Identification et évolution des séquences orthologues par séquençage massif chez les polyploïdes

Julien Boutte

ANNÉE 2015

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Biologie

Ecole doctorale (Vie – Agro – Santé)

présentée par

Boutte Julien

Préparée à l'unité de recherche UMR-CNRS 6553 ECOBIO
Ecologie - Biodiversité - Evolution
Sciences de la Vie et de l'Environnement

Identification et évolution des séquences orthologues par séquençage massif chez les polyploïdes

Thèse soutenue à Rennes
le 03 décembre 2015

devant le jury composé de :

Karine ALIX
MC - INRA-CNRS AgroParis Tech / *rapporteur*

Philippe LASHERMES
DR - IRD Montpellier / *rapporteur*

Dominique LAVENIER
DR - INRIA/IRISA Rennes / *examinateur*

Jonathan WENDEL
PR - Iowa State University / *examinateur*

Malika AINOUCHE
PRU - Université de Rennes 1 / *directrice de thèse*

Armel SALMON
MC - Université de Rennes 1 / *co-directeur de thèse*

Remerciements

Ces travaux ont été réalisés au sein de l'équipe Mécanismes à l'Origine de la Biodiversité (MOB). Cette équipe, sous la responsabilité de Malika Aïnouche, fait partie de l'UMR-CNRS 6553 ECOBIO (Université de Rennes 1), dirigée par Françoise Binet que je remercie. L'ensemble de ces travaux a pu être réalisé grâce au financement de l'Université de Rennes 1 (Contrat de l'Ecole Doctorale Vie Agro-Santé), du Laboratoire International Associé, de « Partner University Funds » avec l'Iowa State University (USA) et le soutien du CNRS (INEE). Je remercie ces institutions pour ces financements qui m'ont permis de réaliser cette thèse. Je remercie les différentes plateformes qui ont rendu ce travail possible : BioGenouest, GenOuest et le Centre Commun d'Ecologie Moléculaire d'ECOBIO.

Je tiens à remercier l'ensemble des personnes qui ont joué un rôle de très près, de près ou de loin dans cette grande aventure. Je remercie en premier lieu Armel Salmon et Malika Aïnouche pour leur accueil, leur encadrement efficace, leur soutien et pour m'avoir offert en plus d'un sujet de thèse aux multiples « challenges », plusieurs années d'expérience en recherche et pédagogie, de multiples opportunités de collaborations et participations à des congrès, pour ces posters et présentations, ces voyages nationaux et internationaux, ces échanges, ces manips, ces ressources, … MERCI.

Je tiens à remercier l'ensemble des personnes ci-dessous, mais également celles que j'aurais pu oublier et qui m'en voudront énormément (je n'en doute pas un instant !) :

Je remercie les différentes personnes qui ont assuré le bon déroulement de cette thèse de l'extérieur (à travers les différents comités de thèse). Un grand merci à Christian Delamarche, Dominique Lavenier, Jonathan Wendel, Mathieu Rousseau-Gueutin. Merci pour votre temps, vos questions, vos remarques et vos encouragements.

Je tiens également à remercier les membres de mon jury de thèse, qui ont accepté d'évaluer mon travail. Je remercie Karine Alix et Philippe Lashermes pour avoir accepté d'être les rapporteurs de ma thèse. Je remercie également Jonathan Wendel et Dominique Lavenier pour avoir accepté d'être dans mon jury de thèse et d'examiner mon travail.

Je remercie Jonathan Wendel pour m'avoir accueilli dans son laboratoire à Ames (Iowa State University, USA) et pour avoir pris du temps pour moi, ainsi que l'ensemble de son équipe pour les opportunités d'échanges autour de leurs travaux sur la génomique des cotonniers polyploïdes (un grand merci à Lei Gong et Guanjing Hu, Corrinne Grover, Joseph Gallagher et Josef Jareczek, Rick Masonbrink, Anna Krush, Simon Renny-Byfield, Xinyu Zhu) pour cet accueil si chaleureux, ces attentions, …. Je remercie Jonathan et Kathleen Wendel pour leur accueil et cette soirée/repas à leur domicile.

Je remercie particulièrement Christian Delamarche, d'avoir accepté d'être mon tuteur et de suivre l'évolution de ce travail, d'assister à mes comités et de m'avoir donné la possibilité d'enseigner la bioinformatique à de nombreux étudiants. Je tiens à exprimer toute ma gratitude envers l'ensemble de l'équipe pédagogique de Biologie Végétale, pour cette expérience enrichissante: Malika, Kader, Armel, Michèle, Cécile, Agnès, Rozène, Abdel. Un grand merci à Kader, mon Neighbor-Joining de bureau ! Merci pour tout, les TPs et TDs de phylogénie, les publications, les discussions, les échanges scientifiques, le partage de vos connaissances sur la phylogénie…

Un remerciement particulier à Benoît Aliaga et Oscar Lima pour le clonage et (re)-séquençage du rDNA et à Mathieu Rousseau-Gueutin et Olivier Coriton pour l'analyse FISH du rDNA. Je remercie également Oscar Lima pour la préparation des banques d'ADN nécessaires pour la capture de séquences.

Je remercie les différents membres du groupe Poly-BNF (soutenu par l'initiative « défis émergents » de l'Université de Rennes 1) pour ces réunions et ces échanges sur la bioinformatique et la polyploïdie: Malika Aïnouche, Armel Salmon, Anne Marie Chèvre, Dominique Lavenier, Gilles Lassalle, Pierre Peterlongo, Claire Lemaitre et Julien Erabit. Je remercie les membres du comité d'organisation du colloque international ICI-SPARTINA 2014: Malika Aïnouche, Armel Salmon, Marie Thérèse Misset, Kader Aïnouche, Mathieu Rousseau-Gueutin, Morgane Gicquel, Hélène Rousseau, Olivier Troccaz, Sandra Rigaud, Tifenn Donguy, Valérie Haubertin, Isabelle Picouays et Valérie Briand. Merci à vous pour cette expérience partagée qui a permis le bon déroulement de cet événement.

Je remercie Eric Petit, Pascaline Le Gouard, Cécile Gracianne et Alice Baudouin pour les différentes collaborations sur les Nématodes (Eric et Cécile) et les Gorilles (Pascaline et

Alice). Je tiens à remercier Cyril Monjeaud et Yvan Lebras pour leur aide sur l'intégration de nouveaux outils et logiciels sur la plateforme Galaxy de GenOuest et pour avoir passé du temps à intégrer le logiciel Newbler sur la plateforme.

Le bon déroulement d'une thèse passe aussi par une administration efficace, composée de personnes toujours souriantes et accueillantes. Merci à Valérie Haubertin, Fabienne Defrance (on l'a eu ce siège !), Sandra Rigaud, Tifenn Donguy, Isabelle Picouays (promis je diminuerai ma consommation d'impression couleur !) et Valérie Briand (merci d'avoir trouvé ces articles si rares !). Je remercie Fouad Nassur et Thierry Fontaine pour les cultures en serre, le rempotage et l'entretien du conviron et Jean-Luc Foulon pour le soutien logistique du conviron. Je remercie également les différentes personnes de l'équipe qui ont mis les mains dans la vase pour le nettoyage du conviron et lors des sorties terrain sur les marais.

Un grand merci à Julie Ferreira de Carvalho. Merci de m'avoir intégré dans tes travaux, merci de tes conseils, discussions, invitations, merci pour tout. Je remercie également David Heijnen, Delphine Giraud et Yannick Namour, qui ont dû subir mon co-encadrement et ont apporté leur contribution dans les recherches sur les Spartines. Je remercie aussi les autres stagiaires que j'ai pu côtoyer durant mon séjour à ECOBIO. Merci aux différents thésards et docteurs de l'équipe et de l'unité (ou autres) pour les discussions, pauses cafés, gâteaux, …, merci à toi Anne Sophie pour les pauses cafés, les discussions, les balades en forêt (enfin plutôt le taxi-forêt), merci Hélène (où sont passés tes gâteaux maisons ?), Jean (nous sommes de vrais petits jardiniers en herbe ! Le rempotage de Lupins et de Spartines n'ont plus aucun secret pour nous !), Xavier (merci pour les parties de Go endiablées et le chocolat), Sarah (courage pour la suite !), Anne Anto (promis je m'occuperai de ce vélo !), Nathalie (c'est pour bientôt !), Kevin (je serai bien passé plus souvent mais ton bureau était vraiment trop loin ^^), Malika, Armand, Erwan, Moez, Rahma, Mounir, Tina, Amina, Guillaume, Julie, Blanca (je ne verrai plus certaines chansons du même oeil), … Un remerciement aux ATERs que j'ai pu côtoyer, merci Paula, Morgane. Je remercie également les différents permanents, chercheurs et enseignants-chercheurs qui ont contribué au bon fonctionnement de cette thèse (apport scientifique et/ou chocolats/gâteaux): Merci Marie-Thérèse, pour les chocolats, les histoires, les résumés de voyages et les photos qui vont avec, pour mon petit surnom, pour le chocolat blanc… Merci

beaucoup. Marie Andrée, Michèle, Rozenne, Agnès : merci pour tout, le chocolat, les livres, les repas, les ballades, le Go…, Merci à Paco, Abdel, Alex, Anne-Marie, Carine, Claire, Pierre, Dominique et Ales. Merci à Sophie, Maude et Alexandra. Merci à Stéphanie Llopis pour les pucerons, nos larves de coccinelles n'auront pas survécu longtemps mais on aura essayé. Merci à toi Louis, pour les petits dépannages, les aides, les poses de câbles RJ45…

Je remercie également les personnes qui ont partagé les bureaux avec moi (et ce n'est pas facile tous les jours), merci à toi Armel de m'avoir accueilli si longtemps, cela m'a fait bizarre de quitter ton bureau après tout ce temps, merci à toi Paula, nous aurons bien rigolé, même si les post-its nous attendent encore dans un placard… Merci à toi Morgane, notre cohabitation n'aura pas duré très longtemps mais ce fut bien sympa. Et merci à toi Hélène, toi qui m'a supporté un petit moment. Toi qui as appris à différencier les moments où je parle à mon ordinateur et les moments où je communique avec un être humain. Il y aurait tant à dire, le lancer de chocolat et de Dragibus de bureau à bureau …, mais lister ce qui se jette dans ce bureau serait trop long, alors je m'arrêterai là.

Je remercie (encore !) Hélène (et Ghislain) et Jean (j'ai hésité à copier-coller la chanson de « La cucaracha ») pour ces soirées, ces cocktails et chansons (et leur aide pour la dernière ligne droite !). Anne Anto (et Jonathan) pour ces soirées cartes arrosées de rhum. C'était des supers soirées et de bons moments… Isabelle, Virginie, merci à vous deux pour ces petites pauses, j'en garderai de bons souvenirs. Et merci aussi pour le reste, le bureau sans « bééeehh » et la logistique sans faille.

Un grand merci à Mathieu et Pauline, merci pour ces repas, ces balades en forêt, ces moments TV, ces séances-peintures et tout le reste. Cette thèse aurait vraiment été différente sans vous.

Je remercie l'ensemble de ma famille qui s'agrandit d'année en année, et merci à vous Maman, Papa, merci pour tout. Merci pour votre soutien, vos attentions, vos visites et vos inquiétudes ! Enfin, je te remercie Elsa. Sans toi il n'y aurait eu ni thèse, ni master de bioinformatique à Rennes. Je te remercie pour tout, ton soutien sans faille, ton aide, tes attentions et ton amour. Je te remercie d'être là pour moi chaque jour, de prêt, de loin ou dans le train qui relie Rennes à Avignon.

Table des matières

Introduction générale

INTRODUCTION GÉNÉRALE

Les génomes eucaryotes sont caractérisés par une redondance plus ou moins importante d'information génétique résultant de la superposition de différents mécanismes, dont la prolifération de séquences répétées (comme les éléments transposables), les duplications individuelles de gènes, de portions de chromosomes (duplications segmentaires) ou les duplications de génomes entiers (polyploïdie). Tous ces mécanismes jouent un rôle important pour l'évolution et l'adaptation des espèces ainsi que dans la formation de nouvelles espèces (Ohno 1970; Sanmiguel and Bennetzen 1998; Wendel 2000).

Chez les plantes, la polyploïdie est un mécanisme prépondérant de spéciation sympatrique qui intervient le plus fréquemment suite à la non-réduction gamétique au cours de la méiose (Ramsey and Schemske 1998). Une espèce tétraploïde aura ainsi le double du nombre de chromosomes de son (ses) parent(s) diploïde(s) et s'en retrouve le plus souvent reproductivement isolée. La polyploïdie peut également intervenir par doublement des chromosomes somatiques (Mallet 2007). La duplication du génome peut avoir lieu au sein d'une même espèce, on parle alors d'**autopolyploïdie** (Ramsey and Schemske 1998). Les génomes autopolyploïdes résultent donc de la duplication d'un jeu de chromosome au sein de la même espèce et vont alors contenir plusieurs génomes **homologues**. La polyploïdisation peut aussi faire intervenir la réunion par hybridation de deux génomes qui ont divergé (chez deux espèces différentes) au cours de leur évolution, on parle alors d'**allopolyploïdie**. Dans ce cas, la nouvelle espèce allopolyploïde va contenir deux ou plusieurs jeux de chromosomes d'espèces différentes appelés génomes **homéologues**.

Si la polyploïdie retient depuis longtemps l'intérêt des botanistes (*e.g.* Stebbins 1950), elle est aujourd'hui reconnue comme un phénomène majeur et récurrent au cours de l'histoire évolutive des espèces (Wendel 2000). Ce phénomène est prédominant chez les plantes où l'on estime que la totalité de ces dernières dérivent d'ancêtres polyploïdes et semble avoir joué un rôle central au sein de toutes les lignées (Jiao et al. 2011; Van de Peer, Maere, and Meyer 2009). Ainsi, tous les génomes nucléaires des plantes actuelles présentent des traces, plus ou moins anciennes, d'événements superposés de duplication. Ces génomes plus ou moins anciennement polyploïdes sont affectés par le processus de

diploïdisation, résultant de la perte de l'une des copies (« Fractionnement », Langham et al. 2004). Ces pertes de copies ne se réalisent pas toujours de manière aléatoire (Zhang 2006; Hittinger and Carroll 2007; Bikard et al. 2009; Lloyd et al. 2014) et peuvent être soumises à différentes pressions de sélections (Blanc and Wolfe 2004). Il existe également de nombreux exemples de polyploïdisation chez les animaux (Mable 2004; Van de Peer, Maere, and Meyer 2009). En effet, de nombreuses espèces polyploïdes ont été recensées chez les insectes (au sein des Coléoptères, Curculionidés, Hyménoptères ou Lépidoptères par exemple), les mollusques, comme l'huître du Pacifique autotétraploïde *Crassostrea gigas*, chez les crustacés et les poissons où la famille des Salmonideae est entièrement composée d'espèces polyploïdes. Ce phénomène est également présent chez les amphibiens, ou environ 30 espèces ont été identifiées comme polyploïdes (Gregory and Mable 2005; Mable 2004). Les espèces polyploïdes sont moins nombreuses au sein des mammifères, deux espèces tétraploïdes de rongeurs (*Tympanoctomys barrerae* et *Pipanacoctomys aureus*) ont néanmoins pu être identifiées (Gallardo et al. 1999; Mares et al. 2000; Gallardo et al. 2004). Beaucoup de plantes cultivées sont des exemples bien connus de spéciation allopolyploïde (blés durs, blés tendres, colza, caféier, cotonnier) et illustrent les opportunités fournies par la duplication du génome dans la sélection de traits d'intérêt pour la domestication. On connait également des cas de spéciation allopolyploïde très récente (de l'ordre d'un siècle) chez les espèces sauvages comme dans les genres *Tragopogon* (Malinska et al. 2011) et *Senecio* (Abbott et al. 2008) chez les Asteraceae, *Cardamine* (Marhold et al. 2009) chez les Brassicaceae, *Mimulus* (Vallejo-Marin 2012) chez les Phrymaceae et *Spartina* (Ainouche, Baumel, and Salmon 2004) chez les Poaceae. Dans la plupart des cas, ce type de spéciation s'est accompagné d'une expansion rapide de l'espèce néopolyploïde qui a pu coloniser une plus large gamme de conditions environnementales, voire devenir envahissante (Ainouche et al. 2008).

Suite à la duplication du génome, les polyploïdes vont présenter plusieurs copies homéologues par locus pouvant évoluer de différentes façons possibles (Ohno 1970; Wendel 2000; Lynch and Force 2000; Adams et al. 2003) : la perte de l'une des deux copies (diploïdisation), l'accumulation de mutations délétères entrainant à une pseudogénisation, ou la modification de l'expression par sous-fonctionnalisation ou néo-fonctionnalisation. Une évolution concertée des copies homéologues par conversion génique peut être

également observée (Wendel, Schnabel, and Seelanan 1995; Lim et al. 2000; Gaeta et al. 2007; Nicolas et al. 2007; Kovarik et al. 2008; Salmon et al. 2009; Gaeta and Pires 2010; Szadkowski et al. 2010; Feliner and Rosselló 2012; Flagel, Wendel, and Udall 2012; Chalhoub et al. 2014). L'évolution de l'expression des gènes mesurée globalement pour un gène donné chez les polyploïdes a fait l'objet de plusieurs études ces dernières années (*e.g.* Ma et al. 2005; Poole et al. 2007; Chelaifa, Monnier, and Ainouche 2010; Chelaifa, Mahé, and Ainouche 2010; Bardil et al. 2011), mais peu de travaux ont pu distinguer l'expression de chaque copie homéologue (*e.g.* Comai 2000; Adams et al. 2003; Adams 2004; Wang 2004; Adams and Wendel 2005; Chen 2007; Flagel and Wendel 2009; Akhunova et al. 2010; Buggs et al. 2010; Combes et al. 2012; Higgins et al. 2012; Ilut et al. 2012; Combes et al. 2013; Yoo, Szadkowski, and Wendel 2013; Akama et al. 2014; Chalhoub et al. 2014).

Les événements de polyploïdisation peuvent également entrainer des changements phénotypiques plus ou moins importants. L'effet « Gigas » (gigantisme) de la polyploïdie sur la taille des cellules ou des organes est bien connu (Lewis 1980). Généralement, les plantes polyploïdes sont plus grandes et plus robustes que leurs parents diploïdes et vont produire des fleurs ou des graines de taille plus importantes (te Beest et al. 2012). Plusieurs études ont ainsi montré une différence morphologique entre les espèces polyploïdes et leurs parents diploïdes et/ou polyploïdes. Des changements phénotypiques ont par exemple été observés au niveau des fleurs et des feuilles chez des espèces allotétraploïdes *d'Arabidopsis* (*Arabidopsis suecica* et des hybrides synthétiques) par rapport aux parents *A. thaliana* (autotétraploïde) et *A. arenosa* (autotétraploïde) (Madlung 2002; Chen 2007). Il a également été montré des différences phénotypiques au niveau des feuilles et des tubercules entre les pommes de terre diploïdes et autotétraploïdes (Stupar et al. 2007). Schranz et Osborn (2000) ont montré des périodes de floraison différentes entre les sous lignées des hybrides synthétiques de colza (*Brassica napus*), ce qui peut avoir contribué à la réussite et la diversification des organismes polyploïdes. Une autre étude menée sur le colza a mis en évidence le lien entre échanges homéologues et variation phénotypique des polyploïdes synthétiques du colza (Gaeta et al. 2007).

Comprendre l'histoire de ces duplications qui ont façonné les génomes actuels, et leurs conséquences sur le fonctionnement des génomes est devenu un champ de recherches important en biologie (Van de Peer, Maere, and Meyer 2009). Néanmoins, l'identification

des copies homéologues au sein de génomes complexes nécessitait des approches lourdes combinant des méthodes de clonage, de séquençage ainsi que des méthodes phylogénétiques permettant d'inférer l'origine des copies détectées. Aujourd'hui, les outils de séquençage massifs tel que le pyroséquençage (développé par Roche-454 Life Science), le séquençage par synthèse (développé par Solexa/Illumina) ou le séquençage par Nanopore (développé par Oxford Nanopore Technologies) par exemple, offrent la possibilité d'explorer le génome d'espèces non-modèles (Metzker 2009). Néanmoins, face à la masse de données disponibles il est nécessaire aujourd'hui de développer des approches bioinformatiques pour répondre à des questions biologiques précises comme la détection des copies homéologues au sein de génomes polyploïdes. En effet, la détection des différentes copies au sein de tels génomes représente un défi, car en plus des copies paralogues (présentes au sein des génomes diploïdes) vient s'ajouter une couche supplémentaire de complexité avec la présence de copies homéologues issues des événements de polyploïdie (Ainouche et al. 2012). Plusieurs études se sont intéressées à la détection des différentes copies au sein de génomes polyploïdes comme le soja, le coton, le blé ou le caféier, néanmoins les stratégies développées ne peuvent être appliquées que sur des espèces dont les parents diploïdes sont connus (Flagel et al. 2008; Ilut et al. 2012; Salmon et al. 2012; Combes et al. 2013; Page et al. 2013; Pfeifer et al. 2014). La détection des différentes copies au sein d'espèces où les parents ne sont pas présents (ou identifié) nécessite le développement de nouveaux outils.

C'est notamment le cas dans le genre *Spartina* qui a subi plusieurs évènements de duplications de génomes au cours de son histoire. Les Spartines ont hérité de la paléoduplication commune aux Poaceae (il y a environ 70 MA). Après avoir divergé des autres Chloridoideae, il y a 12 à 20 MA (Rousseau-Gueutin et al. 2015), ce genre a subi une succession de polyploïdisations (souvent combinées à de l'hybridation) ayant abouti à différents niveaux de ploïdie (tétra- à dodécaploïde) chez les espèces actuelles (Ainouche et al. 2012). Ce genre est notamment connu pour son exemple classique de spéciation récente par allopolyploïdie en Europe, *Spartina anglica*, une jeune espèce envahissante issue de la duplication du génome de l'hybride *Spartina x townsendii* vers 1890, résultant du croisement entre deux espèces hexaploïdes (*Spartina maritima* et *Spartina alterniflora*). Un second croisement naturel entre ces deux espèces hexaploïdes a par ailleurs donné naissance à une autre espèce hybride *S. x neyrautii* dans le sud de la France (vers 1892). Par-

delà l'intérêt que ces espèces suscitent au plan écologique en raison de leur rôle dans la dynamique sédimentaire des marais salés (Strong and Ayres 2013) ; ce système représente un modèle permettant d'explorer les conséquences de l'évolution à court-terme des populations suite à l'hybridation et à la duplication du génome (Ainouche, Baumel, and Salmon 2004). *Spartina anglica* est une espèce allododecaploïde et la détection des différentes copies dupliquées dans ce génome représente un défi. En nous appuyant sur des ressources génomiques et transcriptomiques récemment développées au laboratoire (Ferreira de Carvalho et al. 2012; Ferreira de Carvalho et al. 2013, cette étude) par pyroséquençage Roche-454 (qui génère de longs fragments de séquences plus facile à assembler) et par séquençage de synthèse Illumina (qui fournit une plus grande profondeur de lecture de séquences permettant de corriger les erreurs de séquençage), nous avons cherché à identifier les différentes copies dupliquées ainsi que leur origine évolutive chez ces espèces.

Les objectifs de cette thèse sont plus particulièrement :

1- Développer des outils bioinformatiques (regroupés au sein de pipelines) dans le but de détecter les différentes copies dupliquées (copies homeologues et/ou paralogues). Ces programmes permettent de détecter des copies au sein de génomes hautement dupliqués, sans utiliser de génomes de référence diploïdes. Pour cela nous avons dans un premier temps détecté les différents sites polymorphes tel que les SNPs ou indels puis construit les différents haplotypes (ou copies) à partir de ces sites variables.

2- Construire *de novo* cinq transcriptomes de référence pour les espèces de Spartines étudiées (les parents, hybrides et allopolyploïde) à partir d'une méthode spécifique combinant différents jeux de données NGS. Pour cela nous avons utilisé l'ensemble des données transcriptomiques (RNA-seq) disponible au sein du laboratoire issues des technologies Roche-454 et Illumina. Pour chaque séquence ou contig obtenu, il a ensuite été possible de détecter les différentes copies associées.

3- Développer des outils permettant de déterminer l'origine des copies détectées en se basant sur des approches phylogénomiques et de discriminer l'origine des différentes copies (paralogues ou homéologues) potentiellement présentes dans les génomes hautement redondants du genre *Spartina*.

Ce manuscrit de thèse est structuré de la manière suivante :

I. Un premier chapitre, divisé en deux parties présente (i) les différentes méthodes NGS (Next Generation Sequencing) disponibles à ce jour ainsi que les méthodes et outils de traitement spécifiquement conçus pour les jeux de données massifs et (ii) les méthodes de détection de copies dupliquées au sein d'espèces diploïdes et polyploïdes.
II. Le chapitre 2 décrit le genre *Spartina* au sein de la famille des Poaceae et présente plus particulièrement l'évolution des Spartines européennes.
III. Un troisième chapitre détaille le matériel biologique et les méthodes développées et utilisées au cours de ce travail.
IV. Le chapitre 4 présente les travaux de détection de copies dupliquées à partir de données Roche-454. Ce chapitre s'articule autour de deux parties. La première partie correspond au développement et à la validation de la méthode employée sur l'ADN ribosomique de *S. maritima*. Ces travaux sont présentés sous la forme d'un article (en presse) dans la revue G3 : Genes|Genomes|Genetics sous le titre : « Haplotype detection from next generation sequencing in high ploidy-level species: 45S rDNA gene copies in the hexaploid *Spartina maritima* ». La seconde partie se concentre sur l'application de cet outil sur des données transcriptomiques des cinq espèces étudiées.
V. Le chapitre 5 présente les outils de détection de SNPs (Single Nucleotide Polymophisms) et des différentes copies à partir de données de séquençage par synthèse (Illumina). Les différents résultats sont comparés avec ceux obtenus dans le chapitre 4 et nous avons pu également étudier l'impact des paramètres des outils développés sur les résultats du programme. Les outils développés ont pu être appliqués sur 13 gènes d'intérêt dont les niveaux d'expression ont été étudiés dans des populations naturelles de Spartines (article en cours de préparation). La quatrième partie de ce chapitre, présentée sous la forme d'un article en préparation « Reference transcriptomes and detection of duplicated copies in hexaploid parents, hybrids and

allododecaploid *Spartina* species (Poaceae) » correspond à la construction de 5 nouveaux transcriptomes de référence de Spartines et s'intéresse à la détection de copies et des événements de duplication au sein de ces espèces. La dernière partie de ce chapitre présente la mise en place d'un pipeline permettant de réaliser un grand nombre d'analyses phylogénomiques afin d'étudier l'histoire des haplotypes reconstruits à partir des jeux de donnés Illumina RNA-seq.

VI. Le chapitre 6 présente une conclusion générale de ces travaux de thèse et des différentes perspectives offertes par ce travail.

La partie annexe, présente différents travaux pour lesquels j'ai apporté ma contribution et qui ont donné lieu à trois articles.

Chapitre 1 :

Séquençage à haut débit et détection de copies dupliquées.

CHAPITRE 1: Séquençage à haut débit et détection de copies dupliquées.

INTRODUCTION

Le séquençage de génomes tel que celui d'*Arabidopsis thaliana* (*Arabidopsis* Genome Initiative 2000), du riz (Yu et al. 2005) ou plus récemment du colza (Chalhoub et al. 2014) a permis de révéler l'importance et de préciser l'histoire de nombreux événements de duplications géniques et génomiques au sein des espèces. Les gènes présents en copie unique dans les gamètes s'avèrent peu nombreux, en particulier dans les contextes de polyploïdie récurrente. Il est donc important de définir les relations d'homologie entre ces copies de gènes. Ainsi, les gènes **orthologues** entre deux espèces ('ortho' voulant signifier 'exact') sont des gènes homologues qui ont divergé par spéciation (c'est le cas des copies B1 et C1 ou B2 et C2 représentées dans la Figure 1). Les gènes orthologues permettent donc de retracer l'histoire des spéciations et leur utilisation est privilégiée en phylogénie des espèces. La génomique comparative a introduit le terme de **groupes orthologues** (constitués de **co-orthologues** au sein d'un même génome) faisant référence à un ensemble de gènes homologues qui ont évolué dans différents génomes suite à la spéciation. Les gènes **paralogues** ('para' voulant signifier 'près de' ou 'à côté de') correspondent aux gènes résultant d'une duplication génique (copies B1 et B2 ; Figure 1). Si la distinction entre les copies orthologues et paralogues est conceptuellement simple, son application peut se révéler plus complexe en raison de la dynamique (perte différentielle de copies d'une lignée évolutive à l'autre, divergence fonctionnelle) des copies dupliqués (Gabaldón and Koonin 2013). Dans le cas d'une hybridation interspécifique (suivie ou non de polyploïdie), les copies orthologues chez les espèces parentales sont réunies au sein d'un même noyau et sont dans ce cas appelées **homéologues** (B2 et C2 ; Figure 1). Certaines de ces copies peuvent être perdues au cours de l' évolution ultérieure des allopolyploïdes (processus de diploïdisation ; Figure 1).

Dans ce chapitre, nous présenterons les différentes technologies et méthodes de séquençage disponibles aujourd'hui, puis nous rappellerons les méthodes de détection de copies dupliquées au sein des espèces, ainsi que les méthodes permettant de déterminer l'origine de ces copies. Nous présenterons également les applications des NGS réalisées chez des espèces polyploïdes où l'identification de l'origine des copies dupliquées se présente comme une question majeure pour comprendre l'évolution des génomes et plus généralement pour appréhender l'analyse des génomes eucaryotes redondants.

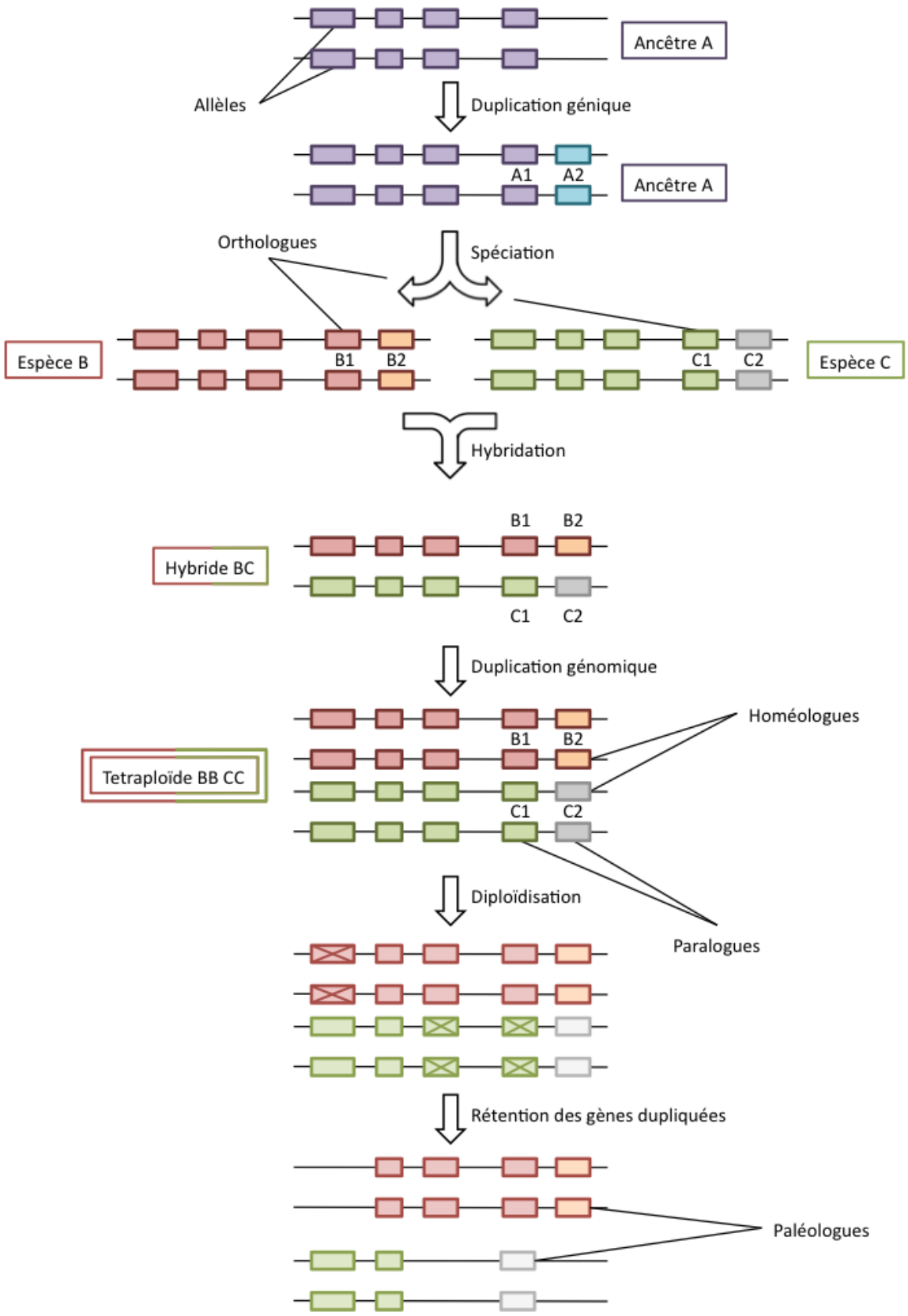


Figure 1: Terminologie et évolution des gènes dupliqués chez l'espèce allotetraploïde BBCC (paralogues : A1 et A2 ; homéologues : B1 et C1 ; B2 et C2). Une perte ou une rétention différentielle des copies dupliquées peut intervenir au cours de l'évolution à long terme des espèces polyploïdes (processus de diploïdisation ; adapté d'après Fortuné 2007; Ha, Kim, and Chen 2009).

PARTIE A. Les NGS, des outils et méthodes pour l'analyse de données génomiques à haut débit.

Les technologies de Séquençage :

Le séquençage de l'acide désoxyribonucléique (ADN) est apparu dans les années 1970 et a révolutionné la biologie moléculaire près de 20 ans après la découverte de la structure de l'ADN par Watson et Crick en 1953 (Watson and Crick 1953). Les deux premières techniques de séquençage d'ADN ont été développées en parallèle par Maxam et Gilbert d'une part (Maxam and Gilbert 1977) et Sanger et ses collaborateurs d'autre part (Sanger, Nicklen, and Coulson 1977). Ces techniques de séquençage ont été utilisées comme base pour l'ensemble des séquenceurs, jusqu'au développement des technologies de séquençage à haut débit « Next Generation Sequencing » (NGS).

La méthode de Maxam-Gilbert (1977) repose sur la dégradation chimique de l'ADN et utilise les réactivités des bases nucléotidiques pour réaliser des coupures sélectives. Il est alors possible de reconstruire la séquence nucléotidique en reconstituant l'ordre des coupures. Les fragments ainsi produits sont déposés sur gel de polyacrylamide et séparés par électrophorèse. L'utilisation de l'autoradiographie permet de détecter les bandes radioactives et de lire la séquence analysée. Un protocole de la méthode de séquençage automatisée de Maxam-Gilbert a été développé en 1994 (Boland et al. 1994) ; néanmoins cette technique de séquençage a été beaucoup moins utilisée que la méthode de Sanger, notamment utilisée pour le séquençage entier du génome humain (Lander et al. 2001).

La méthode de séquençage Sanger est une technique enzymatique dite aux « didésoxynucléosides triphosphates ». Cette méthode fait suite à la première technique de Sanger «plus and minus» développée en 1975 (Sanger and Coulson 1975). Le principe de cette technique est de synthétiser un brin d'ADN, radioactif et complémentaire du brin que l'on veut séquencer. Il est ensuite possible de lire la séquence complémentaire analysée grâce à de l'autoradiographie (Sanger, Nicklen, and Coulson 1977; Étienne and Millot 1998). Plusieurs méthodes ont été développées par la suite pour améliorer la méthode de séquençage Sanger en augmentant les débits de séquençage avec Le principe du « *Dye*

primer » (amorce fluorescente) ou du « *Dye terminator* » (nucléotides fluorescents) (*e.g.* Tagu and Moussard 2006).

L'automatisation du séquençage de l'ADN (utilisant une réaction basée sur la méthode Sanger) a donné la possibilité de séquencer un grand nombre d'échantillon en parallèle (à l'aide de séquenceurs mono-capillaires, puis 8-, 16- et 96-capillaires pour les plus récents) et une diminution du coût. L'arrivée des nouvelles technologies de séquençage ou NGS en 2005 (Margulies et al. 2005) permet aujourd'hui de générer des millions de séquences pour un coût considérablement réduit. Ainsi, le séquençage du génome humain qui avait nécessité 13 ans d'efforts et des moyens très importants (de l'ordre de 3 milliards de dollars) dans les années 2000 (Lander et al. 2001) ne représentait plus que 2 mois d'analyses pour un coût d'un million de dollars en 2008, grâce à l'utilisation de la technologie à haut débit Roche-454 (Wheeler et al. 2008). Aujourd'hui, le coût du séquençage d'un génome humain est de l'ordre de quelques milliers de dollars (Check Hayden 2014b). Ces technologies ont permis un essor considérable du nombre de génomes séquencés (~ 54 000 génomes d'eucaryotes et de procaryotes ; http://www.ncbi.nlm.nih.gov/genome/browse/), quelque soit leur taille.

Il est aujourd'hui possible d'utiliser les technologies NGS pour de nombreuses applications, sur l'ADN ou l'ARN. À partir d'ADN, il est possible d'effectuer du séquençage *de novo* (sans référence) ou du re-séquençage à partir du génome entier ou de l'ADN codant. L'utilisation d'ADN immuno-précipité va permettre l'étude des interactions d'ADN et protéines (séquençage ChIP-Seq) tandis que le séquençage de l'ADN méthylé permettra l'étude de la méthylation de l'ADN (MeDIPSeq, BisulfiteSeq). Le séquençage ClipSeq permet de séquencer l'ARN immuno-précipité pour étudier les interactions ARN/protéines. Il est également possible d'effectuer le séquençage d'ARN non codant, d'ARN messager ou ARNm (séquençage RNASeq) ou encore de micro-ARN (séquençage Small RNASeq). Les méthodes NGS se basent sur 3 principales étapes : la préparation des banques avec une étape de fragmentation de l'ADN et la ligation d'adaptateurs ; une étape d'amplification par PCR (qui est évitée par les techniques les plus récentes) ; et une étape de séquençage en temps réel des fragments obtenus.

Les technologies NGS présentent de nombreux avantages. Outre le coût de séquençage réduit, elles possèdent une très grande sensibilité de détection et produisent des fragments de séquences (lectures ou reads) de plus en plus longs et de plus en plus nombreux. Par exemple, la technologie Roche-454 était capable de produire en un cycle (« run ») 700 Megabases de séquences nucléotidiques d'une longueur de 700 paires de base en 2012 contre 20 Megabases pour une longueur de 100 bp en 2006. Cependant, ces technologies de séquençage à haut débit présentent également plusieurs limites. Les différentes technologies de séquençage NGS (présentées dans les paragraphes suivants) ne sont pas exemptes d'erreurs variant selon la chimie utilisée. Ainsi, des erreurs de séquençage vont être plus particulièrement localisées en début et fin de chaque read, au sein des régions homopolymères (correspondant à des répétitions d'un même nucléotide) et au niveau des insertions/délétions (Tableau 1; Oliphant et al. 2002). L'ensemble de ces erreurs entraine la présence de polymorphismes faux-positifs qu'il est nécessaire de corriger lorsque cela est possible, notamment grâce à la profondeur de séquençage et aux réplicas techniques (Robasky, Lewis, and Church 2014; Sleep, Schreiber, and Baumann 2013; Schulz et al. 2014). Une autre limite de ces technologies est la quantité de données produites par les machines qui représentent des centaines de Gigaoctets de données. Les différentes analyses bioinformatiques augmentent également le nombre de données et confrontent les utilisateurs à des problèmes de stockage de données. S'il est aujourd'hui possible pour un laboratoire de s'équiper de serveurs de stockage à des prix raisonnables (compter environ 4000 euros pour un serveur de stockage de 30 Téraoctets), il sera bientôt moins coûteux de re-séquencer si besoin les échantillons que de conserver les données brutes et traitées.

On compte à ce jour plusieurs technologies NGS, chacune proposant ses avantages et inconvénients (prix, profondeur de séquençage, longueur des lectures (« reads »), biais de séquençage). De plus, la demande étant toujours aussi importante, de nombreuses méthodes de séquençage sont également en cours de développement. Nous présenterons ici les différentes technologies disponibles et en cours de développement, ainsi que leurs caractéristiques propres.

Le Pyroséquençage :

La technologie de Pyroséquençage a été la première technique de séquençage à haut débit, développée en 2005 par Roche Life Sciences. Elle se décompose en 7 étapes et permet de produire jusqu'à 1 million de reads, d'une longueur moyenne de 700 bp par run (Tableau 1). La première étape consiste à fragmenter les différents échantillons qui peuvent correspondre à de l'ADN génomique, des amplicons, des BACs (« Bacterial artificial chromosomes ») ou de l'ADNc (500 ng d'ADN par run). Cette fragmentation aboutit a une banque de petits fragments sur laquelle va être fixé les différents adaptateurs en position 5' et 3', les brins d'ADN sont ensuite séparés. Des billes de streptavidine permettant de fixer l'ADN (par affinité des adaptateurs biotinylés) sont ensuite ajoutées, un seul fragment d'ADN se fixant par bille (en respectant le ratio nombre de molécules de la banque / nombre de billes). Chaque bille est ensuite isolée à l'aide d'une émulsion. La 4ème étape correspond à l'amplification de l'ADN par PCR en émulsion (sur billes) ou emPCR. Les billes sont ensuite déposées sur des plaques de séquençage (Pico-Titer Plate ou PTP) contenant plusieurs millions de puits dont le diamètre ne peut recevoir qu'une seule bille. Il est ensuite possible de réaliser la réaction de pyroséquençage à l'aide des différents réactifs de séquençage, de la sulphurylase et de la luciférase. La plaque multi-puits est ensuite insérée dans le pyroséquenceur où aura lieu le Pyroséquençage et la lecture des reads. Contrairement à la méthode de Sanger, les dNTPs sont ajoutés de façon séquentielle. Lors de la polymérisation et de l'incorporation de chaque nucléotide, un pyrophosphate (PPi) est libéré. La conversion chimique du pyrophosphate avec de l'APS (adénosine 5'-phosphosulfate) sous l'action enzymatique de l'ATP-sulfurylase entraine la formation d'ATP. La réaction enzymatique entre l'ATP et la luciférine (par le biais de la luciférase) va entrainer un signal lumineux spécifique. La lecture de ce signal est traduite en séquence nucléique à l'aide d'un programme informatique dit de « Base Calling » ou « Phred Base Calling » (http://www.454.com; Margulies et al. 2005; Metzker 2009; Voelkerding, Dames, and Durtschi 2009). Si cette technologie a permis l'acquisition de nombreuses données de séquençage pour divers organismes depuis sa commercialisation, elle est aujourd'hui de moins en moins utilisée du fait de l'apparition de technologies à plus haut débit et l'annonce

à l'automne 2013 par la société Roche de l'arrêt de la commercialisation de la chimie du pyroséquençage en 2016.

Le Séquençage par synthèse :

La technologie de séquençage par synthèse (initialement développé par Solexa) est commercialisée par l'entreprise Illumina. Cette technologie générant de courts fragments (« short read ») est capable de séquencer des reads pairés d'une longueur allant jusqu'à 250 bp et permet d'obtenir jusqu'à 1,2 milliard de reads en 60h (Tableau 1). Cette technologie utilise une cellule (« flow-cell ») qui correspond à une lame transparente composée de 8 lignes (« lanes ») individuelles où sont ancrés des oligonucléotides (Voelkerding, Dames, and Durtschi 2009). Dans un premier temps les échantillons d'ADN sont fragmentés en plusieurs centaines de paires de bases et réparés pour générer des extrémités franches en 5'. Une base d'adénine est ajoutée en 3' à l'aide d'un fragment de Klenow pour permettre la fixation des adaptateurs sur les lames. Contrairement à la PCR en émulsion, les échantillons d'ADN sont amplifiés dans la cellule « flow-cell » par ponts d'amplification (« bridge -PCR»). Pour cela, les fragments d'ADN simple brin sont fixés par hybridation aux oligonucléotides déjà ancrés sur la cellule. Ce processus qui se fait de manière aléatoire est suivi de la synthèse des brins d'ADN complémentaires (à l'aide de polymérases). Les deux extrémités des brins d'ADNc sont alors ancrées sur la cellule et forment des ponts qui sont amplifiés et dénaturés à de nombreuses reprises (les brins d'ADN initiaux étant éliminés). Ce processus permet la formation de nombreux ponts correspondant à des groupes ou « clusters » de séquences (un cluster regroupant les séquences d'ADN identiques). Le brin reverse est ensuite clivé et les amorces de séquençage sont hybridés avec des adaptateurs sur les séquences d'ADN. Les différentes séquences peuvent ensuite être séquencée par la méthode *dye terminator* réversible. Lors du séquençage l'incorporation de dNTPs fluorescents (une couleur différente pour chaque nucléotide) et une excitation au laser entraine l'émission d'une fluorescence (pour chaque cluster) qui est capturée et enregistrée. L'analyse d'image par la méthode de « Base Calling » permet d'obtenir les différentes séquences obtenues (Metzker 2009; Voelkerding, Dames, and Durtschi 2009). Il existe aujourd'hui trois types de séquençage

Illumina : le séquençage « Single-End » qui permet de séquencer les fragments d'ADN à partir d'une seule extrémité, le séquençage « Paired-end » qui permet de séquencer, avec une haute qualité, les deux extrémités des fragments d'ADN (d'une longueur inférieure à 1 kb). Les reads obtenus pouvant être alignés par paire, ce type de séquençage permet de détecter plus facilement les réarrangements génomiques, les éléments répétés ainsi que les fusions de gènes et les nouveaux transcrits. Le séquençage « Mate-Pair » permet comme le séquençage « Paired-end » de séquencer les deux extrémités des fragments d'ADN, mais d'une longueur de plusieurs kilobases. Ce type de séquençage peut être utilisé pour un grand nombre d'applications comme le séquençage *de novo*, la finition de génome (en permettant de réduire les zones non couvertes dans le génome et de relier les contigs l'un à l'autre pour créer des scaffolds), la détection de variants structuraux ou l'identification de réarrangements de complexes génomiques par exemple (http://www.illumina.com/).

Tableau 1:Propriétés des différents séquenceurs de nouvelle génération disponibles. D'après http://454.com, http://www.illumina.com, https://www.lifetechnologies.com, http://www.pacificbiosciences.com, Check Hayden 2012; Check Hayden 2014a; Liu et al. 2012; Pareek, Smoczynski, and Tretyn 2011.

Compagnie:	Roche Life Science	Illumina	Life Technologies			Helicos Biosciences	Pacific Biosciences	Oxford Nanopore Technologies
Séquenceur:	454 GS FLX Titanium XL+	HiSeq 2500	5500 SOLiD System	Sanger 3730xl	Ion Torrent PGM system	HeliScope SMS	PacBio RS II	MinION
Méthode de Séquençage:	Pyroséquençage	Séquençage par synthèse	Séquençage par ligation	Terminaison de chaîne	Ion semi-conducteur	Séquençage « Single molécule »	Séquençage « Single molécule », SMRT	Séquençage par Nanopore
Longueur moyenne des reads:	700 bp	2* 250 bp	75 bp * 35 bp ou 2 * 60 bp	400-900 bp	345 bp	55 bp	1000-40000 bp	5400 bp
Précision:	99,997%	99,9%	99,99%	99,999%	99,1%	99%	99% (précision de la séquence consensus)	96%
Nombre de reads par run :	~1 Million	~1,2 Milliard	~1,4 Milliard	-	6,310 Millions	NA	~ 50 000	NA
Nombre de bases par run :	700 Mb	250-300 Gb	90 Gb	690-1600 Kb	2,1 Gb	28 Gb	500 Mb-1Gb	90 Gb
Durée d'un run:	23h	60h	7 jours	1 jour	4-7h	8 jours	30 minutes	15 min
Avantages :	Longueur des reads	Nombre de reads produits	Prix par base séquencée	Longueur des reads	Séquençage rapide	-	Longueur des reads	Longueur des reads / Prix / Séquençage rapide
Inconvénients:	Erreur de séquençage au niveau des homopolymères / Prix du séquençage	Equipement coûteux, concentrations d'ADN importantes	Débit lent, présence de séquences palindromiques	Prix élevé, temps de clonage important	Erreur de séquençage au niveau des homopolymères	Taille des reads / Débit modéré	Débit modéré / Equipement coûteux	Erreur de séquençage / Profondeur de séquençage

Le Séquençage par Ligation :

Le séquençage par Ligation (plateforme SOLiD), proposé par la société Life Technologies utilise une ligase et des amorces marquées pour séquencer les brins d'ADN. Cette technologie permet d'obtenir un grand nombre de séquences (~1.4 milliard) pour un run d'une semaine. Néanmoins si le débit est similaire à celui de la technologie de séquençage par synthèse, la longueur des reads obtenus est moins importante (inférieure à 100 bp ; Tableau 1). Cette technologie utilise une préparation de banque assez proche de celle utilisée par la technologie de Pyroséquençage. Les brins d'ADN sont fixés sur des billes par l'intermédiaire d'adaptateurs préalablement liés aux séquences ; néanmoins les billes ne sont pas séparées dans des puits mais fixées sur des plaques de verre. Des sondes spécifiques sont ensuite liguées pour déterminer la suite de nucléotides à l'aide d'un signal fluorescent. Chaque sonde (de 8 pb) est composée de deux bases spécifiques, 6 bases dégénérées ainsi que de 4 fluorophores différents liés en 5'. Les deux bases spécifiques sont l'une des 16 possibilités de combinaison des nucléotides (TA, TT, TC et TG correspondent à 4 de ces 16 possibilités ; Voelkerding, Dames, and Durtschi 2009). Dans un premier temps les sondes sont hybridées et soumises à la détection par fluorescence, les sondes ne s'étant pas hybridées sont supprimées par une étape de lavage. Les sondes sont ensuite clivées et le cycle (ou « round ») est répété plusieurs fois. Le brin synthétisé est alors dénaturé et une nouvelle amorce est fixée sur la séquence à la position (n-1). Au total, 5 cycles sont effectués, avec un décalage d'une paire de base (positions (n-2), (n-3), etc); cette méthode permettant de lire deux fois chaque nucléotide de la séquence. Les résultats obtenus sont alors compilés pour obtenir la séquence d'ADN initiale (Metzker 2009; Voelkerding, Dames, and Durtschi 2009).

