447	1184	-234259	-235434	1176
69	462547	466174	3628	-477813	-481440	3628	302624	306251	3628	-295096	-298723	3628	-503836	-507463	3628	-292448	-296075	3628	-214678	-218264	3587	-548529	-552169	3641
70	466230	468608	2379	-475389	-477762	2374	306307	308680	2374	-292677	-295045	2369	-501412	-503780	2369	-290019	-292392	2374	-212286	-214635	2350	87588	89950	2363

Genomes	NA			NB			CMS-C			CMS-S			CMS-T			*Zea mays ssp. parviglumis*			*Zea luxurians*			*Zea perennis*		
Fragment number	Start	Stop	Length	Start	Stop	Length	Start	Stop	Length	Start	Stop	Length	Start	Stop	Length	Start	Stop	Length	Start	Stop	Length	Start	Stop	Length
71	468608	469178	571	-474819	-475389	571	308680	309250	571	-292107	-292677	571	-500842	-501412	571	-289449	-290019	571	-211716	-212286	571	93236	93806	571
72	469178	475365	6188	-468637	-474819	6183	309250	315436	6187	-285926	-292107	6182	-494655	-500842	6188	-283262	-289449	6188	-495921	-502051	6131	477913	484071	6159
73	479495	480169	675	138518	139181	664	576695	577369	675	-421447	-422121	675	-446175	-446849	675	-278458	-279132	675	-533910	-534589	680	-564858	-565532	675
74	485070	485524	455	144082	144536	455	582269	582723	455	-416081	-416535	455	152847	153301	455	-273103	-273557	455	-158098	-158544	447	92777	93236	460
75	485542	488309	2768	144554	147321	2768	582741	585508	2768	-413296	-416063	2768	153319	156095	2777	-270318	-273085	2768	-155311	-158080	2770	-89991	-92763	2773
76	490686	492323	1638	149698	151335	1638	587885	589514	1630	194928	196565	1638	158464	160101	1638	-266304	-267941	1638	-220753	-222345	1593	291954	293597	1644
77	492349	498829	6481	151361	157830	6470	589540	596020	6481	196591	203071	6481	160127	166602	6476	-259798	-266278	6481	-260818	-267272	6455	293618	300084	6467
78	498857	505397	6541	157858	164398	6541	596048	602582	6535	203099	209639	6541	166630	173169	6540	-253230	-259770	6541	7824	14337	6514	363339	369878	6540
79	505397	516487	11091	63778	74868	11091	602582	613661	11080	-495494	-506584	11091	173169	184246	11078	604205	615295	11091	14337	25404	11068	369878	381000	11123
80	516489	517424	936	175490	176425	936	613663	614598	936	-494557	-495492	936	-139198	-140133	936	615302	616237	936	25406	26319	914	381002	381931	930
81	517450	520781	3332	176451	179782	3332	614624	617955	3332	-491200	-494531	3332	-135840	-139172	3333	616263	619594	3332	26319	29597	3279	381939	385264	3326
82	520799	521198	400	179800	180199	400	17515	17914	400	-490783	-491182	400	-135423	-135822	400	619612	620011	400	29597	29996	400	385270	385669	400
83	526525	527404	880	185518	186397	880	623687	624566	880	-484571	-485450	880	203073	203952	880	30876	31755	880	35271	36109	839	390994	391869	876
84	527404	545060	17657	186397	204048	17652	624566	642217	17652	-466898	-484553	17656	203952	221594	17643	31755	49411	17657	36109	53616	17508	391869	409484	17616
85	545601	564369	18769	204584	223342	18759	642753	661501	18749	-447602	-466372	18771	222134	240885	18752	49947	68710	18764	-271552	-290097	18546	261248	280017	18770
86	566455	570770	4316	225428	229743	4316	663587	667902	4316	-441218	-445528	4311	242971	247276	4306	70796	75111	4316	-451996	-456260	4265	-229959	-234259	4301
87	570988	572992	2005	229961	231965	2005	668120	670124	2005	-438996	-441000	2005	143055	145072	2018	75329	77333	2005	-206665	-208649	1985	96914	98903	1990
88	573002	574400	1399	231965	233363	1399	670134	671532	1399	-437598	-438996	1399	145077	146475	1399	77348	78746	1399	-205275	-206660	1386	98938	100336	1399
89	574760	575275	516	233723	234238	516	671892	672407	516	-436723	-437238	516	85949	86459	511	79106	79621	516	-204414	-204915	502	100700	101210	511
90	576660	577821	1162	235623	236779	1157	673792	674953	1162	-434177	-435338	1162	514376	515537	1162	81006	82167	1162	-290597	-291758	1162	259579	260740	1162
91	578331	585604	7274	237288	244561	7274	675462	682735	7274	-426395	-433668	7274	516046	523314	7269	82676	89949	7274	309854	317091	7238	349219	356515	7297
92	596336	597967	1632	255283	256914	1632	532373	534004	1632	-155532	-157163	1632	7086	8720	1635	100686	102317	1632	-268250	-269845	1596	-288656	-290279	1624
93	597983	598965	983	256930	257917	988	534020	535007	988	-154529	-155516	988	8736	9718	983	102333	103315	983	-267272	-268250	979	-287664	-288640	977
94	599648	600994	1347	258600	259946	1347	535690	537036	1347	-152500	-153846	1347	10404	11750	1347	103998	105344	1347	-303712	-305037	1326	-285649	-286981	1333
95	601004	605934	4931	259956	264886	4931	537056	541991	4936	-147550	-152480	4931	11760	16698	4939	105354	110284	4931	-298809	-303702	4894	-280727	-285649	4923
96	606612	619957	13346	265564	278908	13345	542669	556013	13345	-133533	-146872	13340	17376	30716	13341	110962	124307	13346	-186628	-199895	13268	105758	119088	13331
97	619957	623339	3383	278908	282290	3383	556048	559422	3375	-130127	-133513	3387	30741	34131	3391	124307	127689	3383	-183264	-186623	3360	119138	122521	3384
98	623364	623487	124	282315	282438	124	559447	559570	124	-129984	-130107	124	34151	34274	124	127714	127837	124	-183137	-183260	124	122541	122664	124
99	623487	625124	1638	282438	284075	1638	559570	561207	1638	-128326	-129954	1629	34274	35916	1643	127837	129474	1638	-181521	-183137	1617	122664	124300	1637
100	625130	625749	620	284081	284700	620	561213	561832	620	-127707	-128326	620	35922	36541	620	129480	130099	620	-180896	-181503	608	124324	124931	608
101	625767	626833	1067	284718	285784	1067	561850	562916	1067	-126623	-127689	1067	36559	37625	1067	130117	131183	1067	-179830	-180896	1067	124931	126001	1071
102	626845	637257	10413	285790	296209	10420	562916	573335	10420	-116188	-126611	10424	37631	48052	10422	131189	141601	10413	-169467	-179830	10364	126037	136424	10388
103	642700	644613	1914	453981	455894	1914	-328189	-330102	1914	-358241	-360154	1914	53495	55408	1914	147044	148957	1914	-254654	-256560	1907	247124	249037	1914
104	644613	645142	530	455894	456423	530	-327660	-328189	530	-357712	-358241	530	55408	55937	530	148957	149486	530	-254105	-254634	530	249052	249581	530
105	645142	649094	3953	456423	460370	3948	-323703	-327655	3953	-353755	-357707	3953	55937	59902	3966	149486	153444	3959	-250207	-254095	3889	249601	253553	3953
106	650213	651653	1441	461489	462929	1441	-321144	-322584	1441	-113260	-114700	1441	487500	488946	1447	154563	156003	1441	256841	258291	1451	326915	328366	1452
107	654689	657357	2669	465969	468637	2669	-315436	-318104	2669	283253	285926	2674	491987	494655	2669	159048	161716	2669	493282	495921	2640	-484071	-486752	2682
108	661709	665753	4045	541921	545969	4049	72171	76206	4036	-340579	-344614	4036	335096	339131	4036	323186	327230	4045	-471529	-475547	4019	301927	305979	4053
109	665827	670972	5146	546043	551183	5141	520463	525618	5156	533573	538713	5141	429042	434197	5156	327304	332449	5146	-53830	-58930	5101	-409707	-414845	5139
110	670982	675722	4741	551193	555929	4737	525628	530368	4741	538723	543468	4746	434207	438932	4726	332459	337199	4741	-125765	-130459	4695	-148202	-152938	4737
111	687349	688338	990	555933	556922	990	64859	65848	990	543472	544461	990	438936	439925	990	666906	667895	990	-124785	-125765	981	-147213	-148202	990
112	688825	691679	2855	557409	560263	2855	66335	69189	2855	544948	547802	2855	440412	443266	2855	668382	671236	2855	-405037	-407853	2817	-182056	-184909	2854
113	694059	694589	531	562643	563173	531	71569	72099	531	550180	550710	531	528837	529367	531	673616	674146	531	1908	2433	526	1914	2444	531
114	694592	695470	879	563176	564054	879	733273	734151	879	550713	551591	879	529370	530248	879	674149	675027	879	2436	3314	879	2447	3325	879
115	696228	701047	4820	564812	569631	4820	734909	739720	4812	552349	557163	4815	531006	535826	4821	675785	680604	4820	534589	539369	4781	565536	570355	4820
Total length			156757			156727			156716			156749			156735			156758			155480			156642

Table S2 – Synteny anchor numbers and their content

Synteny anchor (SA) number	known sequence contained in SA	Duplicated SA
1	backbone DNA sequence fragment #1	no
2	backbone DNA sequence fragment #4	no
3	*nad*1 exon 3	yes
	*nad*1 exon 2	
	*rps*13	
	*trn*fM	
	backbone DNA sequence fragment #6	
	backbone DNA sequence fragment #7	
4	*cob*	yes
	orf99	
	orf147	
	backbone DNA sequence fragment #8	
5	*atp*9	yes
6	*nad*2 exon 2	yes
	backbone DNA sequence fragment #10	
	backbone DNA sequence fragment #11	
7	nad2 exon 1	yes
	backbone DNA sequence fragment #12	
8	backbone DNA sequence fragment #13	yes
	backbone DNA sequence fragment #14	
	backbone DNA sequence fragment #15	
9	*rbc*L°*	yes
	*rpl*23°*	
	*rpl*2°*	
	*trn*H°	
	*ccm*B	
	backbone DNA sequence fragment #16	
10	orf132	yes
	backbone DNA sequence fragment #17	
11	backbone DNA sequence fragment #19	yes
12	*cox*3	yes
	*rps*7	
	backbone DNA sequence fragment #21	
	backbone DNA sequence fragment #22	
	backbone DNA sequence fragment #23	
	backbone DNA sequence fragment #24	
13	*rrn*18	yes
	*rrn*5	
	backbone DNA sequence fragment #25	
	backbone DNA sequence fragment #26	
14	*rrn*26	yes
	backbone DNA sequence fragment #27	
	backbone DNA sequence fragment #28	
	backbone DNA sequence fragment #29	
	backbone DNA sequence fragment #30	
	backbone DNA sequence fragment #31	
	backbone DNA sequence fragment #32	
	backbone DNA sequence fragment #33	
	backbone DNA sequence fragment #34	
	backbone DNA sequence fragment #35	
	backbone DNA sequence fragment #36	
	backbone DNA sequence fragment #37	
	backbone DNA sequence fragment #38	
15	backbone DNA sequence fragment #39	yes
16	*trn*S	yes
	backbone DNA sequence fragment #40	
17	orf117	yes
	backbone DNA sequence fragment #41	
	backbone DNA sequence fragment #42	
	backbone DNA sequence fragment #43	
18	*cox*1	no
	backbone DNA sequence fragment #44	

Synteny anchor (SA) number	known sequence contained in SA	Duplicated SA
28	backbone DNA sequence fragment #56	no
29	backbone DNA sequence fragment #57	yes
30	*nad*4L	yes
	backbone DNA sequence fragment #58	
	backbone DNA sequence fragment #59	
31	*rps*4	yes
	backbone DNA sequence fragment #60	
32	*mtt*B	yes
	backbone DNA sequence fragment #62	
33	backbone DNA sequence fragment #63	yes
	backbone DNA sequence fragment #64	
	backbone DNA sequence fragment #65	
34	backbone DNA sequence fragment #66	yes
35	*rps*2A	yes
	backbone DNA sequence fragment #69	
36	backbone DNA sequence fragment #70	no
37	backbone DNA sequence fragment #71	no
38	*trn*R*°	no
	backbone DNA sequence fragment #72	
39	*atp*4	no
	backbone DNA sequence fragment #73	
40	backbone DNA sequence fragment #74	no
41	backbone DNA sequence fragment #75	no
42	backbone DNA sequence fragment #76	no
43	backbone DNA sequence fragment #77	no
44	*rpl*16	no
	*rps*3 exon 2	
	backbone DNA sequence fragment #78	
45	backbone DNA sequence fragment #80	no
46	*trn*I*	no
	backbone DNA sequence fragment #81	
47	backbone DNA sequence fragment #83	no
	backbone DNA sequence fragment #84	
48	*ccm*C	no
	*trn*L1*	
	orf107	
49	*nad*5 exon 1	no
	*nad*5 exon 2	
	backbone DNA sequence fragment #85	
50	backbone DNA sequence fragment #86	no
51	backbone DNA sequence fragment #87	no
52	*atp*8	no
	backbone DNA sequence fragment #88	
53	backbone DNA sequence fragment #89	no
54	backbone DNA sequence fragment #90	no
55	*nad*7 exon 5	no
	*nad*7 exon 4	
	*nad*7 exon 3	
	*nad*7 exon 2	
	*nad*7 exon 1	
	backbone DNA sequence fragment #91	
56	backbone DNA sequence fragment #92	no
57	backbone DNA sequence fragment #93	no
58	*trn*C	no
	*nad*5 exon 5	
	*nad*5 exon 4	
	*rps*12	
	*nad*3	
	*trn*L2*	
	*trn*S	
	backbone DNA sequence fragment #94	
	backbone DNA sequence fragment #95	

Synteny anchor (SA) number	known sequence contained in SA	Duplicated SA
19	*trn*F°	no
	backbone DNA sequence fragment #45	
20	*trn*P	yes
	*trn*E	
	*trn*I	
	*trn*D	
	*trn*N	
	backbone DNA sequence fragment #55	
21	orf99	yes
	*nad*1 exon 1	
	backbone DNA sequence fragment #54	
22	*nad*2 exon4	yes
	*nad*2 exon 5	
	backbone DNA sequence fragment #53	
23	*nad*2 exon 3	no
	*trn*M	
	*trn*Y	
	*nad*9	
	backbone DNA sequence fragment #46	
24	*ccm*FC exon 2	yes
	*ccm*FC exon 1	
	*trn*K	
	backbone DNA sequence fragment #47	
	backbone DNA sequence fragment #48	
	backbone DNA sequence fragment #49	
25	*nad*4 exon 1	no
	*nad*4 exon 2	
	*nad*4 exon 3	
	backbone DNA sequence fragment #51	
26	*nad*4 exon 4	yes
	backbone DNA sequence fragment #52	
27	*rps*3 exon 1	yes
	backbone DNA sequence fragment #79	