Le Séquençage Ion Torrent :

La méthode de séquençage Ion Torrent, proposée à partir de 2011 par la société Life Technologies est une technologie de séquençage qui n'utilise pas la détection de

fluorescence de nucléotides contrairement aux technologies présentées ci-dessus. La technologie Ion Torrent s'appuie sur un capteur CMOS (« Complementarity Metal-Oxide-Semiconductor ») qui va détecter les ions/protons H^+ libérés lors de la polymérisation de l'ADN. Le capteur CMOS est associé à des multi-puits et va mesurer au sein de chaque puits (un puits contenant un fragment d'ADN à séquencer) les variations de pH (potentiel Hydrogène) ; il est ainsi possible d'identifier en fonction du pH la base incorporée. Les variations de pH sont alors collectées et traitées avec le principe de « base calling ». La préparation de la banque d'ADN se décompose en 4 étapes : la fragmentation de l'ADN avec une sélection des fragments en fonction de leur taille, la ligation des adaptateurs A et P1 en 5' et 3' et la réalisation de 4 à 5 cycles de PCR. La méthode de préparation de la matrice de séquençage est similaire à celle utilisée pour la technologie Roche-454, les fragments d'ADN sont liés à des billes ISP (« Ion Sphere Particles ») par le biais d'une PCR en émulsion (emPCR). Les billes (avec fragment d'ADN) sont alors placées dans les micro-puits et l'ADN séquencé à l'aide du capteur CMOS (Quail et al. 2012; Rothberg et al. 2011; https://www.lifetechnologies.com) . Cette technologie de séquençage à haut-débit permet d'obtenir environ 6 millions de reads par run. Ce séquençage très rapide (de 4 à 7h par run) permet d'obtenir des reads allant jusqu'à 400 bp (Tableau 1).

Le Séquençage Simple Molécule (« Single Molecule ») :

L'une des premières techniques de séquençage d'ADN « Simple Molécule » a été introduite en 2003. Les premières commercialisations de ce système sont proposées à partir de 2007 par la société Helicos biosciences. Cette technologie de séquençage produit des reads de 55 bp pour un débit modéré (Tableau 1). La technique repose sur la technologie tSMS (« true Single Molecule Sequencing ») et n'utilise pas d'amplification par PCR ce qui lui donne l'avantage d'éviter les erreurs de séquençage liés à la PCR. Ainsi, l'échantillon d'ADN est dans un premier temps fragmenté puis lié (à l'aide d'une transférase) à une queue Poly(A) portant une étiquette fluorescente terminale. Les séquences sont ensuite hybridées sur une cellule « flow-cell » et séquencées simultanément dans des réactions parallèles. Le cycle de séquençage correspond à l'extension de l'ADN avec l'un des quatre nucléotides

marqué par fluorescence. Le signal lumineux est capté par le séquenceur à l'aide d'une caméra CCD (« Charge-Coupled Device »). Le clivage du fluorochrome permet au cycle d'élongation suivant de commencer par un autre nucléotide marqué par fluorescence. Ce cycle est répété jusqu'à l'obtention des différentes séquences ou reads (Harris et al. 2008; Pareek, Smoczynski, and Tretyn 2011).

Le Séquençage Simple Molécule en temps Réel (« Single Molecule Real Time ») :

Le séquençage d'ADN simple molécule en temps réel (SMRT) correspond à une autre méthode SMS basée sur le principe de séquençage par synthèse. Cette technologie, développée par la compagnie de biotechnologie Pacific Biosciences, permet d'obtenir jusqu'à 50 000 reads d'une longueur variant de 1 à 40 kb en 30 minutes (chimie P6-C4 ; Tableau 1). La technologie SMRT utilise un ZMW (« zero-mode waveguide ») qui correspond à une structure qui va créer un volume d'observation lumineux assez petit pour observer et identifier les nucléotides d'ADN incorporés sur le brin nouvellement généré par l'ADN polymérase. Cette technologie utilise de petits puits qui contiennent une seule enzyme (ADN polymérase) fixée à la surface d'un ZMW et une seule molécule d'ADN qui va servir de matrice. Chacune des quatre bases d'ADN est fixé à l'un des quatre colorants différents. Lorsqu'un nucléotide est incorporé par l'ADN polymérase, le marqueur fluorescent est clivé et lu par un détecteur (Figure 2 ; Eid et al. 2009; El-Metwally, Ouda, and Helmy 2014).

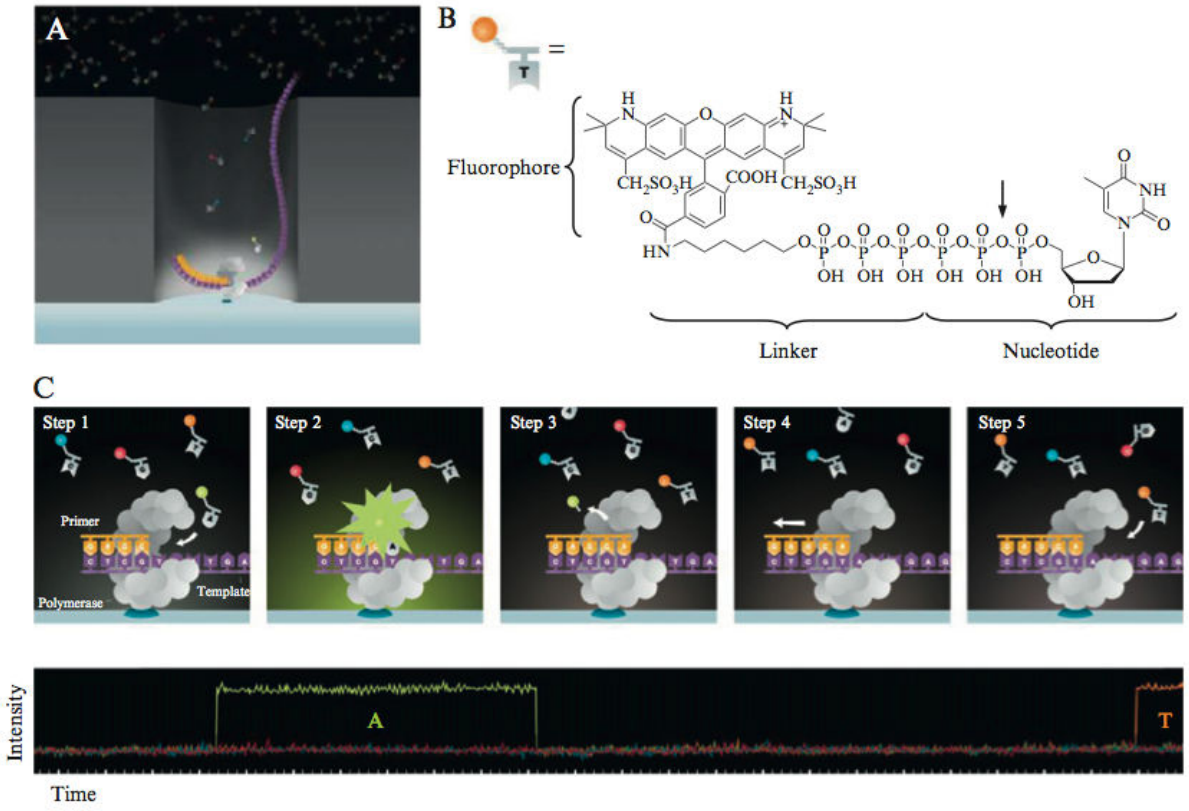


Figure 2: Principe du séquençage d'ADN par SMRT (Single Molecule Real Time). (A) L'ADN polymérase avec l'échantillon d'ADN lié sont immobilisés au fond des nanostructures ZMW. La polymérisation du brin complémentaire est observée en temps réel pour détecter la transformation enzymatique des nucléotides fluorescents liés au fluorophore. (B) Structure moléculaire des nucléotides liés au fluorophore dit « phospho-liés ». La flèche indique le site de clivage provoqué par l'ADN polymérase. (C) Représentation schématique des étapes du séquençage SMRT (en haut) et du tracé de l'intensité fluorescente correspondante. Etape 1 : Le complexe incluant l'échantillon d'ADN, le primer et l'ADN polymérase sont entourés par les nucléotides marqués qui sondent le site actif. Etape 2 : Incorporation d'un nucléotide marqué, avec l'émission de la fluorescence de manière continue. L'identité du colorant fluorescent indique la base incorporée. Etape 3 : La polymérase incorpore le nucléotide dans la chaine d'acide nucléique en clivant le fluorophore lié. La polymérase se positionne sur le site suivant du brin d'ADN. Etape 5 : Le processus est répété (D'après Korlach et al. 2010).

Le Séquençage par Nanopore :

Le séquençage par Nanopore développé depuis 2005 par la compagnie Oxford Nanopore Technologies se présente comme la nouvelle technologie à haut débit pour le séquençage de données génomiques. La commercialisation est aujourd'hui limitée à des programmes d'accès anticipé dans le but de tester la technologie (comme le « MinION Access Programme » ou MAP). Capable de produire jusqu'à 90 Gb en 15 minutes, cette technologie permet d'obtenir des reads d'une longueur supérieure à 5 kb. Néanmoins, cette technologie est encore limitée par une faible profondeur de séquençage et un taux d'erreur important, d'environ 4% (Tableau 1 ; El-Metwally, Ouda, and Helmy 2014). La société Oxford Nanopore Technologies développe plusieurs instruments comme le GridION 2000, le GridION 8000 ou encore le MinION qui est un dispositif USB à usage unique qui permet d'obtenir des reads d'une longueur théorique de 10 kb (voir El-Metwally, Ouda, and Helmy 2014 pour les avantages et inconvénients de chacun de ces instruments). La technologie de séquençage par Nanopore se base sur l'utilisation d'une surface composée de pores d'un nanomètre de diamètre. Le passage de l'ADN à travers le pore va modifier le courant ionique. Chaque type de nucléotide va modifier spécifiquement le courant ionique, la lecture du courant va ainsi permettre la lecture du brin d'ADN. La complexité de cette technique se localise au niveau des surfaces nanopores qui peuvent être constituées de semi-conducteurs ou de protéines tel que l'Alpha hémolysine ou la MspA (Mycobacterium smegmatis porin A ; Figure 3) (El-Metwally, Ouda, and Helmy 2014; Laszlo et al. 2014; Wang, Yang, and Wang 2015).

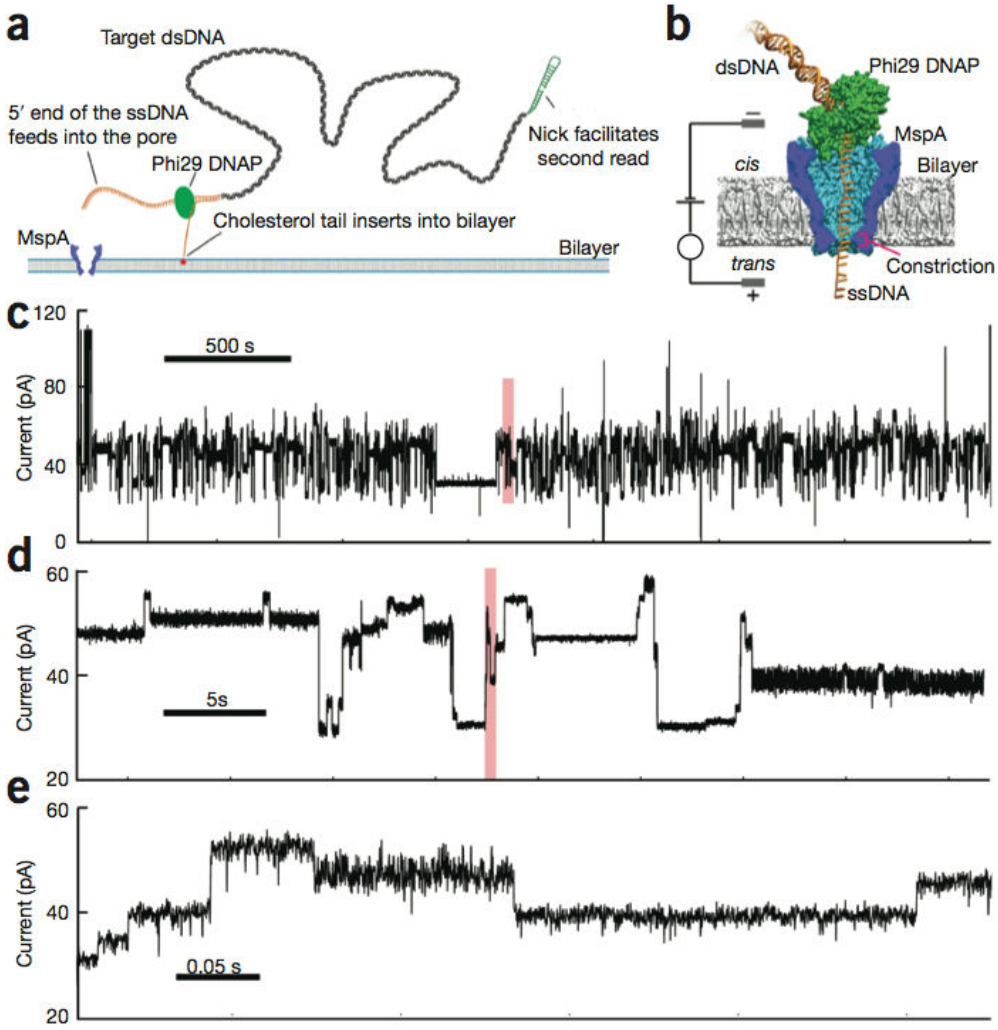


Figure 3 : Méthode de séquençage d'ADN par Nanopore. (a) Préparation de la banque d'ADN, le premier adaptateur en orange contient une queue de cholestérol qui est inséré dans la membrane ce qui permet l'augmentation du taux de capture de l'ADN, tandis que le simple brin en 5' facilite l'insertion dans le pore. Le second adaptateur (en vert) permet la relecture de l'ADN à partir de l'ADN polymérase. (b) La protéine Nanopore MspA est représentée en bleu, la Phi29 ADN polymérase est représenté en vert et l'ADN est identifié par la couleur orange. Une tension est appliquée dans la membrane ce qui entraine la formation d'un courant ionique à travers le pore ce qui permet de déterminer les bases de l'ADN. Le Phi29 permet le passage d'un nucléotide à la fois au sein du pore. (c-e) Représentation des données brutes obtenues pour une fenêtre de temps de 3 000 secondes. Les changements du courant ionique sont induits par le passage de chaque nucléotide indépendamment. Les étapes d et e correspondent à un agrandissement de la fenêtre en rouge et représentent 1% de la fenêtre de temps (d'après Laszlo et al. 2014).

De nouvelles technologies de séquençage à haut-débit sont en cours de développement et se basent sur des technologies similaires au séquençage par Nanopore (présenté ci-dessus) mais également sur des technologies de microscopie, le but étant d'augmenter le débit tout en diminuant le temps de séquençage et le prix. La technologie « Tunnelling currents » utilise la mesure de courants électriques pour déterminer la séquence d'ADN. L'ADN simple brin traverse un tunnel qui mesure le courant électrique propre à chaque base et détermine ainsi la séquence nucléotidique (Di Ventra 2013). Il existe de nombreuses autres méthodes en cours de développement comme le séquençage par hybridation, une méthode non-enzymatique qui utilise une puce à ADN, le séquençage par spectrométrie de masse qui utilise une approche par CE-MS (« Capillary Electrophoresis Mass Spectrometry ; El-Metwally, Ouda, and Helmy 2014), le séquençage par dénaturation (SBD ; Chen and Huang 2009) ou encore le séquençage RNAP (RNA Polymerase). Ce dernier se base sur l'utilisation d'une enzyme RNAP qui est attaché sur une bille de polystyrène et de la molécule d'ADN qui est connectée à une autre bille. Les deux billes sont ensuite placées dans le piège optique où l'information de séquençage sera obtenue par le mouvement de l'acide nucléique et la sensibilité du piège optique (de l'ordre de l'Angstrom). La différenciation entre les quatre types des nucléotides est obtenu en utilisant une approche similaire à la méthode Sanger (El-Metwally, Ouda, and Helmy 2014). D'autres méthodes de séquençage sont en cours de développement comme le séquençage « In vitro High-throughput » ou le séquençage « Microfluidic Sanger » ; néanmoins peu d'informations sont disponibles actuellement sur ces méthodes.

Aujourd'hui, nous disposons d'un large panel de techniques de séquençage à haut débit qui évoluent très rapidement. Il est donc nécessaire de choisir la technologie de séquençage en fonction des objectifs de l'étude menée. Ainsi, pour faciliter l'assemblage de données génomiques ou pour réaliser un séquençage *de novo*, l'utilisation de « longs reads » est répandue alors que l'utilisation de « reads courts », issus de la technologie Illumina par exemple, sont utilisés dans le cadre d'analyses de niveaux d'expression et la correction des erreurs de séquençage de reads longs.

PARTIE B. Le traitement des NGS : méthodes d'assemblage *de novo* et de « mapping ».

L'analyse des données de séquençage de nouvelle génération représente un défi et nécessite différentes approches selon la méthode de séquençage à haut-débit choisie et l'espèce étudiée. Ainsi, une stratégie d'alignement multiple par mapping va être utilisée si l'on dispose d'une référence, comme le génome de l'espèce étudiée ou celui d'une espèce proche par exemple. Ce processus permet d'aligner les différents reads obtenus sur une référence donnée pour ensuite analyser les jeux de données. Il existe de nombreux logiciels et méthodes disponibles en fonction de la nature du jeu de données (génomique, transcriptomique) et de la technologie utilisée (générant des fragments courts ou longs). Si aucune référence n'est disponible, les données seront assemblées entre elles à l'aide d'un assemblage dit *de novo* (sans référence). Tout comme pour la méthode de mapping, différentes méthodes et outils peuvent être utilisés en fonction de la technologie et la nature du jeu de données. Cette approche permet également d'analyser des séquences spécifiques de l'organisme étudié. Nous présenterons ci dessous les différents outils d'assemblages disponibles et leurs spécificités ainsi que les méthodes de mapping pour l'alignement de séquences de courts fragments.

L'assemblage à l'aide d'algorithmes « gloutons » (« Greedy graph »):

Les premiers logiciels d'assemblage de données NGS développés sont basés sur des algorithmes dits « gloutons ». Ces programmes se basent sur un algorithme simple : pour chaque contig ou read on ajoute un autre read (ou contig) présentant une région chevauchante. Ce processus est répété jusqu'au moment où plus aucune opération n'est possible. Chaque itération utilise le chevauchement entre les deux séquences étudiées pour créer la nouvelle séquence et utilise des paramètres d'assemblage comme le pourcentage d'identité. Les algorithmes gloutons sont des algorithmes de graphes implicites qui ne vont considérer que les arêtes présentant un pourcentage d'identité élevé (Figure 4, a.). Ils peuvent également éliminer chaque chevauchement qui a été utilisé pour former un contig

(Miller, Koren, and Sutton 2010). Plusieurs assembleurs ont été développés à partir de ces algorithmes, comme SSAKE (Warren et al. 2007) qui est le premier assembleur de fragments courts ou SHARCGS (Dohm et al. 2007) qui n'utilise que des reads courts non pairés de même longueur et présentant une grande profondeur de séquençage. Plusieurs assembleurs (tel que Newbler (Margulies et al. 2005) et Celera (Myers et al. 2000) ou Newbler et VCAKE (Jeck et al. 2007)) ont également été associés sous forme de pipeline pour des données Solexa/Illumina + 454 ou 454 + Sanger (Goldberg et al. 2006; Reinhardt et al. 2008).

Les assembleurs OLC (« Overlap Layout Consensus ») :

La méthode OLC était traditionnellement utilisée pour les assemblages de données Sanger. Elle a été optimisée pour l'assemblage de génomes et intégrée dans plusieurs logiciels d'assemblage (Miller, Koren, and Sutton 2010) tel que Celera (Myers et al. 2000), CAP ou PCAP (Huang and Yang 2005). Les assembleurs OLC se basent sur l'utilisation de graphes de chevauchement ou « Overlap graph » et se décomposent en 3 étapes. La première étape (« Overlap ») consiste à utiliser un algorithme heuristique permettant de trouver la « graine » (le point de départ de l'assemblage) et d'étendre le processus. Le programme identifie les k-mers (les k-mers d'un read correspondent à l'ensemble des séquences d'une longueur k contenues dans le read ; par exemple les 3-mers du read ACCGTG sont ACC, CCG, CGT et GTG) contenus dans l'ensemble des reads et sélectionne les données présentant les mêmes k-mers. Il est ensuite possible de construire les alignements à partir de ces k-mers qui correspondent à la graine de l'alignement. Cette étape utilise 3 paramètres, la taille des k-mers, la taille minimale de chevauchement entre les reads (en bp) et le pourcentage minimal d'identité nucléotidique entre deux chevauchements de reads. Ces trois paramètres vont directement affecter la qualité de l'assemblage. Ainsi plus ces paramètres seront élevés, plus l'assemblage sera de bonne qualité ; les contigs obtenus seront alors plus courts mais plus précis. La seconde étape de cet algorithme (« Layout ») repose sur la construction et la manipulation des graphes de chevauchement pour obtenir une mise en place approximative des reads. La troisième étape (« Consensus ») correspond à une étape d'alignement multiple de séquence (MSA) qui détermine la séquence consensus

et l'architecture de l'alignement (Figure 4, b.). Cette étape qui nécessite de grandes ressources informatiques est réalisée en parallèle et partagée contig par contig (Miller, Koren, and Sutton 2010).

Certains logiciels OLC ont été adaptés pour analyser des données issues de technologies NGS produisant de longs reads. C'est le cas de Celera, dont la version CABOG (Celera Assembler with Best Overlap Graph) est capable de traiter des données NGS (Miller et al. 2008). Une nouvelle version de Celera (PBrC) vient d'être développée et permet l'assemblage de très longs reads issus de la technologie SMRT (Berlin et al. 2015). Pour cela, Berlin *et al.* ont intégré le processus MHAP (« MinHash Alignment Process ») dans l'assembleur Celera ce qui a permis le ré-assemblage de plusieurs génomes comme celui d'*Arabidopsis thaliana* ou de *Drosophila melanogaster* (Berlin et al. 2015). D'autres logiciels d'assemblage de données SMRT ou d'assemblage hybride Illumina/SMRT sont également disponibles (voir Koren and Phillippy 2015 pour le descriptif de ces logiciels). Le logiciel MIRA (Mimicking Intelligent Read Assembly) se base également sur des algorithmes OLC et permet de réaliser des assemblages hybrides (Chevreux, Wetter, and Suhai 1999). En effet, si ce logiciel a été initialement développé pour assembler des données issues de la technologie Sanger, de nouvelles versions permettent de réaliser des assemblages de données issues de la technologie Roche-454 ou Illumina mais également des assemblages hybrides Sanger/454, 454/Illumina ou encore Sanger/Illumina/454. Une nouvelle version permet à présent de réaliser des assemblages de données Iontorrent ou PacBio (http://sourceforge.net/projects/mira-assembler/). Ce logiciel se décompose en plusieurs étapes : dans un premier temps les reads sont lus et les régions contaminées (correspondant à des vecteurs) sont identifiées et masquées. Les chevauchements potentiels des reads sont ensuite identifiés. L'algorithme de Smith-Waterman est utilisé pour aligner les reads, puis un graphique d'assemblage (« Layout ») est construit. MIRA identifie les régions répétées et les erreurs de séquençage et crée les différents contigs (Abegunde 2010). Même si de nombreux assembleurs de données Roche-454 sont disponibles, la majorité des assemblages de longs fragments sont réalisés à l'aide du logiciel Genome Assembler communément appelé Newbler (Margulies et al. 2005). Ce logiciel OLC, distribué par Roche-454 Life Sciences permet l'assemblage de reads longs en un temps restreint et utilise la profondeur du séquençage pour corriger les erreurs de séquençage présentes dans les données NGS

(lorsque cela est possible). De plus, il peut être aussi bien appliqué sur des données génomiques que sur des données transcriptomiques, grâce à l'utilisation de l'option « -cDNA » (qui permet également de détecter des épissages alternatifs). Deux logiciels utilisent cette méthode d'assemblage OLC : Edena (Hernandez et al. 2008) et Shorty (Hossain, Azimi, and Skiena 2009)) à partir de « short reads » issus de la technologie SOLiD. Néanmoins, devant la quantité de données produite par les technologies NGS de fragments courts comme SOLiD ou Illumina, de nouveaux assembleurs ont été développés en se basant sur les graphes de *de Bruijn*.

L'assemblage à l'aide des Graphes de *de Bruijn* :

Un graphe de *de Bruijn* (GdB) est un graphe orienté qui permet de représenter l'ensemble des chevauchements de longueur *n-1* (les différents k-mers) entre tous les mots (les reads) de longueur *n* ; et ceci pour un alphabet donné (l'ensemble des reads). Les graphes composés des k-mers ne vont pas contenir les reads individuels ni leurs chevauchements et compressent les séquences redondantes. Cet algorithme se base sur le scénario idéal selon lequel les données fournies ne présentent pas d'erreur de séquençage et les k-mers offrent une couverture complète. Le graphe obtenu correspondrait alors à un graphe de *de Bruijn* qui va contenir un chemin *eulérien* (un chemin qui parcourt chaque arête une seule fois ; Figure 4, c.). Pour utiliser cette approche sur des données réelles, la construction du graphe est réalisée à l'aide d'une table de hachage (de complexité constante *O(1)*) qui va rechercher chaque k-mer dans le flux de données (Miller, Koren, and Sutton 2010). Une table de hachage correspond à une structure de données qui permet l'association d'une clé et d'un élément. Pour améliorer les assembleurs et les algorithmes, plusieurs travaux basés sur la résolution des problèmes engendrés par les répétitions génomiques au sein des graphes ont été utilisés (Idury and Waterman 1995; Pevzner 1989). D'après Miller *et al.* (2010) plusieurs facteurs compliquent l'application des graphes de k-mers sur les données d'ADN génomiques : le fait que l'ADN soit double brin, ce qui entraine la présence de structures répétées complexes dans le génome (répétitions en tandem, répétitions inversées), les séquences palindromiques et les erreurs de séquençages.

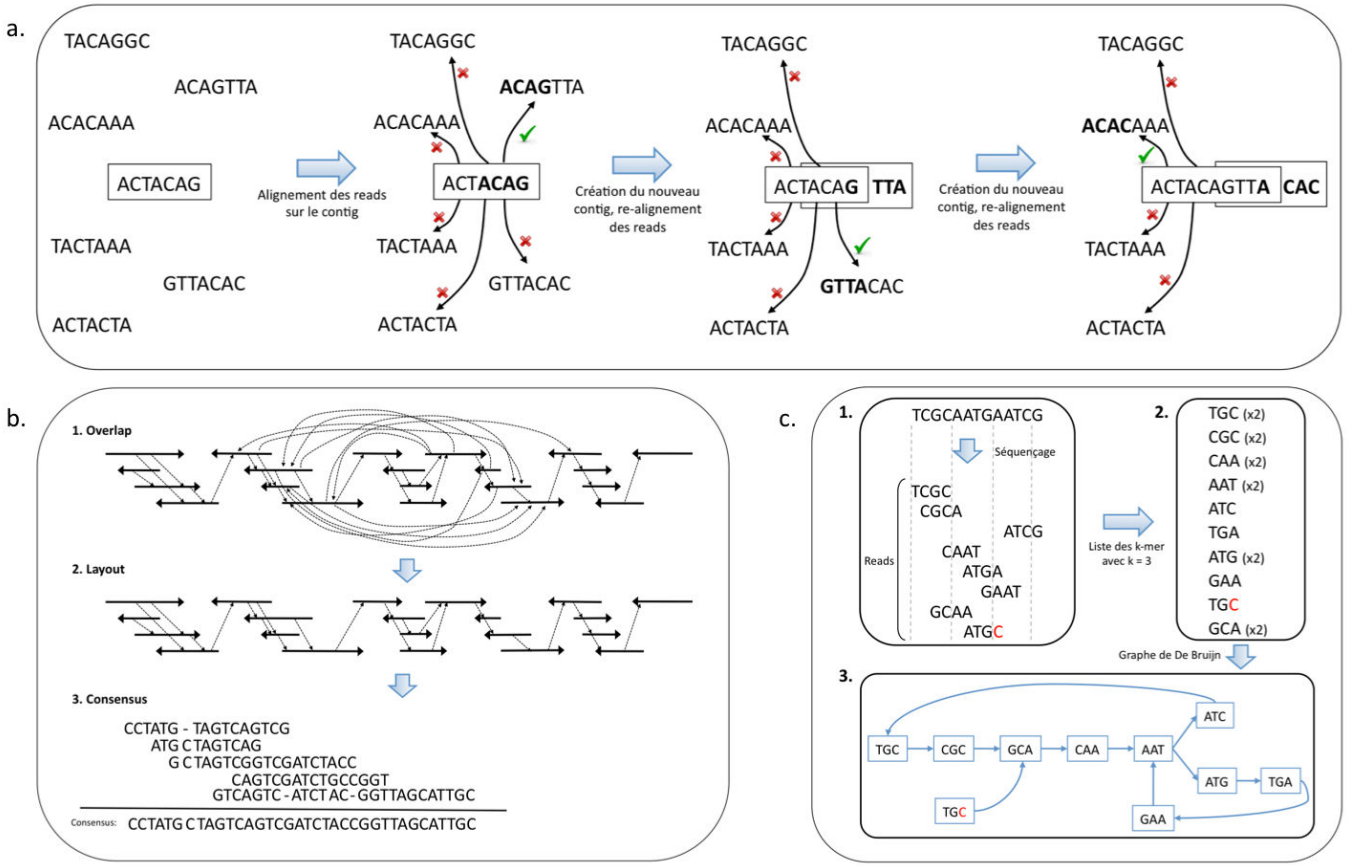


Figure 4 : Exemple de méthodes d'assemblage de jeux de données NGS. (a.) Principe de la méthode d'assemblage d'un algorithme « glouton » : pour chaque contig (ou read) on ajoute un autre read (ou contig) présentant une région chevauchante. L'ensemble des reads sont alignés sur la séquence encadrée, le read présentant le meilleur alignement est alors ajouté sur la séquence encadrée pour former un nouveau contig. Ce processus est recommencé jusqu'à ce que plus aucun read ne puisse s'ajouter au contig formé. Un read non assemblé est alors sélectionné et le processus est relancé. (b.) Représentation des trois étapes d'un assembleur OLC. (1.) le programme recherche les régions chevauchantes entre les différents reads (représentés par des flèches) à l'aide des différents k-mers. (2.) Les graphes de chevauchement sont construits et manipulés pour obtenir une mise en place approximative des reads. (3.) Les différents reads sont alignés à l'aide d'un alignement multiple de séquence (MSA) (adapté d'après Kasahara and Morishita 2006). (c.) Exemple de construction d'un graphe de *de Bruijn*. Les différents reads séquencés sont découpés en séquences (k-mers) de taille k (ici k=3). Les k-mers obtenus sont utilisés dans un graphe de *de Bruijn*. Un k-mer est relié à un autre k-mer s'ils sont identiques sur leur région chevauchante (d'une taille de k-1). Le nucléotide en rouge correspond à une erreur de séquençage.

De nombreux outils d'assemblage se basant sur les graphes de *de Bruijn* ont été développés pour traiter les données NGS produisant de fragments courts ; c'est le cas des logiciels ABySS (Simpson et al. 2009), SOAP *de novo* (Li et al. 2010) ou Velvet (Zerbino and Birney 2008) par exemple. Chacun de ces programmes va présenter des avantages et inconvénients et sera plus ou moins adapté en fonction du jeu de données traité. Pour un jeu de données « single-end », Velvet et ABySS présentent ainsi une moins bonne couverture de séquence par rapport à SOAP *de novo*, mais ABySS présente le taux d'erreur d'assemblage le plus faible. Des tests sur des données pairées indiquent que ABySS présente la meilleure couverture de séquence par rapport à SOAP *de novo* et Velvet qui eux présentent une couverture de séquence similaire. De même que pour les données non pairées, ABySS présente le plus faible taux d'erreur de séquençage. SOAP *de novo* présente le meilleur N50 (correspondant à une mesure de qualité d'assemblage) par rapport aux deux autres programmes, et ceci pour des données présentant de très courts ou courts fragments (35 à 125 bp), riche en GC ou non (36 à 50% ; Lin et al. 2011). Une autre étude menée par Bao *et al.* (2011) a permis de montrer que ABySS était capable de reconstruire des contigs allant jusqu'à 12 804 bp (à partir de données pairées) contre 859 bp et 2 285 bp pour SOAP *de novo* et Velvet, néanmoins la longueur moyenne des contigs obtenus avec ABySS est la plus faible (37,38 bp) par rapports aux deux autres logiciels (>61 bp). Si SOAP *de novo* ne permet pas d'obtenir de contigs de grande taille, il permet en revanche d'obtenir 61,40% de la couverture du génome (contre 17,47% et 5,95% pour Velvet et ABySS) avec la meilleure longueur moyenne des contigs construits (71,36 bp; Bao et al. 2011). D'autres outils ont été développés pour répondre à des besoins spécifiques, comme Minia (Chikhi and Rizk 2012) qui permet d'assembler *de novo* un génome (à l'aide de reads de courts fragments) sans avoir recours à d'importantes ressources informatiques. La grande spécificité de Minia réside dans le fait de réussir à représenter un Graphe de *de Bruijn* exact à l'aide d'une structure de données probabiliste compacte (filtre de bloom) à partir de l'ensemble des k-mers construits. La production des contigs est ensuite réalisée en parcourant une fois chaque nœud du Graphe de *de Bruijn* via un algorithme de parcours en profondeur ou DFS (« Depth First Search »). Les régions localement complexes du Graphe de *de Bruijn* sont parcourues via un algorithme de parcours en largeur (BFS ; « Breadth First Search »), la largeur et la profondeur étant respectivement délimitées à 20 et 500. Si un ou plusieurs chemins d'une région donnée et localement complexe respectent ces paramètres, alors un

chemin est choisi au hasard. Typiquement, dans le cas d'un épissage alternatif, un seul contig sera produit par Minia. Ce programme a été comparé au logiciel Trinity dans le cadre de cette thèse (Chapitre 5, Partie D). Le logiciel Trinity permet la reconstruction *de novo* de jeux de données transcriptomiques à partir de données RNA-Seq. Cet outil est capable de prendre en compte les reads pairées ainsi que les reads brin-spécifiques (« strand specific »). Trinity rend un ensemble de contigs correspondant à l'ensemble des transcrits pleine longueurs incluant les isoformes d'épissage alternatif. Ce logiciel se décompose en trois modules :

- Le module « Inchworm » qui reconstruit à partir des k-mers l'isoforme dominant en pleine longueur ainsi que des séquences partielles avec les k-mers restants.
- Le module « Chrysalis » qui clusterise les contigs se chevauchant sur une longueur de k-1 et construit un Graphe de Bruijn ou GdB (idéalement 1 GdB = 1 gène/locus).
- Le module « Butterfly » qui retrouve les chemins au sein de chaque GdB (aidé par les lectures pairées) et retourne deux isoformes de pleine longueur (un polymorphisme de séquence ou une erreur de séquençage n'entraînant pas la formation de deux contigs).

Il est ainsi possible, selon les jeux de données utilisés de sélectionner le logiciel le plus approprié. Néanmoins, chaque jeu de données présentant ses propres spécificités, des tests préalables sont nécessaires et conseillés avant de sélectionner un assembleur en particulier. La compétition Assemblathon 2 (regroupant 21 équipes de recherche ; Bradnam et al. 2013) a permis de comparer les assemblages réalisés à partir de plusieurs logiciels pour 3 espèces différentes (*Melopsittacus undulatus*, *Maylandia zebra* et *Boa constrictor constrictor*) et d'émettre une série de recommandation pour l'assemblage *de novo* de données NGS : ils suggèrent notamment de tester plusieurs logiciels et paramètres.

Les différentes méthodes de « mapping » :

Les différents logiciels d'alignement multiple par mapping sont dérivés des algorithmes dynamiques d'alignements classiques développés dans les algorithmes de Needleman et Wunsh (1970) ou Smith et Waterman (1981). L'algorithme de Needleman-Wunsh permet l'alignement global de deux séquences contrairement à l'algorithme de Smith et Waterman qui permet l'alignement local de deux séquences (Figure 5, a). Les logiciels d'alignements de séquences tels que BLAST (Basic Local Alignment Search Tool, Altschul et al. 1990; Altschul et al. 1997) ou BLAT (BLAST-Like alignment tool, Kent 2002) permettent de comparer les séquences nucléotidiques et protéiques de manière heuristique à l'aide d'un algorithme d'alignement local dérivé de l'algorithme de Smith et Waterman. Néanmoins ces logiciels ne sont pas adaptés aux jeux de données NGS générant des millions de fragments courts. De nombreux logiciels d'alignement de séquences NGS, capable de traiter rapidement les jeux de données NGS ont ainsi été développés et se basent sur deux types d'algorithmes qui utilisent une table de hachage (Figure 5, b.)ou la transformée de Burrows Wheeler (BWT, Figure 5, c. ; voir Schbath et al. 2012 pour le détail des algorithmes). Nous présenterons ci dessous plusieurs logiciels de mapping parmi les plus utilisés lors de l'analyse de données NGS.

Les logiciels BWA (Burrows Wheeler Aligner, Li and Durbin 2009) et BWA-PSSM (BWA-PSSM) utilisent des algorithmes basés sur la transformée de Burrows Wheeler. L'utilisation de BWA repose sur trois commandes : la première correspond à l'indexation de la référence génomique (bwa index), la seconde (bwa aln) va permettre d'identifier les « hits » de chaque read dans la table des suffixes et la troisième étape convertit les coordonnées du tableau des suffixes en coordonnées de la référence et génère le fichier de sortie d'alignement multiple (bwa samse ; Li and Durbin 2009; Schbath et al. 2012). BWA-PSSM est un logiciel d'alignement probabiliste de fragments courts qui est basé sur l'utilisation de matrices de scores de positions spécifiques ou « Position Specific Scoring Matrices » (PSSM). Cette méthode va permettre d'utiliser des probabilités calculées en fonction du jeu de données et dispose de ce fait d'un panel plus important de paramètres réglables (Kerpedjiev et al. 2014).

Les logiciels Novoalign et NovoalignCS (www.novocraft.com) utilisent des algorithmes basés sur des tables de hachages. Dans le cas d'alignement de séquences, la table de hachage va contenir l'ensemble des séquences d'une longueur de k bp (l'ensemble des k-mers) issues de la séquence de référence. Ainsi pour chaque séquence nucléotidique de k bp (chaque séquence correspondant à une clé) des couples de valeurs vont être associés. Chaque couple de valeurs correspond à l'indice de la séquence et la position du k-mer sur cette dernière. Le logiciel Novoalign permet d'aligner rapidement et précisément les différents reads sur une séquence de référence. Cet outil utilise les données de pairage et de qualités du séquençage lors du mapping des reads. Novoalign se déroule en deux étapes, la première (novoindex) indexe le génome ou la séquence de référence et la seconde (novoalign) aligne les reads contre la séquence de référence. Néanmoins ce logiciel ne permet pas d'obtenir le nombre de mismatchs entre le read et la séquence de référence (Schbath et al. 2012).

Les logiciels de la suite SOAP (Short Oligonucleotide Analysis Package ; SOAP, SOAP2, SOAP3 et SOAP3-dp) se basent sur la transformée de Burrows Wheeler (Li et al. 2008; Li et al. 2009). L'utilisation de SOAP2 repose sur deux étapes, la création de l'index Burrows Wheeler de la séquence de référence, et l'alignement des reads (soap). SOAP2 permet à l'utilisateur de choisir le pourcentage d'identité (via le nombre de mismatchs) entre la séquence de référence et les reads. Néanmoins le logiciel ne prend pas en compte les nucléotides indéterminés ('N') et les remplace par des Guanines ('G') dans les alignements, ce qui entraine l'apparition d'erreurs (Schbath et al. 2012). Le logiciel SOAP est un logiciel robuste pour les alignements présentant un petit nombre d'indels et de mismatchs (de 1 à 3). Cette version est beaucoup plus rapide que le logiciel BLAT et utilise des k-mers de 10 dans la table de hachage (Li et al. 2008). Le logiciel SOAP2 est une version beaucoup plus rapide que SOAP et utilise la transformée de Burrows Wheeler bidirectionnelle pour construire l'index de la référence. Cet algorithme permet de repérer le motif dans les 2 sens de la séquence alignée (Li et al. 2009). La troisième version du logiciel (SOAP3) permet un alignement de séquence encore plus rapide que la deuxième version ; la version SOAP3-dp également optimisée supporte un nombre aléatoire de mismatchs et de gaps (Liu et al. 2012; Luo et al. 2013).

Les logiciels de mapping Bowtie (Langmead et al. 2009) et Bowtie 2 (Langmead and Salzberg 2012) utilisés au cours de cette thèse, se basent également sur la transformée de Burrows Wheeler. L'utilisation de Bowtie repose sur deux commandes, la première (bowtie-build) permet d'indexer la séquence de référence et la seconde (bowtie) génère la liste des alignements à partir de l'ensemble des reads et de la séquence indexée. Le principal défaut de Bowtie est de ne pas prendre en compte les gaps, ce qui ne permet d'aligner qu'un sous ensemble de reads du jeu de données. Néanmoins, la nouvelle version Bowtie 2 prend en compte l'alignement de reads présentant des gaps et des insertions/délétions en bloc. De plus il est possible de régler le pourcentage d'identité minimal voulu entre la séquence et les reads ; ce qui en fait son principal avantage en comparaison des autres logiciels.

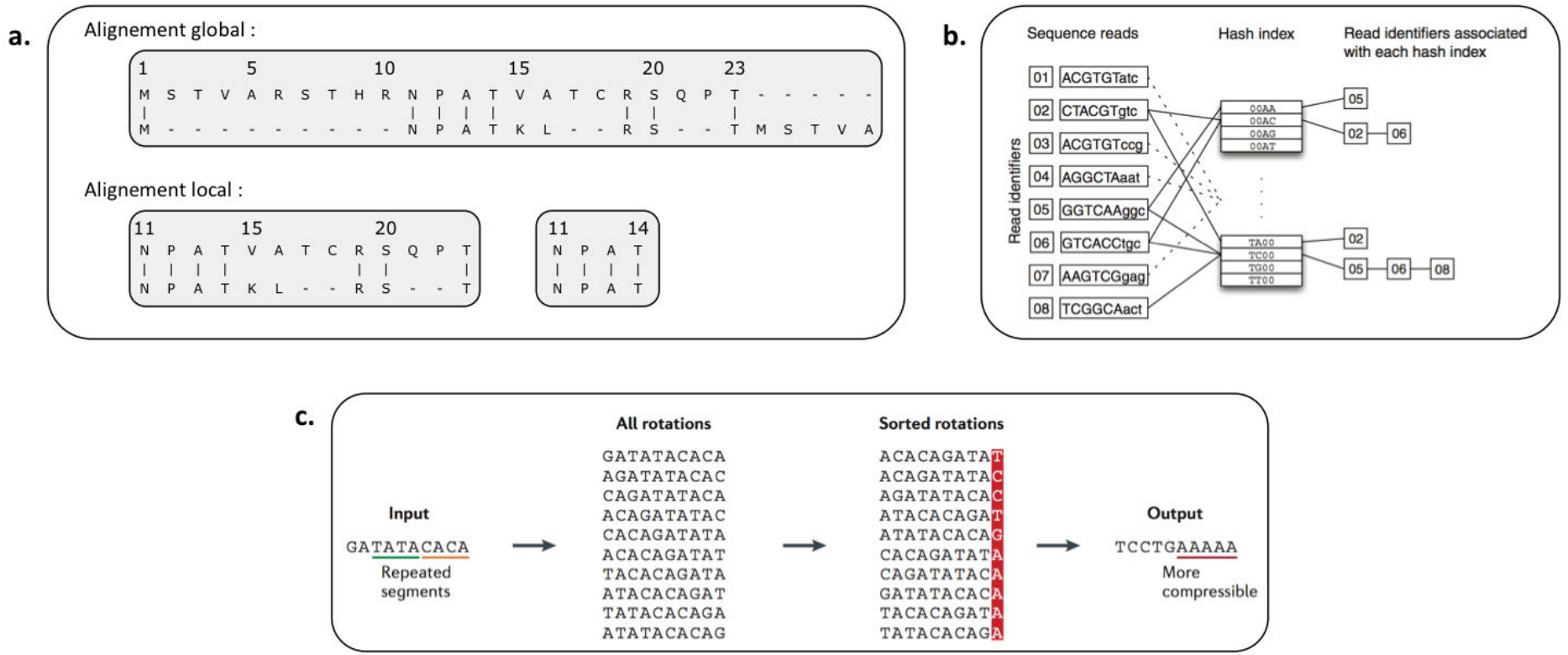


Figure 5 : Exemple de différentes méthodes d'alignement de séquences. (a.) Représentation d'un alignement global qui aligne les 2 séquences sur l'ensemble de leur longueur et d'un alignement local qui aligne les séquences sur la partie commune de plus grande ressemblance. (b.) Représentation schématique d'une table de hachage basée sur la stratégie d'alignement. Les reads avec leurs identifiants associés sont présentés avec les régions qui seront utilisées pour la sélection des graines (en majuscules) et des index de graines (0011 et 1100). Les identifants des reads sont associés avec les graines à l'aide d'une fonction de hachage (un entier unique représente chaque graine). Une fois que la table de hachage est construite pour l'ensemble des reads (ou le génome de référence) fourni en entrée du programme, les données peuvent être scannées à l'aide de cette même fonction de hachage, ce qui permet d'aligner plus précisément sur le génome les sous-ensembles de reads (d'après Flicek and Birney 2009). (c.) Exemple de transformée de Burrows–Wheeler (BWT). Il s'agit de la transformation d'une chaine de caractères, permettant la convertion des séquences hautement redondantes en un format facilement compressible. Dans un premier temps, l'ensemble des permutations possibles de la séquence donnée en entrée est calculé. Chaque position (nucléotide) de la séquence devient la position de début une seule fois. Les séquences permutées sont ensuite ordonnées par ordre alphabétique et la dernière colonne (en rouge) est rendue en sortie sous la forme d'une séquence. Les sous-séquences répétées éventuellement présentes dans la séquence de départ seront regroupées à l'aide de la transformée de Burrows–Wheeler pour aboutir à la répétition d'un seul caractère dans la dernière colonne, dans le but de faciliter la compression de la séquence (d'après Berger, Peng, and Singh 2013).

La comparaison des logiciels d'alignements de reads de courte lecture ne permet pas de définir le meilleur logiciel de mapping. Chacun de ces logiciels va présenter ses avantages et inconvénients en fonction du jeu de données traité, mais également en fonction du résultat attendu par l'utilisateur. En effet, si le but de l'utilisateur est par exemple d'obtenir le maximum de reads alignés sur la séquence, d'aligner les reads en fonction d'un certain pourcentage d'identité ou d'obtenir un alignement de séquences le plus rapide possible, il sera nécessaire d'utiliser différents programmes pour répondre à chacune de ces questions (Hatem et al. 2013; Schbath et al. 2012). Ce constat est identique pour le traitement de données Ion Torrent (dont les lectures sont plus longues que les lectures courtes de type Illumina). En effet, Caboche *et al.* (2014) ont comparé les différents outils d'alignement de reads sur plusieurs jeux de données Ion Torrent simulés et ont pu mettre en évidence que le choix du logiciel dépendait de l'application et du jeu de données analysé. Ainsi, il est par exemple conseillé d'utiliser le logiciel de mapping SMALT pour analyser des motifs répétés au sein de données d'amplicons ou les logiciels SSAHA2, TMAPm SHRiMP2 ou Bowtie 2 pour détecter les importants niveaux de mutations sans utiliser une profondeur de séquençage très importante.

NGS et reconstructions de génomes et transcriptomes :

Le développement des outils et méthodes de séquençage dans les années 70 a permis le séquençage de génomes complets. Le premier génome séquencé a été réalisé par l'équipe de Frederick Sanger à l'aide de la méthode de séquençage Sanger et correspond au génome du virus bactériophage phi X 174 (ΦX174) de la famille des Microviridae. Ce choix a été motivé par le fait que ΦX174 possède un génome très simple, constitué d'une seule molécule d'ADN simple-brin circulaire longue de 5 386 bp (Sanger et al. 1978). Plusieurs autres génomes ont également été séquencés à l'aide de la méthode Sanger, tels que des génomes bactériens (1995 : *Haemophilus influenza ;* 1997 : *Escherichia coli*) ou eucaryotes (1997 : *Saccharomyces cerivisae*). L'utilisation de cette méthode a permis en premier lieu de séquencer les espèces modèles : le premier génome animal (*Caenorhabditis elegans* en 1998) suivi du génome de la première plante séquencée *Arabidopsis thaliana* (Arabidopsis Genome Initiative 2000) et celui de l'Homme (Lander et al. 2001). Le développement des NGS a permis le séquençage d'un nombre très importants d'espèces diploïdes mais

également polyploïdes comme le colza (Chalhoub et al. 2014) ou le blé tendre (The International Wheat Genome Sequencing Consortium (IWGSC) et al. 2014). Aujourd'hui, plus de 6 200 génomes sont séquencés au sein des espèces eucaryotes (Reddy et al. 2015) ; les génomes séquencés à ce jour chez les plantes ainsi que les évènements de polyploïdisation mis en évidence sont présentés dans la Figure 6. Le séquençage de génomes est à présent réalisé selon deux approches distinctes : des approches par ordonnancement hiérarchique qui visent à séquencer des régions génomiques ancrées physiquement, facilitant et optimisant les processus d'assemblage et de scaffolding ; ou des approches de séquençage global (ou « Whole Genome Shotgun » ; WGS) qui séquencent de manière aléatoire le génome et nécessite plus d'efforts d'assemblage et de validation des scaffold obtenus. La majorité des études ont ainsi utilisées des données de séquençage Roche-454 (produisant de longs fragments) et Illumina (pour corriger les erreurs d'insertions/délétions des données Roche-454) combinées à des séquences de BACs (« Bacterial Artificial Chromosome ») comme dans le cas du coton (*Gossypium raimondii* ; Paterson et al. 2012), de l'égilope (*Aegilops taushii* ; Jia et al. 2013), du bananier (*Musa acuminata* ; D'Hont et al. 2012), du colza (*Brassica napus* ; Chalhoub et al. 2014) ou du caféier (*Coffea canephora* ; Denoeud et al. 2014). Certains génomes ont été séquencés à l'aide de données de BACs séquencés et de données Illumina (*Gossypium arboreum* ; Li et al. 2014) ou à l'aide d'une méthode WGS, en utilisant uniquement des données NGS de pyroséquençage et de séquençage par synthèse comme chez le chêne (*Quercus robur* ; Plomion et al. 2015). Le développement récent de la technologie de séquençage Nanopore permet à présent d'obtenir rapidement le génome d'espèces eucaryotes comme celui de la levure *S. cerevisiae* (Goodwin et al. 2015) par l'obtention de long fragments facilitant les assemblages. De plus cette technologie permet d'assembler correctement des régions complexes du génome tel que les ARNr ou les éléments transposables, ce qui n'était pas possible avec des données de séquençage de type Illumina. Ces résultats ouvrent ainsi une perspective intéressante pour l'assemblage et l'analyse de régions génomiques (comme le compartiment répété) ou de génomes plus complexes.