*pseudogenes
°chloroplastic origin

Synteny anchor (SA) number	known sequence contained in SA	Duplicated SA
59	*nad*5 exon 3	no
	*nad*1 exon 5	
	mat-r	
	*rps*1	
	*ccm*FN	
	*trn*Q	
	orf149	
	orf133	
	backbone DNA sequence fragment #96	
	backbone DNA sequence fragment #97	
	backbone DNA sequence fragment #98	
	backbone DNA sequence fragment #99	
	backbone DNA sequence fragment #100	
	backbone DNA sequence fragment #101	
	backbone DNA sequence fragment #102	
60	*atp*1	yes
	backbone DNA sequence fragment #103	
61	backbone DNA sequence fragment #104	yes
	backbone DNA sequence fragment #105	
62	backbone DNA sequence fragment #106	yes
63	*cox*2 exon 1	yes
	*cox*2 exon 2	
	backbone DNA sequence fragment #108	
64	*nad*1 exon 4	yes
	backbone DNA sequence fragment #109	
65	backbone DNA sequence fragment #110	yes
66	backbone DNA sequence fragment #111	yes
67	backbone DNA sequence fragment #112	yes
68	*atp*6	yes
	backbone DNA sequence fragment #113	
	backbone DNA sequence fragment #114	
69	*nad*6	no
	backbone DNA sequence fragment #115	

Figure S3 - Jackknife values according to conserved GSS

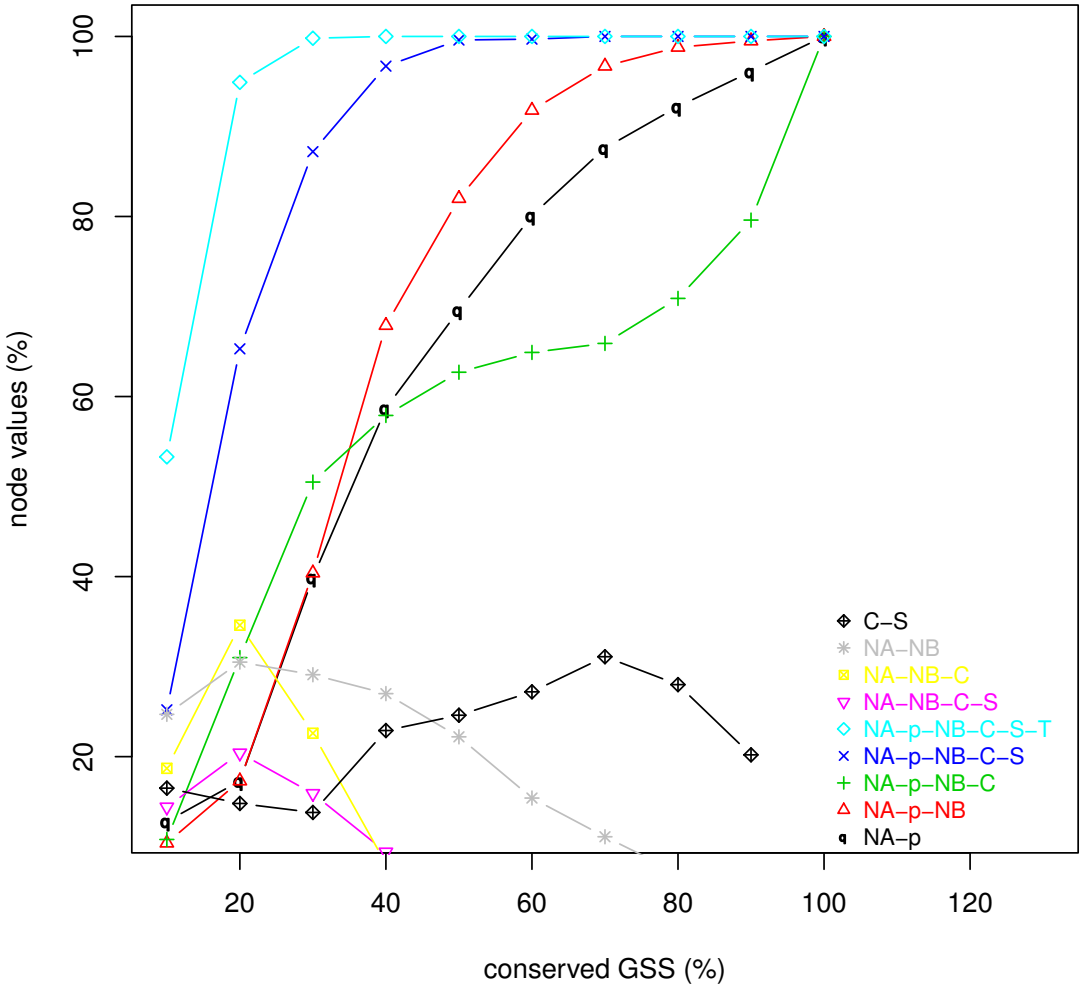


Table S4 – Cluster groups and their content

CLUSTER GROUP	CLUSTER	Synteny anchor number	subcluster
CG1	Cluster 1	3;4	
	Cluster 2	5;6;7	
CG2	Cluster 3	8;9;10	8;9 10
CG3	Cluster 4	12;13;14;15;16;17;18	12;13;14;15;16;17 18
CG4	Cluster 5	19;28;20;21;22;23;24;25;26;27;44;45;46	19 20;21;22;23;24;25;26;28 27;44 45;46
	Cluster 6	47;48	
	Cluster 7	49	
	Cluster 8	42;43	
	Cluster 9	50	
CG5	Cluster 10	39	
CG6	Cluster 11	51;52;53	51;52 53
CG7	Cluster 12	55	
	Cluster 13	54	
CG8	Cluster 14	64;65	
	Cluster 27	66;67	
CG9	Cluster 15	40;41	
CG10	Cluster 16	29;30	
	Cluster 20	63	
CG11	Cluster 17	31;32;33	31;32 33
	Cluster 18	35;36;37	35;36 37
	Cluster 19	38	
	Cluster 28	34	
CG12	Cluster 21	56;57;58	56;57 58
	Cluster 22	59	
CG13	Cluster 23	60;61;62	60;61 62
CG14	Cluster 24	68	
	Cluster 25	69	
CG15	Cluster 26	1	
CG16	Cluster 29	2	
CG17	Cluster 30	11	

Annexes Chapitre 6

Structural and content diversity of the mitochondrial genome in beet. A comparative genomic analysis

Darracq A. 1,2,3,4,5, Varré J.S. 1,4,5, Maréchal-Drouard L. 6, Courseaux A. 1,2,3, Saumitou-Laprade P. 1,2,3, Oztas S. 7, Lenoble P. 7, Vacherie B. 7, Barbe V. 7, Touzet P. 1,2,3

1 Univ Lille Nord de France, F-59000 Lille, France

2 USTL, GEPV, F-59650 Villeneuve d'Ascq, France

3 CNRS, UMR 8016, F-59650 Villeneuve d'Ascq, France

4 USTL, LIFL, F-59650Villeneuve d'Ascq, France.

5 CNRS, UMR 8022, F-59650 Villeneuve d'Ascq, France

6 INRIA Lille-Nord Europe, F-59650 Villeneuve d'Ascq

7 IBMP, UPR CNRS 2357, F-67084 Strasbourg, France

8 CEA–Institut de Genomique-Genoscope, CNRS UMR8030, F-91057 Evry, France

Abstract

Despite their monophyletic origin mitochondrial (mt) genomes of plants and animals have developed over time contrasted evolutionary paths. Animal mt genomes are generally compact and small and exhibit a high mutation rate. Conversely, mt genomes in plants exhibit a low mutation rate, low compactness, subsequent larger sizes and highly rearranged structures. Notably, this variation in structure can be seen at the species level. In the present study, we present the (nearly) whole sequences of 5 new mt genomes in *Beta* genus, 4 from *Beta vulgaris* and 1 from *Beta macrocarpa*, a sister-species, belonging to the same *Beta* section. Adding the two previously sequenced genomes of *Beta vulgaris*, we were able to assess genome complexity diversity at a species level for the first time in a eudicot species. The occurrence of male sterility is an important feature in *Beta vulgaris* breeding system (called gynodioecy), and is expected to affect mt gene/orf content and diversity through the emergence and selection of Cytoplasmic male sterility (CMS) specific mt genes. We therefore compared the two groups of mt sequenced genomes, CMSs versus non-CMSs and showed that they exhibited contrasted nucleotide diversities and divergences when compared with the

outgroup *Beta macrocarpa*. Finally we showed that the comparison of *Beta vulgaris* with monocot *Zea mays* revealed different features in the evolution of mt genomes at the species level.

Introduction

Despite their monophyletic origin (Gray et al. 1999), mitochondrial (mt) genomes of plants and animals have developed over time contrasted evolutionary paths. Animal mt genomes are generally compact and small (around 20kb) and exhibit a high mutation rate. Conversely, mt genomes in plants exhibit a low mutation rate, low compactness, subsequent larger sizes (from 200 to 900 kb, for whole sequenced genomes; Alverson et al. 2010), and highly rearranged structures. Notably, this variation in size and structure is observable at the species level, as illustrates the recent study in maize, where 5 genomes have been totally sequenced (Allen et al. 2007). In maize, mt genome size varies from 535,825 bp to 739,719 bp, mainly due to large duplications. The genomes are highly rearranged when compared with one another, while the level of substitution among species is low, even at the genus level (Darracq et al. 2010). Among the sequenced mt genomes in maize, genomes are associated with cytoplasmic male sterility (CMS), called C, S and T and two are considered as fertile.

CMS is an interesting case of nuclear-cytoplasmic interaction. Described in several crop species (Chase 2007), it is commonly found in wild populations. Species that bear CMS and therefore exhibit a sexual polymorphism, hermaphrodites and females, are called gynodioecious. Wild beet, *Beta vulgaris* ssp. *maritima*, the wild relative of sugar beet, is one of these species. In the species, at least 4 CMSs, Owen, E, G, and H, have been described out of a total of 20 mt types, through the use of RFLP markers (Cuguen et al. 1994). Theoritical models suggest that the maintenance of male-sterile plants in populations imply the relative fitness advantage of being female, for example by producing more seeds through the economy of pollen production, consequently inducing a selection of CMS in population. In wild beet, the study of gynodioecy occurence in populations suggests balancing selection dynamics that favors CMSs when they are rare. When CMS becomes common, it creates a selection pressure that favors the recruitment of the corresponding nuclear restorer alleles, restoring pollen production and generating restored hermaphrodite (Dufaÿ et al. 2007; Dufaÿ et al. 2009). It must be noted that this evolutionary dynamics of gynodioecy is expected to leave signature in mt gene nucleotide diversity (Charlesworth 2002; Touzet and Delph 2009). In addition it raises the question of the events and evolutionary forces that led to the emergence of sterilizing genes. In beet, a phylogeny built from chloroplastic polymorphisms have suggested the independence of the 4 CMSs all derivated from an ancestral fertile cytoplasm. This was corroborated from the polymorphism at an mt intergenic sequence

(Nishizawa et al. 2007). One CMS (Owen/TK81-MS) and one non-CMS (TK81-O) mt genomes have previously been sequenced (Kubo et al. 2000; Satoh et al. 2004). They exhibited a large variation in size and gene order. In the present study, we sequenced 5 additional mt genomes, 2 non-CMSs (A and B), 2 CMSs (E and G) and a mt genome of *Beta macrocarpa*, a sister-species, belonging to the same taxonomic section. We describe the diversity of the mt genome in size, content and structure in this eudicot species, compare it with *Zea mays* mt genome, a monocot. Then we show that the three analyzed CMS genomes belong to a single sterile lineage. This lineage seems to have known an increase of synonymous substitution rate that could have favored the emergence of sterilizing mutations.

MATERIALS AND METHODS

Mitochondrial DNA preparation

For mitogenomes of *Beta vulgaris* ssp *maritima*, maternal progenies were collected in populations from free pollinated plants. Maternal plants were characterized for their mitochondrial according to Cuguen et al. (1994). Mitogenomes A and B are fertile cytoplasms while E and G are sterilizing ones. As an outgroup we used *Beta macrocarpa* that belongs to the same *Beta* section (accession from Morocco, IDBBNR 8549). Mitochondrial DNAs were extracted from roots for *Beta vulgaris* ssp *maritima*, and from leaves for *Beta macrocarpa* using procedures described in Scotti *et al.* (2001).

Library construction, sequencing and finishing

To generate random fragments, the mtDNAs were mechanically sheared and 5 kb generated inserts were cloned into pcDNA2.1 plasmid vector (Invitrogen). Vector DNAs were purified and end-sequenced using dye-terminator chemistries on ABI3730 sequencers until an average of 12 fold coverage for each genome. A pre-assembly was made without repeat sequences as described by Vallenet *et al.* (2008) using Phred/Phrap/Consed software package (www.phrap.com). The finishing steps were achieved by primer walking, transposition and PCR amplifications.

Annotation

A local database was built to facilitate genome annotation. The annotation of genes, tRNAs, rRNAs of new sequenced genomes was conducted using as references the whole sequences of mt genomes from beet (TK81-O [GenBank:BA000009] and TK81-MS [GenBank:BA000024]), *Arabidopsis thaliana* [GenBank:Y08501] and tobacco [GenBank:BA000024], as well as whole sequences of cp genomes of tobacco [GenBank:Z00044], spinach [GenBank:NC_002202]. Edited sites on genes were determined

using the annotation of TK81-O that has beneficiated from an experimental validation (Mower et al. 2006). Annotated tRNAs were validated using the software tRNAScan-SE version 1.23 (Lowe and Eddy 1997).

ORFS with a minimum size of 300 bp were searched and then first compared with known genes of the reference genomes of the database, if they did not match with a known gene in the database, then Blast analyses were conducted on Genbank non redundant database (April 2010). In order to find chimeric ORFs, ORFs were compared with genes, tRNAs, rRNAs from TK81-0 using YASS softwaer (Noe and Kucherov 2005) with an E-value of 0.1. Only matches of 100% were kept. In addition, Blast analyses were conducted with a word size of 16 and an E-value >0.1.

Genome complexity

In order to establish the genome complexity defined as the complete sequence information found in a given genome without duplicates (>500 bp), a YASS analysis was conducted on each genome in order to detect any duplication with a size larger than 500 bp following an E -value < 1e-30 with a score of +1 for matches and -3 for substitutions.

Plastid sequences

Each genome was analyzed through YASS against the spinach chloroplast genome (parameters, E-value of 1e-10; with a score of +1 for matches and -3 for substitutions). We kept sequences larger than 30 bp and with less than 10% of mutation. Note that with this criteria, cp-tRNA like sequences were included for the calculation of the ratio of plastid sequences in the genome.

Phylogeny analyses

In order to compare mitogenomes at the sequence level, all common sequences between the genome complexities were searched using Mauve (seed=9, minimum island=15, maximum backbone gap size=15, minimum backbone gap size=50) (Darling et al. 2004). The sequences were concatenated in mt genome A order for phylogenetic analysis. Neighbor-Joining analyses were performed using BIONJ (Gascuel 1997). Parameters used are bootstrap 1000X and Kimura-2 parameters distance for correction. Maximum likelihood and molecular clock were tested with TREE-PUZZLE (Schmidt et al. 2002) using the nucleotide model of Hasegawa-Kishino-Yano (HKY85) (Hasegawa et al. 1985).

Data analyses

On concatenated 29 protein coding sequences, summary statistics on nucleotide diversity and divergence were calculated using DnaSP version 4.10.9 (Rozas et al. 2003): π_s, the average

number of pairwise synonymous substitutions per synonymous site between CMS or non-CMS *Beta vulgaris* genomes; πa, the average number of pairwise non-synonymous substitutions per non-synonymous site between *Beta vulgaris* genomes; Ks and Ka, the synonymous and non synonymous nucleotide sequence divergences with *Beta macrocarpa*. Estimated size of G-*cox1* was calculated from Expasy's compute pI/MW program (http://scansite.mit.edu/cgi-bin/calcpi). Statistic analyses were conducted with Minitab software 13.20 (Minitab).

RESULTS

Genome size and composition

Using a shotgun-sequencing strategy, similar to the one that succeeded in the whole sequencing of mt genomes in maize (Allen et al. 2007), we were able to achieve the whole sequencing of the two wild fertile mt genomes A and B, and the near completion of CMSs E and G from *Beta vulgaris* ssp *maritima* and a mt genome of *Beta macrocarpa*, a sister species. It was possible for A and B to reconstruct a master circular form, while for CMSs E and G, and *Beta macrocarpa*, each genome was reconstructed in two contigs (one circular, one linear, except macro with two linear) (figure 1). Total genome sizes varied from 341,257 bp for CMS G to 378,457 bp for CMS E, which are closer to the size of the previously sequenced fertile mt genome TK81-O (368,801 bp) than the CMS TK81-MS (501,020 bp). Table 1 summarizes features of the five newly sequenced genomes and the two previously described ones. Over the 7 mt genomes, there is a ratio of 1.47 between the largest and the shortest genomes. The median GC content is 43.89%, similar to the value of other plant mt genomes (Allen et al. 2007). The large variation of size is mainly due to duplications that represent from 4.55% for G to 29.98% for TK81-MS (median = 8.13%) (Pearson coefficient of correlation r = 0.968; p = 0.000). Genome complexity, which is the non redundant genetic content of each genome, is less variable among genomes, from 325,716 bp for G to 357,125 bp for E (size ratio of 1.10 between the largest and the shortest) with a median-value of 335,262 bp. Interestingly genome complexity is not correlated with genome size (Pearson coefficient of correlation r = 0.542; p = 0.209).

Integrated Plastid sequences

The total length of plastid inserted sequences in mt genomes vary from 4202 bp in CMS G to 8123 bp in CMS E, representing from 1.27% to 2.33 of genome complexities (table 1). The size of inserted cp sequences vary from 22 bp in all genomes to 3368 bp in mt genomes B, E and macro (supplementary table 1), with a median size per genome varying from 34 in G to 41 in TK81-O and macro.

Conserved genes

All the analyzed genomes contain the same 29 protein coding genes: 18 genes involved in ATP-generating electron transport : 9 in Complex I (NAD 1, 2, 3, 4, 4L, 5, 6, 7, 9), 1 in Complex III (COB), 3 in Complex IV (COX 1, 2, 3) and 5 in Complex V (ATP 1, 6, 8, 9, orf25); 3 genes involved in biogenesis of cytochrome c (CCM B, FC, FN); 6 genes coding for ribosomal proteins (RPL5, RPS 3, 4, 7, 12, 13) ; one gene involved in independent membrane targeting and translocation system (*tatC*) and one maturase (*mat-r*).

Using TK81-O as reference, a total of 33 edited sites was predicted on protein coding genes. There are 3 cases where ACG is edited to the start codon ATG for genes *atp6*, *nad1* and *nad4L*; 2 cases, CAA (*atp6*) or CGA (*atp9*), are edited into a stop codon, TAA or TGA respectively. The gene *tatC* does not contain a DNA-encoded AUG start codon and does not seem to be edit. Its start codon is AUA. The stop codon is UAA for 16 genes (including edited *atp): atp* 1, 6, 8, *cox* 1, 2, *nad* 1, 2, 3, 4L, 5, 6, 9 -except for one copy of G- , *rpl5* , rps 3, 4, 7, UGA for 10 genes (*atp9* ,ccm B, FC, FN, *cob*, *cox2* -for G and one copy of TK81-MS-, *cox3*, *nad4*, *rps* 12, 13) and UAG for 4 genes (*mat-r*, *nad7*, one copy of *nad9* in G and TK81-MS, *tatC*).

As in Kubo et al. (2000), 20 introns were found for 7 protein coding genes, representing in average 6% of the genome complexity. Six are trans-splicing introns for 3 genes (*nad* 1, 2, 5) and 14 are cis-splicing introns for 6 genes (*nad* 1, 2, 4, 5, 7 and *ccmFC*). It must be noted that contrarily to other plants, *rps3* lacks any intron in beet, even though a second copy of what is homologous to exon1 is found in the beet mt genomes (counted as pseudo-exon in the present study).

The *Beta* genomes contain 3 RNA genes (rrn 5S, 8S and 26S). While *rrn26S* was found in 3 copies in TK81-O, TK81-MS, A and B, it was found in 2 copies in G and macro, and in only one copy in E. This lack of copy is most probably due the remaining gaps of the 3 unfinished genomes.

Eighteen tRNAs were found as well as 5 potential pseudo ones. Among the 18 tRNAs, 11 are native, 6 are cp-like, 1 is of unknown origin: *trnC2* (Kubo et al. 2000). The cp-like in *trnW* is most probably present also in G (in a non integrated contig of 5 kb, data not shown). Among the 5 pseudo-tRNAs as defined by Kubo et al. (2000), cp-like *trnP* and *trnV* were not found in unfinished mt genomes macrocarpa and G, and *trnI* on E.

Polymorphic protein coding genes

We found 19 protein coding genes that were found to be polymorphic among the 7 analyzed genomes (table 3). Overall, 53 mutations were revealed and 1 gene that had a different 5' part

among the genomes (atp6). 23 were specific to G, 14 to TK81-MS, 6 to TK81-O, 3 to E and 1 to macro. 4 variants were shared by TK81-MS and G (then the alternative allele by TK81-O, A, B, E, macro), 1 was shared by TK81-MS, G and E (then the alternative allele by TK81-O, A, B, macro). On one site, an allele unique to TK81-MS, one other allele was shared by A and B, a third one shared by the remaining ones. Overall, 14 were synonymous mutations and 39 were non-synonymous. Among the 39 non-synonymous mutations, one modified the start codon in cox1 of CMS-G, resulting in a shorter coding sequence of 87 bp or a larger one (with an alternative start codon found upstream). Still in CMS G, as previously described by our team (Ducos et al. 2001), one mutation generated a premature stop codon in exon2 of *cox2* and a mutation modified the stop codon of *nad9*, resulting in a longer coding sequence of 42 bp. Three polymorphisms specific to TK81-O (on *nad2* exon4, *rps3*, *rps12*) are expected to be edited resulting on a non modified protein sequence.

As described by Satoh et al (2004), two copies of *cox2* exon2 were found in TK81-MS, with one copy identical to the *cox2* exon2 found in the other genomes, and the other copy being identical in the 158 first nucleotides and then composed of a specific sequence of 506 bp.

Finally, as described by Yamamoto et al. (2005), *atp6* is longer in TK81-MS with an additional 1172 bp upstream. CMS E is found to exhibit also an additional *atp6* 5'-leader sequence with is 88% identical to TK81-MS one and 1% of gap. This specific sequence to CMS-E is identical *atp6* found in to I-12CMS(3), a wild CMS found in Pakistan , suggesting the identity between CMS E and I-12 CMS(3) (Onodera et al. 1999).

ORFS

Out of the total of 235 ORFs (putative coding sequence with a minimum size of 300 bp), 34 were found to overlap genes, 20 to overlap inserted plastid sequences and 13 were chimeric (with at least a 16 bp sequence similar to an mt gene) (supplementary table 2 and table 4 for chimeric ORFs). Considering specific ORFs to CMSs, potential candidate sterilizing genes, we found 28 specific ORFs to TK81-MS, 21 to CMS G, 11 to CMS E. 4 ORFs are shared between TK81-MS and G, 3 to TK81-MS and CMS E, 2 to CMSs E and G, and 3 common to all three CMSs.

In reference to Satoh et al. (2004) where TK81-O and TK81-MS were compared, *orf*317 found in TK81-O but not in TK81-MS, is also found in other genomes, TK81-MS having a smaller corresponding ORF (*orf*221). *Orf*518 found in TK81-O is also found in macrocarpa, and is absent in all other genomes. *Orf*324 and *orf*119c are specific to TK81-MS, *orf*214 are only shared by TK81-O and TK81-MS and *orf*129b and *orf*145 only found in TK81-O but corresponding ORFs (*orf*122b and *orf*176b) described as unique to TK81-MS were found in all other genomes.

Among the specific ORFs found in E, we found *orf*129, described by Yamomoto et al. (2008) as a candidate sterilizing gene of in I-12CMS(3), adding an new argument for both cytoplasm identity or at least a close genetic affinity.

Concerning the ORFs that would be specific or shared among non CMS, it must be noted that due to the incompleteness of E and G genomes, their specificity are not guaranteed. This in mind, 27 ORFs were found to be specific to TK81-O, 10 shared by all genomes but TK81-MS, 5 found in all genomes but TK81-MS and G, 3 specific to TK81-O, A, and macro 9 in all but G, 1 specific to macro, 1 shared by A and macro, 1 specific to A (suppl. table2).

Table 4 summarizes the chimeric ORFs found among the 7 genomes. Only TK81-MS and TK81-O exhibited chimeric ORFs, 2 ORFs for each one.

Repeated sequences

Figure 1 gives a representation of duplicates found in beet mt genomes. Overall, 24 repeats larger than 0.5 kb were detected over the 7 mt genomes (Table 5). TK81-MS, the larger genome is characterized by the two largest duplications, R87 and R51, 87 kbp and 51 kpb large respectively. TK81-O, A and B share a common duplication which is 23 kbp large (R23) which is a sub-fragment of R87. All the other duplications are shorter than 10 kb and are usually unique to a given genome even though they share common sequences with duplications found in other genomes. It can be noted that duplication R6.0 found in 3 copies in A and B, is partially included in R6.2 of TK81-O also found in 3 copies. Interestingly, TK81-MS exhibits also a triplication with R7.6 which also partially includes R6.2 and R6.0. This could be the signature of an ancestral triplication with subsequent rearrangement in the *Beta* mt genomes.

Evolutionary analysis of *Beta* mt genomes

We constructed a concatenated sequence for each *Beta* mt genome composed by the genomic sequence shared by all 7 *Beta* genomes. The consensus sequence length is 267160 bp (including gaps) and, for each genome, the length varies between 266,554 bp (TK81-O) to 266,763 bp (G) (mean of 266,702 bp and median of 266,694 bp). This enabled us to measure pairwise substitution rates (Table 6a) and pairwise indel rates (indels less than 16 bp) (Table 6b) and mutation (substitution and indel) rates (Table 6c) among genomes. Pairwise substitution rates vary from 1.087 substitutions per 10,000 bp (between A and B) to 34.075 substitutions per 10,000 bp (between TK81-MS and CMS G), with a median value of 26.435. When considering small indels, the pairwise indel rates vary from 0.224 (between A and B) to 25.660 (between TK81-O and TK81-MS) indels per 10,000 bp with a median-value of 15.785. Globally, when considering substitutions and small indels, the pairwise mutation rates vary from 1.311 mutations per 10,000 bp (between A and B) to 56.454 mutations per 10,000

bp (between TK81-MS and CMS G) with a median-value of 44.620.

Using the same method used for *Beta* mt genomes, we computed substitution, indel and mutation matrices among 8 published *Zea* mt genomes (Allen et al. 2007; Allen et al. unpublished results; Darracq et al. 2010) (Table 7a, 7b, 7c). Pairwise substitution rates vary from 0.550 to 18.292 substitutions per 10,000 bp, with a median-value of 5.807, a value 5 times smaller than the median substitution rate in *Beta* genomes. Conversely, the pairwise indel rates vary from 2.514 to 164.876 indels per 10,000 bp, with a median-value of 28.460 almost twice as high as in *Beta*'s. Globally, the pairwise mutation rates vary from 3.064 to 182.203 mutations per 10,000 bp, with a median-value of 34.912 that is in the order of magnitude of *Beta*'s.

Focusing on intra-species variation (6 genomes in *Beta vulgaris,* 5 genomes in *Zea mays*) substitution and indel rates are highly correlated in *Beta vulgaris* (r =0.989 p = 0.000) and to a lesser extent in *Zea mays* mt genomes (r = 0.774 p = 0.009). Notably the mean rate of substitution is significantly higher in *Beta vulgaris* (18.9) than in *Zea mays* (4.80) (T = 4.18 p = 0.001 df = 14), while the mean indel (13.66 in *Bv*, 18.78 in *Zm*) and mutation (23.57 in *Bv*, 32.7 in *Zm*) rates are not significantly different among the species (T = -1.53 p = 0.142 df = 20, T = -1.49 p = 0.153 df = 20, respectively).