Les assemblages de données transcriptomiques sont généralement effectués à partir de données Roche-454 assemblées à l'aide du logiciel Newbler (*e.g.* l'eucalyptus : Novaes et al. 2008; le chataîgner : Barakat et al. 2009; l'avocatier : Wall et al. 2009; les Spartines :

Gedye et al. 2010, Ferreira de Carvalho et al. 2012; l'orge : Bedada et al. 2014) . Une étude comparative (Ren et al. 2012) a montré que les logiciels MIRA et Newbler se présentaient comme étant les assembleurs les plus performants dans l'analyse de données transcriptomiques issues de pyroséquençage. Cette étude a également mis en évidence que la totalité des assembleurs testés produisaient 8% de contigs chimériques. La construction de transcriptomes peut également être réalisée à partir de données Illumina, comme cela a été fait pour l'arachide (Chopra et al. 2014), le blé (Duan et al. 2012) ou le piment (Liu et al. 2013). De nombreux logiciels d'assemblage de données RNA-seq sont disponibles, néanmoins la majorité des études menées chez des espèces diploïdes et polyploïdes montrent que le logiciel Trinity se présente comme le meilleur assembleur de données de fragments courts de type Illumina (Chopra et al. 2014; Clarke et al. 2013; Liu et al. 2013). De nouveaux assembleurs continuent d'être développés, comme EBARDenovo, un logiciel capable de détecter les contigs chimériques lors de l'assemblage (Chu et al. 2013). Si de nombreuses études utilisent des données de séquençage Roche-454 et/ou Illumina, ces données sont généralement traitées de manière indépendante. En effet, peu d'études proposent d'effectuer un assemblage hybride à partir de jeux de données Roche-454 et Illumina (Barthelson et al. 2011; Jiang et al. 2011; Sirota-Madi et al. 2010) ; cette méthode permet d'obtenir un nombre de contigs plus importants et de meilleure qualité. L'annotation fonctionnelle des gènes obtenus peut être réalisée à partir de plusieurs approches comme celle combinant un alignement de séquences par blastn et tblastx contre des banques d'EST (« Expressed Sequence Tag ») ou de CDS (Coding DNA Sequences) d'espèces proches et une annotation fonctionnelle basée sur l'ontologie de gènes à l'aide du logiciel BLAST2GO (Conesa et al. 2005). Cette approche a notamment été utilisée pour annoter les premiers transcriptomes de référence de Spartines hexaploïdes (Ferreira de Carvalho et al. 2012). Cette méthode peut être complétée avec l'utilisation de programmes de reconnaissance de domaines protéiques tel que Pfam (Finn et al. 2014). L'annotation fonctionnelle des transcriptomes peut également être réalisée à l'aide de plusieurs autres bases de données protéiques comme Swiss-Prot, COG ou KEGG (Toledo-Silva et al. 2013). Il est aussi possible d'annoter les différents transcriptomes en utilisant des pipelines spécifiques tels que TRAPID qui combine plusieurs étapes d'annotation dont la recherche d'ontologie de gène et la recherche de domaines protéiques à l'aide des logiciels BLAST2GO (Conesa et al. 2005; Götz et al. 2008) et Pfam (Van Bel et al. 2013).

Les méthodes de détection de SNPs (« Single-Nucleotide Polymorphisms ») :

La détection de SNPs représente une part importante de l'analyse des génomes et transcriptomes eucaryotes dans lesquels une diversité génétique et l'existence de plusieurs copies dupliquées sont attendues. L'identification de SNPs au sein de données NGS nécessite l'utilisation d'outils/programmes spécifiques pour parcourir l'ensemble des données et éliminer les faux positifs dus aux erreurs de séquençage. De nombreux outils ont été développés dans ce but, dont les logiciels les plus couramment utilisés : SAMtools (Li et al. 2009), GATK (Van der Auwera et al. 2013), SOAPsnp (Li et al. 2009) et Varscan/Varscan2 (Koboldt et al. 2009; Koboldt et al. 2012). Ces logiciels prennent en entrée de programme un ensemble de reads alignés ou une base de données de variants crées par les assembleurs (comme la suite SOAP par exemple ; Kumar, Banks, and Cloutier 2012). Une étude comparative entre plusieurs logiciels de détection de SNPs a permis de mettre en évidence la nécessité d'utiliser une approche consensus en utilisant plusieurs logiciels dont GATK, SAMtools et CRISP (Pabinger et al. 2013). Lors de cette étude, l'utilisation de GATK, SAMtools et Varscan2 a permis de détecter respectivement ~49 000, ~22 000 et ~34 000 SNPs ainsi que 1 969, 234 et 1 896 indels au sein d'un génome humain. Néanmoins 11 522 SNPs et 431 indels ont été détectés uniquement avec GATK contre 870 SNPs et 414 indels pour Varscan2 (Pabinger et al. 2013). Yu et Sun (2013) ont comparé plusieurs logiciels de détection de SNPs dans des régions faiblement couvertes et ont montré que GATK permettait de détecter un nombre de SNPs vrai positifs plus important que SAMtools et SOAPsnp. De nombreux outils de détection de SNPs continuent à être développés pour permettre la meilleure détection possible au sein des jeux de données NGS. Ainsi, des approches se basant sur le modèle GeMS (« Genotype Model Selection ») ont été développées. Cet outil permet de détecter un nombre de SNPs important dont la très grande majorité est retrouvée par d'autres logiciels tel que SAMtools ou GATK. En effet, sur 4 452 SNPs détectés, seulement 10 SNPs ont été détectés uniquement avec ce programme GeMS (You et al. 2012).

Néanmoins, la majorité de ces outils sont développés pour la détection d'allèles dans un contexte diploïde. La détection de SNPs chez les espèces polyploïdes est particulièrement complexe, en particulier la détection de SNPs homéologues qui permettent l'assignation des

copies hybrides aux copies parentales, comme cela a été réalisée chez le cotonnier (Udall et al. 2006; Flagel et al. 2008; Flagel, Wendel, and Udall 2012; Salmon et al. 2012; Yoo, Szadkowski, and Wendel 2013) ou le caféier (Combes et al. 2012). Pour identifier les différents polymorphismes de séquences présents au sein de ces espèces, il est nécessaire d'utiliser des paramètres spécifiques (comme le pourcentage de présence d'un nucléotide pour considérer un SNP valide) à la qualité associé. Ainsi, de nombreux outils sont développés au cas par cas (Michael and Jackson 2013; Clevenger et al. 2015; Clevenger and Ozias-Akins 2015). L'utilisation de génomes parentaux diploïdes de référence est cruciale dans cette démarche. Lorsque la polyploïdie est ancienne et/ou que les génomes parentaux ne sont pas identifiés, l'inférence des copies homéologues devient plus complexe ; elle nécessite des approches particulières, qui restent très peu développées et que nous explorerons au cours de ce travail.

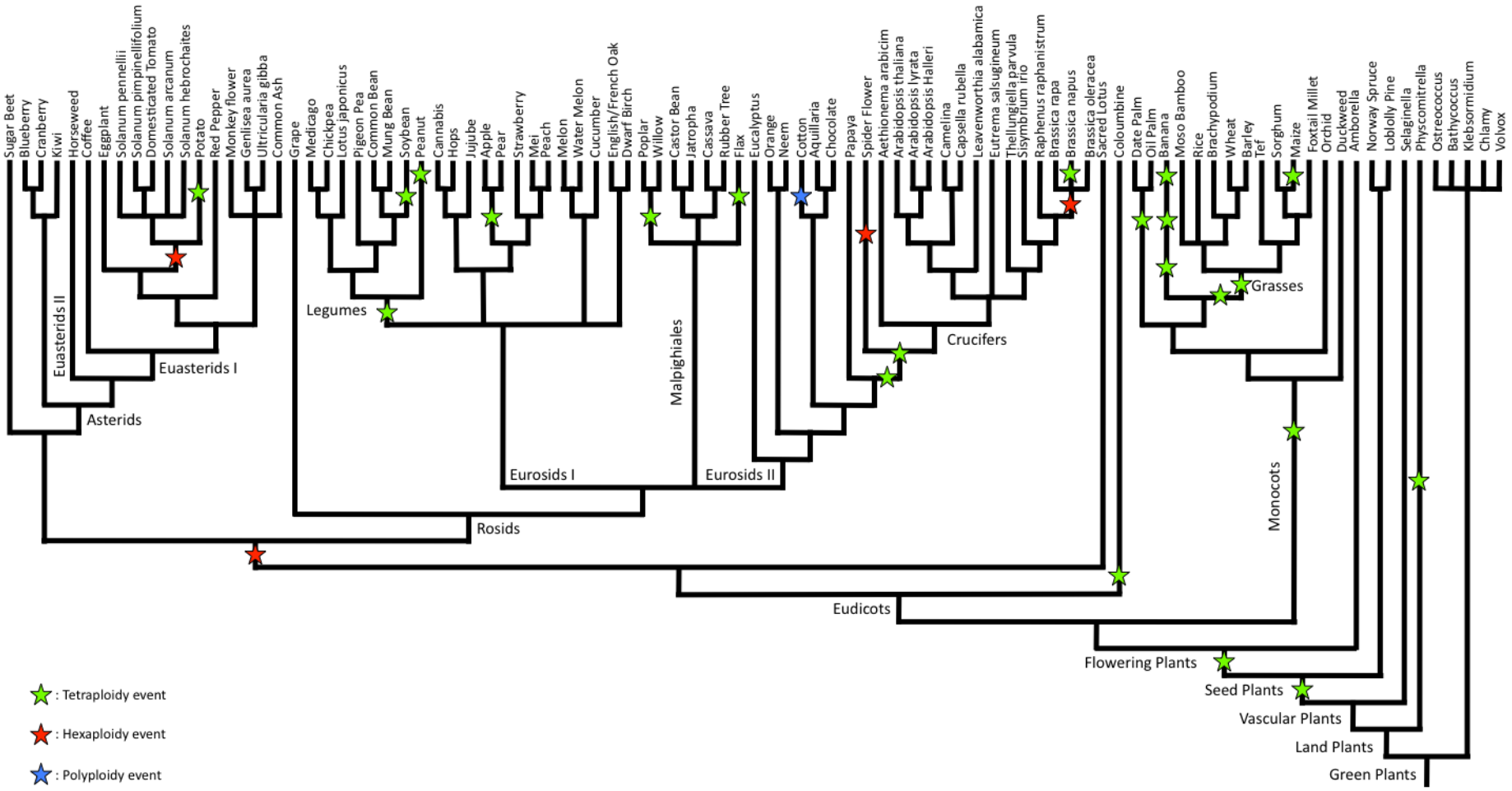


Figure 6 : Arbre phylogénétique représentant les taxons dont les génomes ont été séquencés au sein du règne végétal (au 15/12/2014). Les étoiles représentent les différents évènements de polyploïdisation (Adapté d'après https://genomevolution.org/ et Renny-Byfield and Wendel 2014).

Partie C. Détection de copies dupliquées au sein des espèces :

Méthodes de détection des évènements de polyploïdisation :

La duplication entière du génome (polyploïdie) est un processus biologique récurrent qui représente un mécanisme important dans l'évolution et l'adaptation de nombreuses espèces eucaryotes. Comme nous l'avons souligné en Introduction, il est aujourd'hui établi que la polyploïdie a affecté de façon récurrente toutes les lignées de plantes (Figure 6) où elle représente un mécanisme important dans la formation, l'évolution et l'adaptation des espèces (Mable 2004). L'identification des espèces polyploïdes peut être réalisée à l'aide de différentes techniques de cytogénétique (dénombrement des chromosomes au sein d'une cellule, détection de multivalents qui se forment durant la méiose) ou génétique (nombre d'allèles par locus, estimation de la taille du génome). D'autres critères comme la comparaison de la taille des cellules ont également été traditionnellement utilisés pour identifier les espèces polyploïdes (Mable, Alexandrou, and Taylor 2011). La détection et l'analyse de l'évolution des complexes polyploïdes ont très tôt retenu l'attention des botanistes (Stebbins 1971). En 1976, Raicu résume la démarche d'analyse des complexes polyploïdes en 4 étapes. La première étape correspond au dénombrement des chromosomes puis l'étude des caryotypes des différents individus couplée à l'analyse de la méiose des espèces diploïdes et polyploïdes. La seconde étape correspond à des recherches comparatives (entre les caractères morphologiques, physiologiques et biochimiques) entre les populations présentant des niveaux de ploïdie différents. La troisième étape correspond à la réalisation de croisements entre les différentes espèces dans le but d'étudier la fertilité des hybrides et la méiose de ces individus. La dernière étape étant la reconstitution des polyploïdes par croisement des espèces diploïdes (Raicu 1976). Néanmoins cette étude complexe n'est pas toujours réalisable surtout lorsque les espèces diploïdes ne sont pas connues comme c'est le cas pour les espèces du genre *Spartina* par exemple (Marchant 1968a, 1968b).

Les premières définitions des polyploïdes dans un contexte temporel furent développées par Stebbins (1971), qui distingue trois catégories : les complexes jeunes, les complexes à maturité et les complexes anciens :

- Les complexes jeunes comprennent des espèces diploïdes et polyploïdes dérivés et sont localisés dans des zones de répartition proches ou identiques. Les différentes espèces possédant un patrimoine génétique identique, colonisent le même milieu et rentrent en compétition.
- Les complexes à maturités sont les complexes les plus nombreux. Au sein de ces complexes, les espèces tétraploïdes sont les plus répandues, ce qui oblige les autres espèces diploïdes et polyploïdes à occuper des zones plus réduites.
- Les complexes polyploïdes anciens correspondent à des groupes d'espèces ou l'ensemble des représentants diploïdes ont disparus. Au sein de ces complexes, les espèces présentant de hauts niveaux de ploïdie seront prédominantes (Raicu 1976).

Cet aspect temporel a encore été revisité depuis le développement des approches de génomique qui ont détecté des événements plus anciens de duplication de génome (paleopolyploïdie) passés inaperçus au travers des approches traditionnelles. De nouvelles terminologies ont ainsi été mises en place pour désigner les groupes de gènes dupliqués détectés au sein des génomes. Le terme de « **Paralogon** » a été introduit pour désigner les groupes de gènes dupliqués localisés sur différents chromosomes d'un même génome comme cela a pu être identifié chez l'homme, ou des Paralogons ont été localisés sur les chromosomes 3 et 17 (McLysaght, Hokamp, and Wolfe 2002). Le terme d'« **Ohnologues** » (en référence à Ohno 1974) a été utilisé dans le cas de paires de gènes dupliqués produits par le processus de duplication de génome (Wolfe 2001). Enfin, les gènes « **Paléologues** » correspondent aux gènes retenus après un événement de polyploïdisation ancien (Chapman et al. 2006).

Plusieurs méthodes graphiques de détection des événements de duplication de gènes et de polyploïdisation à partir de données génomiques ont été utilisés ces dernières années. L'analyse à l'aide de **Dot Plot** (Figure 7, a., représentation graphique de résultats de Blast) permet l'identification de régions similaires entre deux séquences d'un même

génome. L'utilisation de cette méthode a permis notamment d'identifier les régions codantes conservées entre les chromosomes d'*Arabidopsis thaliana* (Blanc et al. 2000). Cette méthode permet également de comparer deux espèces entres elles pour identifier les différentes régions homéologues conservées (Schnable, Springer, and Freeling 2011) ou pour reconstruire des génomes ancestraux théoriques. En effet, à partir d'alignements de séquences orthologues, Salse et ses collaborateurs (2008) ont pu identifier les duplications des chromosomes du riz et du blé et identifier les relations de colinéarité entre ces deux génomes et la sous-famille des Panicoideae incluant le maïs et le sorgho. Il a également été possible de proposer plusieurs scénarii possibles de génomes de l'espèce ancestrale des Eudicotylédones d'une part et des Monocotyléones d'autre part (Abrouk et al. 2010). Une seconde méthode graphique, plus récente, permet la visualisation circulaire de données (notamment à l'aide du logiciel Circos (Krzywinski et al. 2009)). Il est ainsi possible d'explorer les relations entre les objets et les positions. Cette **méthode « Circos »** (Figure 7, b.) permet, par exemple, d'observer les relations entre les paires de gènes paralogues ou orthologues. Il est ainsi possible d'identifier des événements de duplication de génomes et de comparer différentes espèces (Jaillon et al. 2007). Une démarche intéressante a été proposée par Guillaume Blanc et ses collaborateurs (Blanc and Wolfe 2004) pour détecter et dater les événements de duplication génomique, en utilisant les **taux de substitution synonymes Ks** (évoluant de façon neutre) entre paires de gènes dupliquées au sein d'un même génome. Suite à un événement de polyploïdisation, l'ensemble des gènes se trouve dupliqué simultanément et la représentation graphique des pourcentages de gènes dupliqués en fonction de leur temps de divergence (estimé à partir des Ks) indique un « pic » de pourcentage de gènes dupliqués correspondant au moment où l'évènement plus ou moins ancien de duplication génomique a eu lieu (détaillé dans la Figure 7, c.) . Cette méthode, initialement utilisée sur le génome paléopolyploïde d'*Arabidopsis thaliana* puis de plusieurs angiospermes (Blanc and Wolfe 2004; Tang et al. 2010) a permis par exemple de mettre en évidence un phénomène de duplication du génome entier au niveau de l'ancêtre commun des plantes à graines (les Spermaphytes), en amont de la divergence Gymnospermes – angiospermes (Jiao et al. 2011).

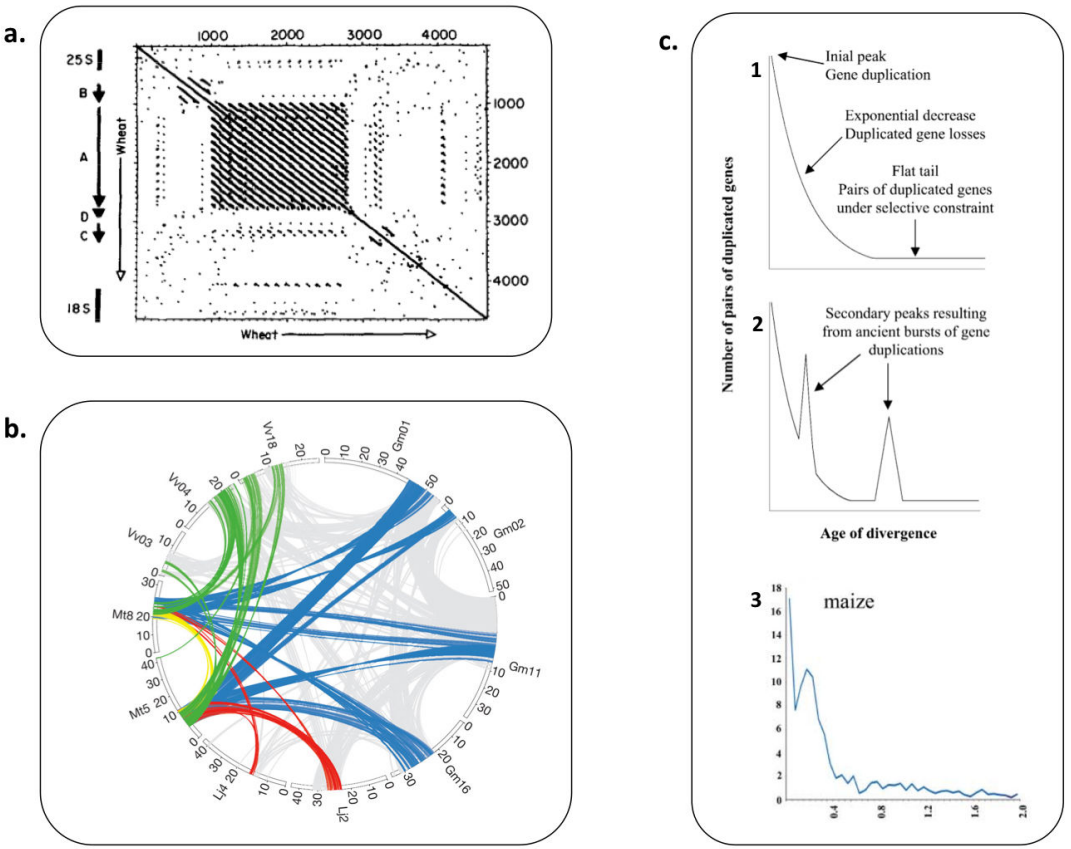


Figure 7 : Exemple de méthodes graphiques permettant la détection de séquences répétées. Représentation d'un (a.) Dot Plot de la région 25S-IGS-18S de l'ADN ribosomique du blé mettant en évidence les 4 régions (A à D) répétées de l'IGS (d'après Barker et al. 1988). (b.) Représentation à l'aide de la méthode « Circos » des relations de synthénies entre *Medicago*, *Glycine*, *Lotus* et *Vitis* (d'après Young et al. 2011). (c) Distribution de l'âge théorique des paires de gènes dupliqués dans un génome. (c.1) Distribution de l'âge des paires de gènes paralogues attendue avec un taux constant de duplication de gènes et de perte de gènes dupliqués. Trois principales phases peuvent être identifiées. Un pic initial représentant la duplication de gènes la plus récente. La distribution chute de façon exponentielle en raison de la perte des gènes dupliqués qui ne sont pas soumis à des contraintes sélectives. La troisième partie (stabilisation de la courbe) correspond aux anciennes paires de copies dupliquées, où les deux gènes évoluent sous contrainte sélective. (c.2) Distribution de l'âge des paires de gènes paralogues attendue pour une espèce ayant subi deux évènements anciens de duplication à grande échelle. La sur-représentation des gènes dupliqués à des périodes correspondants aux duplications à grande échelle donne lieu à deux pics secondaires. (c.3) Application de cette méthode sur le maïs, il est possible d'observer un pic de duplication de gènes correspondant à la duplication de génome survenue chez cette espèce (adapté d'après Blanc and Wolfe 2004).

La détection des copies homéologues chez les espèces polyploïdes se présente donc comme une question essentielle pour retracer l'histoire évolutive des espèces ou identifier le devenir des copies dupliquées. La première analyse permettant de détecter ces copies chez les espèces polyploïdes a été réalisée sur des EST (« Expressed Sequence Tag ») de coton par Udall et ses collaborateurs (2006). L'outil développé utilise les génomes parentaux (*Gossypium arboreum* et *Gossypium raimondii*) du coton et les SNPs spécifiques à chacune de ces espèces pour identifier les différentes copies chez l'espèce allotétraploïde *Gossypium hirsutum*. Pour cela, les différentes régions orthologues entre les 3 espèces sont identifiées et les contigs sont comparés. De nombreuses approches basées sur cette méthode ont ensuite été développées et appliquées sur des jeux de données d'espèces allopolyploïdes. Flagel et ses collaborateurs (2008) ont ainsi pu designer des sondes homéo-specifiques (basés sur les SNPs homéologues) chez les cotons pour mesurer les différents niveaux d'expression de chaque copie sur puce. Il a également été montré à l'aide de ces méthodes que la conversion génique entre les copies homéologues est un phénomène fréquent qui joue un rôle important dans la dynamique des génomes polyploïdes (Salmon et al. 2009; Flagel, Wendel, and Udall 2012). La détection de copies homéologues a été réalisée au sein d'autres espèces allotétraploïdes tel qu'*Arabidopsis* ou il a été possible de montrer que la rétention des copies homéologues dépendait notamment de l'origine des parents (Chang et al. 2010). Cette méthode a également été appliquée sur le blé (Akhunova et al. 2010), le soja (Ilut et al. 2012) et le caféier (Combes et al. 2012). Tennessen et ses collaborateurs (2014) on développé une approche spécifique basée sur l'utilisation d'une référence diploïde combinée à une carte génétique ancrée sur des marqueurs pour identifier les copies homéologues au sein de l'espèce octoploïde du fraisier. Ainsi, il existe aujourd'hui plusieurs outils permettant de détecter les copies homéologues chez des espèces polyploïdes se basant sur des références diploïdes tel que SNiPLoid (Peralta et al. 2013), PolyCat (Page, Gingle, and Udall 2013), BamBam (Page et al. 2014) ou HyLiTE (Duchemin et al. 2014).

Origine évolutive des copies dupliquées:

La présence de copies de gènes dupliquées dans les génomes pose la question de l'établissement de leur origine évolutive. Lors des comparaisons entre espèces : Quelles sont les copies orthologues (divergeant par spéciation) ou paralogues (dérivant de duplications géniques) ? Au sein de chaque génome : les copies dupliquées résultent-elles de duplications géniques individuelles (paralogues), de duplication génomiques récentes (homéologues) ou plus anciennes (paléologues) ?

Des approches comparatives sont nécessaires pour répondre à ces questions. Les comparaisons de cartes génétiques entre espèces ont été utilisées pour inférer l'orthologie des copies (Lagercrantz and Lydiate 1996), en se basant sur les critères de synténie (conservation de l'ordre des gènes sur les chromosomes de différentes espèces). Les analyses phylogénétiques qui se sont développées dans les dernières décennies apportent un éclairage puissant à ces questions. En effet, les topologies des arbres phylogénétiques basées sur des copies orthologues devraient refléter l'histoire des divergences entre espèces. La concordance histoire des gènes – histoire attendue des espèces nous renseigne donc sur l'orthologie des copies. Dans le cas de paralogie ou d'homéologie, la topologie des arbres reflètera l'histoire des duplications ; la position des copies dupliquées sur l'arbre permet ainsi de positionner les événements de duplications géniques par rapport aux évènements de spéciation et d'identifier éventuellement les espèces parentales d'où sont issues les copies dupliquées résultant d'auto ou allopolyploïdie (*e.g.* Doyle et al. 2004). Cependant, l'identification de l'origine des espèces polyploïdes n'est pas triviale, les gènes et génomes étant affectés par des processus évolutifs dynamiques, tel que la perte différentielle de copies dupliquées (phénomène de diploïdisation ; Langham et al. 2004) ; ainsi les phylogénies vont présenter un apport précieux à cet égard.

Plusieurs méthodes phylogénétiques sont disponibles aujourd'hui et peuvent être regroupées en deux grandes catégories : les **méthodes de distances** et les méthodes basées sur les caractères ; parmi ces dernières, on peut distinguer la **méthode du Maximum de Parcimonie** (MP) et les **méthodes probabilistes** telles que le Maximum de Vraisemblance (« Maximum Likelihood » ; ML) ou l'analyse bayesienne (Yang and Rannala 2012).

Les méthodes de distances (ou méthodes phénétiques) se basent sur l'utilisation d'une matrice où sont indiquées les distances entre les différentes séquences (ressemblances ou différences entre chaque paire de séquences), qui peuvent faire intervenir différentes corrections adaptées à la nature des données moléculaires comme par exemple les biais de substitutions multiples, (Jukes and Cantor 1969) ou les biais de transition/transverstion (Kimura 1980).

La construction hiérarchique des groupes de similitudes (« clustering ») fait intervenir l'utilisation de différents algorithmes. Les deux principales méthodes de distances utilisées sont : la méthode de **Neighbor-Joining** (NJ) de Saitou et Nei (1987) et la méthode **UPGMA** (« Unweighted Pair Group Method with Arithmetic mean »). La méthode NJ est généralement préférée en raison de sa prise en compte de taux variables d'évolution moléculaire (ce qui n'est pas le cas de l'UPGMA); l'avantage des méthodes de distances résident dans leur efficacité en terme de temps d'analyse. Toutefois, ces méthodes ne prennent pas en compte l'homologie des caractères. De ce fait, les arbres de distances peuvent biaiser l'interprétation des relations entre organismes (ou entre séquences) lorsqu'ils sont interprétés en tant qu'hypothèse phylogénétique.

La méthode de **maximum de parcimonie** (MP) cherche à maximiser l'homologie des caractères pris en compte en se basant sur la recherche des arbres qui demandent le plus petit nombre de modifications évolutives pour expliquer les différences observées entre les séquences étudiées parmi l'ensemble des arbres possibles. De ce fait, il est possible d'obtenir plusieurs arbres phylogénétiques en sortie d'analyse. La méthode MP utilise uniquement les sites dits informatifs (partage d'états dérivés de caractères homologues ou synapomorphies) contrairement aux méthodes de distances. Ces sites vont permettre d'inférer les liens de plus proche parenté et de discriminer les différentes topologies d'arbres. Un site sera considéré comme informatif si au moins 2 nucléotides sont présents à cette position, chacun étant présent dans au moins deux des différents objets analysés. L'état dérivé (apomorphe) est distingué des états ancestraux (plésiomorphes) par l'introduction dans les analyses de représentants de groupes extérieurs (« outgroups ») au groupe d'objets analysés. Il existe trois méthodes algorithmiques différentes : la **méthode exhaustive**, la **méthode « Branch and Bound »** et la **méthode heuristique**. En fonction du nombre de séquences à analyser et du nombre de sites informatifs, les approches

heuristiques (qui n'explorent qu'une partie des arbres possibles) seront privilégiées. Lorsque plusieurs arbres les plus parcimonieux (de même longueur et de topologie différente) sont obtenus, un **arbre de consensus** qui ne contient que les nœuds internes présents dans l'ensemble (ou la majorité) des arbres obtenus est réalisé. La méthode de parcimonie est l'une des méthodes les plus utilisées mais ne prend pas en compte les corrections des différents modèles d'évolution moléculaire. Cette méthode est également sensible à la variation des taux d'évolution, qui peuvent dans certains cas biaiser les relations obtenues, comme dans le cas du phénomène **d'attraction des longues branches**. Ce phénomène correspond à un artefact qui entraine le regroupement des taxons évoluant rapidement, mais ne reflète pas les liens de parenté (Felsenstein 1978).

Les méthodes probabilistes se basent sur l'utilisation de modèles d'évolution. La principale méthode utilisée est la méthode de **maximum de Vraisemblance** ou ML (« Maximum Likelihood »). Cette méthode se base sur la probabilité d'observer les données étudiées en utilisant un modèle d'évolution spécifique, identifié au préalable. Les méthodes probabilistes sont considérées comme les méthodes les plus fiables mais nécessitent également des ressources informatiques et des temps de calcul plus importants. De nouvelles approches statistiques « Bayésiennes » sont développées ces dernières années. Ces méthodes, similaires à la méthode ML, se basant sur le théorème de Bayes, utilisent des probabilités définies *a priori*. Des méthodes numériques, les chaines de Markov avec la technique de Monte Carlo ou MCMC (*Markov Chain Monte Carlo*) ont été implémentées pour estimer les probabilités postérieures des arbres phylogénétiques (Delsuc and Douzery 2004).

Pour estimer la validité des arbres obtenus avec les différentes méthodes décrites ci-dessus, il est nécessaire de réaliser un test de robustesse. Le test le plus connu et utilisé correspond à **l'analyse par bootstraps** (Felsenstein 1985) qui consiste à réaliser un échantillonnage aléatoire des caractères et de construire les différents arbres phylogénétiques. Ces arbres sont ensuite regroupés par sous-ensemble ce qui permet d'identifier pour chaque nœud, sa fréquence d'apparition (pour revue : Luchetta 2005).

La Phylogénomique : de nouvelles approches pour analyser les jeux de données en masse.

L'identification des relations de parenté au sein de genres sujets à des évènements de polyploïdisations plus ou moins récents est particlulièrement complexe. L'identification des différentes copies dupliquées nécessitaient au préalable des approches lourdes telles que les méthodes de clonage précédant le séquençage de nombreux clones par la méthode Sanger. Par exemple, chez les Spartines polyploïdes, les phylogénies basées sur des gènes nucléaires comme le gène *Waxy* (codant l'amidon synthase) ont nécessité l'amplification et le clonage d'ADN chez plusieurs espèces (Fortune et al. 2007). Cette étude a notamment montré l'existance de paralogues et la rétention différentielle des copies paralogues et homéologues. Toutefois, à partir de telles analyses, il est possible que l'absence d'une copie résulte d'un échantillonnage incomplet des séquences clonées. La profondeur de lecture des séquences fournie par les NGS permet de résoudre ce problème.

Griffin et ses collaborateurs (2011) ont utilisé le pyroséquençage pour identifier les copies paralogues et homéologues dans le genre *Poa* où 3 régions chloroplastiques et deux régions nucléaires ont été séquencés chez 60 individus. L'identification des différentes copies avait dû être réalisée manuellement. Richardson *et al.* (2012) ont pu analyser la phylogénie de 329 échantillons chez *Artemisia tridentata* à l'aide de l'outil automatisé hapHunt, présenté aujourd'hui comme une suite du logiciel BamBam (Page et al. 2014). Récemment une étude menée sur 96 individus du genre *Hordeum* (portant sur 1 région chloroplastique et 12 gènes nucléaires pyroséquencés) a permis d'étudier les relations de parenté de l'ensemble des espèces diploïdes et polyploïdes du genre *Hordeum* (Brassac and Blattner 2015).

Néanmoins, encore peu d'études à ce jour ont été publiées sur les approches de phylogénomique utilisant un nombre important de régions génomiques ou transcriptomiques obtenues à l'aide de données NGS. Une analyse de phylogénomique comparant 22 génomes chloroplastiques de Bambusoideae a récemment été réalisé et a permis de résoudre les relations phylogénétiques au sein de ce groupe de Poaceae (Ma et al. 2014). Le nombre d'analyses phylogénomiques reste cependant limité, probablement en raison des développements particuliers nécessités par le grand nombre de matrices de caractères à prendre en compte et la multitude des topologies ainsi générées à réconcilier.

Pour analyser ces jeux de données en masse, plusieurs approches peuvent être utilisées. Les jeux de données peuvent par exemple être combinés en supermatrices, ce qui facilite l'analyse des résultats. Cette méthode a été appliquée sur des jeux de données de lignées d'Araneae (les araignées) pour résoudre les relations phylogénétiques de cet ordre (Bond et al. 2014). Il est également possible de réaliser l'ensemble des phylogénies pour les gènes étudiés, et de réaliser un arbre des espèces à l'aide de logiciels spécifiques tel que STAR (Liu et al. 2009), PHYLDOG (Boussau et al. 2013) ou ASTRAL (Mirarab et al. 2014). Néanmoins, les arbres utilisés par ces logiciels doivent présenter les mêmes espèces et le même nombre d'OTUs (correspondant aux différents nœuds terminaux ; « Operational Taxonomic Unit »). D'autre méthodes ont été développées pour des annotations sémantiques de données phylogénétiques par exemple (Panahiazar et al. 2013). Pour réaliser la phylogénie d'un grand nombre de gènes ou de supermatrices, le logiciel RAxML (Stamatakis 2014) est couramment utilisé. Cronin et ses collaborateurs (2014) ont utilisé RAxML pour des analyses phylogénomiques au sein des caniformes (Ursidae). Pour cela, ils ont utilisé de nouveaux marqueurs génétiques: des UCE (« Ultra-Conserved Elements ») spécifiquement détectés pour cette étude. Ils ont ainsi pu analyser 1 681 régions d'UCE (concaténées en 13 partitions) représentant une longueur totale de 996 381 bp. Une étude phylogénomique menée chez les tortues a permis de résoudre la phylogénie de cette espèce à partir de 2 381 UCE et du logiciel RAxML (pour une longueur totale de 1 718 154 bp ; Crawford et al. 2015). Tennessen et ses collaborateurs (2014) ont également utilisé ce logiciel pour réaliser les phylogénies de chacun des 7 chromosomes haploïdes du fraisier (*Fragaria*) et ainsi générer une phylogénie des sous-génomes octoploïdes homéologues. McKain et ses collaborateurs (2012) ont publié une étude de phylogénomique (réalisée à l'aide de RAxML) basée sur des données RNA-Seq d'*Agavoideae* (Asparagaceae) où ils ont pu utiliser 12 724 familles de gènes putatives. D'autre études ont également utilisé ce logiciel de phylogénie sur un nombre important de données NGS (*e.g.* Dunn, Howison, and Zapata 2013; Eaton and Ree 2013; Arbizu et al. 2014; Bond et al. 2014; Stephens et al. 2015). Pour les analyses bayesiennes, les logiciels tel que BUCKy (Larget et al. 2010) ou MrBayes (Ronquist and Huelsenbeck 2003) sont le plus souvent utilisés sur les jeux de données massifs (*e.g.* Song et al. 2012; Eaton and Ree 2013).

Récemment, des jeux de données de capture de séquences générés chez les cotons polyploïdes (Grover, Salmon, and Wendel 2012; Salmon et al. 2012) ont été analysés en phylogénie à l'aide des logiciels RAxML et MrBayes, ce qui a permis d'étudier les relations phylogénétiques au sein de ce genre. Il a été ainsi possible à partir de ces méthodes de réévaluer la phylogénie des espèces allotetraploïdes du coton (Grover et al. 2015) à partir de 52 gènes.

La détection et l'identification de l'origine des copies dupliquées au sein d'espèces polyploïdes ne disposant pas d'espèces diploïdes de référence est particulièrement complexe et nécessite une approche particulière. Ces situations sont en fait très répandues dans la nature, et c'est dans cette perspective que les travaux de recherche présentés ici ont été menés ; le but étant de développer des outils de détection de copies de gènes dupliquées sans utiliser de génomes diploïdes. Ces outils ont été appliqués sur des espèces hautement polyploïdes du genre *Spartina* (tétra- à dodécaploide), où aucune espèce diploïde n'est référencée à ce jour.

Chapitre 2 :

Le genre *Spartina*.

CHAPITRE 2: Le genre *Spartina*.

PARTIE A. Le genre *Spartina* au sein des Poaceae.

Le genre *Spartina* appartient à la famille des Poaceae, sous-famille des Chloridoideae. Les Poaceae constituent une famille importante au sein des Monocotylédones, tant au plan du nombre d'espèces (plus de 11 000 espèces regroupées au sein de 771 genres, (Clayton, Harman, and Williamson 2008; Soreng et al. 2015)) que de leur importance aux plans écologiques et économiques. Les Poaceae ont une répartition cosmopolite et occupent de nombreux habitats terrestres, constituant la composante majeure des prairies, savanes ou milieux ouverts anthropisés. Elles regroupent un nombre important d'espèces cultivées à intérêt agronomique majeur comme le riz (*Oryza sativa*), le blé (*Triticum aestivum*), le maïs (*Zea mays*), le sorgho (*Sorghum bicolor*) ou la canne à sucre (*Saccharum officinarum*) par exemple (Grass Phylogeny Working Group II 2012 ; http://www.theplantlist.org). L'histoire évolutive et la phylogénie des Poaceae ont fait l'objet de nombreux travaux (Mathews, Tsai, and Kellogg 2000; Bouchenak-Khelladi et al. 2008; Grass Phylogeny Working Group II 2012; Soreng et al. 2015).

La divergence des Monocotylédones et Eudicotylédones est estimée à la fin du Jurassique il y a environ 140 à 150 millions d'années (Chaw et al. 2004) et celle des Poaceae remonterait quant à elle entre 80 et 85 millions d'années (Prasad et al. 2005; Christin et al. 2014). Au cours de leur étude, Prasad et ses collaborateurs (2005) ont montré que les Poaceae faisaient partie du régime alimentaire des dinosaures durant le Crétacé (il y a 66 à 145 millions d'années). Un évènement de duplication génomique (σ) datant du Jurassique a été mis en évidence chez le riz et le sorgho (Tang et al. 2010). Un autre évènement de paleopolyploïdie (ρ) estimé à 70 millions d'années a également été détecté chez l'ancêtre des Poaceae (Paterson, Bowers, and Chapman 2004; Jiao et al. 2011; Paterson et al. 2012). Les Poaceae ont ensuite divergé en 12 familles dont la majorité est regroupée au sein de deux clades (Soreng et al. 2015) : le clade BEP ou BOP (composé des sous-familles Bambusoideae, Ehrhartoideae/Oryzoideae et Pooideae) et le clade PACMAD (composé des sous-familles Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae et

Danthonioideae), ayant divergé il y a 50 à 70 millions d'années (Paterson, Bowers, and Chapman 2004; Soreng et al. 2015). Au sein du clade PAC, Kim et ses collaborateurs (2009) ont estimé que les sous-familles Chloridoideae et Panicoideae ont divergé il y a 34,6 à 38,5 millions d'années.

Au sein de ces sous-familles, les génomes ont été remaniés de façon importante. On considère que les caryotypes des Poaceae ont évolué à partir d'un nombre de base ancestral de x = 5 (Salse et al. 2008). L'ancêtre commun à l'ensemble des Poaceae posséderait 12 chromosomes, ceci étant dû à une duplication génomique (il y a 50 à 70 millions d'années (Paterson, Bowers, and Chapman 2004)), deux translocations inter chromosomiques et deux fusions de chromosomes (Salse et al. 2008). L'ancêtre commun aux Chloridoideae et aux Panicoideae posséderait quant à lui 10 chromosomes (dû à deux fusions de chromosomes sur les 12 de l'ancêtre commun des Poaceae ; Salse et al. 2008). Certaines sous-familles de Poaceae ont subi un nombre important d'évènements de polyploïdisation et l'on considère aujourd'hui que la famille des Poaceae est l'une des familles où la proportion de polyploïdes « récents » est la plus importante, de l'ordre de 80% (Stebbins 1950). Bien qu'importante par le nombre d'espèces (plus de 1420 espèces recensées selon Clayton, Harman, and Williamson 2008; Watson and Dallwitz 1992), la sous-famille des Chloridoideae à laquelle appartiennent les Spartines est remarquablement peu étudiée, notamment du point de vue génétique et génomique. Le nombre de base chromosomique (x) est variable chez les Chloridoideae (x=6, 8, 9, 10) avec assez peu d'espèces diploïdes (2n=12, 18, 20) répertoriées, mais le nombre chromosomique de nombreuses espèces reste à déterminer (Peterson et al. 2014b). La systématique de ce groupe complexe est longtemps restée problématique (Jacobs 1987; Van Den Borre and Watson 1997). On signalera dans ce domaine les efforts récents de phylogénie moléculaire ayant permis de circonscrire les tribus et positionner les genres (Hilu and Alice 2001; Columbus et al. 2007; Peterson, Romaschenko, and Johnson 2010a; Peterson, Romaschenko, and Johnson 2010b). Le genre *Spartina* forme un groupe monophylétique (Baumel et al. 2002; Fortune et al. 2007). Une phylogénie moléculaire récente de l'ensemble des Chloridoideae (Peterson et al. 2014b) a amené un nouvel éclairage sur les relations entre les différentes lignées et la position du genre *Spartina*, dans la tribu des *Zoisieae (*sous tribu des *Sporobolinae).*

Le clade monophylétique des Spartines s'avère inclus dans le groupe paraphylétique des *Sporobolus,* avec comme lignée sœur un clade constitué des sections *Clandestini* (regroupant des espèces de *Sporobolus* ; 2n=9x= 36, 54, 108) et *Calamovilfa* (regroupant des espèces du même nom ; 2n=9x= 36, 54) (Peterson et al. 2014b). Ces résultats ont conduit Peterson et al. (2014a) à proposer de regrouper l'ensemble de ces sous-clades en un genre *Sporobolus*, très large (comprenant ~220 espèces), incluant le genre *Spartina*, qui deviendrait de ce fait la « section *Spartina* » (Figure 8).

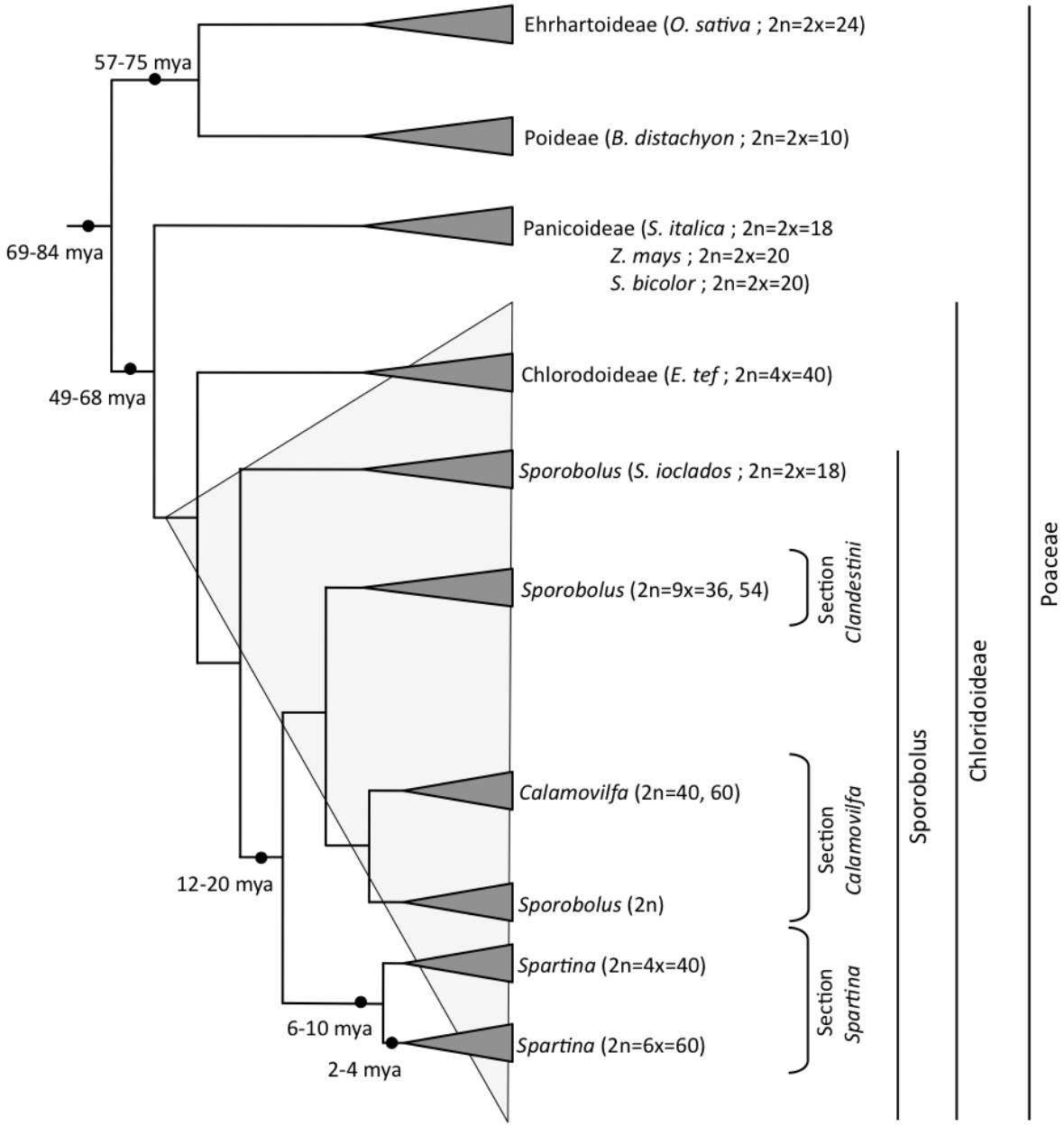


Figure 8 : Phylogénie simplifiée des Poaceae replaçant les Spartines au sein de cette famille (d'après la phylogénie de Peterson et al. 2014b et les datations moléculaires de Rousseau-Gueutin et al. 2015).

Par souci de simplicité et de clarté par rapport à la littérature des Spartines, nous conserverons au clade *Spartina* le rang de genre (en précisant pour chaque espèce l'autorité taxonomique correspondante).

Le genre *Spartina Schreb.* est constitué d'une quinzaine d'espèces vivaces colonisant les marais salés littoraux et parfois l'intérieur des terres. Ce sont des plantes pour la plupart halophytes dont l'évolution est particulièrement marquée par des évènements de polyploïdisation et/ou d'hybridation interspécifique ayant conduit à la formation d'espèces invasives (Ainouche et al. 2003; Ainouche et al. 2008). Le nombre chromosomique de base haploïde (x) est de 10 dans le genre *Spartina* (Marchant 1968b) ; les espèces sont tétraploïdes (2n=4x=40), hexaploïdes (2n=6x=60-62) ou dodécaploïdes (2n=12x=120, 122, 124). Des niveaux heptaploïdes (2n=7x=70) (Fortune et al. 2008), octoploïdes (2n=8x=80) (Kim, Rayburn, and Lee 2010) et nonaploïdes (2n=9x=90) (Renny-Byfield et al. 2010) ont récemment été décrits, correspondant à des hybridations entre espèces de niveaux de ploïdie différents. Aucune espèce diploïde n'est recensée à ce jour au sein de ce genre. Les différentes espèces de Spartines ont évolué en deux lignées principales (Baumel et al. 2002) : un clade tétraploïde (2n=4x=40) constitué principalement des espèces natives américaines et un clade hexaploïde (2n=6x=60). Les données de séquences chloroplastiques suggèrent que ces clades auraient divergé il y a 6 à 10 millions d'années (Figure 9, Rousseau-Gueutin et al. 2015).

Huit à neuf espèces sont regroupées au sein du clade tétraploïde : *Spartina arundinaceae* (Thouars) Carmich. (2n=4x=40) est localisée sur l'île de Tristan de Cunha (Océan Atlantique, Saint-Hélène), l'île Saint-Paul et l'île Amsterdam (Océan Indien, Terres Australes et Antarctiques Françaises) (Mobberley 1956; Marchant 1968a). *Spartina bakeri* Merr. (2n=4x=40, 2C = 1,43 - 1,48 pg) est distribuée sur les côtes de Floride et de Géorgie (Etats-Unis) (Mobberley 1956; Marchant 1968a; Fortune et al. 2008). *Spartina ciliata* (dont le nombre de chromosome n'a à notre connaissance pas encore été analysé) est présente sur les côtes Ouest Atlantiques d'Amérique du sud, au Brésil, en Uruguay et en Argentine (Mobberley 1956). *Spartina cynusoroides* (L.) Roth (2n=4x=40) se trouve sur les côtes Atlantiques des Etats-Unis (excepté en Floride) et sur les côtes américaines du Golfe du Mexique (Mobberley 1956; Marchant 1968a). *Spartina gracilis* Trin. (2n=4x=40) se localise à l'intérieur des terres aux Etats-Unis (Montagnes Rocheuses, Plateau du Colorado), au Canada

(Montagnes Rocheuses) et au Mexique (Mobberley 1956; Marchant 1968a). *Spartina patens* (Aiton) Muhl. (2n=4x=40) se situe sur l'ensemble des côtes Est Américaines, du Golfe du Mexique jusqu'au Golfe du Saint-Laurent au Canada. Cette espèce a été trouvée sur les côtes Est Mexicaines et du Costa Rica ainsi que sur les différentes îles séparant la mer des Caraïbes de l'océan Atlantique. *Spartina patens* a également été introduite sur les côtes Pacifique américaines dans les états de Washington, de l'Oregon et de la Californie (Daehler and Strong 1996) et a été récemment signalée en Angleterre (Hounsome 2013). Cette espèce est morphologiquement très proche de *S. bakeri* (Mobberley 1956; Marchant 1968a). *Spartina pectinata* Link (2n=4x=40) est distribuée dans l'intérieur des terres sur l'ensemble des Etats-Unis et une partie du Canada avec une concentration plus importante sur la côte New Yorkaise et dans le Midwest (Mobberley 1956; Marchant 1968a; Kim, Rayburn, and Lee 2010). *Spartina spartinae* (Trin.) Hitchc. également appelé *Spartina argentinensis* Parodi, est présente en Amérique du Sud (Argentine et Paraguay) et en Amérique du Nord (Costa Rica, Etats-Unis et Mexique). Elle se situe le long des côtes du golfe du Mexique, de la Floride, des côtes Est du Costa Rica et sur les côtes d'Argentine. *S. spartinaea* est également localisée dans les terres au Mexique, en Argentine et au Paraguay (Mobberley 1956). La position phylogénétique de cette espèce au sein du genre n'est pas clairement établie, et varie selon les jeux de données : elle se place en effet alternativement comme une espèce sœur du clade hexaploïde des Spartines dans les phylogénies basées sur le gène nucléaire *Waxy* ou GBSS I (Granule-Bound Starch Synthase I) (Baumel et al. 2002; Fortune et al. 2007) ou comme espèce sœur de toutes les autres Spartines à partir des séquences nucléaires ITS (Peterson et al. 2014a). De plus, les données de séquences chloroplastiques la placent en polytomie au sein du genre (Baumel et al. 2002; Peterson et al. 2014b). *Spartina versicolor* Fabre (Fabre 1849) a été décrite comme une espèce native d' Europe, depuis sa découverte le long du bassin Méditerranéen (France (Fabre 1849; Coste 1906; Jeanmonod and Burdet 1989), Italie (Parlatore 1850), Algérie (Cosson and Durieu de Maisonneuve 1867), Portugal (Daveau 1897)) mais également sur les côtes Atlantiques du Sud-Ouest de la Péninsule Ibérique (Sanchéz-Gullón 2001). En se basant sur les caractères morphologiques, Mobberley (1956) a considéré *S. versicolor* comme synonyme de *S. patens*, en supposant que *S. patens* avait été introduite depuis les côtes Est Américaines où elle est particulièrement abondante. Des analyses phylogénétiques moléculaires récentes (Prieto et al. 2011; Baumel et al. submitted), basées sur des séquences chloroplastiques et nucléaires, ont permis de montrer

qu'aucune différenciation n'existait entre ces deux taxons. Ces résultats, en accord avec les travaux de Mobberley, confirment que *S. versicolor* serait issue de l'introduction de *S. patens* depuis les côtes Nord-Américaines (Figure 10).

Le clade hexaploïde des Spartines contient trois espèces. *Spartina alterniflora* Loisel (2n=6x=62, 2C = 4,33 - 4,36 pg) présente l'une des plus larges distributions dans le genre (Figure 9) : elle se trouve sur l'ensemble des côtes Est Américaines et une partie du Canada (du golfe du Mexique jusqu'à l'embouchure du fleuve Saint-Laurent). *S. alterniflora* est également présente en Amérique du Sud et a été introduite accidentellement en Europe au 19ème siècle ou elle s'est développée en Angleterre et en France (Mobberley 1956; Marchant 1968b; Marchant 1968a; Fortune et al. 2008). *S. alterniflora* a également été introduite en Chine en 1979 où elle est rapidement devenue envahissante (An et al. 2007; Strong and Ayres 2013). *Spartina foliosa* Trin. (2n=6x=62) est localisée sur les côtes Ouest des Etats-Unis (Floride) et sur la côte Ouest Mexicaine (Baja California et Baja California Sur) (Mobberley 1956; Marchant 1968a). *Spartina maritima* (Curtis) Fren. (2n=6x=60 2C = 3,70 - 3,85 pg (Fortune et al. 2008)) est localisée en Europe, le long des côtes Atlantiques et de la Manche, sur les côtes Nord Italiennes de la mer Adriatique ainsi qu'en Afrique (Afrique du Sud, Maroc, Mauritanie et Sénégal) (Mobberley 1956; Marchant 1968a).

Le genre *Spartina* est sujet à de nombreux événements d'hybridation interspécifique et/ou de polyploïdisation. De ce fait, de nombreuses espèces hybrides ont été référencées (pour revue Ainouche et al. 2008; Strong and Ayres 2013). Les données morphologiques ont très tôt suggéré l'existence d'hybrides. Par exemple, en Amérique du Sud, *Spartina longispica* serait un hybride issu du croisement entre *S. densiflora* et *S. alterniflora* et *Spartina x caespitosa* serait une espèce hybride entre *S. patens* et *S. pectinata* (Mobberley 1956). Les données moléculaires ont par la suite permis de documenter de nombreux cas d'hybridation et de préciser l'histoire évolutive des hybrides. Des analyses génétiques et phylogénétiques ont ainsi pu mettre en évidence l'origine heptaploïde de *Spartina densiflora* Brongn. (2n=7x=70, 2C = 4,53 - 4,55 pg). Cette espèce est issue de l'hybridation d'une espèce maternelle tétraploïde proche de *S. arundinaceae* et d'une espèce hexaploïde proche de *S. alterniflora* (Baumel et al. 2002; Ayres et al. 2008; Fortune et al. 2008). *S. densiflora*, originaire de l'Amérique du Sud est aujourd'hui présente dans la Baie de San Francisco, le comté de Grays Harbor (Etat de Washington) mais également en Espagne et en Afrique du

Nord (Ayres et al. 2008; Bortolus 2006). Au cours de ses introductions dans plusieurs régions du monde, *Spartina densiflora* s'est à son tour hybridée avec différentes espèces hexaploïdes : avec *S. maritima* en Espagne (Castillo et al. 2010), en Californie où des hybrides ont été rapportés entre *Spartina densiflora* et *S. foliosa* et/ou *S. alterniflora* (Ayres et al. 2008). La Baie de San Francisco est le théâtre de plusieurs cas d'hybridations, dont les conséquences écologiques ont fait l'objet de nombreuses études (passées en revue par Strong and Ayres 2013). L'introduction de *S. alterniflora* de la côte Est des Etats-Unis vers la Californie a eue pour conséquence la formation récurrente de plantes hybrides qui se propagent dans la Baie de San Francisco. Des backcrosses successifs entre hybrides et *S. foliosa* ont conduit à un processus d'introgression qui « dilue » les génotypes natifs de *S. foliosa*, ce qui complique sérieusement les tentatives d'éradication des plantes introduites et des hybrides, étant donné qu'il devient de plus en plus difficile de distinguer les « hybrides envahissants » des plantes « natives ».

En Europe, *Spartina alterniflora* a été introduite au début du 19^{ième} siècle et s'est hybridée avec l'espèce native *S. maritima* dans le sud-ouest de la France et dans le sud de l'Angleterre. Ces hybridations ont conduit à la formation de deux hybrides F1 stériles (se maintenant toujours aujourd'hui par multiplication végétative) : *Spartina x townsendii* H & J Groves (2n=6x=62) à Southampton en Angleterre et *Spartina x neyrautii* Foucaud (2n=6x=62) au Pays Basque en France. Bien que présentant des morphologies très différentes (ce qui avait au départ suggéré l'idée de croisements réciproques entre les espèces parentales), ces deux hybrides résultent de croisements dans le même sens avec *S. alterniflora* comme parent maternel, donneur du génome chloroplastique (Baumel et al. 2003). La duplication du génome de *Spartina x townsendii* a donné naissance à une nouvelle espèce allododécaploïde, *Spartina anglica* C.E. Hubbard (2n=12x=120,122,124) (Marchant 1968b) (2n=12x=120, 122, 124) observée pour la première fois vers 1890 à Lymington (Gray, Marshall, and Raybould 1991). La naissance de *S. anglica* est un exemple devenu classique pour illustrer les mécanismes de la spéciation allopolyploïde dans les manuels de Biologie Evolutive (*e.g.* Walter and Briggs 1969; Arnold 2008). Cette espèce, très fertile et envahissante a rapidement colonisé l'Ouest de l'Europe et a été introduite sur divers continents (Asie, Australie) où elle s'avère envahissante. Elle est actuellement l'objet de nombreuses tentatives d'éradication dans ces régions (Cottet et al. 2007; Roberts and Pullin

2008). Cette expansion a de nombreuses conséquences écologiques, la Spartine anglaise jouant un rôle déterminant dans la dynamique sédimentaire des marais salés où elle s'installe en position pionnière, accélérant l'accrétion des sédiments et modifiant ainsi les caractéristiques du milieu colonisé (on parle d'espèce « ingénieur d' écosystèmes »). Cette espèce est aujourd'hui considérée comme l'une des 100 espèces les plus invasives et préoccupantes au monde (IUCN 2000). Le succès adaptatif et invasif de *S. anglica* est à mettre en relation avec son métabolisme photosynthétique inhabituellement élevé en C4 adapté aux températures élevées et aux environnements salins ainsi qu'aux multiples conséquences phénotypiques de l'hybridation et de la polyploïdie (pour revues: Ainouche et al. 2008; Ainouche et al. 2012; Ainouche and Wendel 2014). Les populations de *S. anglica* montrent une grande plasticité morphologique (Thompson, McNeilly, and Gay 1991). Au plan physiologique, elle tolère les conditions anoxiques, cette espèce est capable de stocker l'oxygène atmosphérique et de le transporter vers les racines (Maricle and Lee 2002). Lee (2003) a montré que *S. anglica* avait une activité de transport d'oxygène 5 fois supérieure à celle de *S. alterniflora*. Elle présente également une capacité de détoxification des sédiments pollués très importante ce qui en fait un excellent candidat pour les programmes de phytoremédiation (Lee 2003). Des travaux récents au sein de l'équipe suggèrent que *S. anglica* présente de plus grandes capacités de tolérance aux stress par les hydrocarbures aromatiques polycycliques ou HAP (Cavé-Radet 2015).

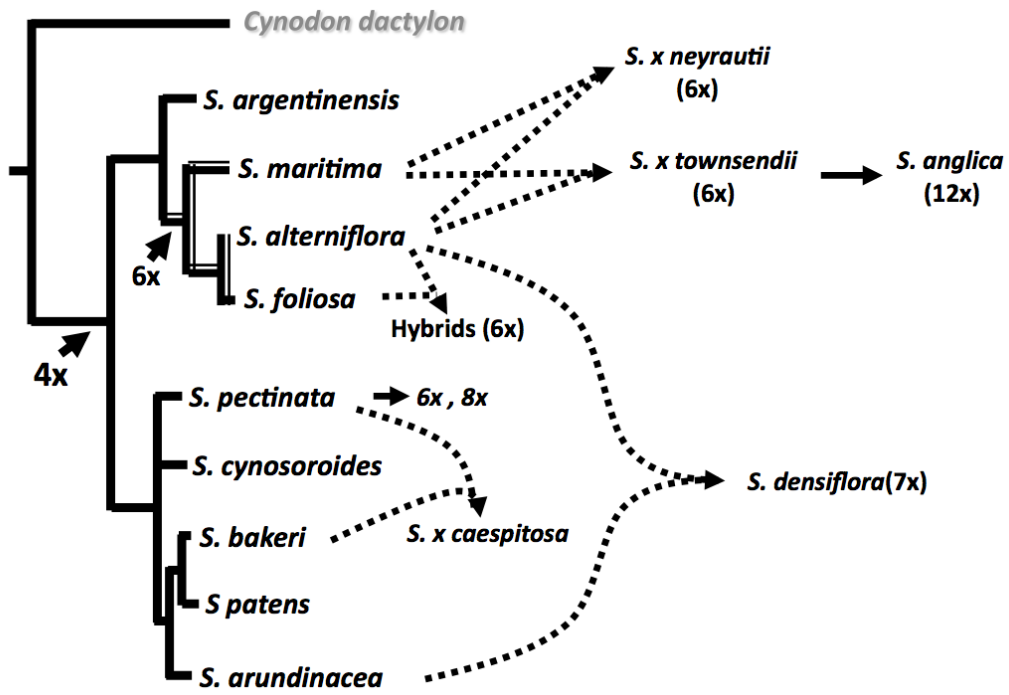


Figure 9 : Phylogénie simplifiée du genre *Spartina*. Cette phylogénie représente les évènements d'hybridations et de polyploïdisations récents (Adapté d'après Ainouche et al. 2012).

PARTIE B : Evolution des hybrides et allopolyploïdes récemment formés en Europe.

Ce travail de thèse concernera plus particulièrement les espèces hexaploïdes, leurs hybrides et l'espèce allopolyploïde formés en Europe au siècle dernier. Les hybridations naturelles récentes et bien documentées au plan historique de *S. maritima* et *S. alterniflora* nous offrent l'opportunité d'explorer les conséquences immédiates de deux événements d'hybridation indépendants (ayant formé les deux hybrides F1 *S. x townsendii* et *S. x neyrautii*) et d'étudier les conséquences à court terme d'un évènement de polyploïdisation récent (chez *S. anglica*) tout en ayant à disposition les génomes parentaux ; ce qui n'est pas le cas chez la plupart des espèces polyploïdes plus ou moins récentes pour lesquelles les parents ont disparu ou ont évolué depuis le temps de l'hybridation.

L'histoire et la génétique évolutive des Spartines hybrides et polyploïdes ont fait l'objet de nombreuses recherches menées notamment au sein du laboratoire d'accueil depuis les 15 dernières années. Plusieurs études ont exploré la **diversité génétique des populations** des Spartines européennes. Baumel et ses collaborateurs (2001) ont mis en

évidence à l'aide de différents marqueurs moléculaires, que les populations de l'espèce allododécaploïde *S. anglica* sont composées d'un génotype multilocus 'majeur'. De plus, ce génotype est identique à l'hybride *S. x townsendii* originaire d'Angleterre et se retrouve dans des régions où la plante a été plus récemment introduite comme en Australie (Ainouche et al. 2010). Il a également été montré que les populations européennes de l'espèce parentale *S. maritima* ne présentaient que très peu de variation génétique. En effet, sur les 98 marqueurs étudiés, seulement un marqueur présente un polymorphisme indiquant la présence de deux génotypes (Yannic, Baumel, and Ainouche 2004). Ces résultats sont en accord avec la biologie de cette espèce qui produit peu de graines et qui se multiplie essentiellement de façon végétative. *Spartina alterniflora* est une espèce allogame qui présente une forte diversité génétique sur les côtes atlantiques américaines (Perkins et al. 2002; Utomo et al. 2009; Blum et al. 2007). Les populations de *S. alterniflora* introduites en Europe (au sud de l'Angleterre, au Pays Basque et dans la rade de Brest en France) montrent une faible diversité génétique pour les marqueurs moléculaires (essentiellement des analyses de fragments multi locus RAPD, ISSR, AFLP) et populations échantillonnées à ce jour. Cependant, quelques marqueurs suggèrent la présence de variants génotypiques entre des échantillons du nord (Bretagne et sud de l'Angleterre) et du sud (Pays Basque) de la France, suggérant des événements d'introduction indépendants (Baumel et al. 2003). Différentes séquences chloroplastiques utilisées s'avèrent identiques entre toutes les populations européennes analysées (Baumel et al. 2003; Gharib 2012) et correspondent à des haplotypes chloroplastiques identifiés dans les populations atlantiques du nord des USA (et plus particulièrement de la région de Boston, Massachussetts). Une certaine diversité génotypique semblerait toutefois exister au sein des populations européennes (Gharib 2012), ce qui reste à confirmer.

Les deux espèces hybrides (*S. x townsendii* et *S. x neyrautii*) et l'allopolyploïde *S. anglica* ont hérité du génome chloroplastique de *S. alterniflora* (Baumel, Ainouche, and Levasseur 2001; Baumel et al. 2003; Ferris, King, and Gray 1997). Les deux hybrides F1 ont donc la même origine (ils ont le même parent maternel, les génotypes de chaque espèce parentale en Angleterre et en France sont très proches), et leurs génotypes sont quasiment identiques sur la base de marqueurs AFLP (A. Salmon et M. Ainouche, données non publiées). Comme on pourrait attendre de la formation récente de ces hybrides, leur

génome hexaploïde est constitué de l'addition des contributions parentales qui restent très peu remaniées (Baumel, Ainouche, and Levasseur 2001; Baumel et al. 2002; Ainouche et al. 2003). Les populations de *S. anglica* se caractérisent par une faible variation génétique inter-individuelle, résultant d'un « goulot génétique » qui intervient suite la spéciation récente (Raybould et al. 1991). Des analyses réalisées sur l'ensemble de l'aire de l'espèce montrent que les différents individus sont constitués d'un génotype majoritaire représentant l'addition des génomes de *S. alterniflora* et *S. maritima* (Baumel et al. 2001). Une étude récente des populations anglaises (Huska et al. in press) montre toutefois la perte de copies de gènes ribosomiques issues du parent paternel (*S. maritima*) chez quelques individus de *S. anglica*, indiquant un processus d'homogénéisation de cette famille multigénique.

Les effets les plus marquants de la spéciation allopolyploïde chez les Spartines se traduisent aux niveaux de la régulation épigénétique et de l'expression des gènes : des changements épigénétiques importants ont été détectés chez les deux hybrides F1 et chez *S. anglica*. Plus de 30% d'altération des profils de méthylation parentaux sont détectés chez *S. x townsendii* et *S. anglica* (Salmon, Ainouche, and Wendel 2005) via la méthode de MSAP (« Methylation Sensitive Amplified Polymorphism »). C'est notamment l'hybridation qui déclenche la majorité des changements qui sont transmis à l'allopolyploïde. Ces altérations de profils de méthylation sont plus importants dans les régions proches d'éléments transposables (Parisod et al. 2009).

Les premières **analyses de transcriptomes** chez les Spartines ont été réalisées à l'aide de puces hétérologues de riz (Chelaifa, Mahé, and Ainouche 2010; Chelaifa, Monnier, and Ainouche 2010) et ont mis en évidence des profils d'expression parentale non additifs chez *S. x townsendii* et *S. anglica*. Cette expression non additive se traduit de différentes manières : pour certains gènes, on note une dominance d'expression parentale, dans laquelle les niveaux d'expression de gènes différentiellement exprimés chez les parents sont similaires à l'un des parents (ici, le plus souvent similaires au parent maternel *S. alterniflora*) chez l'hybride ou l'allopolyploïde. Pour d'autres gènes, une expression transgressive (sur-exression ou sous-expression) des gènes par rapport aux deux parents est observée. Les effets de l'hybridation d'une part et de la duplication du génome d'autre part sont toutefois nuancés : on observe une diminution de la dominance maternelle d'expression entre l'hybride F1 (*S. x townsendii*) et l'allopolyploïde (*S. anglica*), tandis que la surexpression

(transgressive) de gènes augmente chez l'allopolyploïde par rapport à l'hybride F1 (Chelaifa, Monnier, and Ainouche 2010). Les changements observés concernent le niveau « global » d'expression de chaque gène ; la contribution relative des copies de gènes de chaque parent au niveau du transcriptome peut être très variable, allant d'une contribution équivalente, en cas « d'additivité » d'expression parentale, d'une contribution très inégale ou de mise sous silence de certaines copies en cas de « biais d'expression parentale » (Grover et al. 2012). Ces changements ne pouvaient être mesurés avec la technologie employée (puces) et ce d'autant plus que les espèces parentales sont elle-même hexaploïdes et qu'il n'existe pas de Spartines diploïdes de référence.

Ces limites sont aujourd'hui levées grâce aux nouvelles technologies de séquençage à haut débit et les ressources génomiques qui faisaient défaut dans le genre *Spartina* ont commencé à être générées. Un séquençage de transcriptome a été réalisé chez l'espèce tétraploïde *S. pectinata* (Gedye et al. 2010) et des premiers transcriptomes de référence ont été assemblés chez les Spartines hexaploïdes à l'aide de données Roche-454 à partir de différents organes (feuilles, racines) ce qui a permis d'annoter 16 753 gènes chez *S. maritima* et *S. alterniflora* (Ferreira de Carvalho et al. 2012). Une banque BAC (Bacterial Artificial Chromosome) a été construite chez *S. maritima*, et le séquençage de 60 451 extrémités de BACs a permis d'évaluer la composition génomique de cette espèce, où 16.91 % des séquences analysées sont constituées d'éléments transposables (Ferreira de Carvalho et al. 2013). Ces données ont également permis des comparaisons entre Spartines et lignées proches de Poacées, contribuant ainsi à combler les faibles connaissances de l'évolution génomique des Chloridoideae : ainsi, la microsynténie semble conservée entre *Spartina* (Chloridoideae) et *Sorghum* (Panicoideae), avec toutefois des réarrangements macrosynténiques qui peuvent être observés par rapport au sorgho et au riz.

Des compléments de séquençage (génome et transcriptome) en grande profondeur chez les espèces parentales, hybrides et allopolyploïdes de Spartines, permettront d'identifier les copies dupliquées (et leur expression) à chaque locus, et d'évaluer le degré de rétention ou perte des homéologues au cours des évènements successifs de polyploïdisations (anciennes et récentes) ayant affecté la lignée des Spartines. Une analyse phylogénétique réalisée précédemment avait permis d'identifier à l'aide de données de clonage et séquençage Sanger, des pertes différentielles de copies du gène *Waxy* au sein des

espèces parentales hexaploïdes (Fortune et al. 2007). La présence de 3 copies homéologues divergentes chez *S. alterniflora* suggérait une origine allopolyploïde (Fortune et al. 2007), mais la question de l'origine des parents hexaploïdes et celle de l'histoire ancienne du genre représentent des questions encore ouvertes, qui pourront être explorées à l'échelle du génome. Dans le contexte des Spartines (qui ne présentent aucun génome diploïde de référence), une démarche et des outils d'analyse adaptés sont à mettre au point et constituent donc la préoccupation majeure de ce travail.

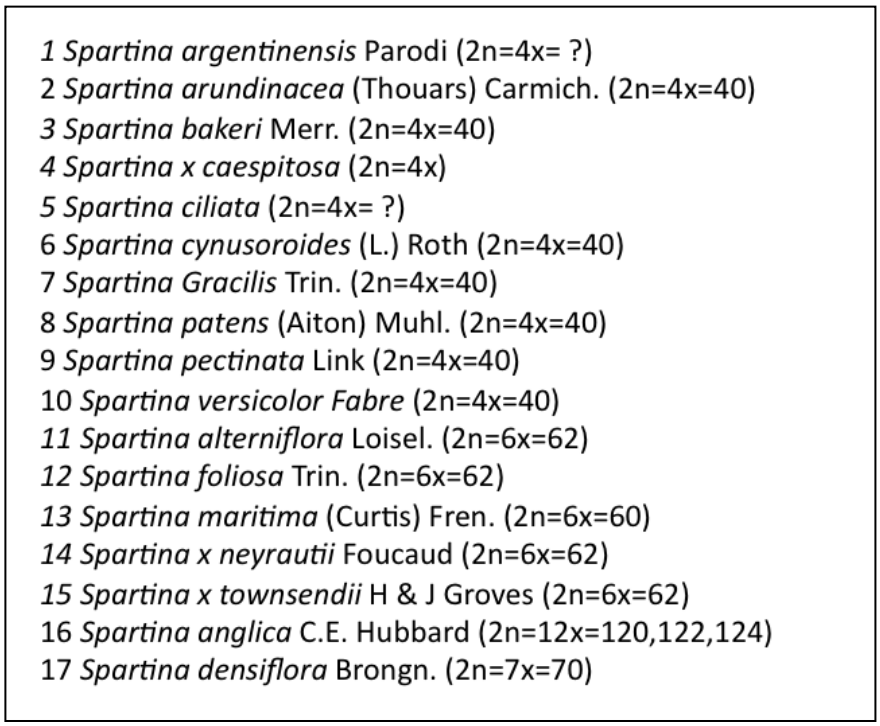
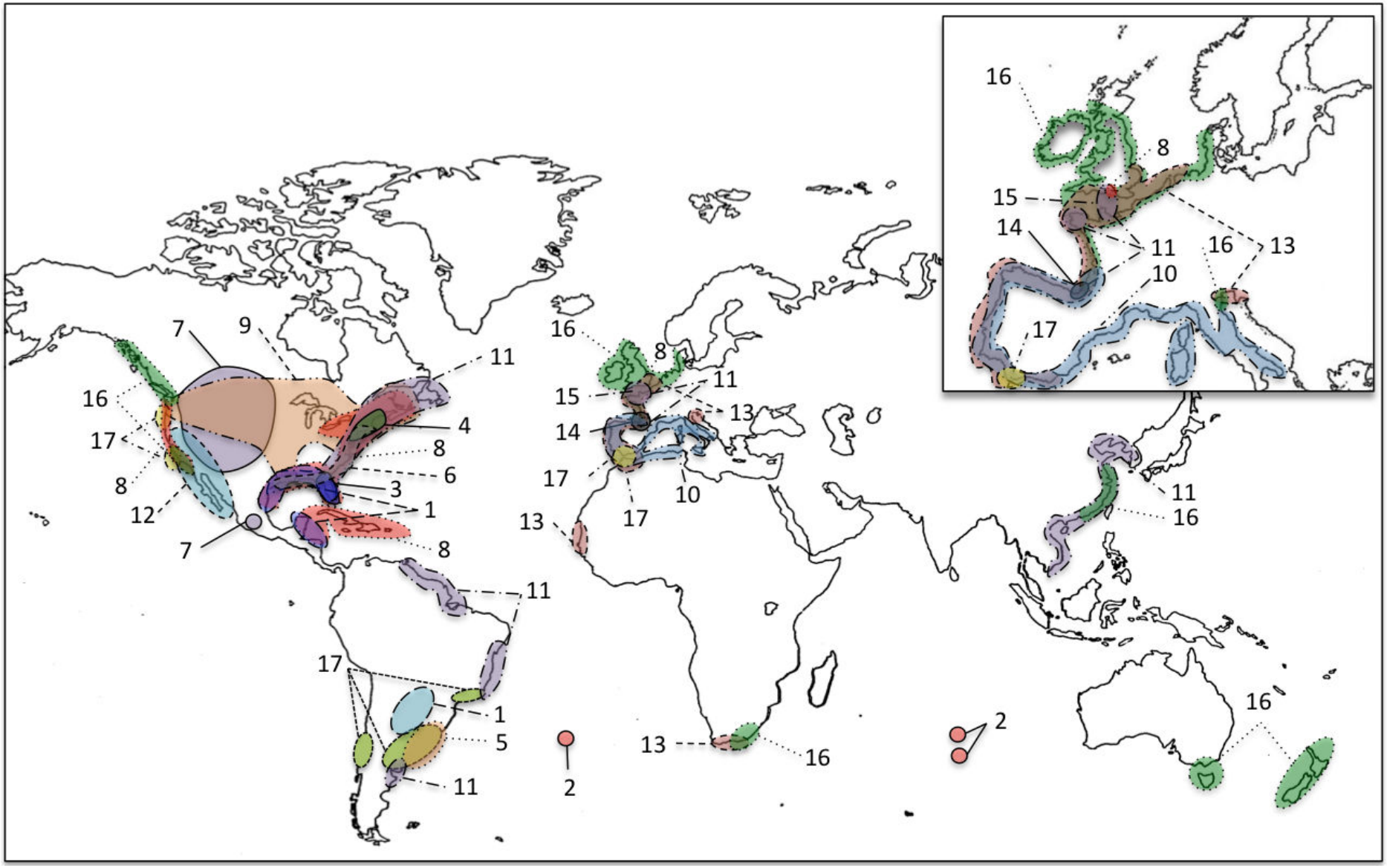


Figure 10 : Distribution géographique de 17 espèces de Spartines.


16
8
15
14
11
16
13
10
17
7
9
11
16
17
8
12
4
8
6
3
1
7
8
11
17
1
5
11
2
16
8
11
15
13
14
17
10
17
13
13
16
2
11
16
16

Chapitre 3 :

Matériel et Méthodes.

Chapitre 3: Matériel et Méthodes.

I. Matériel végétal et ressources génomiques.

i) Matériel biologique :

Les travaux de cette thèse portent essentiellement sur cinq espèces de Spartines : les deux parents hexaploïdes *S. maritima* (2n=6x=60), *S. alterniflora* (2n=6x=62), leurs hybrides F1 *S. x townsendii* (2n=6x=62) et *S. x neyrautii* (2n=6x=62) et l'espèce allododécaploïde *S. anglica* (2n=12x=120, 122, 124) issue du doublement du génome de *S. x townsendii*. Les échantillons de *S. alterniflora* ont été collectés à Landerneau (Finistère, France). *S. maritima* a été collecté sur les sites de la Presqu'île du Verdon (Morbihan, France) et de Noirmoutier (Vendée, France). *S. x townsendii* a été échantillonné à Hythe (Hampshire, Angleterre). Les échantillons de *S. x neyrautii* ont été collectés à Hendaye (Pyrénées Atlantiques, France) et *S. anglica* a été prélevée à Roscoff et dans l'Anse de Goulven (Finistère, France; Tableau 2). Les plantes entières ont été prélevées sur les différents sites avec une motte de sol autour des racines ; puis transplantées dans la serre expérimentale de l'Université de Rennes 1 (Rennes, France) et maintenues dans les mêmes conditions. Les plantes, transplantées dans des pots de 3 à 7 litres (1/3 sable, 1/3 terre et de 1/3 terreau) sont arrosées quotidiennement (environ 6 minutes le matin) à l'aide d'un arrosage automatique. Le cycle d'éclairage est lié à la photopériode naturelle. Les feuilles et racines ont ensuite été récoltées et stockées à -80°C jusqu'à l'extraction d'ARN (Ferreira de Carvalho et al. 2012). Plusieurs espèces de Spartines (représentant les différents niveaux de ploïdie) ont également été utilisées dans des analyses de données génomiques et dans le cadre d'un projet de séquençage ciblé qui permettra d'augmenter la profondeur de séquençage de plusieurs gènes d'intérêt (présenté en Partie 9 de ce chapitre ; Tableau 2). *Sporobolus heterolepis*, une espèce proche de la lignée des Spartines a également été utilisée comme outgroup dans le cadre de la capture de séquence.

Tableau 2: Liste des espèces (et leur provenance) utilisées pour les différentes analyses. * : d'après (Fortune et al. 2007; Ainouche et al. 2008; Ainouche et al. 2012).

Espèce :	Niveaux de ploïdie* :	Provenance des échantillons:	Analyses effectuées :
S. alterniflora	6x	Finistère (France)	- Analyses transcriptomiques - Analyses génomiques - Capture de séquence
S. maritima	6x	Morbihan (France)	- Analyses transcriptomiques - Analyses génomiques - Capture de séquence
		Vendée (France)	- Analyses transcriptomiques (pyroséquençage de banques normalisées)
S. x townsendii	6x	Kent (Angleterre)	- Analyses transcriptomiques - Capture de séquence
S. x neyrautii	6x	Pyrénées-Atlantiques (France)	- Analyses transcriptomiques - Capture de séquence
S. anglica	12x	Finistère (France)	- Analyses transcriptomiques - Capture de séquence
S. bakeri	4x	Floride (USA)	- Analyses génomiques - Capture de séquence
S. gracilis	4x	Bishop (CA, USA)	- Capture de séquence
S. pectinata	4x	Variété commercialisée (var. aureomarginata)	- Capture de séquence
S. spartinae (*S. argentinensis*)	4x	Santa Fe (Argentine ; réf. 3072 T. Colombus)	- Capture de séquence
S. versicolor	4x	Asturies (Espagne)	- Capture de séquence
		Var (France)	- Analyses génomiques
S. foliosa	6x	San Diego (CA, USA)	- Capture de séquence
S. densiflora	7x	Andalousie (Espagne)	- Capture de séquence

ii) Ressources génomiques :

Nous disposons de données génomiques issues des technologies Roche-454 (Ferreira de Carvalho 2013) et Illumina pour l'espèce hexaploïde *Spartina maritima*. L'ADN génomique de *S. maritima* (collectée à la Presqu'île du Verdon dans le Morbihan, France) a été pyroséquencé (un run GS FLX Roche 454 technologie Titanium, 999 229 reads obtenus ; longueur moyenne de 277 bp) à la Plateforme de Génomique Environnementale et Fonctionnelle du réseau Biogenouest (OSUR, Rennes). Les données de séquençage par synthèse (Illumina Hi-Seq) ont été obtenues à partir de banques d'ADN génomique de *S. maritima* (collectée à la Presqu'île du Verdon dans le Morbihan, France) séquencées au Beijing Genomics Institute (BGI ; République populaire de Chine). Nous disposons de 3 librairies pairées de 500 pb ; 800 bp et 2 Kb (mate-paired library) ; le séquençage sur Hi-Seq 2000 (Illumina) a permis d'obtenir respectivement 735,07 ; 566,82 et 298,82 millions de séquences (« reads »). L'ADN génomique de *S. alterniflora*, *S. bakeri* et *S. versicolor* (synonyme de *S. patens*) a également été séquencé par Illumina Hi-Seq (une piste ou « lane » par espèce, librairies pairées à 500 bp), permettant l'obtention respectivement de 177, 162 et 167 millions de séquences.

iii) Ressources transcriptomiques :

L'ARN a été extrait chez les 5 espèces de Spartines étudiées (les deux parents hexaploïdes *S. maritima* et *S. alterniflora*, les deux hybrides *S. x townsendii* et *S x neyrautii* et l'espèce allopolyploïde *S. anglica*) selon le protocole adapté pour les Spartines (Chelaifa, Monnier, and Ainouche 2010; Chelaifa, Mahé, and Ainouche 2010). Des plants de chaque espèce ont été maintenus en conditions contrôlées à la serre de l'Université de Rennes 1 et d'autres ont été échantillonnés sur différents sites selon un gradient d'immersion qui va du bas vers le haut de l'estran afin de maximiser les conditions d'expression du transcriptome en conditions naturelles. Les extractions d'ARNs ont été réalisées sur deux organes (feuilles et racines) comme cela a été décrit dans l'étude de Ferreira de Carvalho et ses collaborateurs (2012).

Les banques d'ADNc non-normalisées de chaque espèce ont été séquencées à l'aide de la technologie Roche-454 au Genoscope - Centre National de Séquençage (Evry, France) et les banques normalisées pour l'espèce *S. maritima* à la Plateforme de Génomique Environnementale et Fonctionnelle du réseau Biogenouest (OSUR, Rennes, France). Le séquençage des jeux de données Illumina a été effectué au Genoscope (Evry, France). Le nombre de reads obtenus pour chaque espèce est indiqué dans le Tableau 3.

Tableau 3: Nombre de reads disponibles pour les différents jeux de données (Roche-454 et Illumina) des cinq espèces du genre *Spartina* étudiées.

Espèces :	Nombre de reads 454 : (Longueur moyenne = 277 bp)	Nombre de reads Illumina : (Longueur moyenne : 108 bp)
S. maritima	984 006	76 985 267
S. alterniflora	495 749	77 321 929
S. x townsendii	322 773	71 358 554
S. x neyrautii	367 577	65 483 843
S. anglica	314 645	60 284 800

iv) Ressources bioinformatiques et logiciels utilisés :

Les différents outils appliqués sur l'ensemble des jeux de données des espèces du genre *Spartina* nécessitent un puissant cluster de calcul et d'importantes ressources de stockage. Pour répondre à cette demande, la plate-forme bio-informatique GenOuest (INRIA-IRISA Rennes Bretagne-Atlantique, Université de Rennes 1, France) du réseau Biogenouest (http://www.genouest.org/), a été sollicitée : elle met à disposition une grappe de calculateurs composée de machines possédant de 32G à 512G de RAM (Random Access Memory). Un gestionnaire de travaux Sun Grid Engine (SGE) est utilisé dans le but d'optimiser la répartition des calculs sur les différents nœuds. SGE est un programme d'équilibrage de charge qui alloue d'une manière optimale les ressources demandées (processeur, mémoire, espace disque) pour l'exécution d'une tâche non interactive. Pour permettre l'intégration des logiciels sous Galaxy, les différents programmes développés ont été vérifiés et validés à l'aide de machines virtuelles spécifiquement créées sous OpenNebula 4.8.0. Ces machines sont hébergées sur le Genocloud (plateforme spécifique dédiée à la création et l'hébergement de machines virtuelles) de la plate-forme GenOuest. Les différents résultats obtenus au cours de ces travaux sont stockés sur le serveur de

l'équipe Mécanismes à l'Origine de la Biodiversité (MOB ; serveur interne au laboratoire) qui dispose d'une capacité de stockage de 30 To (5X6 To ; Serial ATA 7 200tr/min ; RAID 5).

Les différents programmes développés au cours de cette thèse ont été réalisés à l'aide du langage python (v.2.7) et du langage Bash (Bourne-Again shell). Différents logiciels sont utilisés pour permettre la visualisation des données et le traitement des résultats obtenus. La visualisation des données est rendue possible grâce à l'utilisation des logiciels Tablet (Milne et al. 2009) (v. 1.12.03.26) et Jalview (Waterhouse et al. 2009) (v. 2.7). Tablet est un logiciel permettant de lire différents formats de fichiers (.ace, .sam, .bam), et de visualiser les résultats d'assemblage de taille importante issues de données de type NGS (les différents contigs avec leurs reads associés). Le logiciel Jalview, qui permet de visualiser et d'éditer des fichiers fasta de petite taille contenant plusieurs séquences, est utilisé dans les étapes de vérification de construction d'haplotypes. L'analyse des résultats est effectuée grâce au langage de programmation et l'environnement R (v. 2.13.0 ; http://www.r-project.org/; R Development Core Team 2011). Les différentes analyses phylogénomiques ont été réalisées à l'aide des logiciels MEGA (v.5.2.1 ; Tamura et al. 2011), Geneious (v.6.1.6 ; Drummond et al. 2010), et RAxML (v 7.2.8 ; Stamatakis 2014). Les arbres obtenus à partir du logiciel RAxML ont été visualisés à l'aide du logiciel Dendroscope v3.2.2 (Huson and Scornavacca 2012).

II. Méthodes

i) Assemblage et alignement de données NGS :

Assemblage de données transcriptomiques Roche-454 et alignement de données RNA-Seq Illumina

Afin de détecter les différentes copies au sein des données trancriptomiques disponibles, nous avons dans un premier temps assemblé pour chaque espèce les données de pyroséquençage Roche-454 en tenant compte de la qualité de séquençage. Pour cela, nous avons utilisé le logiciel Genome Assembler (Margulies et al. 2005) (v. 2.6, Roche), un assembleur *de novo* de données de séquençage Roche-454. Pour pouvoir détecter les différents haplotypes, l'assemblage *de novo* des données via ce logiciel (communément appelé Newbler) a été réalisé avec les paramètres suivant : -mi 90, -ml (chevauchement minimal en bp) 100 et -cdna. Ces paramètres permettent d'indiquer à l'assembleur que le

pourcentage minimum d'identité entre deux reads doit être de 90%, afin de construire des contigs correspondant aux consensus des différentes copies dupliquées au sein du génome des Spartines, qui serviront de référence pour les alignements de reads, la recherche de polymorphismes (SNPs et indels) et la détection d'haplotypes. Le dernier paramètre permet d'indiquer au logiciel que nous travaillons sur des données transcriptomiques (prise en compte des motifs ARN et de l'épissage alternatif ; Figure 11.A). Dans un deuxième temps, les contigs obtenus ont servi de référence pour aligner (par « mapping ») les données Illumina via l'utilisation du logiciel Bowtie 2 (Langmead and Salzberg 2012) (v. 2.0) qui permet d'aligner les reads de petites tailles sur des séquences de référence. Les paramètres utilisés pour le mapping des reads sont « score-min : G, 52, 8 » pour la fonction logarithmique naturelle $f(x)= 52 + 8 * ln(x)$, ou x correspond à la longueur des reads. En utilisant ces paramètres, seuls les reads (d'une longueur de 108 bp) présentant au moins 89.46% d'identité ont été mappés sur le contig de référence (Figure 11.C). Le choix de cette stratégie et de ces logiciels est motivé par les études comparatives des différents logiciels d'assemblage de données issues de NGS (Miller, Koren, and Sutton 2010; Garg et al. 2011; Fonseca et al. 2012) et par des tests effectués sur nos jeux de données à l'aide des logiciels Bowtie, Bowtie 2, Newbler et Novoalign (www.novocraft.com).

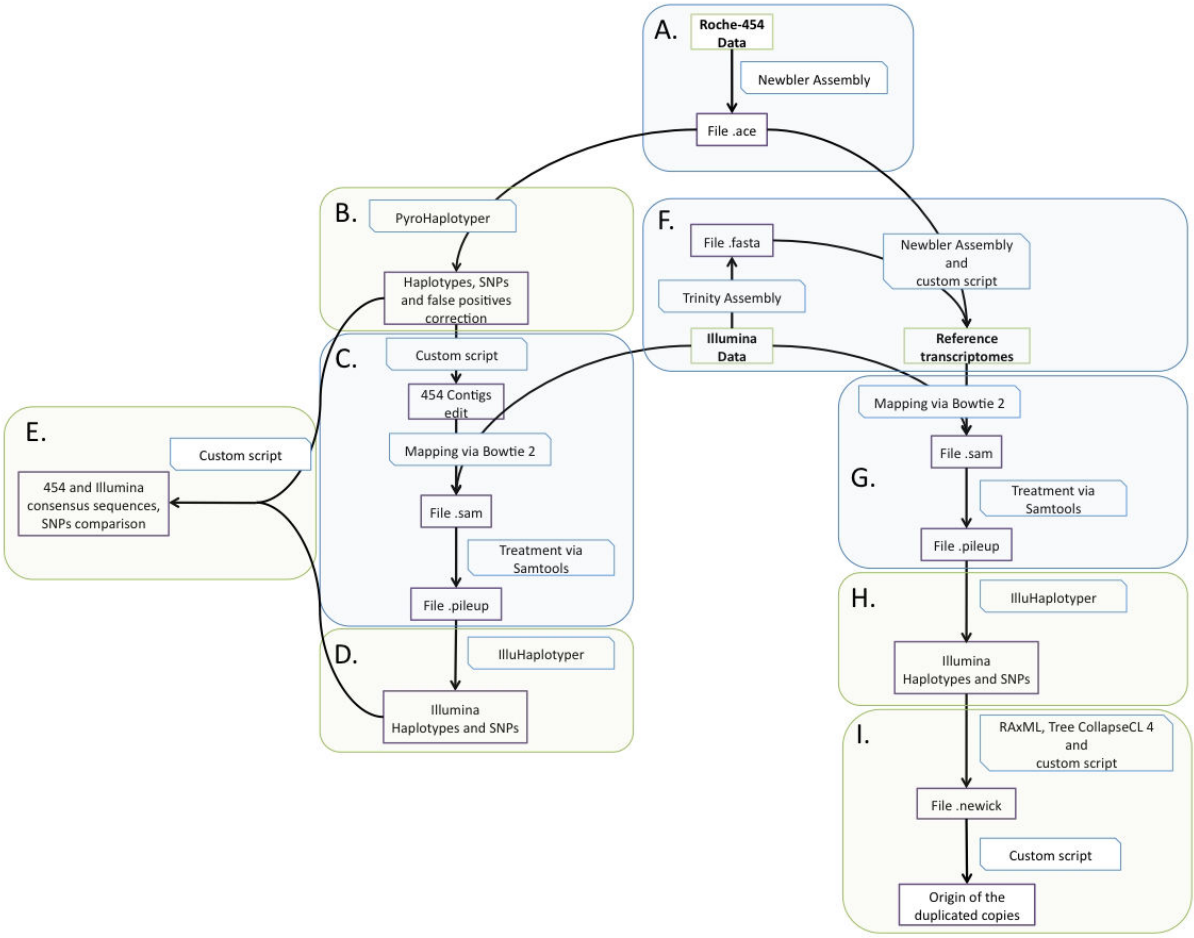


Figure 11 : Représentation globale de la stratégie d'analyse des données NGS (Roche-454 et Illumina) et de la reconstruction d'haplotypes. Les boites vertes correspondent aux jeux de données utilisés pour l'analyse. Les boites bleues correspondent aux programmes développés et aux logiciels utilisés. Les boites violettes indiquent les fichiers générés au cours de chaque étape (et leurs formats). Les cadres A à I représentent les différentes étapes du pipeline global présenté dans ce chapitre. Les cadres bleus correspondent à des étapes d'assemblage ou de mapping de données, les cadres verts correspondent aux étapes d'analyses des données.

Comparaison de logiciels d'assemblage de novo de données RNA-seq Illumina.

L'assemblage des données RNA-Seq Illumina nécessite l'utilisation d'outils adaptés. Pour identifier le logiciel le plus adapté à nos jeux de données, plusieurs comparaisons ont été réalisées entre les différents logiciels disponibles. De plus, une étude comparative des logiciels d'assemblage *de novo* de données de séquençage RNA-seq Illumina Trinity (Grabherr et al. 2011) et Minia (Chikhi and Rizk 2012) a été réalisée en collaboration avec Erwan Scaon, doctorant à l'INRIA-IRISA de Rennes au sein de l'équipe Genscale. Pour réaliser

cette étude, 37 189 145 reads issus du séquençage transcriptomique de feuilles de *Spartina maritima* ont dans un premier temps été assemblés avec chaque logiciel. Nous avons tout d'abord assemblé le jeu de données disponible via le logiciel Minia en utilisant une taille de k-mer de 31 et en ne gardant que les k-mers présent au minimum deux fois. Puis nous avons assemblé les reads via le logiciel Trinity en utilisant une taille de k-mer de 25 et en ne conservant que les k-mers présents deux fois ou plus lors du comptage par la méthode « Jellyfish ». Les données de séquençage ont ensuite été mappées sur les contigs obtenus via l'utilisation de deux logiciels (Bowtie et Bowtie 2 ; Langmead et al. 2009; Langmead and Salzberg 2012) pour obtenir le nombre de reads constituant chaque contig. Les logiciels d'alignements ont été utilisés avec les paramètres par défaut, un deuxième alignement a ensuite été réalisé avec Bowtie 2 en utilisant des paramètres moins stringents (--local, -N 1, --gbar 10, --ma 3). Ces paramètres correspondent à :

--local : Les reads peuvent s'aligner en ayant l'une et/ou l'autre des extrémités étant "soft-clipped" (le read s'aligne partiellement sur la séquence de référence, l'une de ses extrémités en 5' ou 3' ne présentant pas de similitude avec la séquence de référence).

-N 1 : Autorise un mismatch par rapport à la référence lors de l'alignement des graines de 20 nucléotides.

-gbar 10 : N'autorise pas de gap dans les 10 premiers nucléotides de la lecture contre la référence.

--ma 3 : Un bonus de +3 est donné lors d'un match au sein d'un alignement local.

Afin de valider les données obtenues, les contigs ont été alignés contre une base de données incluant 4 espèces de Poaceae (*Brachypodium distachyon*, *Oryza sativa*, *Zea mays* et *Sorghum bicolor*) par tblastx avec une *e-value* réglée à 10^{-06}.

Construction de transcriptomes de référence à partir de données Roche-454 et Illumina

Au cours de ce travail de thèse, nous avons développé une méthode spécifique pour construire un nouveau transcriptome de référence à partir de données Roche-454 et

Illumina. Pour chaque espèce, nous avons dans un premier temps assemblé les données Roche-454 et Illumina indépendamment en prenant en compte les données de qualité de séquençage. Le jeux de données Roche-454 ont été assemblés à l'aide du logiciel Newbler (v. 2.6, Roche) avec les paramètres ml= 80 bp, mi=90% et l'option cdna (Margulies et al. 2005). Le logiciel Trinity (Grabherr et al. 2011), recommandé pour les assemblages de données RNA-Seq Illumina (Chopra et al. 2014; Clarke et al. 2013; Liu et al. 2013) et présentant des résultats de bonne qualité avec nos jeux de données a été utilisé avec les paramètres suivants : taille des k-mers réglé à 25 bp et longueur minimale des contigs fixé à 48 bp. Les contigs obtenus sont ensuite co-assemblés avec le logiciel Newbler (ml= 40 ; mi= 90%). Pour éviter la formation de contigs chimériques durant ce processus, seuls les contigs ayant une longueur supérieure ou égale à 100 bp ont été utilisés. Une fois ces deux étapes réalisées, nous sélectionnons les contigs obtenus lors de l'étape de co-assemblage ainsi que les contigs obtenus avec Newbler et Trinity d'une longueur supérieure ou égale à 40 bp et qui n'ont pas été utilisés durant le processus de co-assemblage. L'ensemble de ces contigs est ensuite assemblé en supprimant les contigs redondants et en maximisant la longueur des contigs chevauchants. Pour réaliser cet assemblage, un self-BLAST a été effectué entre ces contigs. Les contigs strictement inclus dans des contigs de longueur plus importantes ont été supprimés et les contigs se chevauchant de plus de 50 bp et présentant un pourcentage d'identité supérieur ou égal à 90% on été assemblés manuellement à l'aide de scripts développés en langage python. Une dernière étape de self-BLAST a ensuite été réalisée pour vérifier que les jeux de données obtenus ne présentaient plus de redondance. Les paramètres utilisés lors de cette étape sont équivalents aux paramètres minimaux permettant l'assemblage des contigs (longueur de chevauchement minimal : 40 bp et pourcentage minimum d'identité : 90% ; Figure 11.F).

Alignement de données génomiques Illumina

Pour valider le programme développé au cours de cette thèse qui permet de détecter des polymorphismes et de reconstruire des haplotypes à partir de données Illumina, nous avons cherché à valider nos résultats à l'aide de données de clonage et de séquençage

Sanger issues des bases de données de séquences NCBI ou disponible au sein du laboratoire. Pour cela nous avons comparé ces données de clonage à des jeux de données génomiques issues de la technologie Illumina. Les alignements de ces données génomiques ont été réalisés sur des séquences de référence à l'aide du logiciel Bowtie 2 (Langmead and Salzberg 2012). Dans le cadre de l'analyse de l'ADN ribosomique de *S. maritima*, nous avons aligné les reads génomiques Illumina de cette espèce contre la séquence de l'ADNr à l'aide du logiciel Bowtie 2 (« score-min : G, 52, 8 »). Au cours de cette thèse, nous avons également étudié le gène *Waxy* (ou GBSS I pour « Granule Bound Sucrose Synthase I ») précédemment cloné et séquencé à l'aide de la méthode Sanger au sein du laboratoire (Fortune et al. 2007). Pour l'analyse de ce gène, les données génomiques issues de la technologie Illumina pour quatre espèces étudiées (*S. martima*, *S. alterniflora*, *S. bakeri* et *S. versicolor*) ont été mappées à l'aide de Bowtie 2 sur une séquence de référence provenant du génome de *S. maritima*. Les paramètres de mapping ont donc été adaptés en fonction des jeux de données utilisés. Les reads issus de *S. maritima* ont été mappés à l'aide des paramètres suivants : « score-min : G, 52, 8 », ce qui correspond à un alignement de 88.84% pour des reads d'une longueur de 100 bp. Les mappings des reads des trois autres espèces ont été réalisés à l'aide de paramètres moins stringents : « score-min : G, 52, 6 », ce qui correspond à un alignement de séquence de 79.63% (pour des reads de 100 bp).

ii) Annotation fonctionnelle :

L'annotation fonctionnelle des nouveaux transcriptomes de référence construits a été réalisée à l'aide d'une méthode par recherche d'homologie de séquences annotées au sein des bases de données transcriptomiques, similaire à celle précédemment développée au laboratoire (Ferreira de Carvalho et al. 2012) et grâce au logiciel Pfam ; une base de données de protéines qui utilise des alignements multiples de séquences et un profil HMM (Hidden Markov Model ; Finn et al. 2014). L'ensemble des contigs a été alignés à l'aide des algorithmes BLASTn et tBLASTx (avec une *e-value* de 10^{-5} ; Altschul et al. 1997) contre des bases de données construites à partir d'ESTs de Poaceae incluant les séquences de *Oryza sativa*, *Setaria italica*, *Brachipodium distachyon*, *Sorghum bicolor* (www.phytozome.net) et *Zea mays* (les ESTs du maïs ont été obtenus en concaténant deux bases de données

disponibles sur les sites www.phytozome.net et www.plantgdb.org). Seul le Best BLAST Hit (BBH) a été retenu pour obtenir l'annotation fonctionnelle des contigs basées sur l'homologie de séquences. L'ontologie de gènes (GO) a été analysé à partir du logiciel BLAST2GO (Conesa et al. 2005; Götz et al. 2008). Les annotations d'ontologie de gènes des différents contigs assemblés ont été réalisées à partir d'analyses tBLASTx (score de *e-value* minimale réglé à 10^{-5}) contre la base de donnée d'*Arabidopsis thaliana* (téléchargée sur le site TAIR, www.arabidopsis.org). Seul les résultats présentant une *e-value* inférieure ou égale à 10^{-6} avec un maximum de similarité de 55% ou plus ont été retenus. La base de données Pfam 27.0 a été utilisée pour enrichir les annotations des domaines protéiques (les 6 cadres de lectures ont été testés pour chaque contig avec l'option PfamB). Les résultats obtenus à l'aide de Pfam ont été filtrés et seuls les résultats significatifs et présentant une *e-value* inférieure à 10^{-3} ont été conservés. Le choix des paramètres utilisés a été fait d'après les travaux de Finn et ses collaborateurs (2014).