Phylogeny of mt genomes in Beta

Using the matrix of pairwise substitution rates, a phylogenetic tree rooted with macro was built with strong bootstrap values (figure 2). The tree is composed of two clades, one formed by A, B and TK81-O constituting a non-CMS clade, the other one forming a CMS clade with E, G and TK81-MS. Notably, G and TK81-MS form long branches. We assessed whether the evolutionary dynamics was different between CMS and non CMS mt genomes in *Beta vulgaris* ssp *maritima.* On all the 29 protein coding genes (total size of 29344 bp), we assessed the average synonymous and non-synonymous nucleotide diversities (π_S and π_a) within CMS (E, G, TK81-MS) or non-CMS genomes (A, B, TK81-O), and their average synonymous and non-synonymous nucleotide divergence (Ks and Ka) with macrocarpa, as well as individual K_S and K_a for each genome (Table 8).

Neutral divergence and diversity

When considering the average K_S of CMSs and non CMSs with macro, K_S is in average 6.6 higher in CMSs than in non-CMSs. The same phenomenon was observed at the diversity level (π_S): CMSs are 5 times more diverse in average at mitochondrial genes than non CMSs, suggesting a higher rate of synonymous substitution rate in the sterile lineage.

Non-synonymous divergence and diversity

The non-synonymous divergence of mt genes with macrocarpa (K_a) was in average 6 times higher in CMSs than in non-CMSs, but with a similar ratio K_a/K_s. Nucleotide non synonymous diversities were around 5 times higher in CMSs than in non-CMS one, resulting in a similar π_a/π_s among the two groups. Among CMSs, G exhibits a high K_a, and a subsequent K_a/K_s ratio higher than 1 suggesting positive selection occurring on this peculiar genome.

Discussion

In the present study, we present the (nearly) whole sequences of 5 new mt genomes in *Beta* genus, 4 from *Beta vulgaris* and 1 from *Beta macrocarpa*, a sister-species, belonging to the same *Beta* section. Adding the two previously sequenced genomes of *Beta vulgaris*, we were able to assess genome complexity diversity at a species level for the first time in a eudicot species. The occurrence of male sterility is an important feature in *Beta vulgaris* breeding system (called gynodioecy), and is expected to affect mt gene/orf content and diversity through the emergence and selection of CMS specific mt genes (Charlesworth 2002; Touzet & Delph 2009). We therefore compared the two groups of mt sequenced genomes, CMSs versus non-CMSs and showed that they exhibited contrasted nucleotide diversities and divergence when compared with the outgroup *Beta macrocarpa*. Finally we showed that the comparison of *Beta vulgaris* with monocot *Zea mays* revealed different features in the evolution of mt genomes at the species level.

CMS in Beta vulgaris

As no general nomenclature is available for *Beta vulgaris* mt genomes, a secondary output of the present study is the possibility to identify CMSs that had been characterized by independent studies. Consequently, it seems that the CMS that is called E in the present study, following Cuguen *et al.* (1994), corresponds to I12-CMS(3) as called by the Hokkaido's team (Onodera *et al.* 1999; Yamamoto *et al.* 2008). Indeed, CMS E exhibits specific 5'leader atp6 sequence and orf129 that are only found in I12-CMS(3) (Onodera et al. 1999; Yamamoto et al. 2008). CMS E is the most frequent source of CMS in wild beet populations of European Coasts (Dufay et al. 2009; unpublished results), and thus seems to be found on a larger geographical scale as I12-CMS(3) has been found in wild populations from Pakistan. It also addresses the question of the number of CMSs present in beet, where to our knowledge only 4 have been found: E/I12CMS(3), G, Owen/Sv/TK81-MS and H, out of a total of 20 different mt genomes (Desplanque *et al.* 2000), most of them being fertile, like A, B and TK81-O.

Previous studies have tentatively proposed candidates genes of male sterility in beet.

For CMS Owen/Sv /TK81-MS, the CMS that has been widely used in sugar beet breeding, Yamamoto et al. (2005) have shown that among the 4 specific ORFs detected when TK81-MS mt genome was compared with T81-O, only the ORF corresponding to a peculiar 5'-leader sequence of *atp6* (preSatp6), was expressed at the protein level. It codes for a 35 kDa polypeptide that is specific to the Owen CMS. However no effect of nuclear restoration was detected on the size or the amount of the preSATP6 polypeptide, and no transformation experiment has validated the sterilizing effect of preSATP6 (Yamamoto et al. 2008).

For CMS E/I-12CMS(3), Yamamoto et al. (2008) demonstrated that the E-specific *orf*129 was transcribed and coded for a specific 12 kDa polypeptide, that accumulated in mitochondria of flower, root and leaf. Transgenic expression in tobacco of orf129 fused with a mitochondrial targeting-pre-sequence led to male sterile plants, demonstrating the sterilizing effect of ORF129. The question remains on the action of restorer loci on this CMS since no effect on ORF129 abundance has been detected when plants were restored.

For CMS G, in a former study (Ducos *et al.* 2001), we had shown that CMS G exhibited a modified genomic *cox2* sequence that resulted in a truncated protein at the C terminus. In addition, it was shown that the *in vitro* activity of cytochrome *c* oxidase was reduced by 50% in leaves, suggesting a possible effect of the observed mutations on the complex activity. Finally, we were not able to recover complex IV by blue native PAGE in CMS G plants, raising the question of the stability and/or the physico-chemical propriety of the G-complex IV. In the present study, as expected, we found the modified *cox2* sequence but also mutations on *cox1* and *cox3*. For *cox1*, a mutation at the start codon, which is commonly found in the other beet genomes, can potentially result in the translation of a longer protein with an extended N-terminus (660 aa versus 524 aa; estimated weight of 73.5 kDa versus 57.6 kDa) or a shorter one (495 aa; estimated weight 54.2 kDa). It must be noted that a recent study on the same genome confirms that there is only one copy of *cox1* (Kawanishi et al. 2010). We did not detect a long variant on previous SDS-PAGE from *in organello* S-labeled proteins in G-CMS, suggesting that the translated form of G-COX1 might be the shortest one, with a size variation that could not be detected in SDS-PAGE conditions. However in the truncated N-terminus, amino acids are expected to be involved in subunit I/III interface (S10), D-pathway (L19), or subunit I/VIIc interface (A25). In addition, two non synonymous polymorphisms were found: R180/Q, F403/V, two codons not associated to any known function. For *cox3*, one non-synonymous polymorphism was detected leading to an I/L variation on codon 51, with no known associated function. The observed polymorphism of complex IV could be the result of relax in selection, enabling the accumulation on non synonymous mutations, once the sterilizing mutations have been selected through disruption of COX activity. If so, it could imply compensatory mutations on COX nuclear genes in fertility restoration. More generally it addresses the question of the evolution of nuclear-mitochondrial interaction and co-

evolution of protein coding genes involved in the same respiratory complex.

Phylogeny of mt genomes

A phylogeny based on chloroplastic sequences by our group (Fenart et al. 2006) had suggested that the sterile cytoplasms had emerged independently from a non sterile cytoplasm. However the resolution was very low due to the lack of polymorphism. In the present study, through pairwise substitution rate among the mitochondrial genomes, we were able to build a robust phylogenetic tree where it appeared two clusters, one composed of the 3 CMSs and one of the 3 non CMSs. We then tested the two distinct groups had distinct features in particular in terms of nucleotide diversity. CMss seems to exhibit higher synonymous diversity and divergence wit Beta macrocrapa, suggesting a higher mutation rate occurring in the sterile cytoplasm, evolutionary factor that could have favored the emergence of mitochondrial novelties. Interestingly, selection favoring any mutations blocking pollen production seems to have left the signature of positive selection especially on CMS G with the accumulation of non synonymous mutations on protein coding genes in CMS G, in particular those involved in complex IV.

Beta versus *Zea: the mutation burden hypothesis*

When comparing the two cases of intra-species mt genome complexity diversity, one from a monocot species, Zea mays, and the other one from an eudicot, *Beta vulgaris*, common as well as striking different features was observed. In both species, a variation of genome size was mainly due to repeated sequences. Therefore genome complexities were comparable among genomes within a given species. In addition, in both species, as commonly found, gene (or syntenic anchor) order was shuffled. However, it must be noted that while gene contents are similar among species, the median genome sizes and genome complexities were different between the two species. Indeed, median genome size in *Zea* (including all 3 teosinte species) is 569,992 bp, while it is only 367,943 bp in *Beta*. In the same vein, median genome complexity is 518,699 bp in *Zea* while only 341,115 bp in *Beta*. This raises the question of the forces responsible of the difference in size between the two species. The mutation pressure hypothesis has postulated that low mutation rate found in mitochondrial genomes in plants facilitates the accumulation of non coding sequences and hence the overall growth of their genome size, when compared with their animal counterparts (Lynch et al. 2006). The higher substitution rate found among low size *Beta* genomes when compared with large-size *Zea* genomes could be indicative of a difference of mitochondrial mutation rate among the two species. If it is the case, the observation would be congruent with the mutation burden hypothesis which predicts a negative correlation between mutation rate and size of the genomes.

Conclusion

Plant mitochondrial genomes are reputed to evolve slowly in sequence but fast in structure (Palmer and Herbon 1988). It is therefore crucial to analyze genome structure diversity in order to be able to retrace the history of the mt genome at the species level. The main difficulty relies on the occurrence of identical duplicates due to low mutation rate, making difficult the distinction between paralogs and orthologs. We have recently described a new method to deal with duplicates in order to propose evolutionary scenario of rearrangement that led to present mt genomes in *Zea* (Darracq et al. 2010). The same methodology will be applied on the *Beta* mt genomes in order to assess the main rearrangement involved in the shaping of mt genomes in the species.

Acknowledgements

We wish to thank Benjamin Brachi for the scripts in R (Figure 1). This work was funded by a grant from the Agence Nationale de la Recherche (ANR-06-JCJC-0074) and a grant from Région Nord Pas de Calais and the European Community (Arcir PLANT-TEQ6) to PT, a grant from PPF Bioinformatique of University of Lille1 to PT and J-SV, and a PhD fellowship from French Research Ministry to AD.

References

Allen, J.O., Fauron, C.M., Minx, P., *et al.* (16 co-authors). (2007) Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. *Genetics*, **177**, 1173-1192.

Alverson, A.J, Wei, X., Rice, D.W, Stern, D.B, Barry, K. and Palmer, J.D. (2010) Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol. Biol. Evol., **27**, 1436-1448.

Charlesworth, D. (2002) What maintains male-sterility factors in plant populations. Heredity, **89**, 408-409.

Chase, C. D. (2007) Cytoplasmic male sterility: a window to the world of plant mitochondrial-nuclear interactions. Trends Genet., **23**, 81-90.

Cuguen, J., Wattier, R., Saumitou-Laprade, P., Forcioli, D., Mörchen, M., Van Dijk, H. and Vernet, P. (1994) Gynodioecy and mitochondrial DNA polymorphism in natural populations of *Beta vulgaris* ssp *maritima*. *Gen. Sel. Evol,.* **26**, 87-101.

Darling, A.C., Mau, B., Blatter, F.R. and Perna, N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.,* **14**(7), 1394-1403.

Darracq, A., Varré, J-S. and Touzet, P. (2010) A scenario of mitochondrial genome evolution

in maize based on rearrangement events. *BMC Genomics,* **11**, 233.

Desplanque, B., Viard, F., Bernard, J., Forcioli, D., Saumitou-Laprade, P., Cuguen, J. and Van Dijk, H. (2000) The linkage disequilibrium between chloroplast DNA and mitochondrial DNA haplotypes in *Beta vulgaris* ssp. *maritima* (L.): the usefulness of both genomes for population genetic studies. *Mol. Ecol.,* **9**, 141-154.

Ducos, E., Touzet, P. and Boutry, M. (2001) The male sterile G cytoplasm of wild beet displays modified mitochondrial respiratory complexes. Plant J., **26**, 171-180.

Dufaÿ, M., Touzet, P., Maurice, S. and Cuguen, J. (2007) Modelling the maintenance of a male fertile cytoplasm in a gynodioecious population. *Heredity,* **99**, 349-356.

Dufaÿ, M., Cuguen, J., Arnaud, J-F. and Touzet, P. (2009) Sex ratio variation among gynodioecious populations of wild beet: can it be explained by negative frequency-dependent selection? *Evolutio*n, **63**, 1483-1497.

Fénart, S., Touzet, P., Arnaud, J-F. and Cuguen, J. (2006) Emergence of gynodioecy in wild beet (*Beta vulgaris* ssp. *maritima* L.): a genealogical approach using chloroplastic nucleotide sequences. *Proceedings of the Royal Society of London, Series B,* **273**, 1391-1398 .

Gascuel, O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*., **14**, 685-695.

Gray, M.W., Burger, G. and Lang, B.F. (1999) Mitochondrial Evolution**.** *Science*, **283**, 1476-1481.

Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22(2),** 160-174.

Kawanishi, Y. Shinada, H., Matsunaga, M., Masaki, Y. Mikami, T. and Kubo, T. (2010) A new source of cytoplasmic male sterility found in wild beet and its relationship to other CMS types. *Genome*, **53**, 251-256.

Kubo, T., Nishizawa, S., Sugawara, A., Itchoda, N., Estiati, A. & Mikami, T. (2000)The complete nucleotide sequence of the mitochondrial genome of sugar beet (*Beta vulgaris* L.) reveals a novel gene for tRNACys(GCA). *Nucleic Acids Res.,* **28**, 2571-2576.

Lynch, M., Koskella, B. and Schaak, S. (2006) Mutation pressure and the evolution of organelle genomic architecture. *Science*, **311**, 1727-1730.

Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res, 25, 955-964.

Mower JP, Palmer JD. Patterns of partial RNA editing in mitochondrial genes of Beta vulgaris. Mol Genet Genomics 2006, 276:285-293.

Nishizawa, S., Mikami, T. and Kubo, T. (2007) Mitochondrial DNA phylogeny of cultivated and wild beets: relationships among cytoplasmic male-sterility inducing and nonsterilizing cytoplasms. *Genetics*, **177**, 1703-1712.

Noe, L. and Kucherov, G. (2005) YASS: enhancing the sensitivity of dna similarity search. *Nucleic Acids Res*., **33**(2), W540-W543.

Palmer, J.D. and Herbon, L.A. (1988) Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. *J Mol Evol.*, **28**(1), 87-97.

Onodera,Y., Yamamoto,M.P., Kubo,T. and Mikami,T. (1999) Heterogeneity of the atp6 presequences in normal and different sources of male-sterile cytoplasms of sugar beet. J. Plant Physiol. 155, 656-660.

Rozas, J., Sanshez-Delbarrio, J. C., Messeguer, X. and Rozas, R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics, **19**, 2496-2497.