Pour les différentes espèces de Spartines étudiées, il a été possible d'estimer le nombre d'exons de chaque contig. Pour cela, nous avons utilisé le génome du riz ainsi que les fichiers d'annotations au format « .GFF3 » indiquant la localisation des gènes sur le génome (www.phytozome.net). L'ensemble des contigs a été aligné contre le génome du riz via BLASTn. Seuls les alignements d'exons présentant un pourcentage d'identité supérieur ou égal à 70% avec un chevauchement de séquence supérieur ou égal à 60 pb ont été retenus.

iii) Détection de SNPs et reconstruction d'haplotypes à partir de données NGS :

a) Détection de SNPs et reconstruction d'haplotypes à partir de données de pyroséquençage Roche-454 (ou « lectures longues ») :

Le programme « PyroHaplotyper » a été développé à l'aide du langage python (v.2.7) dans le but de détecter les différentes copies dupliquées au sein de jeux de données de pyroséquençage Roche-454. À partir d'un alignement de reads homologues, les différents polymorphismes de séquence (incluant des SNPs et des insertions/délétions en bloc) sont détectés et corrigés pour éliminer les erreurs de séquençage potentielles par l'utilisation de filtres de profondeur minimal et les biais de séquençage propres au pyroséquençage dans

les régions homopolymériques. Les reads partageants les mêmes polymorphismes sont ensuite identifiés et assemblés en un même haplotype. Ce programme a été appliqué sur les données génomiques et plus particulièrement sur l'unité 45S de l'ADN ribosomique de *S. maritima* ainsi que sur l'ensemble des données transcriptomiques Roche-454 des 5 espèces étudiées (Figure 11.B).

Détection de Polymorphismes

Pour détecter les différents SNPs au sein des alignements de séquences (reads), nous avons développé un pipeline qui utilise dans un premier temps le module Ace.py de la suite Biopython (http://biopython.org) et qui permet le parsing des fichiers au format « .ACE ». Dans un deuxième temps, nous détectons et corrigeons les différents sites polymorphes à l'aide des scripts développés. Ce pipeline dispose d'une étape de mapping avec le logiciel GSMapper de la suite Newbler (paramètres par défaut : ml= 40 bp ; mi= 90%). Cette étape permet de réaligner les différents reads et facilite la détection de polymorphismes. Lors de la détection des sites polymorphes, trois types d'erreurs de séquençage pouvant entrainer la formation de SNPs faux-positifs sont corrigés : en début et fin de chaque read, au sein des reads et lors de la détection des SNPs.

Détection des haplotypes

Les différents haplotypes sont construits à partir des reads partageant les mêmes polymorphismes. Pour cela, chaque read est comparé avec les autres reads de l'alignement. Pour chaque paire de reads, les polymorphismes présents dans les régions chevauchantes sont identifiés et comparés. Si tous les polymorphismes sont identiques, alors les deux reads sont assemblés pour construire un haplotype qui sera de taille maximale. Un haplotype correspond donc à une séquence caractérisée par un ensemble de sites polymorphes donnés, correspondant à une même et unique séquence (Figure 12). Néanmoins, pour

chaque alignement, il est nécessaire de définir différentes fenêtres pour compter le nombre d'haplotypes chevauchants. Les fenêtres présentant au moins un SNP sont identifiées pour compter le nombre d'haplotypes alignés localement (Figure 13). Trois fichiers de sortie permettent la visualisation des données pour chaque assemblage sous la forme d'un tableau, d'un fichier fasta et d'un fichier texte.

Paramètres du pipeline

Ce programme dispose de quatre paramètres ajustables par l'utilisateur :

- Le seuil ou « threshold » (option –t en ligne de commande) réglé à 20% par défaut correspond au seuil minimum pour considérer un SNP comme valide. Les SNPs qui présentent au moins deux nucléotides présents plus de 1/5 de fois sont conservés, et ceux présents moins de 1/5 de fois sont supprimés.

- Le bord ou « trimming » (option –s en ligne de commande) correspond au nombre de nucléotides qui ne seront pas pris en compte en début et fin de chaque read. Ce paramètre est réglé à 5 nucléotides par défaut.

- La profondeur ou « depth » (option –d en ligne de commande) réglée à 2 reads par défaut correspond au nombre minimal de reads pour considérer un SNP comme valide.

- Le nombre de SNPs pour assembler deux reads entre eux (option –n en ligne de commande). Ce paramètre réglé à 1 SNP par défaut correspond au nombre minimum de SNPs communs entre deux reads pour pouvoir assembler ces derniers entre eux.

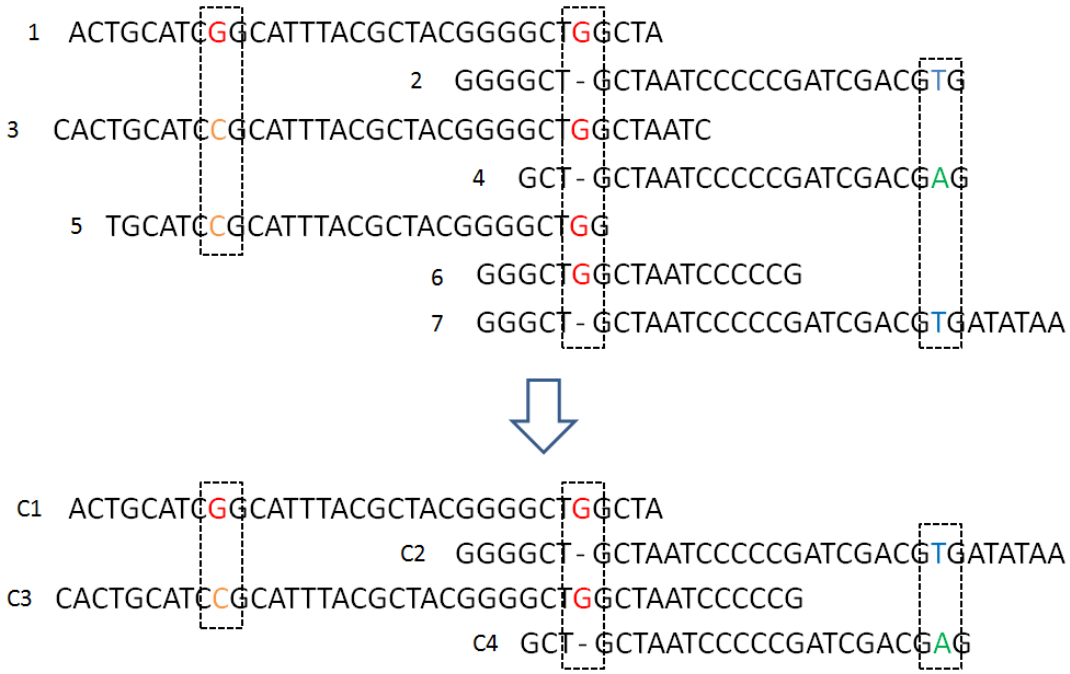


Figure 12 : Exemple d'assemblage effectué par le programme « PyroHaplotyper ». Les reads 3, 5 et 6 présentant un taux de similarité de 100% sont assemblés en un haplotype C3. Les reads 2 et 7 sont assemblés en un haplotype C2. Les reads 1 et 4 ne présentent pas de similitudes avec les autre reads, ils sont considérés respectivement comme les haplotypes C1 et C4.

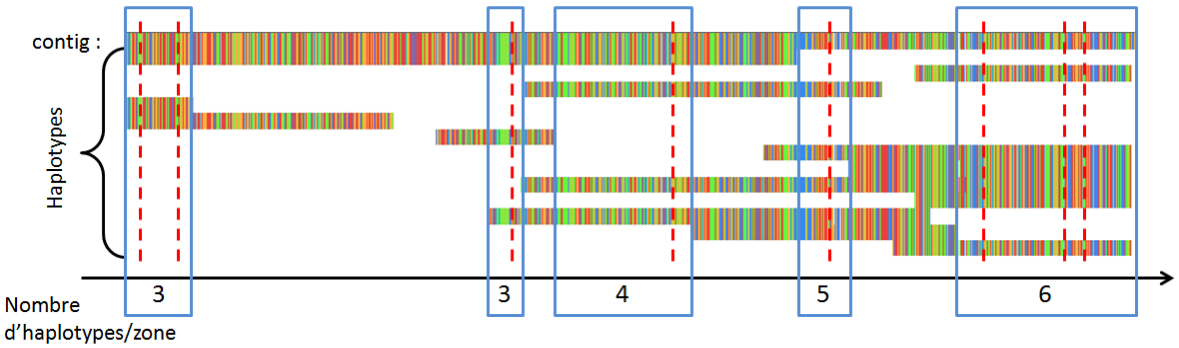


Figure 13 : Représentation des différents haplotypes détectés au sein d'un alignement donné. Les polymorphismes sont présentés par les traits en pointillés rouge, les rectangles bleus représentent les fenêtres contenant au moins un SNP et retenus lors du calcul du nombre d'haplotypes par fenêtre et de la divergence nucléotidique entre haplotypes.

Nous avons également automatisé ce programme à l'aide du langage de programmation Bash, pour pouvoir l'utiliser sur de grands jeux de données, de ce fait nous avons créé un fichier de sortie global contenant des informations sur les différents alignements traités (le nom du contig, sa taille (en bp), le nombre de reads associés au contig, le nombre de reads utilisés pour assembler les différents haplotypes, le nombre de SNPs détectés ainsi que leurs positions respectives, et le nombre d'haplotypes construits).

L'ajout de deux paramètres réglables par l'utilisateur permet de sélectionner la taille minimale d'un alignement pour que celui-ci soit traité par le programme (option –l en ligne de commande ; réglé à 100 bp par défaut) et le nombre minimum de reads (option -n en ligne de commande ; réglé à 10 reads par défaut) dont il doit être composé.

Intégration sous Galaxy du logiciel « PyroHaplotyper » :

Au cours de cette thèse nous avons intégré l'outil « PyroHaplotyper » au sein la plateforme Galaxy de Genouest en collaboration avec Yvan Le Bras (INRIA/IRISA, Genouest) afin de le rendre disponible à la communauté scientifique (Figure 14). Ce travail a fait intervenir Yannick Namour qui a réalisé un stage de Master 1 en bio-Informatique et Génomique (BIG ; Université de Rennes 1) au sein du laboratoire (Namour 2015). Pour cela nous avons développé un descripteur écrit en langage XML (eXtensible Markup Language), qui permet de définir une mise en page des contenus dans des balises personnalisables de type HTML. Ce descripteur, nécessaire pour mettre en ligne le pipeline, permet de créer une interface simple d'utilisation qui propose aux utilisateurs de renseigner leurs propres réglages, de sélectionner leurs données de départ et les guides lors de l'utilisation du logiciel nouvellement installé. Le logiciel développé sous Galaxy dispose de 8 paramètres réglables par l'utilisateur :

- Le chevauchement minimal entre deux reads en nombre de bases (option -ml en ligne de commande, 40 bp par défaut) pour assembler et/ou mapper les reads (option de gsAssembler et gsMapper de Newbler).

- L'identité nucléotidique entre deux chevauchements de reads en pourcentage (option -mi en ligne de commande, 90% par défaut) pour assembler et/ou mapper les reads (option de gsAssembler et gsMapper de Newbler).

- Les 6 paramètres du logiciel « PyroHaplotyper » présentés dans le paragraphe précédent.

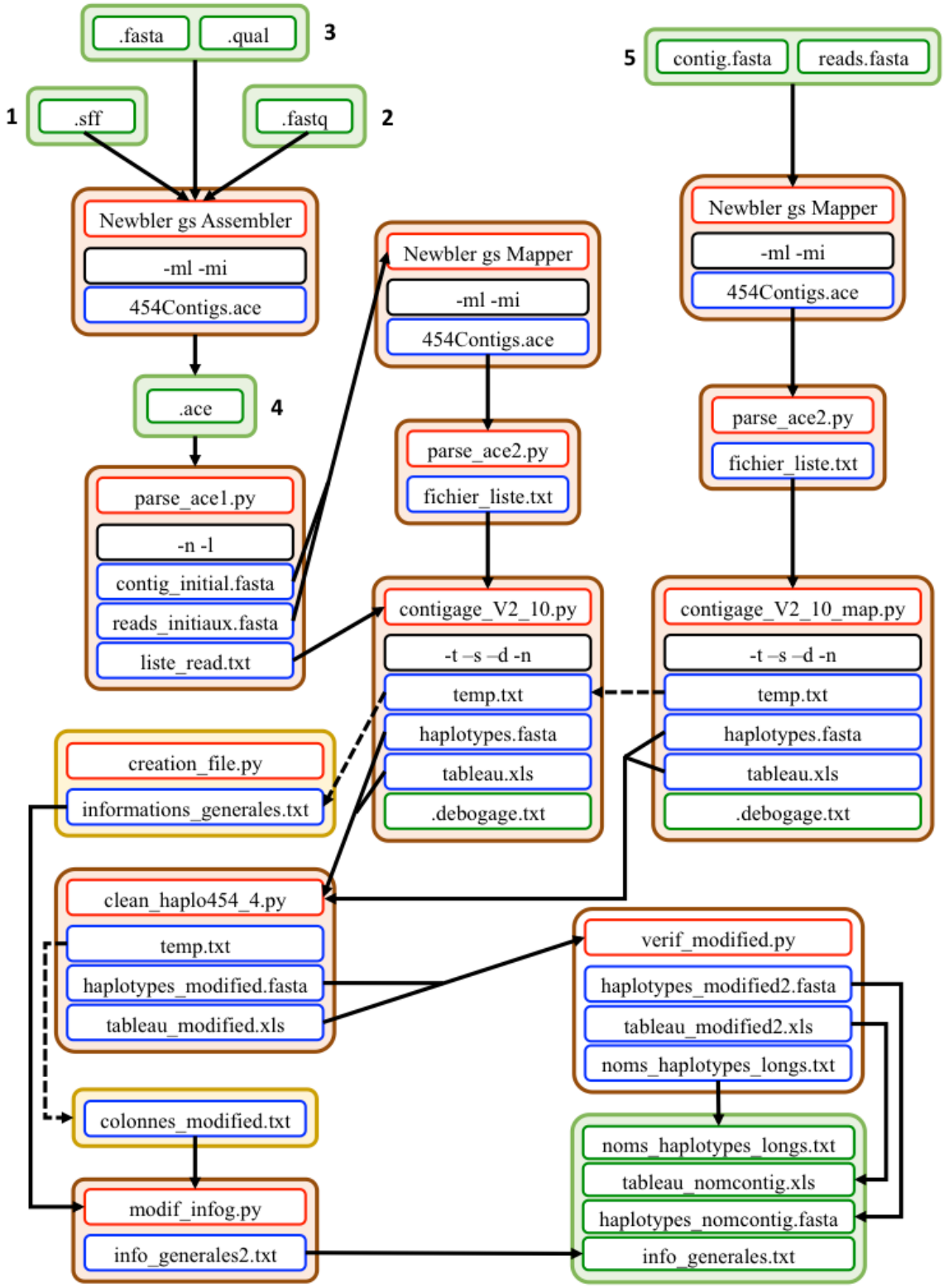


Figure 14 : Présentation du Pipeline « PyroHaplotyper ». Les différents fichiers d'entrées supportés (aux formats « .sff » (1), « .fastq » (2), « .fasta » et « .qual » (3), « .ace » (4) et « .fasta » (5)) sont représentés en vert en début de pipeline. Les programmes utilisés apparaissent en rouge sur fond marron ou sur fond jaune (pour les programmes créant respectivement des fichiers permanents et temporaires). Les paramètres réglables par l'utilisateur lors des différentes étapes du programme sont encadrés en noir. Les fichiers provisoires apparaissent en bleu et les cinq fichiers de résultats finaux récupérés par l'utilisateur apparaissent en vert. Les flèches pleines représentent un transfert de fichier en argument à un programme ainsi que les fichiers renommés. Les flèches en pointillés indiquent les données réécrites à l'intérieur d'un autre fichier.

b) **Détection de SNPs et reconstruction d'haplotypes à partir de données de séquençage Illumina (ou « lectures courtes »).**

Nous avons développé le programme « IlluHaplotyper » (en python v.2.7) afin de détecter les différentes copies dupliquées au sein de jeux de données de séquençage par synthèse de type Illumina (ou « lectures courtes à haute profondeur de séquençage »). Cette approche nécessite une première étape de mapping des données Illumina (sur des séquences de référence de l'espèce considérée) à l'aide du logiciel Bowtie 2. Les différents fichiers au format « .SAM » obtenus lors du mapping sont alors convertis au format « .PILEUP » à l'aide de la suite Samtools (Li et al. 2009) dans le but de faciliter la recherche de polymorphismes au sein des alignements. Le programme « IlluHaplotyper » a ensuite été appliqué sur deux jeux de données : 1) les alignements obtenus lors de l'assemblage des données Roche-454 (Figure 11.C et D) et 2) sur les 5 transcriptomes de référence construits (Figure 11.G et H) pour détecter les différents polymorphismes (SNPs et insertions/délétions) et reconstruire les différents haplotypes.

Détection des polymorphismes

À partir du fichier au format « .PILEUP » nous déterminons dans un premier temps l'état de caractère majoritaire (la nature du nucléotide) pour chaque position, ce qui nous permet de reconstruire le contig ; puis nous calculons la profondeur. Si cette profondeur est supérieure à 30 reads (correspondant au paramètre réglable « profondeur » ou « depth » ; option –d en ligne de commande, 30 reads par défaut), nous recherchons la présence potentielle d'un SNP (ou d'insertions/délétions en bloc). Pour éviter la détection de SNPs faux-positifs dus à des erreurs de séquençage, les nucléotides qui ne sont pas présents au minimum à 2/100 à la position donnée ne sont pas considérés (paramètre réglable « seuil » ou « thresold » ; option –t en ligne de commande, 2/100 par défaut). Un filtre permet de supprimer les erreurs d'alignement dues aux insertions/délétions absentes des fichiers

« .PILEUP » qui créent localement des SNPs faux-positifs. Pour chaque alignement traité, nous disposons de deux fichiers de sortie:

- le premier contient la séquence du contig reconstruite ainsi que la liste des SNPs (Figure 15),
- le second contient la couverture du contig.

Pour comparer les différents contigs traités, un fichier global va contenir pour chaque contig : son nom, sa taille, le nombre de reads associés au contig ainsi que le nombre de SNPs avec leurs positions respectives.

```
>contig37.ref
*ATTGTCGAAATTGTACCTTAGTGACAACAATTTCCAAGGGAGCATATGATGAGTGTTGACA
ATCAAAATTGATCA*GGCTGAGAATTGGGTCTTCACCGGTCAGACCGATCCCACCAATCGATC
TGACCGGTCCAAAGTATGAAATATCTTCAATTGGCCAATTTTAAGCTTTT*TCATGTAC*GCT
CG*TGAGTTAAT*A*AATGCATATTTCTCTC*CCCAAAGTTTATCATC*GATGTGACATTATGA
ATTTAGGGGGAATC*TGAAATCATGTCACACCAGTGACTAA
1     -     39    G     331
77    C     271   G     464
176   A     557   T     12
185   C     47    G     482
191   A     400   G     78
201   C     360   T     75
203   A     14    G     419
220   A     270   C     8
237   G     91    -     38    C     35
267   C     68    G     38
```

Figure 15 : Exemple de fichier de sortie pour un contig de *S. maritima*. Le fichier se décompose en deux parties, la première partie contient le nom de la séquence ainsi que la séquence nucléotidique correspondante. La seconde partie du fichier contient la liste des SNPs détectés. La première colonne indique la position des SNPs au sein de l'alignement. Pour chaque SNP, le type de nucléotide et leur abondance sont indiqués. Par exemple, à la position 185 du contig 37, un SNP de type C/G est présent. À cette position le nucléotide C apparaît 47 fois (8,88%) et le nucléotide G apparaît 482 fois (91,12%).

Détection des haplotypes locaux par fenêtres glissantes

Après l'obtention de la liste des SNPs, nous constituons des sous-listes contenant les sites polymorphes proches. Cette stratégie s'appuie sur le fait que la longueur maximale des reads Illumina à notre disposition est de 108 nucléotides ; deux SNPs éloignés de plus de 120 nucléotides (pour tenir compte de possibles insertions/délétions) ne seront pas contenus sur un même read et ne nous permettront pas d'assembler les différents haplotypes ; ils ne

seront donc pas inclus dans la même sous liste. Pour que ce programme soit applicable sur n'importe quel jeu de données, la longueur des reads est réglable par l'utilisateur à l'aide de l'option « trimming » (-t en ligne de commande ; réglée à 120 bp par défaut). L'étape suivante consiste à détecter les haplotypes à partir de chaque sous-liste créée correspondant à des fenêtres avec une longueur minimale de 240 bp (égal au double de la valeur du paramètre « trimming ») et qui contiennent au moins deux SNPs. Le nombre minimal de SNPs présents au sein de chaque fenêtre peut être réglé à l'aide de l'option « nombre de SNPs pour assembler les haplotypes » ou « Number of shared SNPs» (option –n en ligne de commande, réglé à 2 SNPs par défaut). La borne inférieure de chaque fenêtre correspond à la position du SNP la plus petite (le SNP le plus proche du 5') auquel on soustrait 120 nucléotides, la borne supérieure correspondant à la position de SNP la plus grande (le SNP le plus proche du 3') auquel sont ajoutés 120 nucléotides. Seul les reads contenus entièrement dans cette fenêtre sont pris en compte. Nous détectons et assemblons ensuite les différents haplotypes en utilisant une méthode similaire à celle utilisée par le logiciel « PyroHaplotyper » (Figure 16). Lors de l'assemblage des haplotypes, les reads sont comparés deux à deux.

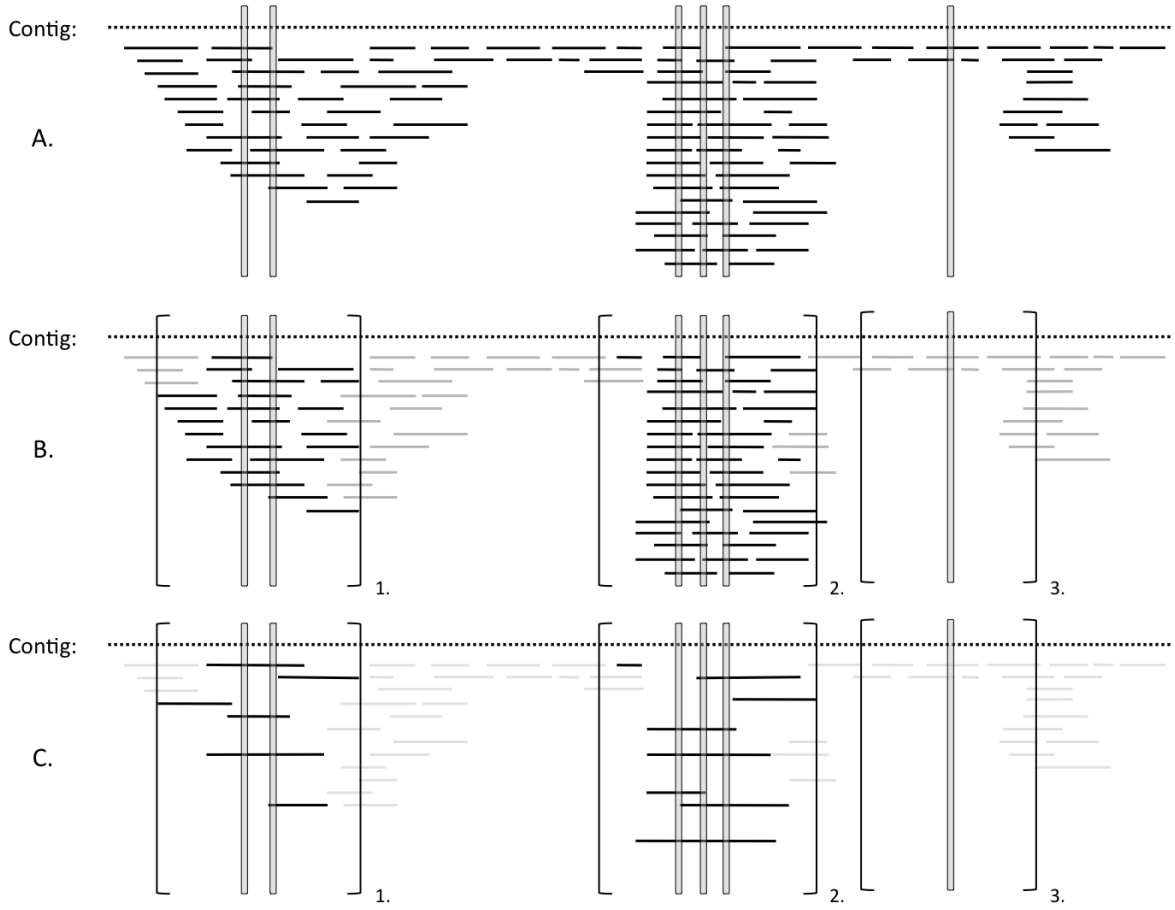


Figure 16 : Exemple de fenêtre sélectionnée pour la construction d'haplotypes. Les boites verticales représentent les polymorphismes. A. Les reads sont alignés contre une référence et les différents polymorphismes sont détectés. Seul les reads totalement inclus dans la fenêtre (et les fenêtres présentant au moins deux SNPs (fenêtre 1 et 2)) sont sélectionnées. C. Les différents haplotypes sont obtenus à partir d'une méthode similaire à celle développée pour les données Roche-454.

Les reads sont assemblés entre eux lorsqu'ils présentent les mêmes polymorphismes sur leurs régions chevauchantes et lorsqu'une autre association de reads n'est pas possible. Cette méthode a l'avantage d'éviter la création d'haplotypes chimériques. Ainsi, lorsque deux choix (ou plus) sont possibles, les reads ne sont pas assemblés entre eux, ce qui augmente le nombre d'haplotypes construits (phénomène en cascade). Pour contourner le problème, nous comptons le nombre maximal d'haplotypes par fenêtre en ne prenant pas en compte les haplotypes « similaires » (un exemple est présenté en Figure 17). Nous obtenons ainsi une liste des différents haplotypes présents. Cette liste est ensuite traitée de la manière suivante : les haplotypes dont la taille ne permet pas de couvrir l'ensemble des SNPs détectés sont supprimés. Nous supprimons également les haplotypes construits à partir d'un seul read.

Deux fichiers de sortie permettent la visualisation des données : un fichier fasta contenant la séquence et les différents haplotypes ainsi qu'un fichier texte contenant la liste des SNPs, les haplotypes et le nombre de reads composant chaque haplotype. Un fichier contenant des informations sur chaque contig traité est également rendu en sortie, pour chaque contig (ou séquence de référence), ce fichier nous renseigne sur :

- le nom du contig (ou séquence de référence), sa longueur,

- le nombre de reads alignés, le nombre total de reads utilisés pour construire les différents haplotypes,

- le nombre total de SNPs ainsi que le nombre total de SNPs utilisés pour assembler les différents haplotypes (et leur position respective),

- le nombre total d'haplotypes, le nombre moyen d'haplotypes par groupe et le nombre d'haplotypes par position (représenté sous la forme de fenêtre).

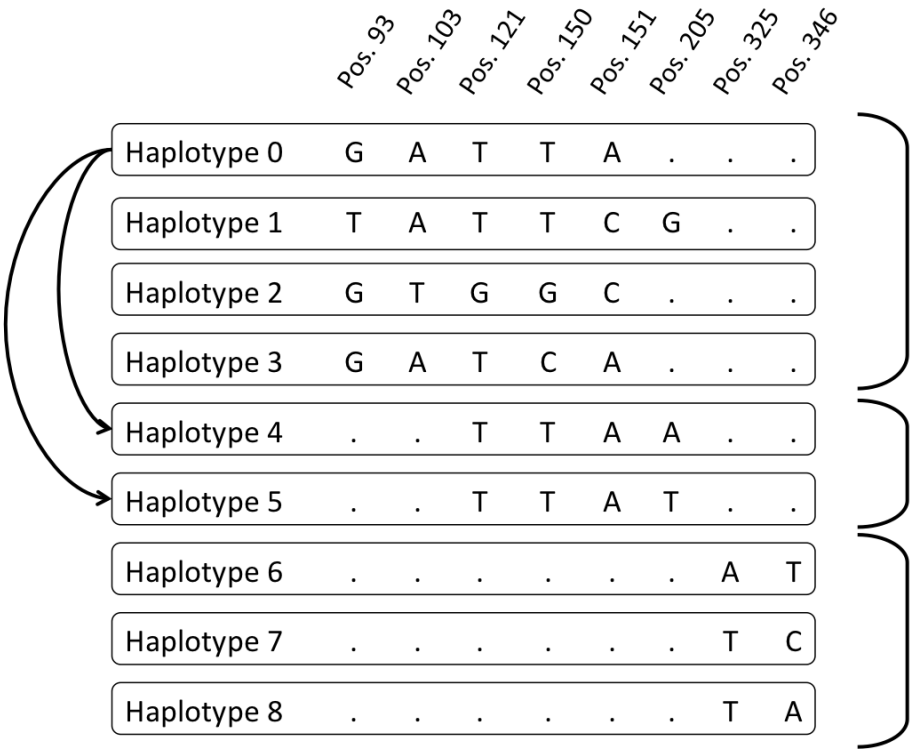


Figure 17 : Exemple du nombre maximum d'haplotypes pour une fenêtre d'un contig issu du transcriptome de *S. maritima* (Smar_2_contig665). Dans cet exemple, l'alignement présente un total de 9 haplotypes, mais seul 6 sont chevauchants. De plus, l'haplotype 4 ou 5 peut correspondre à l'haplotype 0. Au total, nous considérons un nombre maximum d'haplotypes de 6 pour ce contig puisqu'il n'est pas possible d'associer l'haplotype 4 ou 5 à l'haplotype 0.

iv) Impact des paramètres de l'outil « IlluHaplotyper » sur les jeux de données :

Pour étudier la sensibilité et les répercussions directes des outils et paramètres du logiciel « IllHaplotyper » sur les jeux de données, nous avons appliqué le programme développé sur un sous-ensemble du transcriptome de *S. maritima* en faisant varier les différents paramètres du programme. Seuls les alignements présentant au minimum un SNP et ayant un temps de construction des haplotypes par « IlluHaplotyper » inférieur à 10 secondes ont été retenus pour cette étude. Sur les 60 644 contigs issus du transcriptome de *S. maritima*, 18 861 contigs ont pu être ainsi testé. Les valeurs des différents paramètres testés sont présentées dans le Tableau 4. Ce travail a fait intervenir Delphine Giraud qui a réalisé un stage de Master 1 en Bio-Informatique et Génomique (BIG, Université de Rennes 1) au sein du laboratoire (Giraud 2015).

Tableau 4 : Valeurs des différents paramètres testés au cours de l'étude. Les * indiquent les valeurs des paramètres par défaut.

Paramètres :	Option :	Valeurs :					
Seuil	-t	2%*	5%	8%	10%	20%	25%
Profondeur	-d	10	20	30*	40		
Nb SNPs pour assembler	-n	2*	3	4			
Trimming	-s	120 pb*					

v) Création de séquences consensus Roche-454 et Illumina :

Pour chaque contig traité, nous disposons de deux séquences consensus obtenues respectivement à partir des données Roche-454 et Illumina (où les SNPs sont codés). Il est ainsi possible de concevoir la meilleure séquence consensus possible et de comparer les différents polymorphismes obtenus d'une part via l'assemblage des données Roche-454 et d'autre part grâce au mapping des jeux de données Illumina sur les contigs 454. Pour cela, nous alignons les deux séquences consensus obtenues en utilisant un algorithme dynamique similaire à celui développé par Needleman et Wunsch (1970), qui nous assure de trouver

l'alignement de score maximal. En effet, ce programme calcule l'ensemble des alignements possibles entre les deux séquences et rend en sortie l'alignement présentant le meilleur score. Les scores pour les caractères alignés sont les suivants : +2 si les nucléotides sont identiques, +1 si l'un des deux nucléotides est un SNP, -2 si l'on forme un gap et -3 si les nucléotides sont différents. Le choix de ces paramètres permet d'obtenir le meilleur alignement possible en tenant compte des polymorphismes présents dans les alignements. Une fois les deux séquences alignées, nous pouvons créer la séquence consensus de référence en corrigeant les erreurs de séquençage présentes dans la séquence 454 grâce à la séquence Illumina (correction des homopolymères ou de rares polymorphismes grâce à la profondeur assurée par le séquençage Illumina), puis répertorier et comparer les différents sites polymorphes. Deux fichiers sont créés en sortie pour chaque contig, l'un contenant la séquence obtenue grâce aux données Roche-454, la séquence obtenue en utilisant les données Illumina, et la séquence consensus résultant de l'alignement. Le second fichier de sortie contient :

- la liste des SNPs avec leurs positions respectives,

- la nature des nucléotides formant les sites polymorphes,

- l'origine du SNP (SNP présent dans les données Illumina et/ou dans les données Roche-454).

Un fichier de synthèse contenant pour chaque contig son nom, le nombre total de SNPs ainsi que l'origine et la position de chaque SNP est également créé par le programme (Figure 11.E).

vi) Assignation des copies parentales par co-alignements d'haplotypes parentaux et hybrides/allopolyploïdes:

Après avoir détecté les différentes copies (haplotypes) dupliquées au sein de chaque espèce, il est possible d'identifier chez les espèces hybrides (*S. x townsendii* et *S. x neyrautii*) et l'allopolyploïde (*S. anglica*) l'origine parentale (*S. maritima* ou *S. alterniflora*) de chaque haplotype. Pour cela nous identifions pour chaque alignement des espèces hybrides les

régions homologues présents chez les deux parents à l'aide de BLASTn. Les alignements présentant au minimum 90% d'identité et une *e-value* inférieure ou égale à 10^{-6} sont retenus. Les séquences consensus des deux parents et de l'espèce hybride sont ensuite assemblées à l'aide de Newbler (mi = 80%). Il est ensuite possible de mapper les haplotypes des trois espèces sur la nouvelle référence obtenue (mi = 40%) et de comparer les haplotypes de l'espèce hybride aux haplotypes parentaux. Pour identifier l'origine de chaque copie, l'haplotype parental présentant le pourcentage d'identité le plus élevé, la longueur de chevauchement la plus importante et le nombre de SNPs communs le plus grand est associé à l'haplotype de l'espèce hybride. Si deux haplotypes parentaux sont strictement identiques (ou si l'un des parents ne présente pas d'haplotypes ou de polymorphisme pour cette région), la copie de l'espèce hybride n'est pas assigné spécifiquement à un génome parental (Figure 18).

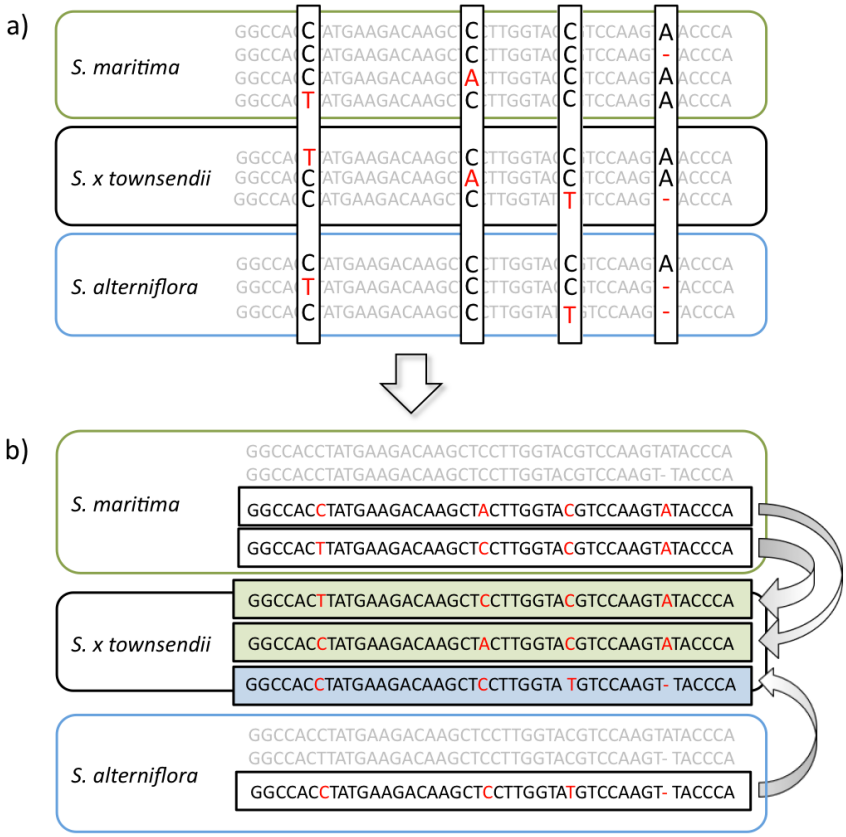


Figure 18 : Identification de l'origine parentale des copies de gènes détectées chez les espèces hybrides. a) Les haplotypes des deux parents et de l'hybride sont alignés. Les différents sites polymorphes présents au sein de l'alignement sont alors identifiés. b) En comparant les différents sites polymorphes, il est possible d'identifier l'origine parentale des différentes copies de l'espèce hybride.

vii) Calcul du ratio KA/KS et datation moléculaire des gènes dupliqués :

Après avoir détecté les différents polymorphismes et reconstruit les copies dupliquées, nous avons calculé le ratio K_A/K_S entre les haplotypes d'un même alignement au sein d'une même espèce polyploïde. Le programme développé prend en entrée un fichier au format « .FASTA » (créé par le programme de détection de SNPs et de construction d'haplotypes) ; ce fichier contient les différents haplotypes détectés ainsi que la séquence consensus. Cette dernière est traduite dans les 6 cadres de lecture et le programme retient uniquement le(s) cadre(s) de lecture présentant le moins de codons stops. Les haplotypes sont ensuite réalignés en fonction de leur position de début et leur longueur. Cette étape est nécessaire pour identifier les différentes fenêtres d'une longueur supérieure ou égale à 120 bp (soit ≥ 30 acides aminés) et présentant au moins 2 SNPs où il sera possible de calculer les taux de substitutions. Les fenêtres contenant des codons stops et/ou des insertions/délétions sont éliminées. Pour chaque fenêtre sélectionnée, nous traduisons les séquences nucléotidiques à partir du meilleur cadre de lecture. Il est ainsi possible de calculer les distances nucléotidiques et protéiques au cours de cette étape. Une fois que les haplotypes ont été traduits, nous pouvons calculer le taux de substitutions synonymes (K_S) et non-synonymes (K_A). Pour cela, nous utilisons les formules développées par Li et ses collaborateurs (1985) et la méthode de Kimura à deux paramètres qui permet de corriger les substitutions multiples et les biais de transitions (Kimura 1980).

Le taux corrigé de transitions de type *i* (A_i) et le taux corrigé de transversions de type *i* (B_i) sont donnés par les formules suivantes :

$$A_i = (1/2) \ln (1/ (1 - 2 P_i - Q_i)) – (1/4) \ln (1 – 2 Q_i)$$

$$B_i = (1/2) \ln (1/ (1 – 2 Q_i))$$

Où :

La proportion de transitions de type *i* : $P_i = S_i / L_i$

Proportion de transversions de type *i* : $Q_i = V_i / L_i$

i = 0-fold, 2-fold, 4-fold

Les valeurs de K_A et de K_S sont obtenues à l'aide des formules suivantes:

$$K_S = (L_2A_2 + L_4A_4 + L_4B_4) / (L_2/3 + L_4)$$

$$K_A = (L_0B_0 + L_2B_2 + L_0A_0) / ((2/3)\ L_2 + L_0)$$

Pour chaque fenêtre étudiée, le programme rend en sortie la longueur de chaque fenêtre, la distance nucléotidique et protéique (identité entre les acides aminés), les ratios K_A, K_S et K_A/K_S ainsi que différentes informations permettant de valider et/ou filtrer les résultats. Nous avons par ailleurs effectué une représentation graphique de la densité des K_S pour détecter les différents évènements de duplications (Blanc and Wolfe 2004). Pour estimer l'âge maximal des évènements de polyploïdisation, nous avons estimé l'âge des pics à partir de l'horloge moléculaire des Poaceae estimée à $6,5*10^{-9}$ substitutions synonymes/site/an (Gaut et al. 1996)

viii) Simulation d'espèces hybrides à partir des jeux de données des espèces parentales :

Pour comparer les résultats de datations moléculaires obtenus chez les espèces hybrides, nous avons simulé deux espèces hybrides à partir des jeux de données Illumina. Dans un premier temps nous avons simulé une espèce allododécaploïde à l'aide de 4 079 contigs, pour cela nous avons identifié les régions homologues des espèces parentales *S. maritima* et *S. alterniflora* à l'aide de BLASTn (pourcentage minimum d'identité : 90% ; *e-value* ≤ 10^{-6}). Les séquences consensus (considérées comme homéologues) ont ensuite été assemblées à l'aide de scripts développés au sein du laboratoire, puis les différents haplotypes des deux espèces parentales (*S. maritima* et *S. alterniflora*) ont été alignés (par mapping) sur ces références à l'aide du logiciel Newbler (option mi = 10%). Les différents alignements ainsi obtenus ont été réalignés avec le logiciel Mafft (option --auto ; Katoh and Toh 2010). Puis les programmes présentés dans le paragraphe précédent (Chapitre 3.7) ont

été appliqués. Dans un deuxième temps nous avons simulé une espèce hexaploïde ; pour cela nous avons sélectionné 317 contigs consensus créés lors de la première simulation sur lesquels nous avons mappé via Bowtie 2 (« score-min : G, 52, 8 ») les données transcriptomiques des deux espèces hexaploïdes parentales *S. maritima* et *S. alterniflora*. Une fois les différents reads alignés, nous avons détecté les différents SNPs et reconstruits les haplotypes à l'aide du programme « IlluHaplotyper » (paramètres par défaut) et appliqué les programmes présentés dans la Partie 7 du Chapitre 3.

ix) Origine des haplotypes construits, analyses phylogénomiques « à haut débit » :

Afin d'explorer à terme l'origine évolutive des différents haplotypes détectés et de discriminer les copies homéologues (résultant d'un événement de polyploïdisation) des copies paralogues (issues de la duplication individuelle de gène), nous avons développé une approche phylogénomique « à haut débit » adaptée aux données NGS.

Dans un premier temps, les régions homologues entre les deux espèces parentales *S. maritima* et *S. alterniflora* ont été identifiées par BLASTn (pourcentage minimum d'identité : 90% ; *e-value* ≤ 10^{-6}). Chaque région homologue a ensuite été alignée à l'aide de BLASTn sur une base de données regroupant les CDS (« Coding DNA Sequences ») de 10 espèces utilisées en outgroup : *Brachipodium dictachyon*, *Oryza sativa*, *Sorghum bicolor*, *Setaria italica*, *Zea mays*, *Musa acuminata*, *Eragrostis tef, Arabidopsis thaliana*, *Phoenix dactylifera* et *Amborella trichopoda*. Seuls les alignements présentant au minimum 80% d'identité sont retenus lors de cette étape. Les séquences consensus des deux espèces du genre *Spartina* (*S. maritima* et *S. alterniflora*) et les différentes séquences homologues des outgroups sont ensuite regroupées au sein d'un même fichier fasta. Les séquences de chaque région homologue ont alors été alignées à l'aide du logiciel Mafft (option --auto ; Katoh and Toh 2010). Pour chaque matrice alignée (correspondant à une région homologue), les haplotypes des espèces parentales sont ajoutés puis la matrice est réalignée via Mafft (option --auto). Les différentes matrices sont ensuite nettoyées à l'aide d'un script développé au sein du laboratoire qui supprime les zones présentant un mauvais alignement (en début et fin d'alignement) ainsi que les éventuelles séquences redondantes. Les

différents alignements obtenus sont ensuite filtrés et seuls les fichiers fasta présentant un pourcentage d'identité multiple supérieur ou égal à 50% (ou supérieur ou égal à 40% et contenant un nombre moyen de gap inférieur à 30%) sont convertis en fichiers au format « .PHYLIP » à l'aide du module Biopython AlignIO (http://biopython.org); les noms des séquences étant au préalable codés en un nom d'une longueur maximale de 10 caractères (propre au format « .PHYLIP »).

Une analyse phylogénomique est ensuite réalisée sur chacune des matrices à l'aide du logiciel RAxML (v.2.7.8), permettant la réalisation de phylogénies pour chaque alignement à l'aide de la méthode probabiliste de maximum de Vraisemblance (ML). La vraisemblance des clusters a été calculée en choisissant le modèle d'évolution GTR (« General Time Reversible ») avec une distribution GAMMA (Rodríguez et al. 1990). En effet, le nombre d'analyses phylogénomiques réalisées ici étant très important, ce modèle d'évolution a été sélectionné puisqu'il se présente comme le modèle le plus général, neutre et indépendant. La robustesse des nœuds a été calculée par la méthode de bootstrap (500 réplicas ; options : –n 1, -m GTRGAMMA, -T 2, -p 1005 et -#500). Pour chaque analyse nous récupérons l'arbre MR (« Majority Rule consensus tree » ; options –T2, -m GTRGAMMA, -J MR) et l'arbre MRE (« Extended Majority Rule consensus tree » ; options –T2, -m GTRGAMMA, -J MRE). Les différents arbres obtenus à l'aide du logiciel RAxML sont ensuite ré-enracinés et seul les nœuds présentant un bootstrap supérieur ou égal à 50% sont conservés. Cette étape est réalisée à l'aide du logiciel TreeCollapseCL 4 (Hodcroft 2015) et des options –t 0 et –b 0.5.

La topologie des différents arbres obtenus a ensuite été étudiée. Pour cela nous appliquons des programmes dédiés à nos jeux de données (relations phylogénétiques entre les Spartines et les outgroups choisis) et utilisant le module Phylo de la librairie Biopython (http://biopython.org) qui permet d'extraire les informations contenues dans les fichiers au format « .NEWICK » (Figure 11.I).

x) Séquence Capture :

A partir des nouveaux transcriptomes de référence construits, nous avons sélectionné un ensemble de gènes d'intérêt en faible nombre de copies correspondant à des gènes

potentiellement impliqués dans la voie du DMSP (Diméthylsulfoniopropionate) chez les Spartines (dont des gènes ayant des activités décarboxylase et amine oxydase). Ce travail est en cours de réalisation dans le cadre de la thèse d'Hélène Rousseau (Equipe MOB, ECOBIO, Université de Rennes 1) : « Evolution des génomes polyploïdes et innovations fonctionnelles : Origine de la production du DMSP au sein des Spartines (Poaceae) ». Une sélection supplémentaire de 17 gènes présents en copie unique (ou en deux copies) chez le riz et étudiés lors d'analyses phylogénétiques a été réalisée (Tableau 5). Pour l'ensemble de ces gènes d'intérêt, nous allons augmenter la profondeur de séquençage par la méthode de Séquence Capture (revue dans Grover, Salmon, and Wendel 2012) afin de détecter toutes ces copies de gènes en utilisant des sondes exoniques (Salmon et al. 2012).

Tableau 5 : Récapitulatif des 17 gènes utilisés pour la Séquence Capture, leur annotation fonctionnelle, ainsi que différentes caractéristiques (le nombre de locus, la localisation chromosomique, la longueur génomique et transcriptomique ainsi que leurs identifiants) des gènes du riz sont indiqués.

Nom du gène :	Annotation :	Nombre de copies chez le riz :	Chromosome nº :	Longueur génomique / trancriptomique (en bp) :	Identifiant :
DMC-1	Meiotic recombination protein DMC1	2	11,12	4227 / 1398	LOC_Os12g04980.1
EF-G	Elongation factor EF-G	1	4	2765 / 2306	LOC_Os04g45490.1
WAXY	Granule-bound starch synthase	1	6	4463 / 2813	LOC_Os06g04200
leafy (LFY)	Regulation of transcription, DNA-dependent	1	4	3264 / 1493	LOC_Os04g51000
GPA1	G-protein alpha subunit	1	5	3640 / 1032	LOC_Os05g26890
PHYA	Phytochrome A	1	3	7954 / 4047	LOC_Os03g51030
PHYC	Phytochrome C	1	3	5059 / 4004	LOC_Os03g54084
RPB2	DNA-directed RNA polymerase II subunit RPB2	1	3	8587 / 2373	LOC_Os03g44484.1
GIGANTEA (GI)	GIGANTEA	1	1	9334 / 4502	LOC_Os01g08700.2
AGT1	Alanine-glyoxylate aminotransferase AGT1	1	8	2756 / 1609	LOC_Os08g39300.1
AroB	3-dehydroquinate synthase	1	9	3438 / 1718	LOC_Os09g36800.1
At103	Rubrerythrin	1	1	2626 / 2011	LOC_Os01g17170.1
Bio2	Biotin synthase	1	8	3931 / 1526	LOC_Os08g42730.1
DET3	V-ATPase subunit C	1	5	4240 / 1665	LOC_Os05g51530.1
HCF136	Alpha/beta hydrolase fold	1	6	5340 / 2330	LOC_Os06g49440.1
Sqd1	NAD dependent epimerase/dehydratase	1	5	3364 / 2290	LOC_Os05g32140.1
DHAR	Dehydroascorbate reductase	1	5	3679 / 1385	LOC_Os05g02530

Chapitre 4 :

Détection de SNPs et construction d'haplotypes à partir de données Roche-454.