Satoh, M., Kubo, T., Nishizawa, S., Estiati, A., Itchoda, N. & Mikami, T. (2004) The cytoplasmic male-sterile type and normal type mitochondrial genomes of sugar beet share the same complement of genes of known function but differ in the content of expressed ORFs. *Mol. Genet. Genomics,* **272**, 247–256.

Schmidt, H. A., Strimmer, K., Vingron, M. and von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502-504.

Scotti, N., Cardi, T. and Maréchal-Drouard, L. (2001) Mitochondrial DNA and RNA isolation from small amounts of potato tissue. *Plant Mol Biol Rep.,* **19**, 1-8.

Touzet, P. and Delph, L.F. (2009) The effect of breeding system on polymorphism in mitochondrial genes of *Silene*. *Genetics,* **181**, 631-644.

Vallenet, D., Nordmann, P., Barbe, V., Poirel, L., Mangenot, S., *et al.* (2008) Comparative Analysis of Acinetobacters: Three Genomes for Three Lifestyles. *PLoS ONE* **3**(3), e1805. doi:10.1371/journal.pone.0001805 .

Yamamoto, M. P., Kubo, T. and Mikami, T. (2005) The 5'-leader sequence of sugar beet mitochondrial atp6 encodes a novel polypeptide that is characteristic of Owen cytoplasmic male sterility. *Mol. Gen. Genom*., **273**, 342-349.

Yamamoto, M. P., Shinada, H., Onodera, Y., Komaki, C., Mikami, T. and Kubo T. (2008) A male-sterility-associated mitochondrial protein in wild beets causes pollen disruption in transgenic plants. *Plant. J.*, **54**(6), 1027-1036.

Figure legends

Figure 1 - *Beta* mitochondrial genome representations.Beta mitogenomes are represented in linear form. Internal boxes represent duplicate regions and external boxes represent genes.

Figure 2 - Phylogenetic tree for *Beta* mitogenomes. Phylogenetic tree was constructed using BIONJ and TREE-PUZZLE. The tree was rooted using *Beta macrocarpa*. Branch lengths are proportional to substitution rates. Bootstrap values (upper values for distance and lower values for likelihood) are reported.

Tables

Table 1 - Portions of *Beta* mitochondrial genomes present as genes and ORFs

	Genomes						
Genome features	**TK81-O**	**TK81-MS**	**A**	**B**	**CMS-E**	**CMS-G**	**macro**
Genomes							
Total genome size	368801	501020	364950	367943	378457 circular : 268616 linear : 109841	341257 circular : 269136 linear : 72121	366580 linear : 266432 linear : 100148
%GC	43.86	43.89	43.91	43.89	43.88	43.92	43.96
Total repeated sequence	34313	150214	29688	37461	21332	15541	19613
Total repeated sequence, % of total genome	9.30	29.98	8.13	10.18	5.64	4.55	5.35
Genome complexity	334488	350806	335262	330482	357125	325716	346997
Genes							

Protein genes	27693	37485	27693	28593	28845	28118	27693
Single-copy protein genes	27693	28839/29142	27693	27693	28845	28118	27693
Cis introns	18727	30641	18749	18749	18748	18746	18749
Single-copy cis introns	18727	18733	18749	18749	18748	18746	18749
rRNA genes	12065	12065	12065	12065	5389	8727	8727
Single copy rRNA genes	5389	5389	5389	5389	5389	5389	5389
tRNA genes	1746	2282	1746	1746	1746	1453	1599
Single-copy tRNA genes	1449	1449	1449	1449	1449	1303	1449
Pseudogenes and pseudo-exons	1180	1177	1180	1180	1180	524	1180
Single-copy pseudogenes and pseudo-exons	1180	1106	1180	1180	1180	524	1180
Coding totals							
Total known genes	41457	51760	41432	42332	35908	38298	37947
Total single-copy genes	34482	35602/35905	34459	34459	35611	34810	34459
Total genes, %	11.24	10.33	11.35	11.51	9.49	11.22	10.35

of total genome							
Single-copy genes, % of complexity	10.31	10.15/10.24	10.28	10.43	9.97	10.69	9.93
ORFs							
Total ORFs	69801	96427	72985	69432	72317	66536	68279
Single-copy ORFs	63147	72944	65218	62973	70850	66536	66704
ORFs % of total genome	18.93	19.25	20.00	18.87	19.11	19.50	18.63
Single-copy ORFs, % of complexity	18.88	20.79	19.45	19.05	19.84	20.43	19.22
Integrated plastid sequences							
Total integrated sequences	7862	7344	7740	7743	8123	4202	7865
Single-copy integrated sequences	7812	6671	7698	7668	8098	4132	7815
% of total genome	2.13	1.47	2.12	2.10	2.15	1.23	2.15
% of complexity	2.33	1.90	2.30	2.32	2.27	1.27	2.27

Table 2 - Comparison of the encoded genes and their transcripts in the mitochondrial

genomes of *Beta*

Product group	Gene	TK81-O	TK81-MS	A	B	CMS-E	CMS-G	macro
Complex I								
	*nad*1 (5 exons)	+	+a	+	+	+	+	+
	*nad*2 (5 exons)	+	+b	+	+	+	+	+
	*nad*3	+	+	+	+	+	+	+
	*nad*4 (3 exons)	+	+	+	+	+	+	+
	*nad*4L	+	+	+	2+	+	+	+
	*nad*5	+	+c	+	+	+	+	+
	*nad*6	+	+	+	+	+	+	+
	*nad*7 (5 exons)	+	2+	+	+	+	+	+
	*nad*9	+	+	+	+	+	+ (2*)	+
Complex III								
	cob	+	+	+	+	+	+	+
Complex IV								
	*cox*1	+	+	+	+	+	+	+
	*cox*2 (2 exons)	+	2+	+	+	+	+	+
	*cox*3	+	+	+	+	+	+	+
Complex V								
	*atp*1	+	+	+	+	+	+	+
	*atp*6	+	+	+	+	+	+	+
	*atp*8	+	+	+	+	+	+	+
	*atp*9	+	+	+	+	+	+	+
	*orf*25 (*atp*4)	+	+	+	2	+	+	+
Cytochrome-c-biogenesis								
	*ccm*B (*ccb*206)	+	2+	+	+	+	+	+
	*ccm*FC (2 exons) (*ccb*438)	+	+	+	+	+	+	+
	*ccm*FN	+	+	+	+	+	+	+

(*ccb*577)							
Ribosomal proteins							
*rpl*5	+	+	+	+	+	+	+
*rps*3	+	2+	+	+	+	+	+
*rps*4	+	+	+	+	+	+	+
*rps*7	+	2+	+	+	+	+	+
*rps*12	+	+	+	+	+	+	+
*rps*13	+	2+	+	+	+	+	+
Other proteins							
mat-r	+	+	+	+	+	+	+
Pseudogenes							
*sdh*4	+	+	+	+	+	+	+
petG cp-like	+	+	+	+	+	(*)	+
*rps*7 cp-like	+	+	+	+	+	(*)	+
*rps*3 exon 1	+	-	+	+	+	+	+
Sec-independant membrane targeting and translocation system							
tatC	+	+	+	+	+	+	+
Ribosomal RNAs							
*rrn*5S	+	+	+	+	+	+	+
*rrn*18S	+	+	+	+	+	+	+
*rrn*26S	3+	3+	3+	3+	+	2+	2+
Transfert RNAs							
native							
*trn*C1-GCA	ψ	2ψ	ψ	ψ	ψ	ψ	ψ
*trn*E-UUC	+	2+	+	+	+	+	+
*trn*F-GAA	+	+	+	+	+	+	+
*trn*G-GCC	+	+	+	+	+	+	+
*trn*I-CAU	+	+	+	+	+	+	+
*tnr*P-UGG	+	+	+	+	+	+	+

	*trn*Q-UUG	+	+	+	+	+	+	+
	*trn*S-GCU	+	+	+	+	+	+	+
	*trn*S-UGA	+	2+	+	+	+	+	+
	*trn*Y-GUA	+	2+	+	+	+	+	+
	*trn*K-UUU	+	2+	+	+	+	+	+
	*trn*fM-CAU	4+	5+	4+	4+	4+	3+	3+
	chloroplast-like							
	*trn*D-GUC	+	+	+	+	+	+	+
	*trn*H-GUG	+	2+	+	+	+	+	+
	*trn*I-CAU	ψ	ψ	ψ	ψ	-	ψ	ψ
	*trn*N-GUU	+	2+	+	+	+	+	+
	*trn*M-CAU	+	2+	+	+	+	+	+
	*trn*P-UGG	ψ	ψ	ψ	ψ	ψ	-	ψ
	*trn*S-GGA	+	+	+	+	+	+	+
	*trn*V-GAC	+	+	+	+	+	-	+
	*trn*W-CCA	+	+	+	+	+	(*)	+
	Origin unknown							
	*trn*C2-GCA	2+	+	2+	2+	2+	+	+

+, present; -, absent, ψ pseudogene, whole-gene copy numbers > 1 are given

a : 2 copies of exons 1, 2, 3 and 5

b : 2 copies of exons 3, 4 and 5

c : 2 copies of exons 1, 2 and 3

(*): found in unassembled contig

Table 3 - Substitutions and indels within maize mitochondrial genes

Genes	Position in TK81-O genome	Genomes						
		TK81-O	**TK81-MS**	**A**	**B**	**CMS-E**	**CMS-G**	**macro**
*atp*1	1386	gaT→D	gaT→D	gaT→D	gaT→D	gaT→D	gaG→E	gaT→ D
*atp*6			First 1171 bp specific to Sv			First 1177 bp specific to E		
	264	gtT→V	gtT→V	gtT→V	gtT→V	gtG→V	gtT→V	gtT→V
	489	ccT→P	ccC→P	ccT→P	ccT→P	ccT→P	ccT→P	ccT→P
*ccb*438 exon 1	306	ttA→L	ttA→L	ttA→L	ttA→L	ttA→L	ttC→F	ttA→L
*ccb*438 exon 2	288	cCa→P	cCa→P	cCa→P	cCa→P	cCa→P	cAa→Q	cCa→P
*cox*1	1-3	AtG→M	AtG→M	AtG→M	AtG→M	AtG→M	TtT→Fa	AtG→M
	14	gTt→V	gTt→V	gTt→V	gTt→V	gTt→V	gAt→D	gTt→V
	131	cGa→R	cGa→R	cGa→R	cGa→R	cGa→R	cAa→Q	cGa→R
	153	ggT→G	ggC→G	ggT→G	ggT→G	ggT→G	ggT→G	ggT→G
	966	atC→I	atA→I	atC→I	atC→I	atC→I	atC→I	atC→I
	1179	gcA→A	gcG→A	gcA→A	gcA→A	gcA→A	gcA→A	gcA→A
	1206	atC→I	atA→I	atC→I	atC→I	atC→I	atC→I	atC→I
	1207	Ttt→F	Ttt→F	Ttt→F	Ttt→F	Ttt→F	Gtt→V	Ttt→F
*cox*2 exon 2	376	tTa→L	tTa→Lb	tTa→L	tTa→L	tTa→L	tGa→*	tTa→L
*cox*3	151	Att→I	Att→I	Att→I	Att→I	Att→I	Ctt→L	Att→I
mat-r	750	aaT→N	aaT→N	aaT→N	aaT→N	aaG→K	aaT→N	aaT→N

	1086	gtC→V	gtA→V	gtC→V	gtC→V	gtC→V	gtC→V	gtC→V
	1215	aaT→N	aaG→K	aaT→N	aaT→N	aaT→N	aaT→N	aaT→N
*nad*1 exon 1	7	Ata→T	Ata→T	Ata→T	Ata→T	Ata→T	Gta→V	Ata→T
*nad*1 exon 3	160-161	CGt→R	GCt→A	GCt→A	GCt→A	GCt→A	GCt→A	GCt→A
*nad*2 exon 4	14	ccC→P	ccT→P	ccT→P	ccT→P	ccT→P	ccT→P	ccT→P
	74	atA→I	atA→I	atA→I	atA→I	atA→I	atC→I	atA→I
*nad*4L	17	tAt→Y	tAt→Y	tAt→Y	tAt→Y	tAt→Y	tTt→F	tAt→Y
	19	Ttt→F	Ttt→F	Ttt→F	Ttt→F	Ttt→F	Gtt→V	Ttt→F
*nad*5 exon 1	13	Atc→I	Atc→I	Atc→I	Atc→I	Ctc→L	Atc→I	Atc→I
*nad*5 exon 4	3	Aat→N	Cat→H	Aat→N	Aat→N	Aat→N	Cat→H	Aat→N
*nad*7 exon 1	15	atC→I	atC→I	atC→I	atC→I	atC→I	atG→M	atC→I
*nad*9	59	aAa→K	aAa→K	aAa→K	aAa→K	aAa→K	aCa→T	aAa→K
	66	atA→I	atC→I	atA→I	atA→I	atA→I	atA→I	atA→I
	74	tCa→S	tCa→S	tCa→S	tCa→S	tCa→S	tTa→L	tCa→S
	118	Caa→Q	Aaa→K	Caa→Q	Caa→Q	Caa→Q	Caa→Q	Caa→Q
	261	cgG→R	cgC→R	cgC→R	cgC→R	cgC→R	cgC→R	cgC→R
	262	Cta→L	Gta→V	Gta→V	Gta→V	Gta→V	Gta→V	Gta→V
	318	ccA→P	ccA→P	ccA→P	ccA→P	ccA→P	ccG→P	ccA→P
	525	ttT→F	ttT→F	ttT→F	ttT→F	ttT→F	ttG→L	ttT→F
	559	Cgt→R	Cgt→R	Cgt→R	Cgt→R	Cgt→R	Ggt→G	Cgt→R
	557	Taa→*	Taa→*	Taa→*	Taa→*	Taa→*	Gaa→Ec	Taa→*
*orf*25	463	Cac→H	Cac→H	Cac→H	Cac→H	Cac→H	Aac→N	Cac→H
*rps*3	85	Agt→S	Ggt→G	Agt→S	Agt→S	Agt→S	Ggt→G	Agt→S
	106	Ctc→L	Atc→I	Atc→I	Atc→I	Atc→I	Atc→I	Atc→I
	755	tCc→S	tTc→F	tTc→F	tTc→F	tTc→F	tTc→F	tTc→F

	1106	aTa→I	aTa→I	aTa→I	aTa→I	aTa→I	aGa→R	aTa→I
	1232	aTa→I	aGa→R	aTa→I	aTa→I	aTa→I	aGa→R	aTa→I
	1240	Gct→A	Cct→P	Gct→A	Gct→A	Gct→A	Cct→P	Gct→A
*rps*4	103	Aag→K	Gag→E	Aag→K	Aag→K	Aag→K	Aag→K	Aag→K
	527	cTg→L	cGg→R	cTg→L	cTg→L	cTg→L	cTg→L	cTg→L
	573	cgC→R	cgC→R	cgC→R	cgC→R	cgC→R	cgA→R	cgC→R
	745-746	TAt→Y	GAt→D	TCt→S	TCt→S	TAt→Y	TAt→Y	TAt→Y
*rps*7	198	gtC→V	gtA→V	gtC→V	gtC→V	gtA→V	gtA→V	gtC→V
*rps*12	269	tCg→S	tTg→L	tCg→S	tCg→S	tCg→S	tCg→S	tCg→S
	326	gAt→D	gGt→G	gAt→D	gAt→D	gAt→D	gAt→D	gAt→D
tatC	323	aGa→R	aGa→R	aGa→R	aGa→R	aGa→R	aGa→R	aTa→I
	567	tcT→S	tcC→S	tcT→S	tcT→S	tcT→S	tcT→S	tcT→S

a : Undefinied start codon (begining 408 bp before or 87 bp after)

b : two copies, one is identical to Nv, the other have the first 198 bp identicals and 506 bp unique

c : leading to supplementary 42 bp

* : stop codon

Table 4 - Chimeric ORFs

ORF name	**TK81-O**	**TK81-MS**	**A**	**B**	**CMS-E**	**CMS-G**	**macro**	Gene fragments present
*orf*100e	262102-262404							
*orf*105a	19539-19856		8023-8340	117299-117616	99299-99616a		90547-90864a	59bp *matK*
*orf*105d		57459-57776 249975-250292	287788-288105	32176-32493	86329-86646b	91046-91363a	152611-152928a	27bp *rrn18*
*orf*105e					140735-141052a	29317-29634a		34bp de *atp9*
*orf*107a	72786-73109		319702-	64090-64413	54410-54733b	122961-	184529-	27bp,25b

			320025			123284a	184852a	p *atp8*
*orf*119a	117425-117784	485856-486215	153118-153477	19402-19761	7506-7865b	65812-66171a	139831-140190a	20 bp plastid *ycf2*
	359394-359753		275014-275373	144555-144914				
orf62	157959-158448	228055-228544	87868-88357	322031-322520	195890-196379a	52895-53384b	64246-64735a	20 bp plastid *ycf2*
		388149-388638						
*orf*221t		213770-214435						274pb cox2 exon1
		402259-402924						
*orf*246t	272854-273594							74bp rsp3 exon 1 ; 24bp,30b p rps3 exon 2
*orf*273			188701-189522	183003-183824	70563-71384a	70539-71360b	67251-68072b	74bp *rps3* exon 1
*orf*281			177940-178785	172242-173087	81300-82145a	77988-78833b		30bp *atp8*
*orf*282			284458-285306	28846-29694				27bp *nad9*
*orf*317t	170862-171815		100788-101741	334951-335904	182507-183460a	37990-38943b	77166-78119a	274bp *cox2* exon 1

a : first contig, b : second contig

Table 5 - Large repeats (larger than 0.5 kb) in *Beta* mitochondrial genomes

	Genomes					
Repeats	**Nv**	**Sv**	**A**	**B**	**CMS-E**	**CMS-G**
R87	-, R23, R9.8, R6.2	+, R7.6	-, R23, R6.5, R6.0	-, R23, R8.3, R6.5, R6.6, R6.0	-, R6.8	-, R7.6, R7.2, R1.44
R51	-	+	-	-	-	-
R23	+, R9.8, R6.2	-, R7.6, R4.85	+, R6.5, R6.0	+, R6.5, R6.0	-, R6.8, R6.6, R0.60	-, R7.6
R9.8	+, R6.2	-, R7.6	-, R6.5, R6.0	-, R6.5, R6.0	-, R6.8, R6.6	-, R7.6, R7.2, R1.44
R8.3	-, R0.64, R0.58	-	-, R0.64	+	-, R0.64	-
R7.6	-, R6.2	+, 3x	-, R6.5, R6.0	*-, R6.5, R6.0*	-, R6.8, R6.6	+, R7.2, R1.44
R7.2	-	-	-	-	-, R6.8, R6.6	+, R1.44
R6.8	-, R6.2	-	-, R6.5, R6.0	-, R6.5, R6.0	+	-, R1.44
R6.7	-	-	-	-	+	-
R6.6	-, R6.2	-	-, R6.5, R6.0	*-, R6.5, R6.0*	+	-
R6.5	-, R6.2	-	+, R6.0	+, R6.0	-	-
R6.2	+, 3x	-	-, R6.0	-, R6.0	-	-
R6.0	-	-	+, 3x	+, 3x	-	-
R5.94	-	-	-	-	-, R0.60	-
R4.85	-	+	-	-	-	-
R4.62	-	-	-	-	-	-
R4.48	-	-	-	-	-	-
R3.10	-	-	-	-	-	-
R1.44	-	-	-	-	-	+, 3x
R1.37	-	-	-	-	-	-
R1.26	-	-	-	-	-	-

R0.75	-	-	-	-	-	+
R0.73	-, R0.58	-	-	-	-	-
R0.64	+	-	+	-	+	-

Repeats (R) are indicated by numbers indicating lengths

+ repeat present in mitogenome, - repeat absent in mitogenome

- italic : repeat partially incluse in other one

3x : three copies of the repeat

Table 6a - Substitutions per 10,000 bp between pairs of *Beta* genomes

Genomes	TK81-MS	A	B	CMS-E	CMS-G	macro
TK81-O	28.812	4.724	4.837	5.662	31.301	4.911
TK81-MS	-	26.734	26.547	26.435	34.075	26.471
A	-	-	1.087	3.337	28.759	2.587
B	-	-	-	3.224	28.834	2.474
CMS-E	-	-	-	-	29.248	3.374
CMS-G	-	-	-	-	-	28.827

Table 6b– Indels per 10000 bp between pair of *Beta* genomes

Genomes	TK81-MS	A	B	CMS-E	CMS-G	macro
TK81-O	25.660	7.874	7.724	8.249	22.904	7.949
TK81-MS	-	19.948	19.797	19.873	22.379	20.022
A	-	-	0.224	1.199	15.860	0.824
B	-	-	-	1.049	15.785	0.674
CMS-E	-	-	-	-	16.311	1.199
CMS-G	-	-	-	-	-	16.006

Table 6c – Mutations per 10,000 bp between pairs of *Beta* genomes

Génomes	TK81-MS	A	B	CMS-E	CMS-G	macro
TK81-O	54.473	12.598	12.561	13.911	54.205	12.860
TK81-MS	-	46.683	46.345	46.309	56.454	46.494

A	-	-	1.311	4.537	44.620	3.412
B	-	-	-	4.274	44.620	3.1496
CMS-E	-	-	-	-	45.559	4.574
CMS-G	-	-	-	-	-	44.833

Table 7a - Substitutions per 10000 bp between pair of *Zea* genomes

Genomes	NB	CMS-C	CMS-S	CMS-T	parvi	lux	per
NA	2.490	2.481	4.884	4.621	0.550	15.539	14.055
NB	-	3.712	5.915	5.699	2.514	16.764	15.112
CMS-C	-	-	5.652	5.341	2.443	16.526	15.017
CMS-S	-	-	-	7.206	4.980	18.292	16.568
CMS-T	-	-	-	-	4.645	17.002	15.517
parvi	-	-	-	-	-	15.564	14.056
lux	-	-	-	-	-	-	4.888

Table 7b – Indels per 10000 bp between pair of *Zea* genomes

Genomes	NB	CMS-C	CMS-S	CMS-T	parvi	lux	per
NA	5.866	12.211	13.169	24.638	2.514	160.760	107.125
NB	-	14.058	14.776	26.440	6.370	162.017	108.131
CMS-C	-	-	19.736	26.418	12.382	159.949	110.943
CMS-S	-	-	-	30.479	14.102	163.911	109.753
CMS-T	-	-	-	-	24.761	164.876	114.108
parvi	-	-	-	-	-	161.228	107.158
lux	-	-	-	-	-	-	114.122

Table 7c – Mutations per 10000 bp between pair of *Zea* genomes

Genomes	NB	CMS-C	CMS-S	CMS-T	parvi	lux	per
NA	8.356	14.629	18.053	29.259	3.064	176.300	121.180
NB	-	17.770	20.692	32.140	8.885	178.782	123.243
CMS-C	-	-	25.388	31.759	14.826	176.475	125.961
CMS-S	-	-	-	37.685	19.082	182.203	126.321
CMS-T	-	-	-	-	29.407	181.879	129.626
parvi	-	-	-	-	-	176.793	121.214
lux	-	-	-	-	-	-	119.011

Table 8 Synonymous and non-synonymous nucleotide diversity and divergence with *Beta macrocarpa* of CMS and non-CMS mt genomes

mt_genes (29344 bp)	**πs**	**πa**	**πa /πs**	**K_s**	**Ka**	**Ka /Ks**
CMS	**0.00124**	**0.00112**	**0.916**	**0.00073**	**0.00069**	**0.944**
E				0.00063	0.00035	0.551
G				0.00067	0.00095	1.416
TK81-MS				0.00174	0.0007	0.401
non CMS	**0.00021**	**0.0002**	**0.945**	**0.00011**	**0.00012**	**1.102**
A				0	0.00005	/
B				0	0.00005	/
TK81-O				0.00032	0.00025	0.787