Chapitre 4 : Détection de SNPs et construction d'haplotypes à partir de données Roche-454.

Introduction et démarche générale

Le développement des outils NGS permet d'explorer le génome d'espèces pour lesquelles nous ne disposons pas forcément de ressources génomiques préalables (Metzker 2009). Les technologies actuelles les plus utilisées sont complémentaires. Ainsi les technologies de pyroséquençage Roche-454 (Margulies et al. 2005), qui permet d'obtenir jusqu'à un million de reads d'une longueur moyenne de 700 bp et de séquençage par synthèse (commercialisé par Illumina) se présentent comme candidates pour identifier les copies dupliquées des génomes particulièrement complexes et hautement redondants tel que les génomes d'espèces polyploïdes. Les biais inhérents au pyroséquençage (saturation des signaux dans les régions homopolymèriques, profondeur de séquençage relativement faible), peuvent être corrigés par l'utilisation de technologies générant des fragments plus courts (limitant la longueur des haplotypes détectés) mais à une plus forte profondeur de séquençage (comme le séquençage Illumina).

Plusieurs outils de détection de copies dupliquées à partir de lectures (ou « reads ») NGS ont été développés ces dernières années et appliqués sur des espèces polyploïdes (Udall et al. 2006; Flagel et al. 2008; Salmon et al. 2009; Ilut et al. 2012; Combes et al. 2013; Page, Gingle, and Udall 2013; Pfeifer et al. 2014; Tennessen et al. 2014). Néanmoins les stratégies développées ne peuvent être appliquées que sur des espèces dont les parents diploïdes sont connus. La détection des différentes copies au sein des espèces où les génomes parentaux ne sont pas disponibles nécessite ainsi le développement d'outils spécifiques. Dans ce chapitre, nous développerons un outil de détection de copies dupliquées à partir de données Roche-454.

Cet outil a tout d'abord été mis au point sur les gènes ribosomiques codant l'unité 45S. Cette famille de gène a été choisie pour son évolution particulièrement dynamique chez les espèces polyploïdes et son utilisation fréquente en phylogénie moléculaire pour reconstruire l'histoire des espèces (Álvarez and Wendel 2003). Cette étude fait l'objet d'un

article (en presse) dans la revue Genes|Genomes|Genetics. Nous avons exploré cette région particulièrement redondante au sein des génomes et validé notre outil de détection de SNPs et de reconstruction d'haplotypes développé pour des alignements de fragments longs tel que les technologies Roche-454 (Margulies et al. 2005), Sanger (Sanger and Coulson 1975), Nanopore (Wang, Yang, and Wang 2015) ou SMRT (Eid et al. 2009). Les profondeurs relatives des polymorphismes et des haplotypes ont également été validées à l'aide d'un jeu de données de séquençage Illumina du génome de *S. maritima*.

La partie B de ce chapitre présente dans un premier temps, une application de ce programme sur un pool de gènes d'intérêt écologique analysés à partir des premiers transcriptomes de référence des espèces hexaploïdes de Spartines construit par Ferreira de Carvalho et al. (2012). La seconde partie se concentre sur l'application des outils développés sur les assemblages de données Roche-454 de 5 espèces de Spartines (*S. maritima, S. alterniflora, S. x townsendii, S. x neyrautii* et *S. anglica*) représentant jusqu'à 19 380 contigs.

Cet outil a également été appliqué pour analyser des jeux de données de type amplicon chez des espèces animales : des Nématodes à kystes (en collaboration avec Cécile Gracianne et Eric Petit de l'UMR ESE, INRA, Rennes) et le Gorille (en collaboration avec Alice Baudouin et Pascaline Le Gouard de l'UMR-CNRS Ecobio, Université de Rennes 1) afin d'identifier les différentes copies et allèles pour un ensemble de gènes d'intérêts de chaque population.

Ce programme est en cours d'intégration sur la plateforme Galaxy de GenOuest (Namour 2015) pour permettre sa diffusion à la communauté scientifique.

PARTIE A : Etude de l'ADN ribosomique de *S. maritima* et validation du programme de construction d'haplotypes.

A travers cet article, nous avons appliqué le programme de détection de SNPs et de construction d'haplotypes que nous avons appelé « PyroHaplotyper », sur un contig obtenu à partir d'un assemblage *de novo* de reads Roche-454 codant pour l'unité 45S de l'espèce hexaploïde *Spartina maritima*. La région ITS (comprenant le gène 5.8S flanqué par les espaceurs ITS1 et ITS2) a fourni un très bon signal phylogénétique pour élucider les relations

entre les différentes espèces de Spartines (Baumel et al. 2002) ; néanmoins l'unité 45S n'a encore jamais été analysée dans son ensemble. De plus, le degré d'homogénéisation des copies dans le génome des espèces parentales hexaploïdes n'est pas connu.

L'assemblage du contig codant pour l'unité 45S de l'ADN ribosomique de *S. maritima* a été obtenu à l'aide d'une approche de clustering (Novák, Neumann, and Macas 2010) dans le cadre d'une collaboration avec le laboratoire de Jiri Macas (Institute of Plant Molecular Biology, České Budějovice, République Tchèque). Les reads génomiques de *S. maritima* ont ensuite été alignés (par mapping) par le logiciel Newbler sur cette première référence afin d'optimiser le nombre de reads alignés. L'annotation des différentes régions codantes a été réalisée par recherche d'homologie à l'aide des gènes 45S du riz et du maïs. Le contig obtenu (8 456 bp) contient l'ensemble des régions codantes (18S, 5.8S et 25S), les deux espaceurs internes transcrits (ITS-1 et ITS-2), l'espaceur externe transcrit (5'-ETS) localisé en amont du 18S et une partie de l'IGS (espaceur intergénique). Les régions codantes et non codantes de l'unité 45S de *S. maritima* ont été comparées avec les unités 45S du riz et du maïs ; les identités de séquences, les pourcentages en GC et les régions variables entre les domaines du 25S sont discutés.

Pour valider les polymorphismes utilisés pendant l'étape de reconstruction des haplotypes ainsi que les différents haplotypes détectés, certaines régions ont été clonées et re-séquencées à l'aide de séquençage Sanger et d'amorces spécifiques. Nous avons également aligné 1 263 153 de reads pairés issus de la technologie Illumina sur la séquence de référence de l'unité 45S à l'aide du logiciel Bowtie 2 dans le but de valider les polymorphismes détectés. Vingt neuf sites polymorphes ainsi que 11 haplotypes ont ainsi été validés.

L'application de ce programme a permis de mettre en évidence l'homogénéité intragénomique des régions codantes et des espaceurs internes transcrits (ITS). Une importante variabilité intra-génomique a été détectée dans les régions contenant l'espaceur intergénique et l'espaceur externe transcrit. Ces résultats sont en accord avec les variations inter-individuelles ou inter-spécifiques présentées dans la littérature. L'analyse cytogénétique à l'aide de la méthode FISH (Fluorescent *In Situ* Hybridization, réalisée en collaboration avec Olivier Coriton de la Plateforme de Cytogénétique moléculaire, UMR

IGEPP, Centre INRA de Rennes) a mis en évidence la présence d'une paire de signaux d'ADN ribosomique au sein du génome de *S. maritima*. Au sein de ce génome hexaploïde où trois copies homéologues dupliquées sont attendues, nous observons donc une perte des loci homéologues.

Ce travail nous a permis de valider le programme de détection de polymorphismes et d'haplotypes. Ce logiciel a été appliqué dans un contexte particulier de polyploïdie, où aucun génome de référence diploïde n'est disponible pour faciliter la détection des copies héritées des parents. Il peut ainsi être appliqué sur n'importe quelle espèce polyploïde, présentant un haut niveau de ploïdie mais également sur des espèces diploïdes pour détecter les différents allèles au sein de données de type amplicon par exemple. Le programme développé représente une ressource bio-informatique particulièrement importante pour détecter les copies homéologues dans le contexte des Spartines qui présentent des niveaux de ploïdie variables (4x, 6x, 7x, 9x et 12x), et permettra ainsi de mieux comprendre leur origine et leur évolution.

Haplotype Detection from Next-Generation Sequencing in High-Ploidy-Level Species: 45S rDNA Gene Copies in the Hexaploid *Spartina maritima*

Julien Boutte,* Benoît Aliaga,* Oscar Lima,* Julie Ferreira de Carvalho,* Abdelkader Ainouche,* Jiri Macas,† Mathieu Rousseau-Gueutin,*,‡ Olivier Coriton,‡ Malika Ainouche,* and Armel Salmon*,[1]

*UMR CNRS 6553 Ecobio, OSUR (Observatoire des Sciences de l'Univers de Rennes), University of Rennes 1, Bât 14A Campus Scientifique de Beaulieu, 35 042 Rennes Cedex, France, †Biology Centre ASCR, Institute of Plant Molecular Biology, Branišovská 31, České Budějovice, CZ-37005, Czech Republic, and ‡UMR Institut de Génétique, Environnement et Protection des Plantes, Institut National de la Recherche Agronomique, BP35327, 35653 Le Rheu Cedex, France

ABSTRACT Gene and whole-genome duplications are widespread in plant nuclear genomes, resulting in sequence heterogeneity. Identification of duplicated genes may be particularly challenging in highly redundant genomes, especially when there are no diploid parents as a reference. Here, we developed a pipeline to detect the different copies in the ribosomal RNA gene family in the hexaploid grass *Spartina maritima* from next-generation sequencing (Roche-454) reads. The heterogeneity of the different domains of the highly repeated 45S unit was explored by identifying single nucleotide polymorphisms (SNPs) and assembling reads based on shared polymorphisms. SNPs were validated using comparisons with Illumina sequence data sets and by cloning and Sanger (re)sequencing. Using this approach, 29 validated polymorphisms and 11 validated haplotypes were reported (out of 34 and 20, respectively, that were initially predicted by our program). The rDNA domains of *S. maritima* have similar lengths as those found in other Poaceae, apart from the 5′-ETS, which is approximately two-times longer in *S. maritima*. Sequence homogeneity was encountered in coding regions and both internal transcribed spacers (ITS), whereas high intragenomic variability was detected in the intergenic spacer (IGS) and the external transcribed spacer (ETS). Molecular cytogenetic analysis by fluorescent *in situ* hybridization (FISH) revealed the presence of one pair of 45S rDNA signals on the chromosomes of *S. maritima* instead of three expected pairs for a hexaploid genome, indicating loss of duplicated homeologous loci through the diploidization process. The procedure developed here may be used at any ploidy level and using different sequencing technologies.

KEYWORDS
poaceae
duplication
paralogy
polyploidy
bioinformatics

Gene and genome duplications play an important role in the diversification of eukaryotic functions and the formation of new species (Ohno 1970; Wendel 2000). These common phenomena contribute to sequence heterogeneity at homologous loci in all plant (eukaryotic) genomes. Detection of the duplicated copies is essential to investigate species history and the evolutionary dynamics of genes. In this work, we focus on a multigene family widely used in evolutionary genetics, coding for ribosomal RNA (45S rDNA) in the context of a highly duplicated genome, the hexaploid *Spartina maritima* (Curtis) Fern. (Poaceae, Chloridoideae).

Polyploidy or whole genome duplication is a process that has played a major role in the evolution and the adaptation of many eukaryotes in both animals (Mable 2004; Van de Peer *et al.* 2009) and plants (Jiao *et al.* 2011; Levy and Feldman 2002; Soltis *et al.* 2009). Whole genome duplication is most often due to the formation of unreduced gametes during meiosis in the same species (autopolyploidy) or in an interspecific hybrid (allopolyploidy) (Ramsey and Schemske 1998). Autopolyploid species contain several homologous genomes, whereas allopolyploid species contain more or less divergent duplicated homeologous genomes.

doi: 10.1534/g3.115.023242
Manuscript received September 30, 2015; accepted for publication October 23, 2015; published Early Online November 3, 2015.

Supporting information is available online at www.g3journal.org/lookup/suppl/doi:10.1534/g3.115.023242/-/DC1
[1]Corresponding author: UMR CNRS 6553 Ecobio, University of Rennes 1, France. E-mail: armel.salmon@univ-rennes1.fr

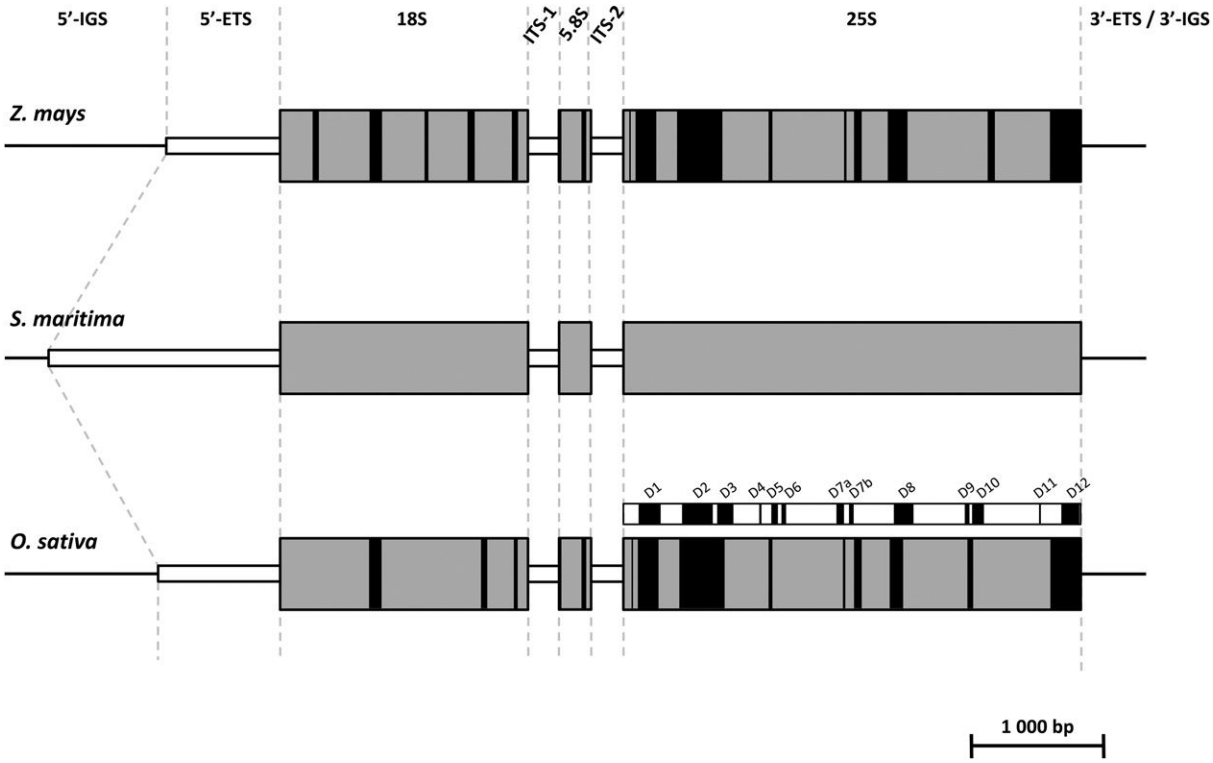


Figure 1 Schematic representation of the 45S ribosomal DNA in *Z. mays*, *S. maritima*, and *O. sativa*. The lines correspond to the intergenic spacers, the gray boxes correspond to the coding regions (18S, 5.8S, and 25S), and the white boxes correspond to the external and internal transcribed spacers (ETS, ITS-1, and ITS-2). Black boxes represent the regions presenting less than 90% identity between either *Z. mays* and *S. maritima* or *O. sativa* and *S. maritima*. The rice expansion segments (D1 to D12, represented in black) detected by Hancock and Dover (1988) and Kuzoff *et al.* (1998) are indicated above the rice 25S. The figure is drawn to scale.

The grass genus *Spartina* Schreb. belongs to the subfamily Chloridoideae (a poorly investigated group in genomics), where it forms a monophyletic lineage embedded in the paraphyletic *Sporobolus* genus, which recently led some authors to consider taxonomical inclusion of *Spartina* in genus *Sporobolus* (Peterson *et al.* 2014). The *Spartina* clade is particularly affected by recurrent polyploidization and/or interspecific hybridization events (Ainouche *et al.* 2008; Strong and Ayres 2013). It contains 13–15 species ranging from tetraploid (2n = 4x = 40) to dodecaploid (2n = 12x = 120, 122, 124) levels, with a basic chromosomes number x = 10. So far, no diploid *Spartina* species has been identified (Ainouche *et al.* 2012). The different *Spartina* species have evolved into two major lineages (Baumel *et al.* 2002): a tetraploid clade mainly consisting of American native species and a hexaploid clade, including New World and Old World species. Within this hexaploid clade, *Spartina maritima*, native to European Atlantic coasts and *Spartina alterniflora* Loisel., native to American east coasts, have naturally hybridized twice in the early 19th century following introductions of the American species in Europe. Hybridization between these species led to the formation of two sterile F1 hybrids, *Spartina x townsendii* in the Bay of Southampton (England) (Groves and Groves 1880) and *Spartina x neyrautii* in the Basque region (France) (Foucaud 1897). Genome duplication of *S. x townsendii* gave rise to the highly fertile and invasive allopolyploid *Spartina anglica* C. E. Hubbard (2n = 12x = 120,122,124) (Marchant 1968). *S. anglica* colonized Western Europe as well as various regions (*e.g.*, Australia and China). This expansion has many ecological effects, with this species playing an important role in the salt marsh sediment dynamics. *S. anglica* is able to colonize mudflats and to accelerate sediment accretion, thus altering the characteristics of the colonized habitats (referred to as an "ecosystem engineer" species) (Ainouche *et al.* 2008 and references therein). Genome evolution analysis of *S. anglica* compared to its parents (*S. maritima* and *S. alterniflora*) offers a special opportunity to understand the genomic and adaptive mechanisms associated with the formation of a new species in the wild (Ainouche *et al.* 2012). With this perspective, knowledge of the parental genomes is essential.

Our study focused on the European parental species *S. maritima* (2n = 60), which has a genome size that is estimated to be 2C = 3.8 pg or ~3700 MB (Fortune *et al.* 2008). This hexaploid species, probably of hybrid origin, is expected to contain three duplicated homeologous genomes (Fortune *et al.* 2007) that may have diverged in the past 10 mya (Rousseau-Gueutin *et al.* 2015). Sequence heterogeneity is then expected as a result of successive whole-genome duplications. This may be more complicated in gene families where both paralogs and homeologs may be encountered, such as the 45S ribosomal gene family, which is known for its highly dynamic evolution, most particularly in polyploids (see below). Because these genes are commonly used in phylogenetic studies, distinguishing paralogs and homeologs are critical.

The 45S rDNA unit is composed of a high number of transcription units (TU) per genome (>500 in plants) arranged in tandem repeats on one or several loci (Rogers and Bendich 1987; Prokopowich *et al.* 2003). Each unit contains three coding regions (18S, 5.8S, and 25S/26S) separated by two internal transcribed spacers (ITS-1 and ITS-2). The 18S and 25S coding regions are flanked by two external transcribed spacers or ETS (Figure 1) (Schaal and Learn 1988; Poczai and Hyvönen 2010). The 5′-ETS is subdivided into three regions: the ETS region I, which contains the TATA box corresponding to the transcription initial site

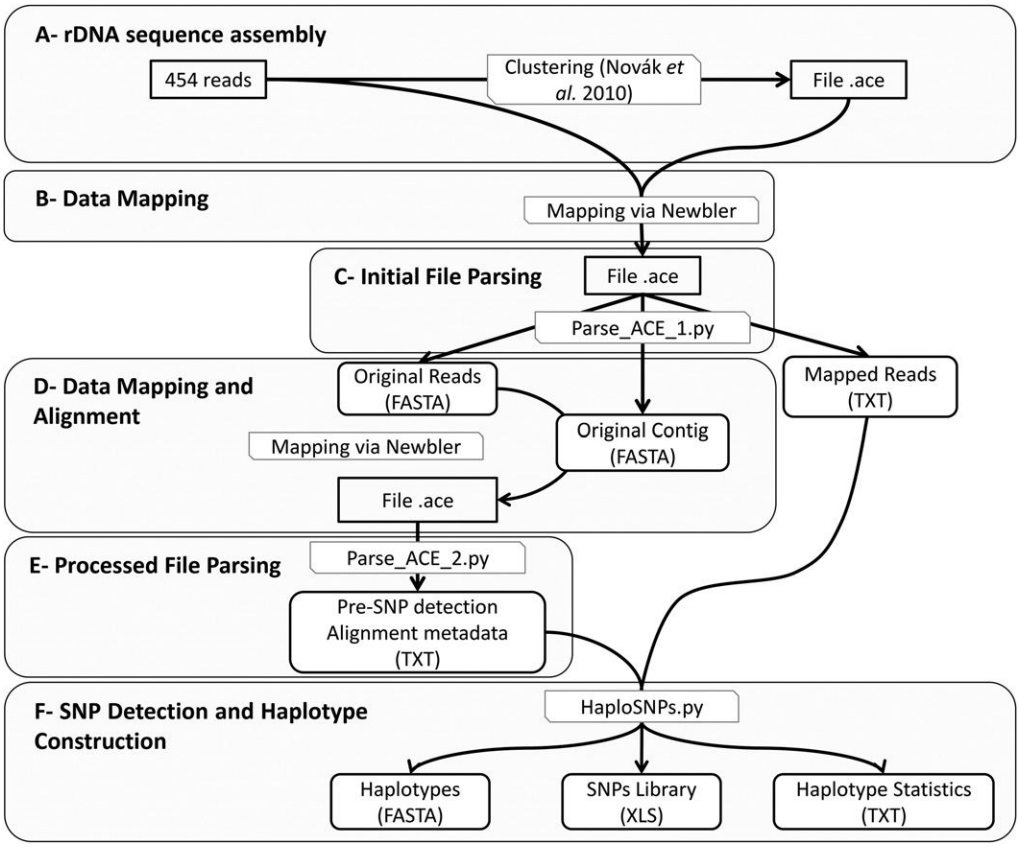


Figure 2 Overall workflow for haplotype construction using Roche-454 data.

(TIS); the ETS region II, which includes several and highly variable subrepeats; and the ETS region III, which is adjacent to the 18S coding region (Volkov *et al.* 1996). Each TU is separated from the other by an intergenic spacer (IGS) that contains the two ETS bordering the nontranscribed spacer region (NTS) composed of several repeats of 80 bp. The number of repeats and the length of each vary considerably among species (Poczai and Hyvönen 2010). The different repeats are known to undergo a process of homogenization by gene conversion leading to the observation of "concerted evolution" of these ribosomal genes (Nei and Rooney 2005). This homogenization limits the number and divergence of paralogous copies in this gene family, which has promoted its use in molecular phylogenies. The differential rates of variation along the transcriptional unit (the coding regions are more conserved than the spacers) allow phylogenetic inferences at different levels of the taxonomic hierarchy (Alvarez and Wendel 2003). In allopolyploid species, it was shown that the process of concerted evolution also affects homeologs, resulting in a preferential retention of one of the parental repeat copies, as demonstrated in polyploid cottons (Wendel *et al.* 1995) or tobacco (Kovarik *et al.* 2008). This process can occur rapidly after allopolyploid speciation, as evidenced in natural populations of allotetraploid *Tragopogon* formed in the past 80 years (Koh *et al.* 2010). In *Spartina*, the ITS region (including the 5.8S gene flanked by the ITS-1 and ITS-2) has been used to elucidate phylogenetic relationships among the different species (Baumel *et al.* 2002; Fortune *et al.* 2007), but studies on the evolutionary dynamics of the paralogous and homeologous copies are lacking in this system. The recently formed *Spartina* hybrid (*S. x townsendii*) and its allododecaploid derivative (*S. anglica*) exhibit parental additivity of the ITS regions inherited from their hexaploid parents (Baumel *et al.* 2001), and recent investigations in natural populations suggest that interlocus homogenization and/or homeolog loss may also occur in this region (D. Huska, I. J. Leitch, J. Ferreira de Carvalho, A. R. Leitch, A. Salmon, M. Ainouche, and A. Kovarik, unpublished data).

Next-generation sequencing technologies (*e.g.*, pyrosequencing or sequencing by synthesis) offer powerful tools to generate suitable sequence read depth especially in nonmodel species where genomic resources were previously lacking (El-Metwally *et al.* 2014; Metzker 2009). Detection of different copies expected in these genomes may be challenging, because individual gene duplications (resulting in paralogs in both diploid and polyploid genomes) usually add an additional layer of complexity (Ainouche *et al.* 2012). Detection and analysis of homeologs is central to polyploidy research, and several studies have focused on the detection of different copies in polyploid genomes such as in *Glycine* (Ilut *et al.* 2012), *Gossypium* (Flagel *et al.* 2008; Salmon *et al.* 2010), *Coffea* (Combes *et al.* 2011), or *Triticum* (Akhunova *et al.* 2010) species. In these systems, the diploid parental (or related) representatives are known and can be used to identify duplicated homeologs in the allotetraploids. The strategy developed in these studies is to assemble NGS datasets using parameters adapted to optimize recovery of paralogous and homeologous copies. Contigs obtained are then compared with diploid parental genomes using species-specific polymorphic sites.

The detection of different copies in polyploid species where the parents are not identified or extinct still requires the development of adapted tools. The goal of this study is to develop a bioinformatic pipeline to detect the different copies within a set of NGS reads from a highly polyploid species without any reference parental diploid genome. We developed this pipeline to explore the heterogeneity of the coding and noncoding regions of the highly repeated 45S unit of the hexaploid *Spartina maritima*.

Table 1 Identification of the different rDNA regions in *S. maritima* after comparisons with *Oryza sativa* and *Zea mays*

Species		5′-ETS	18S	ITS-1	5.8S	ITS-2	25S
S. maritima	Length (in bp)	1754	1812	226	167	206	3391
	start-end region	392–2145	2146–3957	3958–4183	4184–4350	4351–4556	4557–7947
	GC%	50.9	50	53.5	52.7	49.5	55.4
O. sativa	Length (in bp)	966	1812	198	167	215	3377
	GC%	73.0	51.3	72.7	58.1	79.1	59.4
	GenBank IDs	—[a]	X00755.1	AF169230	AF169230	AF169230	M11585.1
	Blastn results (Identity, e-value)	—	97%, 0.0	—	94%, 2e-69	—	93%, 0.0
Z. mays	Length (in bp)	834	1809	213	164	220	3385
	GC%	70.0	51.0	70.4	56.7	73.2	58.7
	GenBank IDs	X03989	NR_036655.1	AF019817	AF019817	AF019817	NR_028022.2
	Blastn results (Identity, e-value)	—	96%, 0.0	—	94%, 2e-68	—	93%, 0.0

The length, position, and GC content of each domain are presented. Genbank accession numbers of the *O. sativa* and *Z. mays* sequences used, as well as the percentage of identity and e-value between these and the *S. maritima* sequences, are mentioned.
[a] 5′-ETS was detected using *O. sativa* genome (v. 204).

The strategy was to: (i) identify single nucleotide polymorphisms (SNPs) and indels among reads from the hexaploid genome and (ii) assemble reads based on shared polymorphisms (after removing putative sequencing errors) to distinguish the different copies from reads. Polymorphisms were validated using comparisons with Illumina sequence data sets and by cloning and Sanger (re)-sequencing. The method presented here can be used to identify duplicated copies from any sequenced regions of polyploid or diploid genomes.

MATERIALS AND METHODS

Plant material sequencing

Samples from *Spartina maritima* were collected at the Etel river estuary (Morbihan, France). Plants were transplanted and maintained in controlled conditions in the greenhouse (University of Rennes 1, France). Total genomic DNA was isolated from fresh leaf tissue with the Nucleospin Plant II (Macherey-Nagel) extraction kit following the manufacturer's instructions. One run of *S. maritima* DNA was sequenced at the Functional and Environmental Genomics platform (Biogenouest, OSUR Rennes, France) using a 454 GS FLX Titanium pyrosequencer (Life Sciences, Roche) that generated 999,229 reads with an average length of 377.0 bp. Illumina genomic reads (172,528,550 reads of 100 bp length; 500-bp paired ends) from *S. maritima* (from the same population) were also used for mapping and validation or correction of the detected SNPs. Genome coverage and 45S rDNA copy number estimates are presented in Supporting Information, Table S1 for both Roche-454 and Illumina Whole Genome Shotgun sequencing.

Assembly and annotation of the 45S rDNA region

We first analyzed Roche-454 genomic reads of *S. maritima* (Figure 2A) using the computational pipeline developed by Novák *et al.* (2010). This pipeline uses graph-based clustering of sequence reads sharing mutual sequence similarities (minimum percent identity: 90%; minimum overlap: 55% of the shorter sequence length) to detect groups of frequently overlapping reads representing genomic repeats. In addition, it performs contig assembly within identified clusters and provides information aiding in repeat annotation. The cluster representing *S. maritima* 45S rDNA was identified by similarity searches against various available 45S rDNAs of Angiosperms, and an 8456 bp-long contig was retrieved from the assembled reads. Roche-454 genomic reads were then mapped on this contig to increase the number of reads. This first mapping procedure (Figure 2B) was performed using the GSMapper tools from Newbler (ml = 100 bp; mi = 95%). Annotation of the different 45S domains was performed by aligning the *S. maritima* contig with *Oryza sativa* and *Zea mays* 45S genes available on the NCBI database (GenBank IDs: X00755.1, AF169230, M11585.1 and X03989, NR_036655.1, AF019817, NR_028022.2, respectively) using BLASTn with default parameters for the alignment algorithm Megablast (Table 1) (Altschul *et al.* 1997). The *Oryza sativa* genome was downloaded from www.phytozome.net (v. 204) to locate the 5′-ETS region. To identify this region, we followed the approach of Barker *et al.* (1988), who identified the repeat unit of the rDNA intergenic region and the transcription initial site.

SNPs detection and sequencing error removal

To detect SNPs, we developed a pipeline allowing the parsing of ".ACE" alignment files (Figure 2C) using the Ace.py program from biopython (http://biopython.org/) and custom python script for editing homopolymer-driven false-positive SNPs. This step produced three files containing original reads, the original contig, and the list of the original reads (Figure 2D). This pipeline uses a mapping step with GSMapper (v.2.6, Roche; default parameters: ml = 40 bp; mi = 90%) to enhance alignments and ease SNP detection. The newly generated ".ACE" file is parsed to extract the aligned reads (Figure 2E). To remove false-positive SNPs caused by sequencing errors (within reads, and at the first or last position), the first five and the last five nucleotides of each read, as well as nucleotides with frequency lower than 20% were not considered. SNPs present at least in one-fifth of aligned reads were then considered true positives. SNPs present in homopolymeric region (more than four identical nucleotides) were not considered. The nature of the detected polymorphisms and their frequency are stored as output results (Figure 2F). Visual checking of the detected and not considered false positives was done using the Tablet software (Milne *et al.* 2009) for ".ACE" alignment files and with the Jalview software (Waterhouse *et al.* 2009) for ".FASTA" files.

Haplotype detection

Haplotypes were detected after the SNP detection by identifying and assembling reads sharing the same SNPs. This assembly was performed by comparing each read to all the others and by screening all the SNPs present in overlapping regions. If all the SNPs are identical in this region, the considered reads are assembled to create a haplotype with a maximum size. A unique haplotype is then a sequence characterized by multiple polymorphic sites corresponding to the same and unique

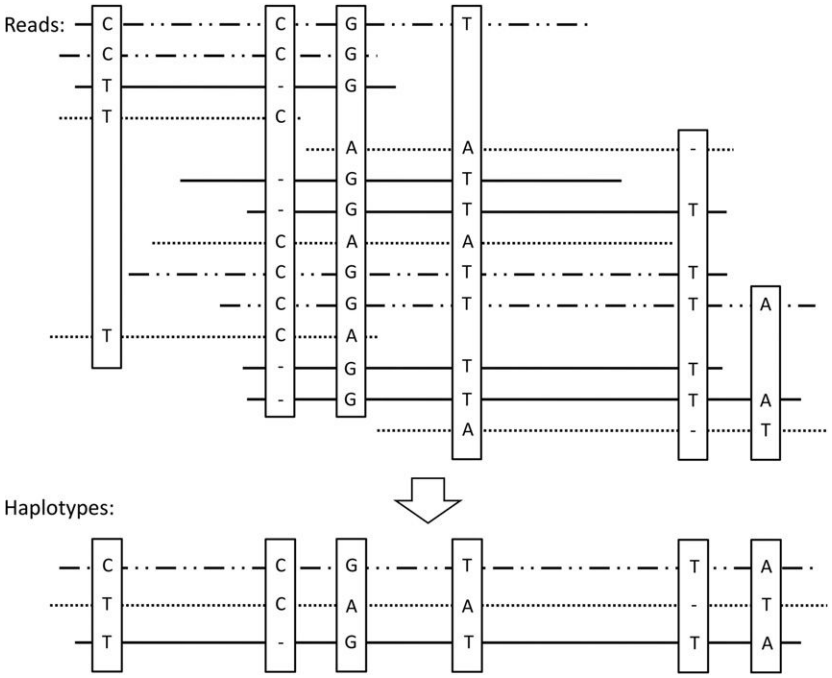


Figure 3 Representation of a read assembly. After SNP detection and false-positive correction, all reads displaying the same SNP (represented with the same line) are assembled to create a haplotype that corresponds to a consensus sequence of maximal size.

allele (Figure 3). However, for each alignment, it was necessary to define different windows corresponding to local alignments for counting haplotypes. The windows exhibiting at least one SNP were selected to count the number of locally aligned haplotypes. Two output files allow the dataset visualization: (1) a ".FASTA" file that contains the consensus sequence with the different aligned haplotypes and (2) a tabulated file describing SNPs positions and their nature. A third output file is created during these processes and contains information about the assembly, the name of the contig, the initial and final lengths of the contig obtained after the correction of sequencing errors, the number of reads used to assemble the contig, and the different haplotypes. This file contains the number of SNPs and their position, the number of haplotypes, and their coverage (mean, SD, and C.I. of 95%). The last information contained in this file is the number of haplotypes for each window.

The SNP and haplotype detection programs described above were incorporated into a single pipeline that includes four adjustable parameters: (i) the read trimming length corresponding to the number of nucleotides at the beginning and the end of each read not to consider in the SNP detection step; (ii) the minimum read depth corresponding to the number of reads required to detect SNP from; (iii) the SNP detection threshold corresponding to the minimal proportion of reads displaying a putative polymorphism; and (iv) the minimal number of shared SNPs to assemble two reads into a haplotype. The minimum depth of the analyzed rDNA contig was 61. Consequently, the minimum read depth parameter was fixed to 60. The other default parameters used in our pipeline were as follows: read trimming length = 5; SNP detection threshold = 20; number of shared SNPs = 1.

Cloning and sequencing of polymorphic 45S rDNA regions

For validating the SNP and haplotype detection pipeline we developed, a subset of the detected SNPs and haplotypes was targeted for molecular cloning and (re)-sequencing (Sanger method). We have designed four pairs of primers using Primer 3, v. 1.1.4 (Rozen and Skaletsky 2000) (Table 2) flanking variable regions detected with pyrosequencing (5′- IGS/ETS, ITS-1, 25S and 3′- ETS/IGS). All amplifications were carried out in a 50 μL reaction mixture containing 45 μl of Platinum PCR SuperMix High Fidelity (Invitrogen), 2 μl of each 5 μM primer, and 1 μl of DNA (50 ng). The PCR amplification conditions were as follows: 2 min of DNA denaturation at 94° followed by 30 cycles of 30 sec at 94°, 20 sec at 58°, and 30 sec at 68° for each cycle, followed by 8 min final extension at 68°. PCR products were purified with the NucleoSpin Gel and PCR Clean-up (Macherey-Nagel). Cloning reaction was performed with TOPO TA Cloning Kits for sequencing (Invitrogen) using 2 μl of PCR purified products, 1 μl of salt solution (1.2 M of NaCl, 0.06 M of $MgCl_2$), 2 μl of water, and 1 μl of plasmid vector pCR4Blunt-TOPO (Invitrogen). The reactions were incubated for 5 min at room temperature. Transformations were realized using electro-competent *Escherichia coli* DH5α and Gene Pulser Xcell Electroporation System (BioRad); 18 μl of water was added to 6 μl of TOPO cloning reaction. Transformations were done in 0.2 cm cuvettes with 2 μl of DNA added to 40 μl of ElectroMax DH5α Competent Cells (Invitrogen) incubated 1 min on ice and pulsed at 2.5 kV, 25 μF, and 200 Ω. After the addition of SOC medium (1 ml), the cell suspension were incubated at 37° for 1 hr; 25–100 μL of each transformation culture was plated onto LB Agar with ampicillin (100 mg/ml) and incubated overnight in liquid medium at 37° (LB Broth and 100 mg/ml of ampicillin). Plasmids were purified with the Pure Yield Plasmid Miniprep System (Promega) and sequenced using T7 and T3 primers to sequence the samples (16–24 clones for each region) with 3730XL DNA sequencer technology (Sanger method) from both ends. For each region, the different sequences were aligned with MAFFT (v 6.864b) (Katoh and Toh 2010). All sequences were cleaned (plasmid deletion) and the sequencing errors were visually corrected using chromatographs.

SNP validation and haplotype estimation with Illumina data

To compare pyrosequencing and cloning SNPs to Illumina SNPs, we mapped 1,263,153 Illumina paired reads with an average read depth of 7886 (SE = 273.84) on the rDNA sequence consensus of *S. maritima* using Bowtie2 (score-min: G, 52, 8) (Langmead and Salzberg 2012).

Table 2 Primers designed to amplify and sequence the 45S ribosomal DNA in *S. maritima*

Primer	Sequence	Primer pair	Amplicon Size	Region
ASribo_1_FP1	ACACGACTGGGTTTAGTCCG	FP1/RP1	579	5′- IGS/ETS
ASribo_1_RP1	AGGCCAGGTTTAGTCCGTTT			
ASribo_2_FP1	AAACGGACTAAACCTGGCCT	FP1/RP2	720	5′-ETS
ASribo_2_RP2	CTATTTTCAGAGGGGGAGGG			
ASribo_5_FP1	TGTCGTGACCCAAACAAAAA	FP1/RP2	725	ITS-1/5.8S/ITS-2
ASribo_5_RP2	CGATTCTCAAGCTGGGCTAC			
ASribo_6_FP1	AGACATTGTCAGGTGGGGAG	FP1/RP2	752	25S
ASribo_6_RP2	AAAGGCCACTCTGCCACTTA			

These "score-min" parameters were included in the minimum score function $f(x)= 52 + 8 * ln(x)$, where x corresponds to the read length. This function corresponds to a minimum of 87.06–90.30% of identity between the mapped reads and the reference sequence for reads with a length of 80–120 bp. The ".SAM" file created by Bowtie2 was converted in a ".PILEUP" format using the Samtools software suite (Li *et al.* 2009). We have detected the different SNPs within the Illumina data using custom python scripts (minimum read depth = 30; SNP detection threshold = 2, corresponding to nucleotides that are present more than two out of 100 times per position). To estimate the relative presence of each haplotype, Illumina reads were mapped on previously detected haplotypes (~100 bp for each region including 5′-ETS/IGS, 18S, ITS-1 and 25S) using Bowtie2 (score-min: L, 100, 0; $f(x)= 100 + 0 * x$ where x = length of read). Results obtained were filtered using custom python scripts to identify the number of reads mapped with 100% of identity.

Chromosome preparation and fluorescence *in situ* hybridization

In situ hybridization was performed on mitotic chromosomes from *S. maritima* roots. Root tips with a length of 0.5–1.5 cm were treated in the dark with 0.04% 8-hydroxiquinoline for 2 hr at 4°, followed by 2 hr at room temperature to accumulate metaphases, then fixed in ethanol–acetic acid (3:1, *v/v*) for 48 hr at 4°, and stored in ethanol 70% at −20° until required. After washing in 0.01 M enzyme buffer (citric acid–sodium citrate, pH 4.5) for 15 min, the prepared roots were digested in a solution of 5% Onozuka R-10 cellulase (Sigma) and 1% Y23 pectolyase (Sigma) at 37° for 30 min. The root tips were then washed with distilled water for 30 min. A root tip was transferred to a slide and macerated with a drop of 3:1 fixation solution. After air-drying, slides with good metaphase chromosome spreads were stored at −20°. Fluorescence *in situ* hybridization was carried out using the ribosomal probe pTa 71 (Gerlach and Bedbrook 1979), which contained a 9-kb *Eco*RI fragment of rDNA repeat unit (18S-5.8S-25S genes and spacers) isolated from *Triticum aestivum*. The pTa 71 probe was labeled by random priming with biotin-14-dUTP (Invitrogen, Life Technologies). Chromosome preparations were incubated in RNase A (100 ng/μl) and pepsin (100 mg/ml) in 0.01 M HCl, and fixed with paraformaldehyde (4%). Chromosomes were denatured in a solution of 70% formamide in 2X SSC at 70° for 2 min, dehydrated in an ethanol series (70%, 90%, and 100%), and air-dried. The hybridization mixture, consisting of 50% deionized formamide, 10% dextran sulfate, 2X SSC, 1% SDS, and labeled probe (200 ng per slide), was denatured at 92° for 6 min and transferred to ice. The denatured probe was placed on the slide and *in situ* hybridization was carried out overnight in a moist chamber at 37°. After hybridization, slides were washed for 5 min in 50% formamide in 2X SSC at 42°, followed by several washes in 4X SSC-Tween. The biotinylated probe was immunodetected by Texas Red avidin DCS (Vector Laboratories). The chromosomes were mounted and counterstained in Vectashield (Vector Laboratories) containing 2.5 μg/ml 4',6-diamidino-2-phenylindole (DAPI). Fluorescence images were captured using a CoolSnap HQ camera (Photometrics, Tucson, AZ) on an Axioplan 2 microscope (Zeiss, Oberkochen, Germany) and analyzed using MetaVue (Universal Imaging Corporation, Downingtown, PA).

Data availability

The S. *maritima* rDNA consensus sequence and haplotype sequences are available at NCBI under the KT874468 to KT874488 accession numbers.

RESULTS

Analysis and annotation of the 45S rDNA region

The contig obtained with the clustering approach was 8456 nucleotides long; it was assembled from 3219 reads with an average length of 377.0 bp. Genomic DNA reads of *S. maritima* were then mapped to the contig with GSMapper (with the following parameters: ml = 100 bp; mi = 95%) (Margulies *et al.* 2005), resulting in a total of 4014 aligned reads with an average length of 506.8 nucleotides. This step allowed us to build a 45S rDNA reference sequence totaling 8464 bp. The different regions of the 45S rDNA of *S. maritima* were annotated by homology searches using rDNA sequences from Poaceae species (*Zea mays*, *Oryza sativa*) available in GenBank (NCBI) and by motif searches for the 5′-ETS region (Table 1). The beginning of the 5′-ETS region was identified using the TATA-box (5′-TATATTAGGGGG-3′) motif at position 395, corresponding to the expected plant TATA-box motif (5′-TATA(G)TA(N)GGGGG-3′), as found in several species such as *Zea mays*, *Oryza sativa*, or *Arabidopsis thaliana* (Fan *et al.* 1995; Zentgraf *et al.* 1998). To confirm this position, a dot plot of the 5′-IGS/ETS region against itself was performed. Regions 2 and 3 of the 5′-ETS (Volkov *et al.* 1996, 1999) were also detected. We found that regions 2 and 3 of the 5′-ETS totaled 1600 bp. This confirms the total length of the 5′-ETS of *S. maritima* (1754 bp), which is approximately two-times larger than the 5′-ETS of *Z. mays* (825 bp; X03989) and *O. sativa* (966 bp). The *S. maritima* rDNA contig covers the whole coding region, spanning the 18S, 5.8S, and 25S, the ITS-1, ITS-2, 5′-ETS (Figure 1), and part of the IGS (5′-IGS: 391 bp; 52.9 GC% and 3′ETS/3′-IGS: 517 bp and 42.6 GC%) (Table 1). Interspecific comparisons of the three coding rDNA regions between maize and *Spartina* and between rice and *Spartina* indicate that nucleotide variation is distributed all along these regions (Figure 1).

SNP validation in coding and noncoding regions

Using the mapped Roche-454 data, 34 SNPs or indels (insertion/deletion) were identified. These polymorphisms were localized both in coding (18S, 25S) and noncoding regions (5′-IGS, 5′-ETS, ITS-1 and 3′-ETS/IGS). Cloning and resequencing were performed and confirmed 10 of the 34 polymorphic sites (two substitutions, four variants including substitution and indel at the same position, and four indels)

that had been detected using the mapping procedure. Twenty-nine SNPs and indels (including nine polymorphic sites found by cloning and (re)-sequencing) were subsequently validated using Illumina data (Table 3). Within the coding region, one indel and four SNPs were detected and validated using the Roche-454 and lllumina data in the 18S and 25S domains, respectively. In the noncoding regions, a substitution, the variant presenting substitution and indel at the same position, and two indels were validated using Illumina method in the 5′-IGS. In the 5′-IGS/ETS noncoding region, a total of 21 polymorphisms were detected using the mapped Roche-454 data, including nine substitutions and 12 indels (two of 4 bp) (Table 3). In this region, all the identified substitutions (except one G/T, position 1235) were validated using Illumina data. Within the ITS-1 region, the Roche-454 and Illumina data detected the same SNP (A/T/-) that was validated by (re)-sequencing. However, the indel was only found in the pyrosequencing data. In the 3′-ETS/IGS noncoding region, one indel using the Roche-454 data was not validated by Illumina data. Thus, the cloning-sequencing data as well as the Illumina reads substantiated 29 of the 34 polymorphisms previously identified in the Roche-454 reads. Four of the five invalidated variants were included in homopolymeric regions. The last indel was not found in the Illumina data, but cloning and sequencing data revealed variation in the chromatograph, suggesting either a sequencing error or a substitution.

Table 3 Nucleotide polymorphisms detected within coding and noncoding regions of *Spartina maritima* rDNA using the three methods (mapping method, cloning-sequencing Sanger method, and Illumina method)

Region	Mapping Method	Variants Validated	
		Cloning-Sequencing	Illumina
5′-IGS	6 (4)	3	4
5′-ETS	21 (21)	4	21
18S	1 (1)	—	1
ITS-1	1 (1)	1	1
25S	4 (2)	2	2
3′-IGS/ETS	1 (0)	—	0

The numbers between brackets correspond to the number of variants selected for haplotype reconstruction. Cloning and sequencing were not performed in the 18S and 3′-IGS/ETS regions.

Haplotype detection

After rDNA domain annotations and SNP detection and validation, the number of haplotypes for each region of the 45S unit was determined. In this study we retained only the Roche-454 haplotypes constructed with a minimum of 10× read depth. The five indels not validated by cloning and Illumina data were not considered. From the 29 polymorphic sites retained, 20 haplotypes with a mean length of 1197.70 bp (σ = 938.32; C.I.$_{95\%}$ = 786.47–1608.93) were constructed using our program. These haplotypes were constructed from 69.35 reads on average (σ = 66.24; C.I.$_{95\%}$ = 40.32–98.38). The 20 different haplotypes were localized on the different rDNA domains as shown in Figure 4. We did not identify any polymorphism in the 5.8S coding region, the internal transcribed spacer 2 (ITS-2), and the 3′ intergenic spacer (IGS). The 5′-IGS/ETS region was the most variable; we detected four variants in the 5′-IGS and 21 variants in the 5′-ETS. These SNPs allowed the construction of 13 haplotypes in this region (Figure 4, Block I). One indel was detected in the 18S coding region indicating the presence of two haplotypes (Figure 4, Block II). Within the ITS-1 we detected only one SNP indicating the presence of three haplotypes (Figure 4, Block III). Within the 25S, four haplotypes were constructed from two SNPs (Figure 4, Block IV). Finally, 20 haplotypes were obtained in total.

Haplotype validation

To validate the developed program, the constructed Roche-454 haplotypes were compared with haplotypes obtained using the cloned sequences (16X to 24X both ends). Four regions were selected for these comparisons (5′-IGS/ETS, 5′-ETS, ITS-1/5.8S/ITS-2, and 25S). In the 5′-IGS/ETS noncoding region, seven polymorphisms were validated by cloning, including the indel block of 2 bp, one substitution, two deletions, and two substitutions/insertions/deletions. Using these polymorphisms, it was possible to validate three of nine Roche-454 constructed haplotypes. Cloning depth could explain the low number of haplotypes validated. In the 5′-ETS noncoding region, only one polymorphism was validated by the cloning approach, which confirms one haplotype (two haplotypes were detected with this variant using Roche-454 data) (Figure 4, Block I). In the ITS-1/5.8S/ITS-2, cloning-sequencing data validated the variant detected with the program, and two of three detected haplotypes were validated (Figure 4, Block III). In the 18S coding region, the two haplotypes detected with the program were validated using Illumina data (Figure 4, Block II). In the 25S coding region, the two SNPs detected using our program (developed on Roche-454 data) were found with cloning/(re)-sequencing data. Four haplotypes were found with the program and three were validated (Figure 4, Block IV).

Haplotype proportion estimation

Using the number of Illumina reads mapped on the different haplotypes, it was possible to estimate the relative proportion of each haplotype. For each rDNA region, the number of reads per haplotype was estimated (Figure 5). In the 5′-IGS/ETS, 12 haplotypes were compared, including three haplotypes presenting a high proportion (more than 17.8% of the rDNA copies). Three haplotypes presented a very small number of estimated copies (between 0.01% and 2.50%). The estimated proportions for the six other haplotypes detected in this region were 6.92% on average. In the 18S coding region, two haplotypes were present in very different frequencies. The first one represented only 1.28% of the rDNA copies and the second one was overrepresented (98.72% of rDNA copies). In the ITS-1, two haplotypes were highly frequent (46.45% and 51.49% respectively). The third haplotype detected in the ITS-1 was represented in a low copy number (2.06%). In the 25S coding region, two haplotypes were approximately equally frequent (39.84% and 40.66%), and two others were less represented (4.86% and 14.65%). Finally, in the different regions studied, we found between one and three copies highly represented, which could indicate that these copies present a high number of repetitions in the *S. maritima* genome.

rDNA location on *S. maritima* chromosomes

Fluorescence *in situ* hybridization (FISH) was performed from *S. maritima* somatic metaphase chromosomes (counterstained with DAPI) to identify the physical position of 45S rDNA arrays. The chromosomal locations of 45S rDNA arrays in *S. maritima* are shown in Figure 6. Hybridization signals were consistently observed on two chromosomes indicating the presence of one pair of 45S rDNA loci.