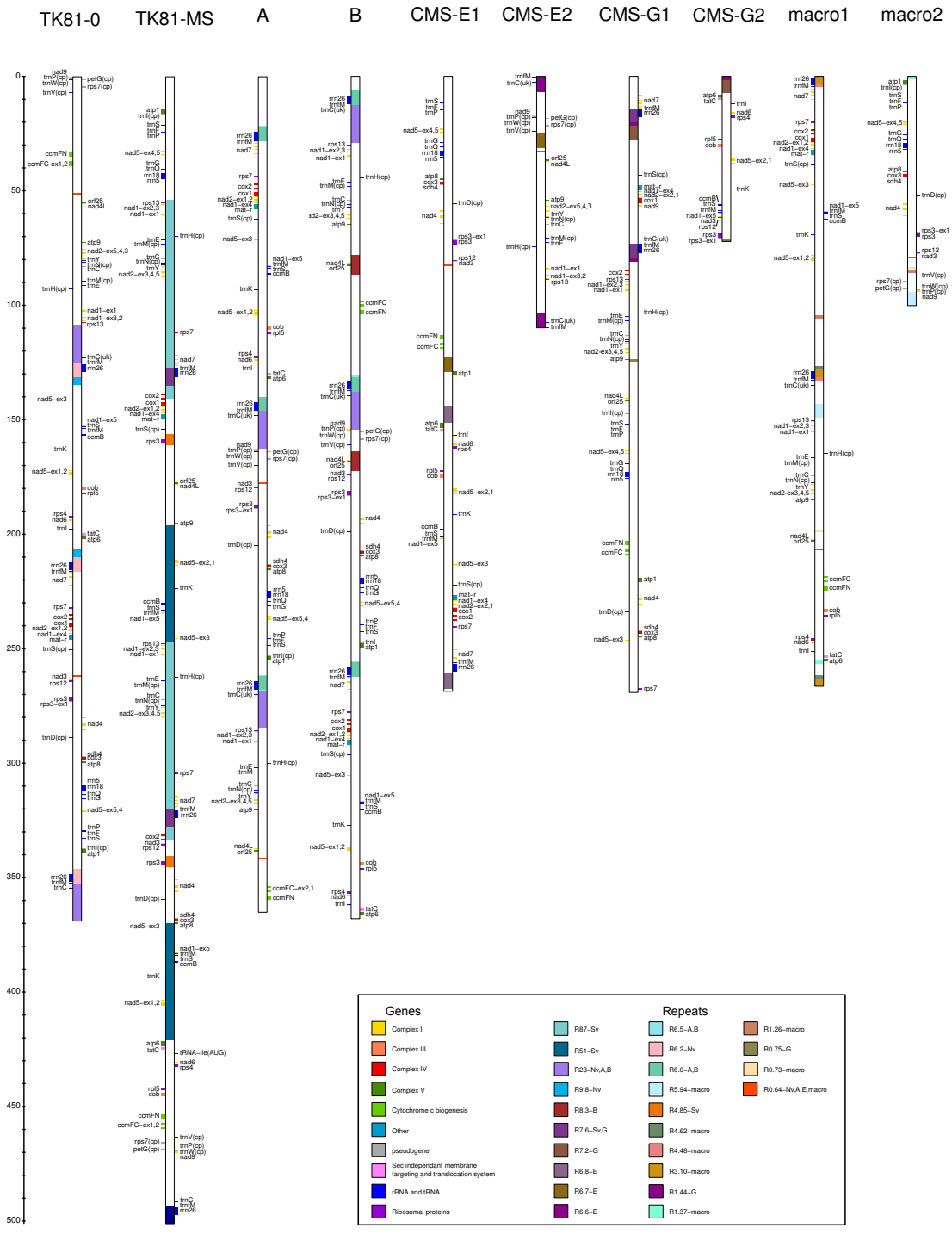


Figure 1

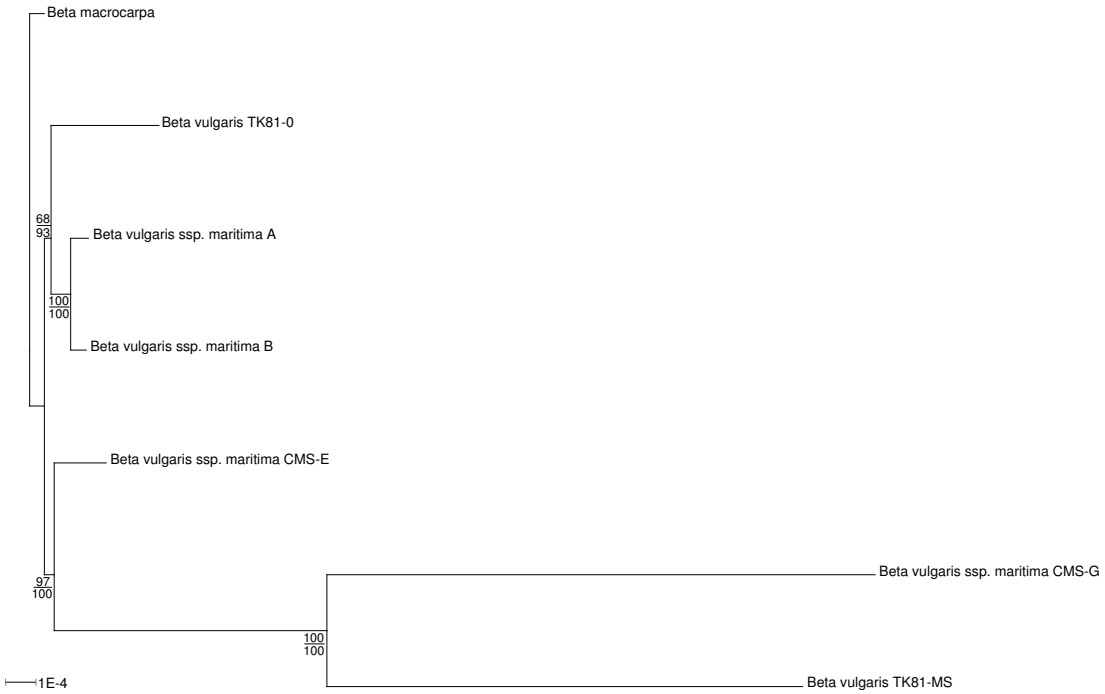


Figure 2

Table S1 – Chloroplastic insertion in Beta mitogenomes

Nicotiana tabacum		TK81-O			TK81-MS			A			B			CMS-E			CMS-G			macro		
start	stop	start	stop	length	start	stop	length	start	stop	length	start	stop	length	start	stop	length	start	stop	length	start	stop	length
1	76	92807	92882	76	69592	69667	76	299919	299994	76	44307	44382	76	343057	343132	76	103182	103257	76	164746	164821	76
					262108	262183	76															
7480	7509													222078	222107	30	43154	43183	30			
29103	29181	288735	288813	79	359263	359341	79	204599	204677	79	198901	198979	79	55408	55486	79	234054	234132	79	318528	318606	79
30303	30327	54291	54315	25	131401	131425	25	23947	23971	25	8070	8094	25	260278	260302	25	18081	18105	25	293	317	25
		129282	129306	25	176956	176980	25	141850	141874	25	82889	82913	25	304722	304746	25	77319	77343	25	128531	128555	25
		211899	211923	25	323917	323941	25	263746	263770	25	133223	133247	25	153431	153455	25	141923	141947	25	203321	203345	25
		347872	347896	25	497427	497451	25	338498	338522	25	168821	168845	25				278037	278061	25	263620	263644	25
		200849	200873	25				130797	130821	25	258046	258070	25							254425	254449	25
											364960	364984	25									
32209	32273													361146	361210	65						
34189	34216													222080	222107	28	43156	43183	28			
35199	35220	91074	91095	22	71382	71403	22	301709	301730	22	46097	46118	22	341321	341342	22	104972	104993	22	166536	166557	22
					263898	263919	22															
38972	38993	12601	12622	22				175132	175153	22										348052	348073	22
39012	39159	234047	234194	148	109408	109555	148	45849	45996	148	280012	280159	148	238251	238398	148	265301	265448	148	22227	22374	148
					301924	302071	148															
42326	42364	5590	5628	39	464338	464376	39	168119	168157	39	159556	159594	39	291123	291161	39				355009	355047	39
43990	44116	250238	250369	132	153981	154112	132	62044	62175	132	296207	296338	132	222076	222173	98	43152	43249	98	38422	38553	132
					232898	232920	23							200734	200756	23	327860	327882	23			
46328	46349	25782	25803	22	206987	207008	22	2075	2096	22	111351	111372	22	105543	105564	22	195739	195760	22	84599	84620	22
					409686	409707	22															
46944	46985	329553	329582	30	24046	24088	43	245426	245455	30	239726	239755	30	14616	14658	43	154841	154883	43	277737	277779	43
48331	48931	256881	257485	605				17084	17688	605	126360	126964	605	86764	87368	605	249442	250046	605	99619	100223	605
50858	50935	89132	89209	78	73269	73346	78	303596	303673	78	47984	48061	78	339378	339455	78	106859	106936	78	168423	168500	78
					265785	265862	78															
51592	51626	161133	161167	35	225332	225366	35	91048	91082	35	325211	325245	35	193166	193200	35	320292	320326	35	67426	67460	35
					391328	391362	35															
55397	55428	73798	73829	32	196474	196505	32	318982	319013	32	63370	63401	32	324038	324069	32	122241	122272	32	183809	183840	32
					420189	420220	32															
60069	60090													995	1016	22						
64617	64791	1241	1411	171	468558	468734	177	163769	163939	171	155206	155376	171	286773	286943	171				359220	359397	178
65143	65227	833	917	85	469058	469142	85	163361	163445	85	154798	154882	85	286365	286449	85				359721	359805	85
77837	77863																309447	309473	27			
84041	84088													364194	364241	48						
86178	86601	324678	325081	404	28550	28964	415	240551	240954	404	234851	235254	404	19119	19533	415	159345	159759	415	282240	282654	415
86448	88358	324958	326873	1916	26769	28611	1843	240831	242746	1916	235131	237046	1916	17338	19180	1843	157564	159406	1843	280459	282301	1843
91071	91095				94904	94928	25							137827	137851	25	32514	32538	25			
					287420	287444	25															
94478	97852	3942	7307	3366	463104	466034	2931	166470	169837	3368	157907	161274	3368	289474	292841	3368				353329	356696	3368
94903	94973	178039	178109	71	207190	207260	71	107969	108039	71	342132	342202	71	176209	176279	71	300828	300898	71	84347	84417	71
					409434	409504	71															
97994	98017	153027	153050	24	233460	233483	24	82929	82952	24	317092	317115	24	201296	201319	24	328422	328445	24	59307	59330	24
					383211	383234	24															
98031	98055				42332	42356	25							32901	32925	25	173126	173150	25	296021	296045	25
99277	99313	309711	309747	37				225586	225622	37	219887	219923	37									
99377	99408				44041	44072	32							34611	34642	32	174836	174867	32	297731	297762	32
99556	99603																187074	187121	48			
100611	100636	162949	162974	26	393149	393174	26	92873	92898	26	327036	327061	26							69251	69276	26
101571	101759	247973	248123	151	151709	151860	152	59778	59929	152	293941	294092	152	224288	224470	183	45364	45546	183	36156	36307	152
														93653	93799	147				106508	106654	147
103354	103398	126813	126857	45	128932	128976	45							257809	257853	45	15612	15656	45			
					321448	321492	45										74850	74894	45			
					494958	495002	45															
105958	106037	80994	81073	80										331240	331319	80						
118175	118215	5588	5628	41	464338	464378	41	168117	168157	41	159554	159594	41	291121	291161	41				355009	355049	41

Table S2 – ORF in Beta mitogenomes

ORF	TK81-O		TK81-MS		A		B		CMS-E		CMS-G		macro	
	start	stop	start	stop	start	stop	start	stop	start	stop	start	stop	start	stop
orf99a	321479	321779	31862	32162	237352	237652	231652	231952	22431	22731	162657	162957	285552	285852
orf99b			335604	335904	179881	180181	174183	174483	79903	80203	330861	331161	343023	343323
orf99c			183772	184072										
orf99d			62164	62464										
			254680	254980										
orf99e									5662	5962				
orf99f									371514	371814				
orf99g°	38740	39040	458925	459225	353780	354080	98106	98406	118508	118808			218604	218904
orf100a	61677	61980	184343	184646	330831	331134	75222	75525	311909	312212	134089	134392	195665	195968
orf100b	133186	133489												
	207714	208017												
orf100c	195685	195988												
orf100d											191315	191618		
orf100e	262101	262404												
orf100f°	264036	264339												
orf100g			232517	232820	83591	83894	317754	318057	200353	200656	327479	327782	59969	60272
			383873	384176										
orf100h°			333205	333508										
orf100i			423455	423758										
orf100j											210954	211257		
orf101a			99486	99792										
			292002	292308										
orf101b			210480	210786										
			405907	406213										
orf101c*									363899	364205				
orf102a	234550	234859			46352	46661	280515	280824	237585	237894	264635	264944	22730	23039
orf102b	255869	256178												
orf102c													112528	112837
orf102d			161200	161509										
orf102e			225436	225745	90668	90977	324831	325140	193270	193579	320396	320705	67046	67355
			390948	391257										
orf103a	40990	41302	461179	461491	351517	351829	95842	96154	120760	121072			216340	216652
orf103b	99428	99740	62731	63043	293058	293370	37446	37758	349680	349992	96321	96633	157885	158197
			255247	255559										
orf103c	299512	299824	53821	54133	215381	215693	209683	209995	44391	44703			307511	307823
orf103d									4489	4801				
orf104a	70921	71236			321573	321888	65964	66279	321159	321474	124829	125144	186403	186718
orf104b	94418	94733												
orf104c	104688	105003												
orf104d											188614	188929		
orf105a	19538	19856			8022	8340	117298	117616	99298	99616			90546	90864
orf105b	312967	313285	40356	40674	228841	229159	223141	223459	30925	31243	171150	171468	294046	294364
orf105c	314330	314648	38993	39311	230203	230521	224503	224821	29562	29880	169788	170106	292683	292998
orf105d			57458	57776	287787	288105	32175	32493	354944	355262	91045	91363	152610	152928
			249974	250292										
orf105e									140734	141052	29316	29634		
orf105f									366392	366710				
orf106a	10980	11301			173511	173832							349372	349693
orf106b	109356	109677	470956	471277	161226	161547	152663	152984	284230	284551	57741	58062	147939	148260
	367499	367820			283122	283443	27510	27831					361618	361939
orf106c			471220	471541	160962	161283	27246	27567	283966	284287	58005	58326	147675	147996

ORF	TK81-O		TK81-MS		A		B		CMS-E		CMS-G		macro	
	start	stop	start	stop	start	stop	start	stop	start	stop	start	stop	start	stop
orf107a	72785	73109			319701	320025	64089	64413	323025	323349	122960	123284	184528	184852
orf107b°	79554	79878	82602	82926	312930	313254	57318	57642	329796	330120	116188	116512	177757	178081
			275118	275442										
orf107c					336789	337113	81180	81504	306130	306454			201612	201936
							167112	167436						
orf107d											336279	336603		
orf108a	183645	183972	440310	440637	113578	113905	347741	348068	170342	170669	294961	295288	237206	237533
orf108b	323801	324128	29513	29840	239674	240001	233974	234301	20082	20409	160308	160635	283203	283530
orf108c°	102232	102559	59905	60232	290234	290561	34622	34949	352488	352815	93492	93819	155062	155389
			252421	252748										
orf109a	145537	145867	240644	240974	75434	75764	309597	309927	208483	208813			51812	52142
			375719	376049										
orf109b			102854	103184										
			295370	295700										
orf109c	137055	137385	202405	202735	66934	67264	301097	301427	216983	217313			43312	43642
			413958	414288										
orf110a	162208	162541												
orf110b*	329395	329728	23912	24245	245268	245601	239568	239901	14482	14815	154707	155040	277603	277936
orf110c			137579	137912					145071	145404	83497	83830		
			330095	330428					266457	266790				
orf110d			246316	246649							244834	245167		
			370044	370377										
orf110e°			333437	333770										
orf111a	98372	98708	63763	64099	294090	294426	38478	38814	348624	348960	97353	97689	158917	159253
			256279	256615										
orf111b			135843	136179					146804	147140	81761	82097		
			328359	328695					264721	265057				
orf111c			340455	340791					75012	75348				
orf111d			350843	351179										
orf111e°	72270	72606	194945	195281			64592	64928	322510	322846	123457	123793	185031	185367
orf112a°	181471	181810												
orf112b	255897	256236			18338	18677	127614	127953	88012	88351	250696	251035	100867	101206
orf112c									366963	367302				
orf112d			7972	8311										
orf113a	58002	58344	180668	181010	334467	334809	78858	79200	308234	308576	137725	138067	199290	199632
							164790	165132						
orf113b											37136	37478		
orf113c											88192	88534		
orf114a*	5953	6298	463667	464012	168482	168827	159919	160264	291486	291831			354338	354683
orf114b	26998	27343	447176	447521	534	879	109810	110155	106759	107104	196955	197300	230308	230653
orf114c	200393	200738	423528	423873	130341	130686	364504	364849	153568	153913	278174	278519	253969	254314
orf114d	341851	342196	89201	89546	257724	258069	252024	252369	133308	133653	223491	223836	257598	257943
			281717	282062										
orf115a	164511	164859	221632	221980	94437	94785	328600	328948	189462	189810	316588	316936	70815	71163
			394713	395061										
orf115b°	199519	199867	424399	424747	129467	129815	363630	363978	154439	154787	279045	279393	253095	253443
orf115c*	326699	327047	26594	26942	242572	242920	236872	237220	17163	17511	157389	157737	280284	280632
orf116a	57289	57640	179956	180307	335171	335522	79562	79913	307521	307872	138429	138780	199994	200345
							165494	165845						
orf116b°	171786	172137	213159	213510	101712	102063	335875	336226			306803	307154	78090	78441
			403183	403534										

ORF	TK81-O		TK81-MS		A		B		CMS-E		CMS-G		macro	
	start	stop	start	stop	start	stop	start	stop	start	stop	start	stop	start	stop
orf116c			97259	97610							34877	35228		
			289775	290126										
orf116d			183878	184229										
orf117°	104910	105264												
orf117b*			469002	469356										
orf118													111230	111587
orf119a	117424	117784	485855	486215	153117	153477	19401	19761	276121	276481	65811	66171	139830	140190
	359393	359753			275013	275373	144554	144914						
orf119b	157967	158327	228176	228536	87876	88236	322039	322399	196011	196371	323137	323497	64254	64614
			388157	388517										
orf119c			14013	14373										
orf119c*			465262	465622										
orf120a			104278	104641										
			296794	297157										
orf120b	149058	149421	237090	237453	78958	79321	313121	313484	204926	205289			55336	55699
			379240	379603										
orf121	229769	230135	113471	113837	41569	41935	275732	276098	242313	242679	225	591	17947	18313
			305987	306353										
orf122	84596	84965	77512	77881	307839	308208	52227	52596	334842	335211	111097	111466	172666	173035
			270028	270397										
orf122b			244750	245119	71288	71657	305451	305820	212590	212959	246364	246733	47666	48035
			371574	371943										
orf122c			83205	83574										
			275721	276090										
orf123	253746	254118												
orf124	339169	339544			255042	255417	249342	249717	130626	131001	220809	221184	267787	268162
orf125a	246266	246644	150001	150379	58071	58449	292234	292612	225798	226176	46874	47252	34449	34827
orf125b*			69518	69896	299845	300223	44233	44611	342827	343205	103108	103486	164672	165050
			262034	262412										
orf126*	12616	12997											347676	348057
orf128*			94913	95300					137454	137841	32523	32910		
			287429	287816										
orf129a	12942	13332												
orf129b	141279	141669												
orf129c	153573	153963												
orf131	305095	305491	48151	48547	220967	221363	215269	215665	38721	39117	178946	179342	301841	302237
orf132a									6850	7249				
orf132b									267514	267913				
orf133a	14841	15243												
orf133b	323076	323478	30163	30565	238949	239351	233249	233651	20732	21134	160958	161360	283853	284255
orf133c°			57152	57554	287481	287883	31869	32271	355166	355568	90739	91141	152304	152706
			249668	250070										
orf134a°	282620	283025	353140	353545	198479	198884	192781	193186	61200	61605	227935	228340	324320	324725
orf134b	136980	137385												
orf134c°	240472	240877	144206	144611	52275	52680	286438	286843	231567	231972	52643	53048	28653	29058
orf135a	328582	328990	24651	25059	244455	244863	238755	239163	15220	15628	155446	155854	278341	278749
orf135b											257319	257727		
orf136a			108294	108705										
			300810	301221										
orf136b	41185	41596	461374	461785	351223	351634	95548	95959	120955	121366	211140	211551	216046	216457
orf136c°	298558	298969												
orf137					20804	21218	130080	130494	90478	90892	253162	253576	103333	103747

ORF	TK81-O		TK81-MS		A		B		CMS-E		CMS-G		macro	
	start	stop	start	stop	start	stop	start	stop	start	stop	start	stop	start	stop
orf138	223355	223772	119836	120253	35157	35574	269320	269737	248674	249091	6580	6997	11535	11952
			312352	312769										
orf139			428160	428580										
orf140	182586	183009	441274	441697	112518	112941	346681	347104	171306	171729	295925	296348	236146	236569
orf141			103176	103602										
			295692	296118										
orf143a	257827	258259			16315	16747	125591	126023	85989	86421	248673	249105	98844	99276
orf143b*	326240	326672			242113	242545	236413	236845	17538	17970	157764	158196	280659	281091
orf143c*			94913	95345					137409	137841				
			287429	287861										
orf144											191767	192202		
orf145a†	141323	141761												
orf145b			101459	101897										
			293975	294413										
orf145c									143920	144358				
orf145d					28368	28806	262531	262969	255442	255880	13341	13779	4746	5184
orf146*	6179	6620			168708	169149	160145	160586	291712	292153			354016	354457
orf147a			105204	105648										
			297720	298164										
orf147b	149518	149962	236549	236993	79418	79862	313581	314025	204385	204829			55796	56240
			379700	380144										
orf148	115569	116016	484000	484447	154885	155332	21169	21616	277889	278336	63956	64403	141598	142045
	361161	361608			276781	277228	146322	146769						
orf152a	81658	82117	80360	80819	310687	311146	55075	55534	331904	332363	113945	114404	175514	175973
			272876	273335										
orf152b	18650	19109			8769	9228	118045	118504	98410	98869			91293	91752
orf153													117997	118459
orf155a			7529	7997										
orf155b°	218340	218808	124805	125273	30139	30607	264302	264770	253641	254109	11540	12008	6517	6985
			317321	317789										
orf160													68557	69040
orf162	157959	158448	228055	228544	87868	88357	322031	322520	195890	196379	323016	323505	64246	64735
			388149	388638										
orf165	41185	41683												
orf166	292290	292791	362824	363325	208160	208661	202462	202963	51423	51924	237615	238116	314543	315044
orf167a			193405	193909										
orf167b			99864	100368							37514	38018		
			292380	292884										
orf167c°			333207	333711										
orf169a			45625	46135	223379	223889	217681	218191	36195	36705	176420	176930	299315	299825
orf169b											329387	329897		
orf169c°	192952	193462	430803	431313	122896	123406	357059	357569	160847	161357	285453	285963	246524	247034
orf170a	159741	160254												
orf170b									143042	143555				
orf174a*			463572	464097										
orf174b					18396	18921	127672	128197	88070	88595	250754	251279	100925	101450
orf175°	284948	285476	355468	355996	200807	201335	195109	195637	58749	59277	230263	230791	321869	322397
orf176a†			244703	245234	71173	71704	305336	305867	212543	213074	246249	246780	47551	48082
			371459	371990										
orf176b											38673	39204		

ORF	TK81-O		TK81-MS		A		B		CMS-E		CMS-G		macro	
	start	stop	start	stop	start	stop	start	stop	start	stop	start	stop	start	stop
orf177a			8480	9014										
orf177b											259213	259747		
orf178a			101023	101560										
			293539	294076										
orf178b°	234901	235438	138636	139173	46703	47240	280866	281403	237006	237543	84548	85085	23081	23618
			331152	331689										
orf184a			348505	349060										
orf184b*									902	1457				
orf185			156967	157525										
			341018	341576										
orf187a	269453	270017			185299	185863	179601	180165	74221	74785			337341	337905
orf187b			246956	247520					357716	358280				
orf189°	38582	39152	458767	459337	353668	354238	97994	98564	118350	118920	208545	209115	218492	219062
orf190	289698	290271												
orf192	260938	261517			176776	177355			82729	83308			345849	346428
orf193°	181751	182333	441950	442532	111683	112265	345846	346428	171982	172564	296601	297183	235311	235893
orf198a	246219	246816	149954	150551	58024	58621	292187	292784	225626	226223	46702	47299	34402	34999
orf198b°			85132	85729										
			277648	278245										
orf199	95311	95911	66560	67160	296887	297487	41275	41875	345563	346163	100150	100750	161714	162314
			259076	259676										
orf202	293660	294269							49945	50554			313065	313674
orf203									371833	372445				
orf204*	325602	326217	27424	28039	241475	242090	235775	236390	17993	18608			281114	281729
orf208°			458710	459337	353668	354295	97994	98621	118293	118920	208488	209115	218492	219119
orf211	307507	308143												
orf214	175167	175812	209485	210130										
			406563	407208										
orf215	107679	108327												
orf216			6912	7563										
orf217a	291552	292206	362086	362740	207422	208076	201724	202378	52008	52662	236877	237531	315128	315782
orf217b			100312	100966							37962	38616		
			292828	293482										
orf217c°	298962	299616	369494	370148	214831	215485	209133	209787	44599	45253			307719	308373
orf221			213769	214435										
orf224a			360228	360903	205564	206239	199866	200541	53845	54520	235019	235694	316965	317640
orf224b											186363	187038		
orf227°			368817	369501	214154	214838	208456	209140	45246	45930	243608	244292	308366	309050
orf234*											158129	158834		
orf237									369004	369718				
orf238*					175147	175864							347340	348057
orf245	308042	308780												
orf246	272853	273594												
orf247°											277257	278001		
orf249					223906	224656								
orf251°	200902	201655	422590	423343	130850	131603	365013	365766	152651	153404			254478	255231
orf256					105094	105865	339257	340028	178382	179153			81472	82243
orf265a			136237	137035					145948	146746	82155	82953		
			328753	329551					265115	265913				
orf265b			44859	45657			218159	218957	35429	36227	175654	176452	298549	299347
orf266									370742	371543				
orf268*			26594	27401										
orf270	237642	238455	141376	142189	49445	50258	283608	284421	233989	234802			25823	26636
orf273					188700	189522	183002	183824	70562	71384	339674	340496	333682	334504
orf279													149289	150129
orf281					177939	178785	172241	173087	81299	82145			344419	345265
orf282					284457	285306	28845	29694						
orf284			474801	475656										
orf288°	282620	283487	353140	354007	198479	199346	192781	193648	60738	61605	227935	228802	323858	324725

orf352	290409	291468			206279	207338	200581	201640	52746	53805	235734	236793	315866	316925
orf360											184932	186015		
orf393*	324674	325856	27785	28967	240547	241729	234847	236029	18354	19536	158580	159762	281475	282657
orf399°	172972	174172	211124	212324	102898	104098	337061	338261	180149	181349	304768	305968	79276	80476
			404369	405569										
orf409	114300	115530												
	361647	362877												
orf435											238985	240293		
orf477											145347	146781		
orf496a†			172066	173557			86309	87800	299839	301330				
orf496b			364194	365685	209530	211021	203832	205323						
orf518†	49342	50899			341918	343475							206741	208298
orf575°	238500	240228			50303	52031	284466	286194	232216	233944			26681	28409
orf598			360943	362740										
orf670	117103	119116	485534	487547	151785	153798	18069	20082	274789	276802	65490	67503	138498	140511
	358061	360074			273681	275694	143222	145235						
orf764	58852	61147	181518	183813	331664	333959	76055	78350	309084	311379	134922	137217	196498	198793
orf774			481636	483961	155371	157696	146808	149133	278375	280700	61592	63917	142084	144409
					277267	279592								
orf864°											53292	55887		

† known as transcribed

° supreposed to gene

* superposed to cp fragment

Liste des publications et communications

Publications internationales

- Structural and content diversity of the mitochondrial genome in beet. A comparative genomic analysis,
Darracq A, Varré J-S, Maréchal-Drouard L, Courseaux A, Saumitou-Laprade P,Oztas S, Lenoble P, Vacherie B, Barbe V, Touzet P,
en préparation.

- A scenario of mitochondrial genome evolution in maize based on rearrangement events,
Darracq A, Varré J-S, Touzet P,
BMC genomics 2010 ;11 :233.

- Genetic architecture of zinc hyperaccumulation in Arabidopsis halleri : the essential role of QTL x environment interactions,
Frérot H, Faucon MP, Willems G, Godé C, Courseaux A, **Darracq A**,Verbruggen N, Saumitou-Laprade,
New Phytologist 2010 ;doi : 10.1111/j.1469-8137.2010.03295.x.

- Linkahe and association mapping reveals genetics for growth-related traits in *A. thaliana*,
Faure N, Brachi B, Villoutreix R, **Darracq A**, Bergelson J, Roux F,
en préparation.

Séminaires et Communications

- Scénarios d'évolution du génome mitochondrial chez *Zea mays* basés sur les évènements de réarrangements,
Darracq A,
GDR Génomique des Populations CNRS 1928, Paris (France) 2009, communication.

- A Study of Genomic Rearrangements in Maize Mitochondrial Genomes,
Darracq A, Varré J-S, Touzet P,
The 17th annual meeting of the Society for Molecular Biology and Evolution (SMBE), Iowa City (États-Unis), June 3-7 2009, poster.

- A Study of Genomic Rearrangements in Maize Mitochondrial Genomes,
  Darracq A, Varré J-S, Touzet P,
  JOBIM 2009, Nantes (France) 2009, poster.
- Evolution of the mitochondrial genome in beet. A comparative genomic study at the intra-specific level,
  Darracq A, Varré J-S, Courseaux A, Maréchal-Drouard L, Touzet P,
  XXth International Congress of Genetics, Berlin (Allemagne), 12-17 juillet 2008, poster.
- Évolution du génome mitochondrial. Approche par génomique comparative au niveau intraspécifique,
  Darracq A,
  Rencontres PPF, Lille, 2008, communication.
- Évolution du génome mitochondrial de *Beta vulgaris*. Approche par génomique comparative,
  Darracq A,
  Rencontres PPF, Lille, 2007, communication

Résumé

L'étude de l'évolution des génomes peut être abordée par différentes stratégies. Généralement, les analyses reposent sur les polymorphismes de séquences. Cependant, il existe des génomes dont le taux de mutation est très faible et dont la principale source de polymorphisme provient de l'arrangement différent de leurs gènes le long des chromosomes. Les événements de réarrangements chromosomiques deviennent alors les seuls marqueurs utilisables pour retracer l'évolution de ces génomes. Nous nous sommes intéressés dans ce travail à l'analyse de l'évolution des génomes mitochondriaux d'espèces végétales au niveau de leur structure. En effet, ces génomes sont caractérisés par un faible taux de mutation et un taux élevé de réarrangements. Cette étude s'est portée à un niveau intraspécifique afin de limiter le nombre de réarrangements à analyser et sur deux espèces : *Zea mays*, le maïs, et *Beta vulgaris*, la betterave. Il s'avère, qu'en plus du polymorphisme de structure, ces génomes contiennent un grand nombre d'éléments dupliqués. Or les outils d'analyse d'événements de réarrangements ne permettent pas d'inclure les événements de duplication autrement qu'en distinguant les paralogues des orthologues, ce qu'il est particulièrement difficile à réaliser ici, du fait que les dupliqués sont identiques en séquence. Nous avons ici établi une stratégie basée sur l'hypothèse que les éléments dupliqués proviennent de duplications en tandem, permettant la reconnaissance, le tri et la distinction des éléments dupliqués. Cette méthode nous a conduits à proposer une histoire évolutive basée sur des réarrangements congruente avec les phylogénies de séquences. Les comparaisons entre génomes mitochondriaux de maïs et betteraves nous ont permis de montrer que des mécanismes évolutifs différents sont à l'origine de la diversité génomique observée. Nous avons également observé des différences évolutives entre les génomes à un niveau intraspécifique soulevant le problème d'échantillonnage lorsque l'on veut comparer des génomes à un niveau interspécifique.

Mots clefs : bio-informatique, génomes mitochondriaux de plantes, rearrangements, génomique comparative

Abstract

Several methods can be used to study genome evolution. Most of the time, genome evolution is studied through nucleotide sequence polymorphism. However, in some species, mutation rate is low and polymorphisms are mainly caused by chromosomal rearrangements. In such a case, chromosomal rearrangement is the only informative marker to study genome evolution. In this study, we focused on plant mitochondrial genome evolution at the structural level. Plant mitochondrial genomes have been described as highly rearranged, but no study has been conducted on their rearrangement evolution. We chose to analyze the diversity of plant mitochondrial genomes at the intraspecific level to work on a short evolutive scale, limiting rearrangement events among genomes. The study was conducted on two species : *Zea mays* and *Beta vulgaris*. Moreover, besides structural polymorphisms, plant mitochondrial genomes contain large number of duplicated elements which are not taken into account by rearrangement tools if orthologous and paralogous relations are not established. Based on the hypothesis that the duplicated elements were caused by tandem duplication events, we proposed a new approach to find, sort and differentiate duplicated elements. This method led to phylogenies based on rearrangement events consistent with phylogenies based on nucleotide sequences. The comparison of genome evolution between maize and beet allowed us to show the existence of different evolution histories and mechanisms between these two species. We also observed evolutionary differences at the intraspecific level, raising the question of sampling strategy when genomes are compared at the interspecific level.

Keywords : bioinformatics, plant mitochondrial genomes, rearrangements, comparative genomics