DISCUSSION

In this study, we developed a program for SNP and haplotype detection. This program was applied on Roche-454 pyrosequencing data corresponding to the 45S rDNA genes of *S. maritima*. The objective was to detect the different rDNA sequences (designated as haplotypes) in this hexaploid species where both paralogous and homeologous copies are expected for this gene family. Results were validated by cloning and

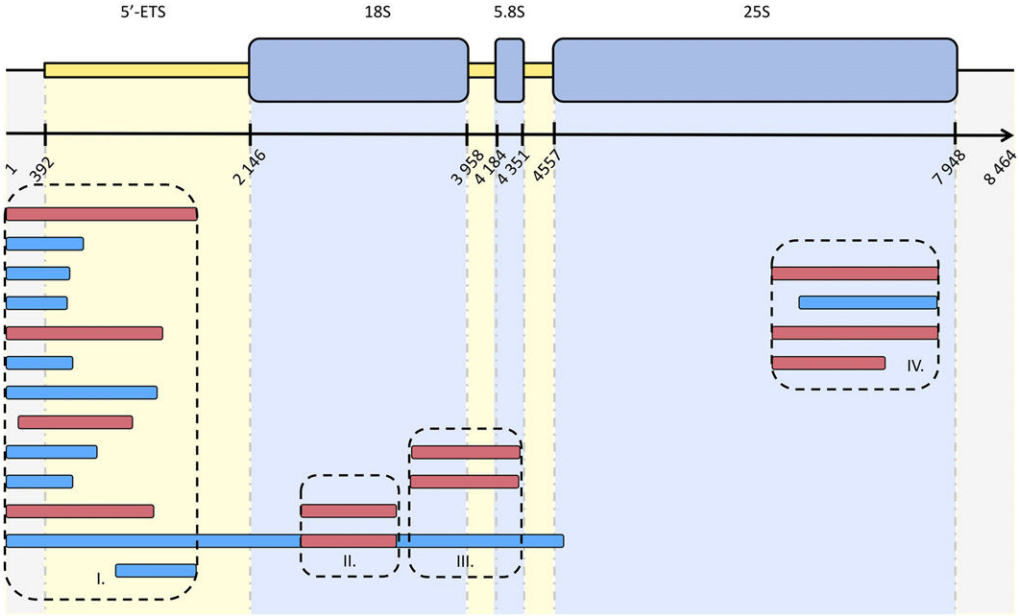


Figure 4 Schematic representation of haplotypes detected in *S. maritima* rDNA using our developed program (mapping data) and cloning data. The blue boxes represent the haplotypes detected only by the developed program. The red boxes represent the haplotypes detected by the developed program and validated by cloning and (re)-sequencing or Illumina data (*i.e.*, haplotypes in the block II). The different blocks (I to IV) correspond to regions of rDNA presenting at least two haplotypes.

(re)-sequencing and also from Illumina sequencing datasets; the results allowed the estimation of the relative proportion of haplotype variation. To date, very few programs or pipelines have been developed to distinguish duplicated copies from next-generation sequencing data without diploid parents as a reference. SNiPloid (Peralta *et al.* 2013), BamBam (Page *et al.* 2014), or HyLiTE (Duchemin *et al.* 2014) were developed for haplotype detection in polyploid species using comparisons between the polyploids and their diploid progenitors. The POLiMAPS program developed by Tennessen *et al.* (2014) used a related sequenced diploid genome combined with dense genetic map to infer homeologs in octoploid *Fragaria*. Our program will be particularly useful in nonmodel systems having experienced repeated duplication events and for which diploid reference genomes are not available.

Number of rDNA loci in *S. maritima*

Molecular cytogenetic analysis by FISH revealed the presence of one pair of 45S rDNA signals on the chromosomes of *S. maritima* instead of the three expected pairs for a hexaploid genome. The number of rDNA loci varies greatly among flowering plants of the same ploidy level. For instance, only one locus was identified in diploid *Medicago* species (Rosato *et al.* 2008), whereas five loci were detected in diploid *Brassica rapa* species (Maluszynska and Heslop-Harrison 1993). In polyploids, the number of rDNA loci can be either retained (in recent or old polyploids) or deeply decreased, most probably as part of the diploidization process (Leitch and Bennett 2004). For example, three rDNA loci were identified in the hexaploid bread wheat (AABBDD), corresponding to the sum of the loci present in the parental genomes [two in the allotetraploid *Triticum turgidum* (AABB) and one in the diploid *Aegilops tauschii* (DD)] (Badaeva *et al.* 1996; El-Twab 2007). Similarly, the number of loci detected in the recent *Tragopogon* allotetraploids is additive with respect to their diploid parental species (Kovarik 2005). In contrast, in other polyploid genera, such as *Fragaria*, a decrease in the number of rDNA loci during the diploidization process is observed, with only six loci in the decaploid *F. iturupensis* (Liu and Davis 2011), whereas related diploid species exhibit three loci. In the hexaploid *S. maritima*, the presence of one locus indicates rDNA loci have been lost since its formation, following evolution of the tetraploid and hexaploid *Spartina* lineages that occurred in the past 6–10 mya (Rousseau-Gueutin *et al.* 2015). To our knowledge, this is the first 45S locus number reported in *Spartina*, and it would be interesting to examine this number in other polyploid *Spartina* species. Such rDNA locus loss can occur rapidly, as in the recent allotetraploid *Brassica napus*, which formed approximately 8000 years ago (Chalhoub *et al.* 2014), where six loci are retained instead of the seven expected from the progenitor species (five in *B. rapa* and two in *B. oleracea*) (Snowdon *et al.* 1997).

In silico detection of rDNA domains

The intraindividual homogeneity of ribosomal genes was so far mainly examined across plant species by cloning and sequencing, a relatively time-consuming procedure that limits the number of sequences that can be sampled in the genome. Massively parallel sequencing now provides the opportunity to assess more accurately the intraindividual sequence polymorphism (Matyášek *et al.* 2012). Furthermore, in rDNA genes, most studies have focused on the ITS region (Alvarez and Wendel 2003; Poczai and Hyvönen 2010) or the 18S and 25S coding regions. For the first time, all the repeat unit and spacers (IGS and ETS) in *Spartina* are reported here. Analysis of the contig corresponding to the 45S ribosomal DNA unit of *S. maritima* enabled us to annotate the different regions of the transcriptional unit and to detect the different copies for each region (IGS/ETS; 18S; ITS-1; 5.8S; ITS-2; and 25S) using an approach relying on a low-stringency mapping of NGS reads.

Detection of the different domains was performed by sequence alignments of 45S rDNA available in the databases of two Poaceae species (maize and rice). Lengths of the rDNA coding regions and ITS of *S. maritima* are similar to those found in two Poaceae. However, the percentage of GC of the two ITS and 5.8S coding regions is very low in comparison with other Poaceae (Table 1). When examining the GC content of the ITS region in various grass species, we found that some Chloridoideae lineages related to *Spartina* in the Zoysieae, Cynodonteae, and Eragrostideae tribes similarly exhibit less than 60% GC content (Table S2). In more distantly related Chloridoideae tribes such as Triraphideae and Centropodieae (Soreng *et al.* 2015), as well as other grass subfamilies, the GC content is greater (up to 79.1% in rice ITS-2) (Table S2). Other monocot lineages (*e.g.*, *Prospero autumnale*, Hyacinthaceae) also exhibit higher GC content. Although information

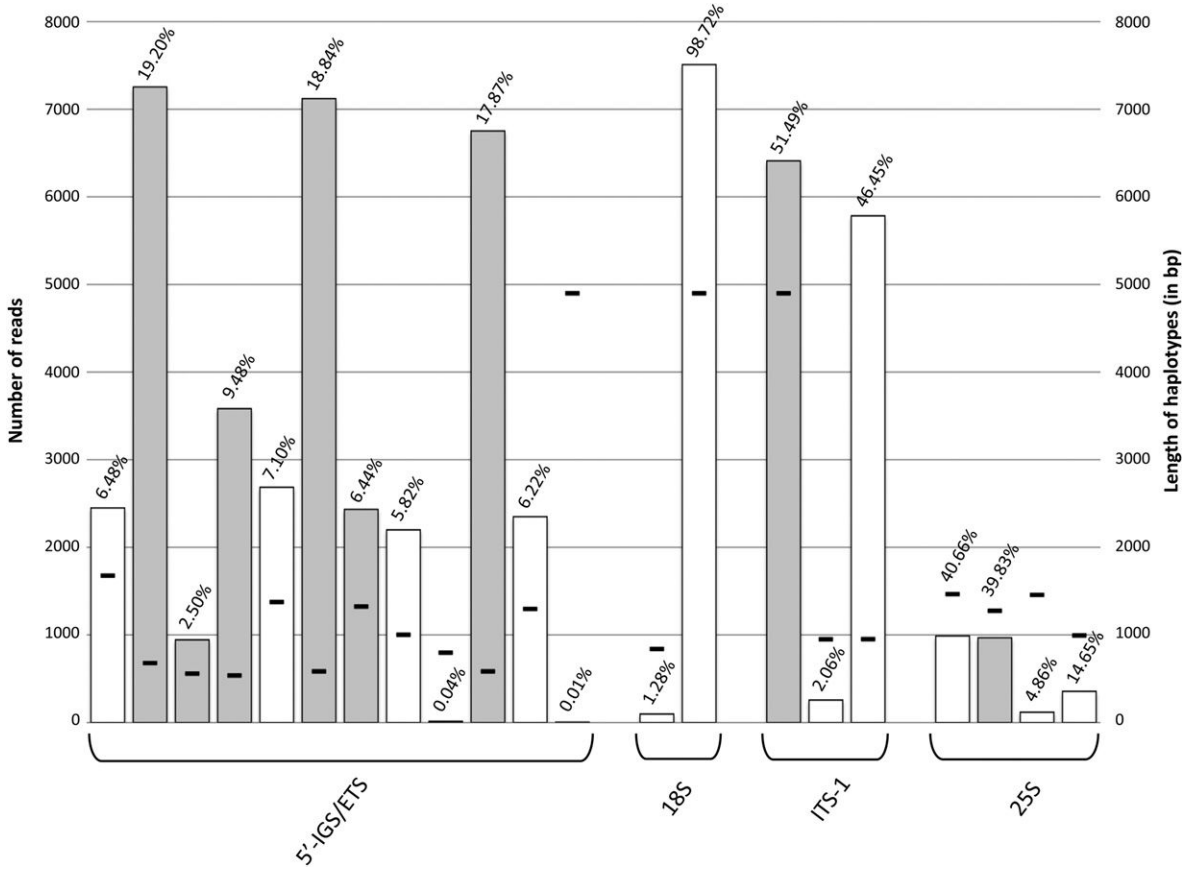


Figure 5 Number of reads mapped with 100% of identity on the different haplotypes for each region (5′-IGS/ETS, 18S, ITS-1, and 25S). The relative proportion of haplotype is presented above each histogram. White bars represent haplotypes validated by either Sanger sequences or Illumina data (18S coding region only). Black lines represent the length of the haplotypes constructed by our program. Some haplotypes are spanning two or more subregions.

about more monocot representatives is needed, these results examined in the light of recent grass phylogenies (Soreng *et al.* 2015) suggest that GC content decrease occurred in the Chloridoideae lineage following the split between the basal Centropodieae–Triraphideae tribes and the other more derived tribes including *Spartina*.

Within 25S, variable regions were identified when comparing *S. maritima* with *Zea mays* and *Oryza sativa* (Figure 1). These regions mostly correspond to previously described expansion segments (Hancock and Dover 1988) representing variable 26S subdomains in plants (Kuzoff *et al.* 1998). The 5′-ETS detected in *Oryza sativa* (966 bp) has a similar length compared to the 5′-ETS of *Zea mays* (825 bp) identified by Toloczyki and Feix (1986). However, this spacer is much larger in *S. maritima* (1754 bp), with lower GC content (50.9%) than in other Poaceae (70.0% and 73.0% for maize and rice, respectively). These results are consistent with the literature showing that the size of 5′-ETS is variable across species, even within genera. Volkov *et al.* (1996) found different lengths of the 5′-ETS between two tobacco species: *Nicotiana sylvestris* and *Nicotiana tomentosiformis*, which have, respectively, an ETS size of 1444 bp and 2172 bp. The size difference between these two ETS is explained by the number of subrepeats present in region 2 of the ETS, where five and 10 subrepeats are present in *N. sylvestris* and *N. tomentosiformis*, respectively (Volkov *et al.* 1996). In the *Fabaceae* family, *Lupinus luteus* has a small 5′-ETS with a size of 487 bp (Mahé *et al.* 2010). It would be interesting to identify and compare the 5′-ETS in other hexaploid and tetraploid *Spartina* species as well as in related Chloridoids to explore the amplitude and origin of this size variation. These studies would verify whether this heterogeneity is widespread in the polyploid *Spartina* genus, or whether it is a characteristic of the hexaploid *Spartina maritima*.

Polymorphism of coding and noncoding regions within the *S. maritima* rDNA locus

Thirty-nine polymorphic variants and 20 haplotypes were detected in *S. maritima* rDNA using our developed program. Thirty-four SNPs and 11 haplotypes were validated by cloning and Sanger sequencing or with Illumina data. The 5.8S coding region exhibited less heterogeneity than 18S and 25S coding regions. Indeed, no polymorphism was encountered in the 5.8S and only one polymorphic site (one deletion) was encountered in the 18S coding region, highlighting their high sequence homogeneity. The 25S unit also presents little variation, with two SNPs (variants presenting substitution and indel at the same position) detected using pyrosequencing data and validated by the other methods. Although most coding regions are highly conserved, some variation may be observed in the 18S region. Such results are in accordance with those of Mentewab *et al.* (2011), who detected a deletion of 270 bp within *Arabidopsis thaliana* 18S, or those obtained in animals (*e.g.*, *Anguilla*, for which a specific SNP was found) (Frankowski and Bastrop 2010). Similarly, the intraspecific variation of the 18S and 25S detected in *S. maritima* accords with the results obtained in two *Oenothera* species

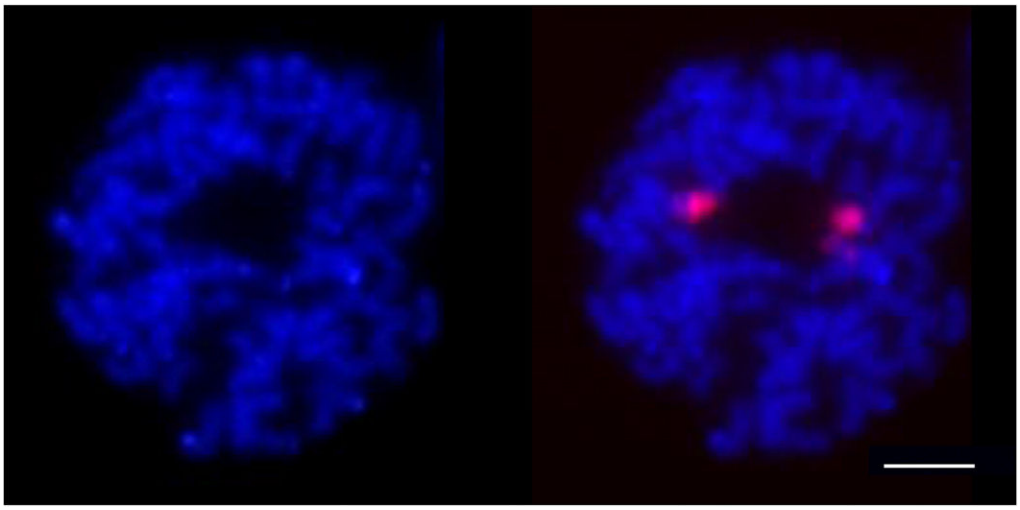


Figure 6 Fluorescence *in situ* hybridization (FISH) of the 45S rDNA in *S. maritima* on somatic metaphase chromosomes counterstained with DAPI (blue), in the absence (left) or presence of the red rDNA probe (right). Bar, 5 μm.

using cloning and sequencing methods (Seo *et al.* 2013), for which 17 and 10 SNPs were identified in the 18S of *O. odorata* and *O. laciniata* and 11 and 13 SNPs in the 25S regions, respectively.

In the noncoding regions, the internal transcribed spacers of *S. maritima* exhibit less heterogeneity than the IGS/ETS regions, without any polymorphism in ITS-2 and only one substitution/deletion in ITS-1. The high polymorphism encountered in the IGS/ETS region (25 variants in the 5′-IGS/ETS) is consistent with the literature, which indicates that intergenic regions evolve more rapidly than the ITS region, which is most particularly prone to homogenization (Alvarez and Wendel 2003; Kovarik *et al.* 2008). The high IGS/ETS variation we detected at the intraindividual level is in accordance with other studies at the interindividual or intraspecific levels (Poczai and Hyvönen 2010). Indeed, ETS regions evolve very quickly, 1.5-times faster than ITS regions (Bena *et al.* 1998); IGS regions also evolve rapidly (Chang *et al.* 2010). The number of detected haplotypes in the IGS/ETS region (13) exceeds the number of expected homeologous copies in a hexaploid species, which suggests incomplete homogeneization within the rDNA locus as also evidenced in diploid *Nicotiana* species (Matyášek *et al.* 2012). Our finding of a single rDNA locus in *Spartina maritima* mentioned above supports this hypothesis.

Like coding regions, the ITS region in *S. maritima* contains two haplotypes that are highly represented. The ITS region might be as strongly selected as the flanking coding regions, but the rapid evolution of the ITS region at the interspecific level suggests that this region is rather subject to more rapid homogenization (within individual genomes) than the other regions. Interspecies variation is then most likely more important than intraindividual (intragenomic variation), which explains why this region is useful in phylogenetic studies aiming at reconstructing organismal (*e.g.*, species) history (Baumel *et al.* 2002 for *Spartina*). In the 5′-IGS/ETS, three haplotypes are present in high proportion and two haplotypes are less represented, which could indicate that the homogenization process is underway in these regions.

Impact of the parameters on polymorphism and haplotype detection

The number of SNPs and haplotypes that may be detected in our study is indeed limited by the employed technology (*i.e.*, 454 read depth) and the parameters used to avoid false positives resulting from sequencing errors. Thus, the number of actual rDNA variants and their relative proportion (or copy number) in the *S. maritima* genome is most likely underestimated. In our developed program, several parameters can be tuned for relaxing or constraining read mapping, SNP detection, and haplotype construction. For example, the choice to not consider nucleotides present in less with a frequency lower than 20% at a specific position is in line with the procedure usually used when analyzing polyploid genomes. In their analysis of allotetraploid cotton EST (Expressed Sequence Tag) assemblies, Udall *et al.* (2006) have chosen to set this parameter to 25%. Tennessen *et al.* (2014) have chosen to adjust this parameter to 12.5% (one-eighth frequency) to detect variants in octoploid *Fragaria* species. Our parameter is less stringent than the parameter used for homeo-SNP detection (fixed to 40%) in allotetraploid cotton (Page *et al.* 2013a,b). This lower value allowed us to detect putative homeologous and paralogous copies, as expected for highly repeated rDNA arrays. Using this parameter, 34 variants were detected and 29 were validated. Four of these variants are localized in homopolymeric regions of three or four repeats. The developed program does not consider the homopolymeric regions of at least five repeats. These results could indicate that variants in homopolymeric regions of three or four repeats correspond to false positives and should also not be considered. In the internal transcribed spacer and the coding region, it was possible to validate most of the haplotypes constructed by the developed program (seven haplotypes validated on nine). In the noncoding region, the number of haplotypes, their relative estimated proportion, and the cloning depth explain the number of haplotypes not validated by the cloning method. In fact, several haplotypes localized in the 5′-IGS/ETS built with the developed program were also detected and validated using cloning/(re)-sequencing in the *Spartina* hybrids between *S. maritima* and *S. alterniflora* (*S. x townsendii* and *S. x neyrautii*; unpublished data). New and emerging single molecule sequencing methods that are producing longer reads should provide precious information regarding the assembly of longer (or even whole-length) 45S rDNA haplotypes.

In summary, our program enables detection of SNPs and haplotypes within NGS read datasets. Haplotype construction and validation are based on three methods/technologies [Roche-454, cloning, and (re)-sequencing Sanger method and Illumina approaches] in the context of polyploidy in a nonmodel species where diploid reference parental genomes are not available. The program developed in this study could then be used for any polyploid species (including high ploidy levels) with or without reference genome sequences but also on diploid species (*e.g.*, for allele detection from amplicon data sets). In the polyploid *Spartina* genus, which contains various ploidy levels (4×, 6×, 7×, 9×, and 12×), such a program will be useful to detect duplicated homeologous genes and to increase our understanding of their origin and evolutionary fate.

ACKNOWLEDGMENTS

We thank two anonymous reviewers for helpful comments on the manuscript. This work was supported by the International Associated Laboratory "Ecological Genomics of Polyploidy" supported by CNRS (INEE, UMR CNRS 6553 Ecobio), University of Rennes 1, Iowa State University (Ames, USA), and the Partner University Funds (to M. A., A. S., J. B.). The analyses benefited from the Molecular Ecology Platform (UMR CNRS 6553 Ecobio) and Genouest (Bioinformatics) facilities. Author J. Boutte benefited from a PhD scholarship from the University of Rennes 1.

LITERATURE CITED

Ainouche, M. L., H. Chelaifa, J. Ferreira de Carvalho, S. Bellot, A. K. Ainouche et al., 2012 Polyploid evolution in *Spartina*: Dealing with highly redundant hybrid genomes, pp. 225–243 in *Polyploidy and Genome Evolution* edited by E. Douglas, Springer, Berlin, Heidelberg.

Ainouche, M. L., P. M. Fortune, A. Salmon, C. Parisod, M.-A. Grandbastien *et al.*, 2008 Hybridization, polyploidy and invasion: lessons from *Spartina* (Poaceae). Biol. Invasions 11: 1159–1173.

Akhunova, A. R., R. T. Matniyazov, H. Liang, and E. D. Akhunov, 2010 Homoeolog-specific transcriptional bias in allopolyploid wheat. BMC Genomics 11: 505.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Álvarez, I., and J. F. Wendel, 2003 Ribosomal ITS sequences and plant phylogenetic inference. Mol. Phylogenet. Evol. 29: 417–434.

Badaeva, E. D., B. Friebe, and B. S. Gill, 1996 Genome differentiation in *Aegilops*. 2. Physical mapping of 5S and 18S–26S ribosomal RNA gene families in diploid species. Genome 39: 1150–1158.

Barker, R. F., N. P. Harberd, M. G. Jarvis, and R. B. Flavell, 1988 Structure and evolution of the intergenic region in a ribosomal DNA repeat unit of wheat. J. Mol. Biol. 201: 1–17.

Baumel, A., M. L. Ainouche, and J. E. Levasseur, 2001 Molecular investigations in populations of *Spartina anglica* CE Hubbard (Poaceae) invading coastal Brittany (France). Mol. Ecol. 10: 1689–1701.

Baumel, A., M. L. Ainouche, R. J. Bayer, A. K. Ainouche, and M. T. Misset, 2002 molecular phylogeny of hybridizing species from the genus *Spartina* Schreb. (Poaceae). Mol. Phylogenet. Evol. 22: 303–314.

Bena, G., M.-F. Jubier, I. Olivieri, and B. Lejeune, 1998 Ribosomal external and internal transcribed spacers: combined use in the phylogenetic analysis of *Medicago* (*Leguminosae*). J. Mol. Evol. 46: 299–306.

Chalhoub, B., F. Denoeud, S. Liu, I. A. P. Parkin, H. Tang *et al.*, 2014 Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science 345: 950–953.

Chang, K.-D., S.-A. Fang, F.-C. Chang, and M.-C. Chung, 2010 Chromosomal conservation and sequence diversity of ribosomal RNA genes of two distant *Oryza* species. Genomics 96: 181–190.

Combes, M.-C., A. Cenci, H. Baraille, B. Bertrand, and P. Lashermes, 2011 Homeologous gene expression in response to growing temperature in a recent allopolyploid (*Coffea arabica L.*). J. Hered. 103: 36–46.

Duchemin, W., P.-Y. Dupont, M. A. Campbell, A. R. Ganley, and M. P. Cox, 2014 HyLiTE: accurate and flexible analysis of gene expression in hybrid and allopolyploid species. BMC Bioinformatics 16: 8.

El-Metwally, S., O. M. Ouda, and M. Helmy, 2014 Next generation sequencing technologies and challenges in sequence assembly, Springer, New York, NY.

El-Twab, M. H. A., 2007 Physical mapping of the 45S rDNA on the chromosomes of *Triticum turgidum* and *T. aestivum* using fluorescence in situ hybridization for chromosome ancestors. Arab J. Biotechnol. 10: 69–80.

Fan, H., K. Yakura, M. Miyanishi, M. Sugita, and M. Sugiura, 1995 In vitro transcription of plant RNA polymerase I-dependent rRNA genes is species-specific. Plant J. 8: 295–298.

Flagel, L., J. Udall, D. Nettleton, and J. Wendel, 2008 Duplicate gene expression in allopolyploid *Gossypium* reveals two temporally distinct phases of expression evolution. BMC Biol. 6: 16.

Fortune, P. M., K. A. Schierenbeck, A. K. Ainouche, J. Jacquemin, J. F. Wendel *et al.*, 2007 Evolutionary dynamics of Waxy and the origin of hexaploid *Spartina* species (Poaceae). Mol. Phylogenet. Evol. 43: 1040–1055.

Fortune, P. M., K. Schierenbeck, D. Ayres, A. Bortolus, O. Catrice *et al.*, 2008 The enigmatic invasive *Spartina densiflora*: A history of hybridizations in a polyploidy context. Mol. Ecol. 17: 4304–4316.

Foucaud, 1897 Un *Spartina* inédit. Ann. Soc. Sci. Nat. Char. Inf. 32: 220–222.

Frankowski, J., and R. Bastrop, 2010 Identification of *Anguilla anguilla* (*L.*) and *Anguilla rostrata* (Le Sueur) and their hybrids based on a diagnostic single nucleotide polymorphism in nuclear 18S rDNA. Mol. Ecol. Resour. 10: 173–176.

Gerlach, W. L., and J. R. Bedbrook, 1979 Cloning and characterization of ribosomal RNA genes from wheat and barley. Nucleic Acids Res. 7: 1869–1885.

Groves, H., and J. Groves, 1880 *Spartina x townsendii* Nobis. Rep. Bot. Soc. Exch. club Br. Id. 1–37.

Hancock, J. M., and G. A. Dover, 1988 Molecular coevolution among cryptically simple expansion segments of eukaryotic 26S/28S rRNAs. Mol. Biol. Evol. 5: 377–391.

Ilut, D. C., J. E. Coate, A. K. Luciano, T. G. Owens, G. D. May *et al.*, 2012 A comparative transcriptomic study of an allotetraploid and its diploid progenitors illustrates the unique advantages and challenges of RNA-seq in plant species. Am. J. Bot. 99: 383–396.

Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr *et al.*, 2011 Ancestral polyploidy in seed plants and angiosperms. Nature 473: 97–100.

Katoh, K., and H. Toh, 2010 Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics 26: 1899–1900.

Koh, J., P. S. Soltis, and D. E. Soltis, 2010 Homeolog loss and expression changes in natural populations of the recently and repeatedly formed allotetraploid *Tragopogon mirus* (Asteraceae). BMC Genomics 11: 97.

Kovarik, A., 2005 Rapid concerted evolution of nuclear ribosomal DNA in two *Tragopogon* allopolyploids of recent and recurrent origin. Genetics 169: 931–944.

Kovarik, A., M. Dadejova, Y. K. Lim, M. W. Chase, J. J. Clarkson *et al.*, 2008 Evolution of rDNA in *Nicotiana* allopolyploids: a potential link between rDNA homogenization and epigenetics. Ann. Bot-London 101: 815–823.

Kuzoff, R. K., A. J. Sweere, D. E. Soltis, P. S. Soltis, and E. A. Zimmer, 1998 The phylogenetic potential of entire 26S rDNA sequences in plants. Mol. Biol. Evol. 15: 251–263.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359.

Leitch, I. J., and M. D. Bennett, 2004 Genome downsizing in polyploid plants. Biol. J. Linn. Soc. Lond. 82: 651–663.

Levy, A. A., and M. Feldman, 2002 The impact of polyploidy on grass genome evolution. Plant Physiol. 130: 1587–1593.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Liu, B., and T. M. Davis, 2011 Conservation and loss of ribosomal RNA gene sites in diploid and polyploid *Fragaria* (Rosaceae). BMC Plant Biol. 11: 157.

Mable, B. K., 2004 "Why polyploidy is rarer in animals than in plants": myths and mechanisms. Biol. J. Linn. Soc. Lond. 82: 453–466.

Mahé, F., H. Pascual, O. Coriton, V. Huteau, A. Navarro Perris *et al.*, 2010 New data and phylogenetic placement of the enigmatic Old World lupin: *Lupinus mariae-josephi* H. Pascual. Genet. Resour. Crop Ev. 58: 101–114.

Maluszynska, J., and J. S. Heslop-Harrison, 1993 Physical mapping of rDNA loci in *Brassica* species. Genome 36: 774–781.

Marchant, C., 1968 Evolution in *Spartina* (Gramineae). II. Chromosomes, basic relationships and the problem of *Spartina x townsendii* agg. Biol. J. Linn. Soc. Lond. 60: 381–409.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader *et al.*, 2005 Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.

Matyášek, R., S. Renny-Byfield, J. Fulneček, J. Macas, M.-A. Grandbastien *et al.*, 2012 Next generation sequencing analysis reveals a relationship between rDNA unit diversity and locus number in *Nicotiana* diploids. BMC Genomics 13: 722.

Mentewab, A. B., M. J. Jacobsen, and R. A. Flowers, 2011 Incomplete homogenization of 18S ribosomal DNA coding regions in *Arabidopsis thaliana*. BMC Res. Notes 4: 93.

Metzker, M. L., 2009 Sequencing technologies—the next generation. Nat. Rev. Genet. 11: 31–46.

Milne, I., M. Bayer, L. Cardle, P. Shaw, G. Stephen *et al.*, 2009 Tablet–next generation sequence assembly visualization. Bioinformatics 26: 401–402.

Nei, M., and A. P. Rooney, 2005 Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. 39: 121.

Novák, P., P. Neumann, and J. Macas, 2010 Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11: 378.

Ohno, S., 1970 *Evolution by Gene Duplication*. Springer-Verlag, New York.

Page, J. T., A. R. Gingle, and J. A. Udall, 2013a PolyCat: A resource for genome categorization of sequencing reads from allopolyploid organisms. G3 (Bethesda) 3: 517–525.

Page, J. T., M. D. Huynh, Z. S. Liechty, K. Grupp, D. Stelly *et al.*, 2013b Insights into the evolution of cotton diploids and polyploids from whole-genome re-sequencing. G3 (Bethesda) 3: 1809–1818.

Page, J. T., Z. S. Liechty, M. D. Huynh, and J. A. Udall, 2014 BamBam: genome sequence analysis tools for biologists. BMC Res. Notes 7: 829.

Peralta, M., M.-C. Combes, A. Cenci, P. Lashermes, and A. Dereeper, 2013 SNiPloid: A utility to exploit high-throughput SNP data derived from RNA-Seq in allopolyploid species. Int. J. Plant Genomics 2013: 1–6.

Peterson, P. M., K. Romaschenko, Y. H. Arrieta, and J. M. Saarela, 2014 A molecular phylogeny and new subgeneric classification of *Sporobolus* (Poaceae: Chloridoideae: Sporobolinae). Taxon 63: 1212–1243.

Poczai, P., and J. Hyvönen, 2010 Nuclear ribosomal spacer regions in plant phylogenetics: problems and prospects. Mol. Biol. Rep. 37: 1897–1912.

Prokopowich, C. D., T. R. Gregory, and T. J. Crease, 2003 The correlation between rDNA copy number and genome size in eukaryotes. Genome 46: 48–50.

Ramsey, J., and D. W. Schemske, 1998 Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annu. Rev. Ecol. Syst. 29: 467–501.

Rogers, S. O., and A. J. Bendich. 1987 Ribosomal RNA genes in plants: variability in copy number and in the intergenic spacer. Plant Mol. Biol. 9:509:520.

Rosato, M., M. Castro, and J. A. Rossello, 2008 Relationships of the woody *Medicago* species (Section Dendrotelis) assessed by molecular cytogenetic analyses. Ann. Bot-London 102: 15–22.

Rousseau-Gueutin, M., S. Bellot, G. E. Martin, J. Boutte, H. Chelaifa *et al.*, 2015 The chloroplast genome of the hexaploid *Spartina maritima* (Poaceae, Chloridoideae): Comparative analyses and molecular dating. Mol. Phylogenet. Evol. 93: 5–16.

Rozen, S., and H. Skaletsky, 2000 Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. 132: 365–386.

Salmon, A., L. Flagel, B. Ying, J. A. Udall, and J. F. Wendel, 2010 Homoeologous nonreciprocal recombination in polyploid cotton. New Phytol. 186: 123–134.

Schaal, B. A., and G. H. Learn, 1988 Ribosomal DNA variation within and among plant populations. Ann. Mo. Bot. Gard. 75: 1207–1216.

Seo, J. H., H. G. Bae, D. H. Park, B. S. Kim, J. W. Lee *et al.*, 2013 Sequence polymorphisms in ribosomal RNA genes and variations in chromosomal loci of *Oenothera odorata* and *O. laciniata*. Genes Genomics 35: 117–124.

Snowdon, R. J., W. Köhler, and A. Köhler, 1997 Chromosomal localization and characterization of rDNA loci in the *Brassica* A and C genomes. Genome 40: 582–587.

Soltis, D. E., V. A. Albert, J. Leebens-Mack, C. D. Bell, A. H. Paterson *et al.*, 2009 Polyploidy and angiosperm diversification. Am. J. Bot. 96: 336–348.

Soreng, R. J., P. M. Peterson, K. Romaschenko, G. Davidse, F. O. Zuloaga *et al.*, 2015 A worldwide phylogenetic classification of the Poaceae (Gramineae): Phylogenetic classification of the grasses. J. Syst. Evol. 53: 117–137.

Strong, D. R., and D. R. Ayres, 2013 Ecological and evolutionary misadventures of *Spartina*. Annu. Rev. Ecol. Evol. Syst. 44: 389–410.

Tennessen, J. A., R. Govindarajulu, T.-L. Ashman, and A. Liston, 2014 Evolutionary origins and dynamics of octoploid strawberry subgenomes revealed by dense targeted capture linkage maps. Genome Biol. Evol. 6: 3295–3313.

Toloczyki, C., and G. Feix, 1986 Occurrence of 9 homologous repeat units in the external spacer region of a nuclear maize rRNA gene unit. Nucleic Acids Res. 14: 4969–4986.

Udall, J. A., J. M. Swanson, K. Haller, R. A. Rapp, M. E. Sparks *et al.*, 2006 A global assembly of cotton ESTs. Genome Res. 16: 441–450.

Van de Peer, Y., S. Maere, and A. Meyer, 2009 The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. 10: 725–732.

Volkov, R., S. Kostishin, F. Ehrendorfer, and D. Schweizer, 1996 Molecular organization and evolution of the external transcribed rDNA spacer region in two diploid relatives of *Nicotiana tabacum* (*Solanaceae*). Plant Syst. Evol. 201: 117–129.

Volkov, R. A., N. V. Borisjuk, I. I. Panchuk, D. Schweizer, and V. Hemleben, 1999 Elimination and rearrangement of parental rDNA in the allotetraploid *Nicotiana tabacum*. Mol. Biol. Evol. 16: 311–320.

Waterhouse, A. M., J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton, 2009 Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189–1191.

Wendel, J. F., 2000 Genome evolution in polyploids. Plant Mol. Niol. 42: 225–249.

Wendel, J. F., A. Schnabel, and T. Seelanan, 1995 Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). Proc. Natl. Acad. Sci. USA 92: 280–284.

Zentgraf, U., R. Velasco, and V. Hemleben, 1998 Different Transcriptional Activities in the Nucleus, pp. 131–168 in Progress in Botany, edited by H.-D. Behnke, K., J. W. Esser, U. Kadereit, Lüttge, and M. Runge. Springer, Berlin.

Communicating editor: A. H. Paterson

Table S1: Estimation of the number of rDNA copies using Roche-454 and Illumina Whole Genome Shotgun datasets.

S. maritima: C= 1.85 Gb NOR = 1	Roche-454 data:	Illumina data:
Number of reads (mean length):	999,229 (377.0 bp)	172,528,550 (100 bp)
Coverage:	0.2x	9.3x
Length of rDNA contig:	8,464 bp	
Number of rDNA reads:	4,014	1,263,153
Copy number estimation:	878	1,600

Table S2: GC percentages in the 5.8S coding region and ITS-1, ITS-2 non-coding regions in various Monocotyledon. Genbank accession numbers for rDNA sequences used in this analysis are mentioned.

			Species	GenBank Accession numbers	GC%		
					ITS-1	5.8S	ITS-2
Hyacinthaceae			*Prospero autumnale*	KF873575	75.9	58.1	77.1
Poaceae	**Oryzoideae**		*Oryza sativa*	AF169230	72.7	58.1	79.1
	Pooideae		*Triticum aestivum*	AY346120	62.2	60.1	61.1
			Brachypodium sylvaticum	AJ608155	62.8	59.0	68.5
			Catabrosa werdermannii	EU792333	59.2	57.9	57.5
			Brachypodium distachyon	L11578	64.1	59.8	67
	Aristidoideae		*Aristida gypsophila*	GU359267	69.5	56.7	71.4
	Panicoideae		*Panicum virgatum*	AM404348	54.2	55.8	59.9
			Sorghum nitidum	DQ131134	66.3	57.3	68.5
			Zea mays subsp. Mexicana	AF019817	70.4	56.7	73.2
			Digitaria sanguinalis	AM404347	58.2	55.8	64.6
			Chasmanthium latifolium	GU359319	70.9	56.1	74.2
	Chloridoideae	Centropodieae	*Ellisochloa rangei*	JQ3455167	70.7	57.9	75.0
		Triraphideae	*Neyraudra reunaudiana*	GU359124	67.3	56.7	67.0
		Eragrostideae	*Cottea pappophoroides*	GU359237	58.8	56.1	58.1
		Cynodonteae	*Eragrostiella leioptera*	GU359305	54.3	51.8	57.1
		Zoysieae	*Zoysia japonica*	GU359196	56.1	54.3	52.2
			Sporobolus ioclados	KM010430	50.0	52.4	50.0
			Calamovilfa gigantea (Sporobolus curtissianus)	KM010317	57.4	53.0	49.8
			Sporobolus heterolepis	KM010426	51.6	52.4	50
			Spartina gracilis	AF019844	57.1	51.8	47.9
			Spartina patens	AJ489795	56.2	51.8	49.5
			Spartina maritima	KT874468	53.5	52.7	49.5

PARTIE B: Détection d'haplotypes à partir d'assemblages de novo de données de pyroséquençage Roche-454.

I. Détection d'haplotypes pour 4 gènes d'intérêt :

Le programme « PyroHaplotyper » a été appliqué dans un premier temps sur 4 gènes d'intérêt issus des premiers transcriptomes de référence de Spartines hexaploïdes. Ces travaux ont été publiés dans la revue *Heredity* (Ferreira de Carvalho et al. 2012).

Dans les génomes hexaploïdes de *S. maritima* et *S. alterniflora*, jusqu'à trois paires de copies (plus ou moins divergentes) peuvent être attendues selon le degré de rétention des copies dupliquées (Fortune et al. 2007). Quatre alignements correspondant à des gènes d'intérêts (une phosphoénolpyruvate carboxykinase (PEPCK), un domaine HECT, un domaine codant pour une boîte homéotique et une protéine de choc thermique (HSP)) ont été analysés pour identifier des séquences homéologues à partir des différents sites polymorphes et des haplotypes reconstruits. Au sein de ces alignements, trois à quatre haplotypes exprimés peuvent être détectés pour chaque gène des deux espèces étudiées. L'analyse du polymorphisme est illustrée dans le Tableau 6 pour une région de 200 bp du gène codant pour le domaine HECT. Au sein de cette fenêtre, 7 haplotypes (addition des haplotypes des 2 espèces) ont été alignés. Six polymorphismes ont été identifiés, quatre sont communs à *S. maritima* et *S. alterniflora* et deux sont spécifiques à chaque espèce. Les polymorphismes communs permettent de distinguer deux haplotypes divergents (où les six sites polymorphes sont différents) présents chez les deux espèces et un (chez *S. maritima*) ou deux (chez *S. alterniflora*) moins divergents (présentant un à deux nucléotides de différence). Bien que le nombre de sites polymorphes détectés au sein des différents alignements soit variable, le même schéma est retrouvé : nous identifions pour chaque espèce deux haplotypes divergents et un ou deux haplotypes moins divergent au sein de chaque alignement.

Tableau 6 : Présentation des polymorphismes ayant permis l'assemblage des reads au sein de l'alignement du gène codant pour le domaine HECT chez *S. maritima* et *S. alterniflora* (Table 3; Ferreira de Carvalho et al. 2012).

HECT domain-containing protein, expressed

S. alterniflora contig 03059 (length = 3961, reads = 85)						
Nucleotide position	1034	1085	1100	1119	1130	1167
Haplotype 1	C	T	C	A	A	T
Haplotype 2	T	T	T	A	A	C
Haplotype 3	T	C	T	A	A	C
Haplotype 4	C	T	C	G	C	T
S. maritima contig 02799 (length = 4294, reads = 127)						
Nucleotide position	1344	1352	1371	1382	1401	1419
Haplotype 1	C	C	G	C	C	T
Haplotype 2	T	C	G	C	C	T
Haplotype 3	T	T	A	A	A	C

Analysis of a 200-bp window, including two species-specific polymorphic sites (positions 1304, 1085 in *S. alterniflora* and positions 1344 and 1401 in *S. maritima*) and four polymorphic sites shared between the two species. These shared polymorphic sites are vertically aligned in the table.

II. Détection d'haplotypes pour l'ensemble des jeux de données transcriptomiques Roche-454 de cinq espèces de Spartines :

Nous avons par la suite appliqué le programme « PyroHaplotyper » sur l'ensemble des jeux de données transcriptomiques Roche-454 disponibles pour chaque espèce étudiée (*S. maritima*, *S. alterniflora*, *S. x townsendii*, *S. x neyrautii* et *S. anglica*). Les paramètres du programme ont été réglés de manière à ce que les contigs composés de moins de 10 reads et/ou de taille inférieure à 100 bp soient supprimés. Les quatre autres paramètres disponibles n'ont pas été modifiés et sont utilisés avec leurs valeurs par défaut (-t 20, -s 5, -d 2 et -n 1) lors de la détection de copies au sein de l'unité 45S de l'ADNr de *S. maritima*. Les nombres de contigs analysés pour chaque espèce (de taille supérieure à 100 bp, composé d'au moins 10 reads et présentant au minimum un SNP) sont présentés dans le Tableau 7.

Tableau 7 : Nombre total de contigs obtenus par espèce, nombre de contigs retenus pour les analyses avec leur taille moyenne et le nombre moyen de reads qui ont permis l'assemblage du contig (σ = écart-type et $IC_{95\%}$ = intervalle de confiance à 95%). Les contigs retenus pour les analyses correspondent aux contigs de taille supérieure à 100 bp, composé d'au moins 10 reads et présentant un SNP ou plus.

Espèce :	Nombre de contigs issus de l'assemblage:	Nombre de contigs retenus :	Taille moyennes des contigs retenus (en pb) :	Nombre moyen de reads par contig retenu :
S. maritima	19 380	13 712	966,21 (σ=804,73 ; $IC_{95\%}$=952,69 – 979,74)	80,26 (σ=487,39; $IC_{95\%}$=72,07 – 88,45)
S. alterniflora	8 964	5 699	1292,05 (σ=1046,82 ; $IC_{95\%}$=1264,77 – 1319,33)	106,80 (σ=626,32 ; $IC_{95\%}$=90,48 – 123,12)
S. x townsendii	6 874	3 933	834,81 (σ=976,45 ; $IC_{95\%}$=804,14 – 865,47)	71,65 (σ=274,10 ; $IC_{95\%}$=93,04 – 80,26)
S. x neyrautii	5 673	2 772	863,02 (σ=679,94 ; $IC_{95\%}$=837,47 – 888,56)	93,36 (σ=648,65 ; $IC_{95\%}$=68,99 – 117,74)
S. anglica	3 610	1 508	718,17 (σ=701,26 ; $IC_{95\%}$=682,30 – 754,04)	206,95 (σ=942,31 ; $IC_{95\%}$=158,75 – 255,16)

i) Détection de polymorphismes :

Nous avons analysé le nombre de sites polymorphes détectés par le programme « PyroHaplotyper » pour les différentes espèces. Le nombre moyen de sites polymorphes pour les alignements est de 2,66 (pour *S. maritima)*, 2,75 (pour *S. alterniflora*), 3,12 (pour *S. x townsendii*), 3,19 (pour *S. x neyrautii*) et de 2,74 (pour *S. anglica*) SNPs pour 100 bp. Le nombre de polymorphismes observés chez les deux espèces hexaploïdes est similaire (test de Student ; p-value > 0,05). De même, nous avons mis en évidence un nombre similaire de SNPs chez les deux hybrides (test de Student ; p-value > 0,05) qui est, comme attendu, plus important que le nombre de SNPs détectés chez les parents (test de Student ; p-value < 0,001).

ii) Détection d'haplotypes :

Les alignements de séquences sélectionnés comptent en moyenne 5,27 (pour *S. maritima*), 5,84 (pour *S. alterniflora*), 5,84 (pour *S. x townsendii*), 5,32 (pour *S. x neyrautii*) et 4,80 (pour *S. anglica*) haplotypes de longueur comprise entre 150 et 950 bp. Le nombre moyen d'haplotypes par alignement obtenu pour les deux espèces *S. maritima* et *S. x*

neyrautii est similaire (test de Student ; p-value > 0,05). De même, le nombre moyen d'haplotypes par contig est similaire pour les deux espèces *S. x neyrautii* et *S. anglica* (test de Student ; p-value > 0,05). Le nombre moyen d'haplotypes détectés chez *S. alterniflora* et *S. x townsendii* est également similaire (test de Student ; p-value > 0,05). Les jeux de données étudiés ne permettent pas de dégager une tendance entre les 5 espèces, aucune espèce ne présentant un nombre moyen d'haplotypes exprimés statistiquement différents par rapport aux autres espèces. Les haplotypes construits chez l'espèce hexaploïde *S. alterniflora* sont de taille supérieure par rapport aux haplotypes obtenus chez les quatre autres espèces étudiées (test de Student ; p-value < 0,001). Les tailles des haplotypes construits chez les deux hybrides *S. x townsendii* et *S. x neyrautii* sont similaires (test de Student ; p-value > 0,05) (Figure 19).

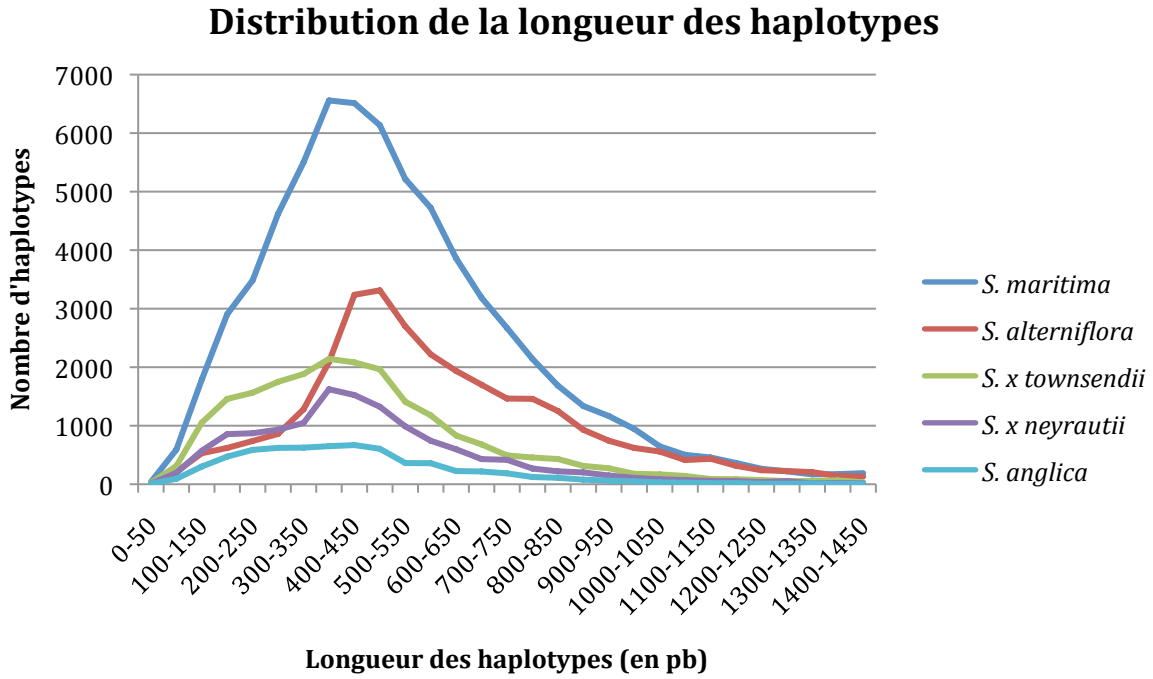


Figure 19 : Distribution de la longueur des haplotypes construits avec le programme "PyroHaplotyper" sur les données Roche-454 pour les cinq espèces étudiées.

Nous avons également calculé le nombre d'haplotypes pour chaque alignement local au sein de chaque espèce (Figure 20). Le nombre d'alignements présentant entre 2 et 4 haplotypes (représentant entre 70,03% et 81,70% de l'ensemble des alignements selon les espèces étudiées) d'une part et 5 ou 6 haplotypes d'autre part (représentant entre 10,34%

et 16,56% selon les espèces étudiées) présentent les mêmes proportions pour les cinq espèces étudiées (test exact de Fisher ; p-value > 0,05).

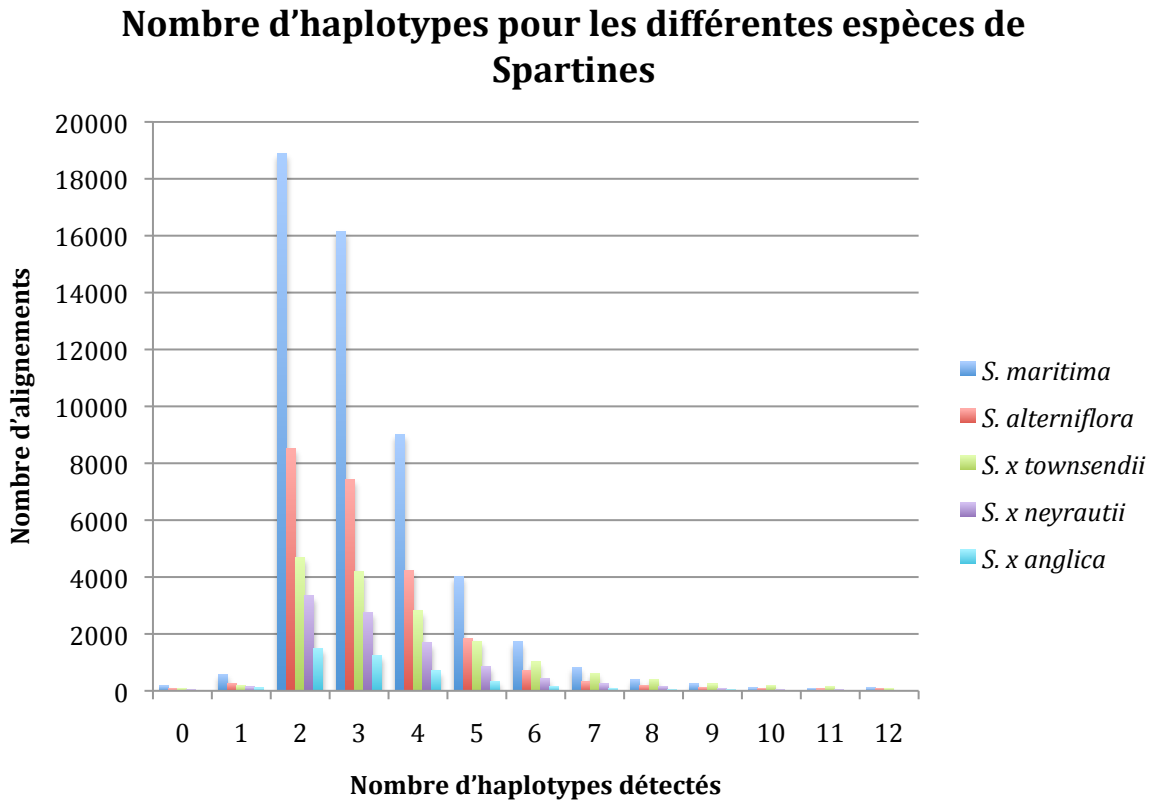


Figure 20 : Distribution du nombre d'alignements en fonction du nombre d'haplotypes détectés pour chaque espèce à partir des données Roche-454.

iii) Divergence des haplotypes détectés :

Nous avons comparé les divergences nucléotidiques obtenues entres les haplotypes pour les cinq espèces (nombre de nucléotides différents sur la plus longue distance commune entre les deux haplotypes traités) et nous observons le même type de données pour les 5 espèces étudiées. Nous pouvons observer un décalage du plateau (ou pic) de certaines courbes, les alignements présentant 5 ou 6 (et à un degré moindre 4) haplotypes contiennent des haplotypes moins divergents (Figure 21).

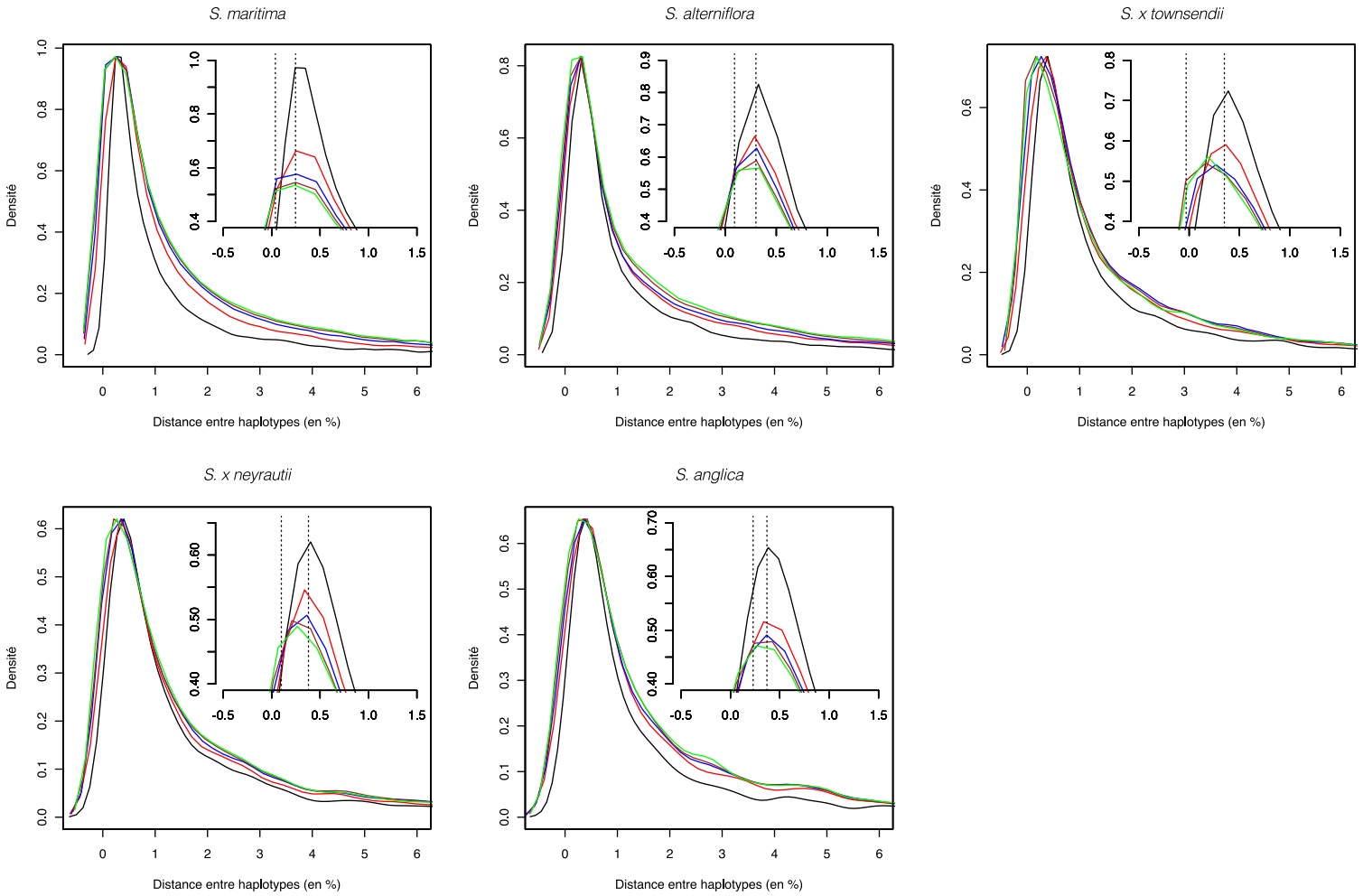


Figure 21 : Représentation de la divergence nucléotidique entre haplotypes (par paires) au sein de chaque espèce étudiée. Les courbes représentent la densité de distribution des distances au sein des alignements locaux présentant 2 (en noir), 3 (en rouge), 4 (en vert), 5 (en marron) ou 6 (en vert) haplotypes. Les graphiques situés en haut à droite représentent un agrandissement du graphique de densité. Les traits verticaux en pointillés représentent les débuts de pallier (ou pics) des différentes courbes.

iv) Détection des copies homéologues au sein des génomes hybrides et allopolyploïdes :

Après avoir reconstruit les différents haplotypes, nous avons identifié l'origine parentale de chaque haplotype au sein des espèces hybrides (*S. x townsendii* et *S. x neyrautii*) et de l'espèce allododécaploïde *S. anglica* en utilisant les polymorphismes communs aux différentes copies dupliquées (voir le Chapitre 3. Partie II.5 pour la stratégie et le détail de la méthode utilisée). Ainsi, nous avons identifié l'origine des copies au sein de 1 314, 1 140 et 668 alignements correspondant respectivement aux jeux de données de *S. x townsendii*, *S. x neyrautii* et *S. anglica* (Tableau 8). Pour les trois espèces étudiées, le nombre de copies assignées à *S. maritima* d'une part, et *S. alterniflora* d'autre part est similaire (test exact de Fisher ; p-value > 0,05). Le nombre de copies assignées à l'espèce paternelle *S. maritima* représente 51 à 57% des copies, tandis que la proportion des copies assignées à *S. alterniflora* est comprise entre 38 et 43% selon les espèces. Le nombre de copies non assignées représente une faible proportion des haplotypes (de 4,5 à 5,4%).

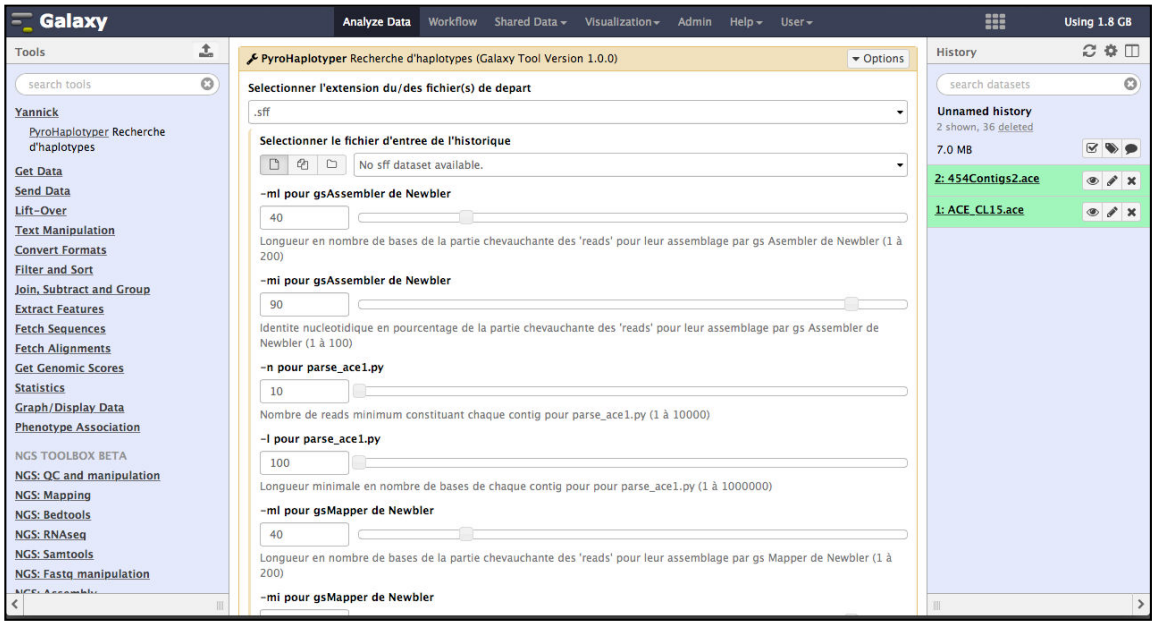
Tableau 8 : Identification de l'origine parentale des haplotypes présents chez les espèces hybrides et allopolyploïde pour un sous ensemble de contigs. Les copies sont non assignées si les copies parentales sont identiques ou si l'une des copies parentales est absente de l'alignement.

			Assignation des haplotypes hybrides et allopolyploïdes :		
	Nombre de contigs (et fenêtres) :	Nombre total d'haplotypes hybrides :	*S. maritima*	*S. alterniflora*	Haplotypes non assignés
S. x townsendii	1 314 (6 237)	36 641	18 775	15 881	1 985
S. x neyrautii	1 140 (6 035)	31 723	17 571	12 734	1 418
S. x anglica	668 (3 338)	15 952	9 069	6 093	790

PARTIE C : « PyroHaplotyper », un outil intégré sur Galaxy.

Les programmes développés ici permettent ainsi de détecter différentes copies dupliquées et d'analyser les génomes des plantes et en particulier les génomes des espèces polyploïdes pour lesquels nous ne disposons pas de références diploïdes. Ces programmes ont été dans un premier temps intégrés au sein d'un pipeline écrit en langage BASH appelé « PyroHaplotyper » (Namour 2015). Pour mettre à disposition « PyroHaplotyper » à la communauté scientifique, ce logiciel est en cours d'intégration comme outil sur le serveur Galaxy de la plateforme GenOuest en collaboration avec Yvan Le Bras (INRIA/IRISA, Genouest). Pour ce faire nous avons développé un pack contenant un fichier XML qui inclut l'ensemble des commandes nécessaires à la création de l'interface du logiciel ainsi que plusieurs descripteurs. En effet, cet outil prenant en entrée plusieurs formats de fichiers (« .ace », « .sff », « .fastq », «.fasta + .qual », « .fasta +.fasta ») il a été nécessaire de développer un descripteur spécifique à chaque type d'entrée de fichiers, ce qui permet d'enchainer les opérations nécessaires pour détecter les différentes copies. Ce pipeline dispose de nombreux paramètres réglables par l'utilisateur : les longueurs de chevauchements pour l'assemblage des reads, l'identité nucléotidique, la longueur des contigs, le nombre de reads les composants, le seuil d'identification d'un SNP, la profondeur, le trimming ou le nombre de SNPs nécessaires pour l'assemblage (Figure 22).

a



b

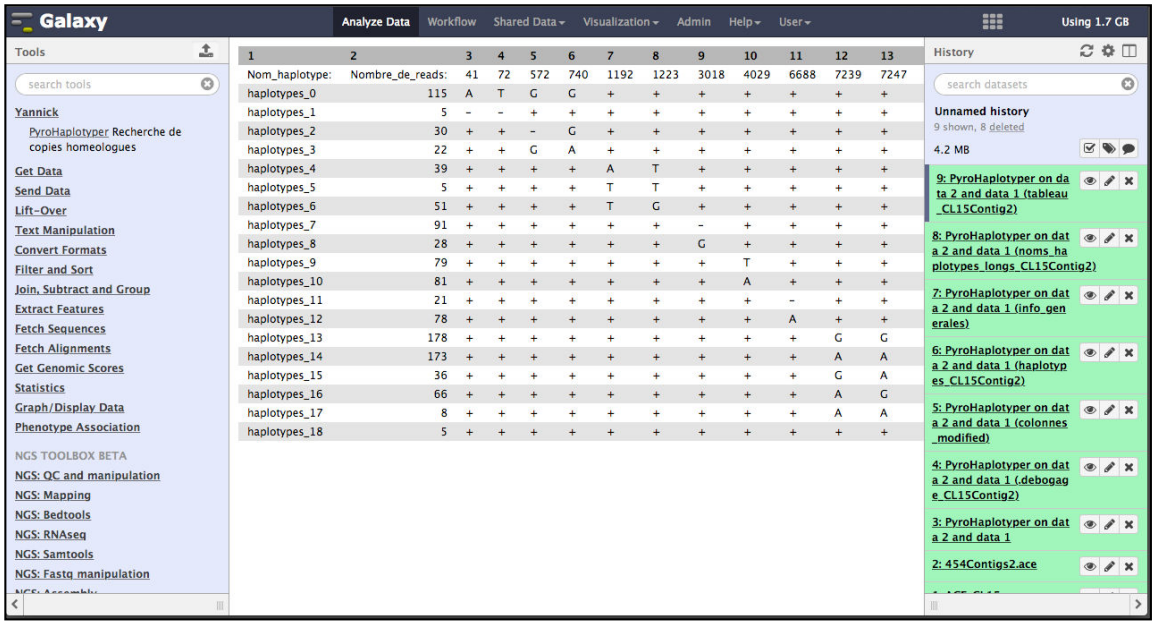


Figure 22 : Captures d'écran de l'interface Galaxy du logiciel « PyroHaplotyper ». a) Capture d'écran des paramètres réglables du logiciel « PyroHaplotyper » lors de la sélection d'un fichier d'entrée au format « .ace ». A gauche, le menu présentant l'ensemble des outils disponibles; au milieu l'interface de l'outil; à droite l'historique avec les fichiers d'entrées et de sorties. Les deux premières valeurs concernent le logiciel gsMAPPER de Newbler (-ml et -mi) et les quatre valeurs suivantes (-t, -s, -d et -n) correspondent aux paramètres du programme de détection de SNPs et de reconstruction d'haplotypes. b) Capture d'écran d'un fichier de résultat généré par « PyroHaplotyper » dans l'historique de l'interface de Galaxy. Les fichiers créés apparaissent dans la partie historique à droite. Le fichier de sortie présente les différents haplotypes construits ainsi que les SNPs détectés sous la forme d'un tableau.

Le logiciel « PyroHaplotyper » a été testé et validé au cours de cette thèse pour des données de séquençage de type Roche-454 ou Sanger qui génèrent des fragments longs. Les jeux de données NGS des nouvelles technologies de séquençage de type Nanopore (Wang, Yang, and Wang 2015) ou SMRT (Eid et al. 2009) par exemple, qui génèrent des reads de grandes tailles (> 1 kb) peuvent également être traités par cet outil.

Ce logiciel peut également être utilisé en génétique des populations pour détecter différents allèles au sein d'un ensemble de séquences (générées par exemple par des approches de type « amplicons ») dans le but de mieux comprendre l'origine d'individus ou de prédire l'évolution des populations (Dufresne et al. 2014). Dans la perspective de proposer un outil générique, nous avons testé « PyroHaplotyper » sur des jeux de données de type amplicon chez des espèces animales : des Nématodes à kystes et le Gorille. En collaboration avec Cécile Gracianne et Eric Petit (UMR ESE, INRA, Rennes), nous avons appliqué ce programme sur des jeux de données provenant de différentes populations de deux nématodes (*Heterodera schachtii* et *Heterodera betae*) ravageurs des betteraves. Le but de cette approche était d 'identifier les différentes copies et allèles pour un ensemble de gènes d'intérêt de chaque population. Les données d'amplicons ont été séquencées pour environ 50 populations (échantillonnées de la région de Gibraltar au sud de la Suède). Les premiers résultats montraient une hétérogénéité de la profondeur de séquençage pour les différents gènes ciblés. De nouveaux amplicons ont été produits et sont en cours de séquençage. Nous avons également appliqué le logiciel « PyroHaplotyper » sur des jeux de données amplicon du groupe de gènes codants pour le Complexe Majeur d'Histocompatibilité (MHC) chez le Gorille (*Gorilla gorilla gorilla*). Ce travail, qui s 'inscrit dans le cadre d'une étude des populations des Gorilles touchés par Ebola, est une collaboration avec Alice Baudouin et Pascaline Le Gouard (UMR-CNRS Ecobio, Université de Rennes 1).

Discussion

L'application du programme « PyroHaplotyper » sur des reads longs issus de la technologie de pyroséquençage Roche-454 nous a permis de détecter différentes copies dupliquées au sein de jeux de données génomiques et transcriptomiques de Spartines.

Diversité des copies paralogues et diploïdisation de l'ADNr 45S :

L'analyse de l'unité 45S codant l'ARN ribosomique de *Spartina maritima* a mis en évidence un phénomène de diploïdisation au sein de ce génome hexaploïde. En effet, l'analyse cytogénétique moléculaire réalisée par FISH a mis en évidence la présence d'une paire de loci de l'unité 45S chez *S. maritima* contre 3 paires attendues pour un génome hexaploïde. Ce résultat est en accord avec différentes études menées chez les espèces polyploïdes où le nombre de locus d'ADNr peut être conservé (chez les polyploïdes récents ou anciens) ou peut diminuer de manière importante par diploïdisation (Leitch and Bennett 2004).

Nous avons également montré que les longueurs des régions codantes et des ITS de *S. maritima* sont similaires à celles du riz et du maïs. Néanmoins, le pourcentage en GC des deux ITS et de l'unité codante 5.8S est plus faible en comparaison avec les autres Poaceae ; ces régions présentent une variabilité interspécifique importante. L'espaceur externe transcrit (5'-ETS) que nous avons pu détecter chez *Oryza sativa* (966 bp) a une longueur similaire à l'espaceur de *Zea mays* (825 bp) identifié par Toloczyki et Feix (1986) alors que celui identifié chez *S. maritima* présente une longueur deux fois supérieure (1 754 bp). Ces résultats sont en accord avec la littérature qui montre que la taille des 5'-ETS peut être variables au sein d'un même genre (Volkov et al. 1996), et donc entre les espèces de Poaceae.

Une perspective intéressante de ce travail sera d'étudier l'unité 45S des autres espèces de Spartines tétraploïdes et hexaploïdes afin d'identifier le nombre de loci de chaque espèce. Il sera alors possible de savoir si la perte de locus de l'unité 45S est propre à

S. maritima, aux espèces hexaploïdes ou à l'ensemble de ce clade. Il sera également intéressant de déterminer le nombre de loci présents au sein du génome de *S. anglica* pour voir si ce nombre correspond à une addition du nombre de loci des parents comme cela a pu être observé chez des espèces allopolyploïdes comme le blé ou les *Tragopogon* (Badaeva, Friebe, and Gill 1996; El-Twab 2007; Kovarik 2005) ou si l'hybridation et/ou la duplication du génome a entrainé une perte de loci comme cela a pu être observé chez les fraisiers ou le colza par exemple (Liu and Davis 2011; Snowdon, Köhler, and Köhler 1997).

La majorité des différents sites polymorphes et copies dupliquées identifiés au sein de ces jeux de données ont été validés à l'aide de données génomiques issues : 1) de clonage et re-séquençage par la méthode Sanger et 2) de séquençage par synthèse Illumina. Un total de 29 sites polymorphes (sur 34) et 11 haplotypes (sur 20) ont été retrouvés et confirmés ce qui permet de valider le programme « PyroHaplotyper ». A l'aide de ce programme, nous avons mis en évidence l'homogénéité des régions codantes et des espaceurs internes transcrits (ITS) qui semblent donc plus particulièrement concernés par l'évolution concertée. Ceci confirme l'intérêt de ces régions pour les phylogénies interspécifiques puisqu'elle représente des marqueurs peu variables au niveau intra individuel, mais discriminants entre les organismes. En revanche, une importante variabilité intra-génomique a été détectée dans les régions contenant l'espaceur intergénique (IGS) et l'espaceur externe transcrit (ETS).

Détection d'haplotypes et recherche de copies homéologues au sein des jeux de données transcriptomiques (Roche-454) :

Une détection d'haplotypes a tout d'abord été réalisée sur 4 gènes d'intérêts dans le cadre de l'étude des premiers transcriptomes de référence des Spartines hexaploïdes (Ferreira de Carvalho et al. 2012). Il a été ainsi possible d'identifier jusqu'à 4 haplotypes (dont deux plus divergents) par gène et par espèce ce qui pourrait potentiellement indiquer la présence de deux copies homéologues exprimées par locus.

Nous avons ensuite analysé les haplotypes détectés sur l'ensemble des jeux de données Roche-454 disponibles au sein du laboratoire pour les cinq espèces étudiées ; soit respectivement 13 712 contigs et 5 699 contigs pour les deux espèces parentales *S. maritima* et *S. alterniflora*, 3 933 contigs et 2 772 contigs pour les hybrides *S. x townsendii* et *S. x neyrautii* et 1 508 contigs pour l'allopolyploïde *S. anglica*. Nous avons détecté pour chacun de ces contigs, un nombre moyen de sites polymorphes variant de 2,75 à 3,19 SNPs pour 100 bp. Nous pouvons constater que les deux hybrides présentent un nombre plus important de SNPs que les deux parents hexaploïdes et l'allododécaploïde. Comme attendu pour des espèces hybrides, les deux espèces *S. x townsendii* et *S. x neyrautii* présentent un polymorphisme intragénomique plus important. Le nombre de sites polymorphes pour 100 bp détecté chez *S. anglica* est moins important que chez les deux hybrides. Cela peut s'expliquer par la taille des jeux de données et le nombre de contigs traités chez l'espèce allopolyploïde (1 508 contigs contre 3 933 et 2 772 contigs chez les deux hybrides).

La détection des différents SNPs au sein des jeux de données nous a permis d'assembler différentes séquences polymorphes ou haplotypes. Pour l'ensemble des espèces étudiées les haplotypes construits à partir des jeux de données Roche-454 ont une longueur variant entre 200 et 1 100 bp. Nous pouvons constater que les haplotypes des deux espèces parentales hexaploïdes *S. maritima* et *S. alterniflora* présentent une taille plus importante due à la meilleure qualité des reads disponibles (d'une longueur moyenne de 286 bp pour *S. alterniflora* et 463 bp pour *S. maritima*).

Nous avons identifié le nombre d'haplotypes au sein de chaque alignement local et calculé la divergence nucléotidique entre ces différents haplotypes. Pour l'ensemble des espèces étudiées, la majorité des alignements locaux présentent entre 2 et 4 haplotypes. Les deux espèces hexaploïdes, elles-mêmes probablement d'origine hybride (Fortune et al. 2007), contiennent chacune trois génomes homéologues dupliqués plus ou moins divergents. Nous pouvons observer pour ces deux espèces que le nombre de copies exprimées varie entre 2 et 4 (et à un degré moindre 5 et 6), plusieurs hypothèses peuvent expliquer ces résultats : les trois copies homéologues et leurs allèles respectifs ne s'expriment pas toutes au sein du génome, ou alors nous sommes en présence de copies identiques. Pour les régions où nous avons détecté plus de 6 haplotypes, nous sommes probablement en présence des copies homologues (orthologues ou homéologues) et de

copies paralogues peu divergentes. Nous avons observé des pics de densité différents en comparant les différentes régions des alignements présentant de 2 à 6 haplotypes. Les alignements présentant 5 et 6 haplotypes présentent des séquences moins divergentes entre elles par rapport aux alignements présentant 2 et 3 haplotypes. Cette information pourrait indiquer que nous sommes en présence de copies peu divergentes (des allèles orthologues) et de copies plus divergentes, probablement homéologues.

Les différentes copies dupliquées détectées au sein des hybrides (*S. x townsendii* et *S. x neyrautii*) et de l'allopolyploïde *S. anglica* ont été assignées à l'un des deux parents hexaploïdes. Pour les trois espèces, un nombre similaire d'haplotypes assignés respectivement à *S. maritima* et *S. alterniflora* peut être observé. Le nombre de copies non assignées représente 4,5 à 5,4% des haplotypes. Le nombre d'haplotypes assignés à *S. maritima* est plus important pour les trois espèces, ces résultats peuvent s'expliquer par la nature des jeux de données initiaux. En effet, nous disposons de deux fois plus de reads issus de banques normalisées et non-normalisées pour l'espèce parentale *S. maritima* ce qui nous a permis de mieux reconstruire les différents haplotypes au sein des alignements.

Intégration de l'outil « PyroHaplotyper » sur Galaxy :

Le développement d'outils capables de détecter les copies dupliquées au sein des génomes polyploïdes hautement redondants permet d'analyser le devenir des copies dupliquées au sein de ces organismes complexes. Il est ainsi possible de mieux appréhender les événements de spéciation et d'étudier l'évolution fonctionnelle (contrôle épigénétique, mises sous silence, compartimentation de l'expression) et structurale (maintien, perte, accumulation de mutations, conversions géniques) des génomes redondants. La validation de ce programme à l'aide de plusieurs technologies et son application sur des jeux de données de différentes espèces animales et végétales nous permettent d'être confiants quant à la qualité des haplotypes détectés par le programme « PyroHaplotyper ». En effet, la détection d'haplotypes est un pré requis pour toute analyse phylogénétique, et dans le cas de l'utilisation de données NGS à plus large échelle, pour des approches de phylogénomique. De plus, ce logiciel devient le premier logiciel capable de détecter des haplotypes à partir de

données brutes de pyroséquençage Roche-454 à être intégré dans Galaxy. D'autres logiciels développés comme Hylite (Duchemin et al. 2014), PolyCat (Page, Gingle, and Udall 2013) et PolyDog (Page and Udall 2015) effectuent un travail similaire mais ne sont soit pas intégrables en tant qu'outils au sein de Galaxy (logiciels non libres de droit), soit ne détectent pas eux même les SNPs et ont besoin de bibliothèques de polymorphismes existantes. Les outils comme SniPloid (Peralta et al. 2013) ou PolyCat (Page, Gingle, and Udall 2013), sont principalement axés sur des organismes modèles tétraploïdes et utilisent des génomes diploïdes pour détecter les différentes copies dupliquées par polyploïdisation. De ce fait, ces logiciels ne sont pas appropriés pour la détection de copies chez des espèces à haut niveau de ploïdie.

Chapitre 5 :

Détection de SNPs et construction d'haplotypes à partir de données de séquençage Illumina.

Chapitre 5: Détection de SNPs et construction d'haplotypes à partir de données de séquençage Illumina.

Introduction et démarche générale :

Les nouvelles technologies de séquençage à haut débit (NGS) évoluent très rapidement depuis la commercialisation de la technologie de pyroséquençage Roche-454 (2005) aux techniques de séquençage de « Molécule Unique » (la technologie SMRT développé par Pacific Biosciences qui utilise la chimie P6-C4 en 2014 ; la technologie Minion de Nanopore en tests depuis 2015). De nouvelles technologies telles que la technologie de séquençage par synthèse permettent d'obtenir un nombre important de séquences de courtes lectures. La technologie « short read » Illumina par exemple, est capable de séquencer jusqu'à 1,2 milliard de reads pairés d'une longueur allant jusqu'à 250 bp (http://www.illumina.com). Néanmoins, il est nécessaire de développer de nouveaux outils adaptés à ces jeux de données pour construire de nouveaux transcriptomes de référence ou pour détecter les copies homéologues au sein des génomes hautement redondants. Pour répondre à cette dernière problématique, nous avons développé le logiciel « IlluHaplotyper », un outil de détection de polymorphismes et de reconstruction d'haplotypes correspondant à des copies dupliquées au sein d'alignements de reads de courts fragments.

La première partie de ce chapitre est consacré au test des différents paramètres du logiciel développé afin de déterminer les paramètres optimaux nous permettant de détecter les copies dupliquées présentes au sein des jeux de données.

Ce programme a ensuite été testé sur des alignements de séquences Illumina alignées au préalable sur des contigs Roche-454 précédemment obtenus (Chapitre 4). Nous avons alors comparé les polymorphismes détectés au sein des jeux de données de pyroséquençage (obtenus à l'aide de « PyroHaplotyper ») et Illumina (obtenus avec « IlluHaplotyper ») pour un sous-ensemble de contigs (Partie B). Les polymorphismes et haplotypes détectés par ce programme ont pu être validés à l'aide de données de clonage

(ADN et ADNc) obtenues pour plusieurs gènes d'intérêt étudiés au sein du laboratoire (Partie C ; Fortune et al. 2007; Julie Ferreira de Carvalho 2013).

Nous avons assemblé cinq nouveaux transcriptomes de référence qui ont été obtenus à partir de données Roche-454 et Illumina pour les espèces *S. maritima*, *S. alterniflora*, leurs hybrides (*S. x townsendii* et *S. x neyrautii*), et l'allopolyploïde *S. anglica*. Ces transcriptomes enrichissent les premiers transcriptomes de référence construits à partir de données de séquençage Roche-454 pour les espèces parentales hexaploïdes *S. maritima* et *S. alterniflora* (Ferreira de Carvalho et al. 2012). L'application du pipeline « IlluHaplotyper » nous a permis de détecter des haplotypes au sein des jeux de données des différentes espèces étudiées ; les différentes copies des espèces hybrides et allopolyploïde ont ensuite été assignées à l'un des parents hexaploïdes (Partie D). Nous avons pu mesurer la divergence entre les différents haplotypes (comparaisons des taux de substitutions synonymes), et identifié les « pics » de divergence résultants de différents évènements de duplications génomiques. Ces résultats sont inclus dans un article en préparation « Reference transcriptomes and detection of duplicated copies in hexaploid parents, hybrids and allododecaploid *Spartina* species (Poaceae) » présenté ci-dessous.

PARTIE A : Etude de l'impact des paramètres du logiciel « IlluHaplotyper » sur un jeu de données :

Le logiciel « IlluHaplotyper » développé ici permet de détecter les différents sites polymorphes et reconstruit les haplotypes au sein d'alignements de courts fragments. Ce logiciel dispose de 4 paramètres ajustables (seuil de détection de SNPs, profondeur de séquençage, nombre de SNPs pour assembler un haplotype, « trimming »), ce qui permet à l'utilisateur d'adapter le programme en fonction du jeu de données traité.

I- Paramètres de mapping :

Nous avons cherché dans un premier temps à déterminer dans quelle mesure les paramètres d'alignement de séquences (par mapping) pouvaient avoir un impact sur les

résultats. Plusieurs logiciels de « mapping » ont été testés pour réaliser des alignements de reads peu stringents (80 et 90% d'identité) afin de co-aligner les reads provenant de différentes copies dupliquées. Ces analyses ont été réalisées sur le nouveau transcriptome de référence (Partie D) de *S. maritima* composé de 60 644 contigs et sur lequel nous avons aligné par mapping l'ensemble des reads Illumina de *S. maritima* (Tableau 9) à l'aide des logiciels Bowtie 2 (Langmead and Salzberg 2012), BWA (Li and Durbin 2009), SOAP2 (Li et al. 2009) et Novoalign (www.novocraft.com). Les résultats présentés dans le Tableau 9, permettent de montrer que le logiciel Bowtie 2 aligne significativement plus de reads divergents (au moins 8,52% de plus que le deuxième meilleur logiciel) sur le transcriptome de référence par rapport aux autres logiciels testés (BWA, SOAP2 et Novoalign). De plus, une diminution du pourcentage d'identité de mapping de ce logiciel n'entraine pas une augmentation significative du nombre de reads alignés (38,49% contre 37,46% de reads alignés sur le transcriptome de *S. maritima* pour des mapping respectivement à 80% et 90% d'identité).

Tableau 9 : Nombre de reads alignés sur le transcriptome de référence de *S. maritima* (60 644 contigs) en fonction des logiciels et paramètres testés. Le nombre initial de reads alignés est de 76 985 267. * : le pourcentage minimum de mapping du logiciel SOAP2 est de 96%.

Logiciel testé:	Paramètres de Mapping:	Formule:	Valeur de mapping:	Nombre de reads alignés:	Pourcentage de reads alignés:
Bowtie 2	G, A52, B8	$f(x) = A + B * ln(x)$ où x = longueur des reads	min id: 89,46%	28 837 359	37,46%
	G, A52, B6		min id: 80,1%	29 632 592	38,49%
BWA mem	Par défaut (A=1, B=4)	0,75*exp(-log(4)*B/A)	max mismatch: 0,067	23 076 210	29,97%
	A=0.8, B=4		max mismatch: 0,353	0	0,00%
	A=2, B= -11		max mismatch: 20,57	22 081 143	28,68%
BWA aln	n22	-	max mismatch: 22%	0	0,00%
SOAP2	Par défaut	-	*	16 039 372	20,83%
	v20			16 039 372	20,83%
Novoalign	Par défaut	-	de 0 à 99	20 177 881	26,21%
	t80		de 0 à 80	16 069 670	20,87%

II- Paramètres de détection de polymorphismes et d'haplotypes :

Nous avons ensuite étudié la sensibilité et les répercussions directes des outils et paramètres du logiciel sur les jeux de données pour déterminer la valeur optimale de chaque paramètre et les valeurs critiques de chaque option. Une partie de ce travail a été réalisée par Delphine Giraud, au cours de son stage dans le cadre du Master 1 Bio-Informatique et Génomique (BIG, Université de Rennes 1) encadré au cours de cette thèse (Giraud 2015). Nous avons choisi d'évaluer les différents paramètres du logiciel « IlluHaplotyper » sur des alignements de séquences obtenus à 90% de similarité *via* Bowtie 2. En combinant les différentes valeurs des paramètres testés, le logiciel « IlluHaplotyper » a été appliqué 60 fois (Tableau 10) sur un sous-ensemble (18 861 contigs) du nouveau transcriptome de référence de *S. maritima*.

Tableau 10 : Valeurs des différents paramètres testés au cours de l'étude. Les * indiquent les valeurs des paramètres par défaut.

Paramètres :	Option :	Valeurs :					
Seuil	-t	2%*	5%	8%	10%	20%	25%
Profondeur	-d	10	20	30*	40		
Nb SNPs pour assembler	-n	2*	3	4			
Trimming	-s	120 pb*					

i) Seuil de détection des SNPs :

A partir des fichiers de sortie du programme, nous avons observé l'impact du paramètre « seuil » (« threshold ») sur les résultats. Ce paramètre permet de fixer un seuil de détection des polymorphismes ; ainsi, pour une position donnée, si deux nucléotides différents sont présents à plus de n% (où n correspond à la valeur du seuil), alors le logiciel considère le SNP comme valide. Les résultats obtenus montrent une **diminution du nombre de SNPs détectés par alignement lorsque le seuil augmente**. En effet, avec un seuil à 2% le logiciel détecte 14,85 SNPs en moyenne tandis qu'avec un seuil à 25%, 6,88 SNPs sont détectés (Figure 23, a). En revanche, le nombre de reads utilisés par le programme pour détecter les différents SNPs et reconstruire les haplotypes est similaire pour l'ensemble des

valeurs du seuil (Figure 23, b ;~173,73 reads utilisés pour chaque alignement en moyenne (σ = 332,38 ; $IC_{95\%}$ = 173,10 – 174,38)). Le nombre moyen d'haplotypes est de 7,74 avec le seuil réglé à 2% puis diminue progressivement jusqu'à atteindre une valeur de 4,03 avec un seuil à 25% (Figure 23, c).

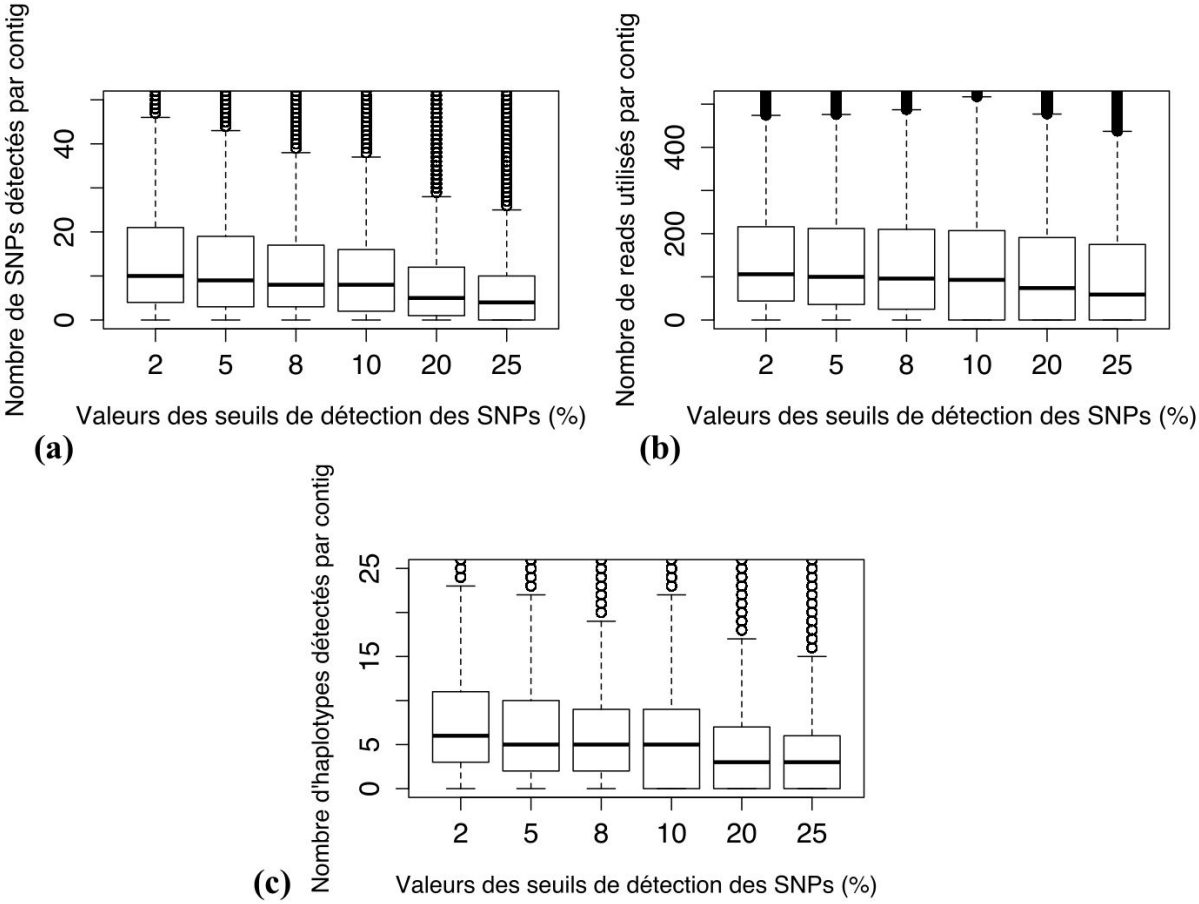


Figure 23 : Nombre de SNPs détectés (a), de reads utilisés (b) et d'haplotypes reconstruits (c) en fonction des valeurs du paramètre « seuil » défini.

ii) Profondeur de détection des SNPs :

Nous avons ensuite étudié l'effet du paramètre « profondeur » (« depth ») sur les résultats. Le logiciel va rechercher les SNPs uniquement dans les régions où la profondeur en reads est supérieure à la valeur indiquée par ce paramètre. **Le nombre de SNPs détectés par le logiciel diminue au fur et à mesure que la profondeur fixée augmente**. En moyenne,

16,89 SNPs par alignement sont détectés par le logiciel avec un paramètre de profondeur égal à 10 alors qu'avec une profondeur fixée à 50, 7,61 SNPs par alignement sont détectés (Figure 24, a). Le nombre de reads utilisés par « IlluHaplotyper » pour construire les haplotypes est similaire quand la profondeur testée est égale à 10, 20 et 30 reads (~200,59 reads utilisés en moyenne ; σ = 346,72 ; $IC_{95\%}$ = 198,57 – 202,61) mais connaît une diminution lorsque la profondeur est égale à 50 (170,85 reads utilisée en moyenne ; Figure 24, b). De la même façon, le nombre d'haplotypes construits par le logiciel atteint un plateau avec des profondeurs de 10, 20 et 30 (~7,14 haplotypes sont détectés en moyenne par contig ; σ = 6,73; $IC_{95\%}$ = 7,10 – 7,18). Cette valeur diminue à 4,74 haplotypes lorsque la profondeur est réglée à 50 reads (Figure 24, c).

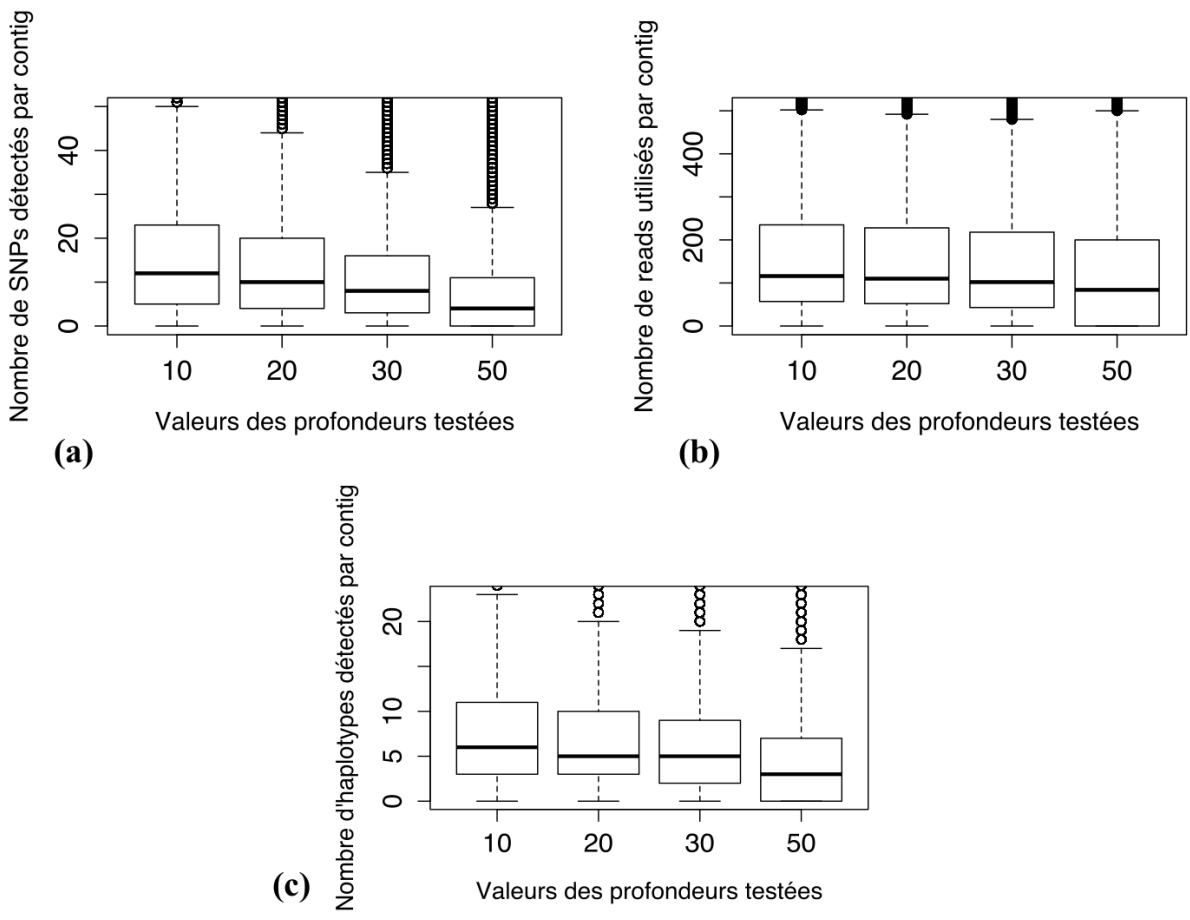


Figure 24 : Nombre de SNPs détectés (a), de reads utilisés (b) et d'haplotypes reconstruits (c) en fonction des valeurs du paramètre « profondeur » défini.

iii) Nombre de SNPs pour assembler les haplotypes :

L'analyse de l'impact du **paramètre fixant le nombre de SNPs minimal pour assembler** les reads entre eux (« nombre de SNPs pour assembler les haplotypes » ou « Number of shared SNPs») ne montre pas d'influence sur le nombre de SNPs détectés, comme attendu, la détection des sites polymorphes étant indépendante de ce paramètre. Pour chaque alignement, 11,17 SNPs sont détectés en moyenne, quelle que soit la valeur du paramètre (Figure 25, a ; σ = 11,52 ; $IC_{95\%}$ = 11,11 – 11,22). Une variation du nombre de reads utilisés et du nombre d'haplotypes construits était attendue, ce paramètre ayant un impact direct dans la construction des haplotypes. Néanmoins, **le nombre de reads utilisés et le nombre d'haplotypes** formés par le logiciel « IlluHaplotyper » **restent constants** pour les 3 valeurs du paramètre testées, en moyenne 173,45 reads sont utilisés (Figure 25, b ; σ = 331,68 ; $IC_{95\%}$ = 172,33 – 174,58) et 5,94 haplotypes sont construits (Figure 25, c ; σ = 6,63; $IC_{95\%}$ = 5,92 – 5,96). Ce paramètre ne semblant pas jouer sur les résultats d'un point de vue global, nous avons analysé son impact alignement par alignement. Cette étude nous a permis d'observer que la totalité des alignements étudiés ont une valeur de SNPs détectés et utilisés identique quelle que soit la valeur du paramètre testé. En comparant les différents nombres d'haplotypes construits par le logiciel, nous pouvons constater que 88% des contigs ont un nombre d'haplotypes construits égal quelle que soit la valeur de ce paramètre.

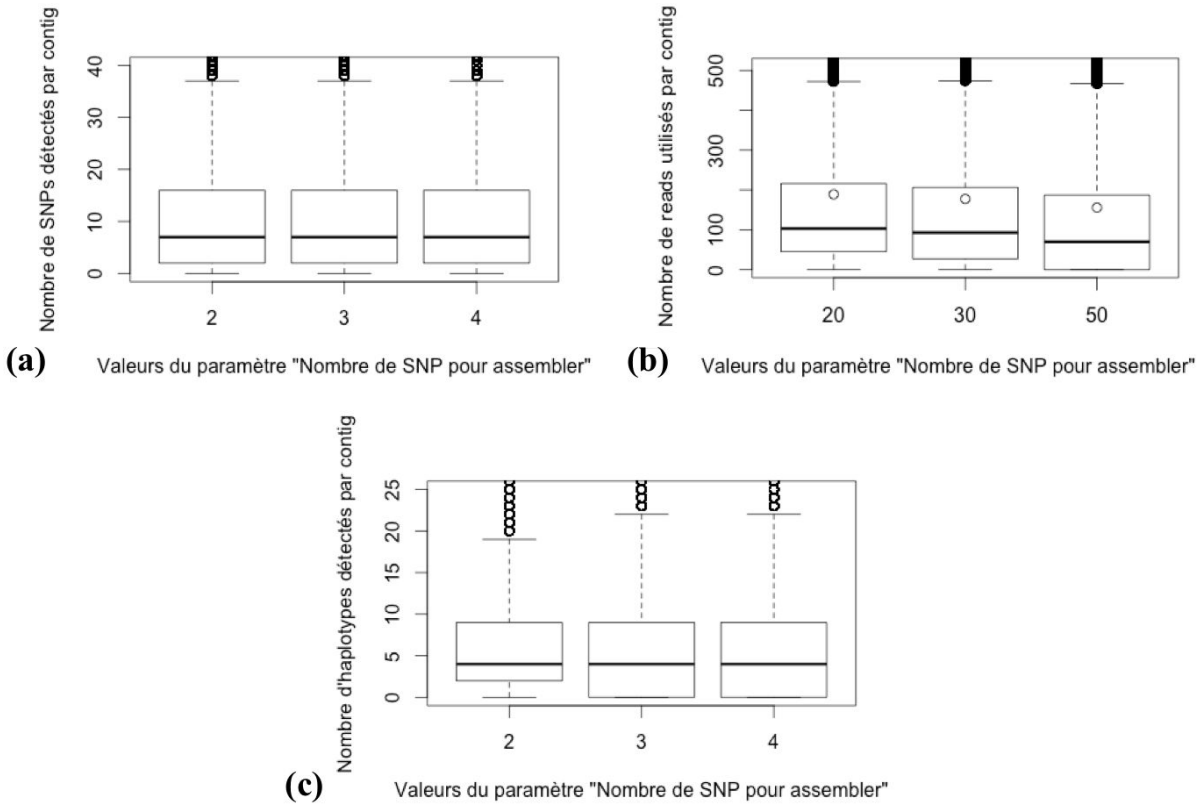


Figure 25 : Nombre de SNPs détectés (a), de reads utilisés (b) et d'haplotypes reconstruits (c) en fonction des valeurs du paramètre « Nombre de SNPs pour assembler » défini.

iv) Impact des paramètres sur le temps de calcul :

Nous avons également cherché à étudier l'impact des paramètres sur le temps de calcul. Pour cela, nous avons évalué le temps de calcul du programme pour chaque paramètre testé (Figure 26). Les courbes représentant les différents paramètres de seuil testés ont des profils similaires. Le temps total d'exécution est en moyenne de 33h31min (σ = 11h14min ; $IC_{95\%}$ = 24h32min – 42h30min ; Figure 26, a). Une distribution similaire pour les 3 valeurs du paramètre « nombre de SNPs pour assembler les reads » avec un temps global d'exécution de 60h16min en moyenne peut également être observé (σ = 49min ; $IC_{95\%}$ = 59h21min – 61h11min ; Figure 26, c). **Le paramètre « profondeur » apparaît comme le paramètre qui influence le plus les temps de calcul** (Figure 26, b). En effet, si la distribution du temps CPU (Central Processing Unit) est similaire pour les valeurs 20, 30 et 50, le profil de la courbe pour la valeur 10 est très différent. Ceci est confirmé par le temps total d'exécution qui varie fortement en fonction de la valeur du paramètre « profondeur »

(41h57min pour une profondeur de 10 contre respectivement 25h57min, 21h41min et 17h35min pour une profondeur de 20, 30 et 50).

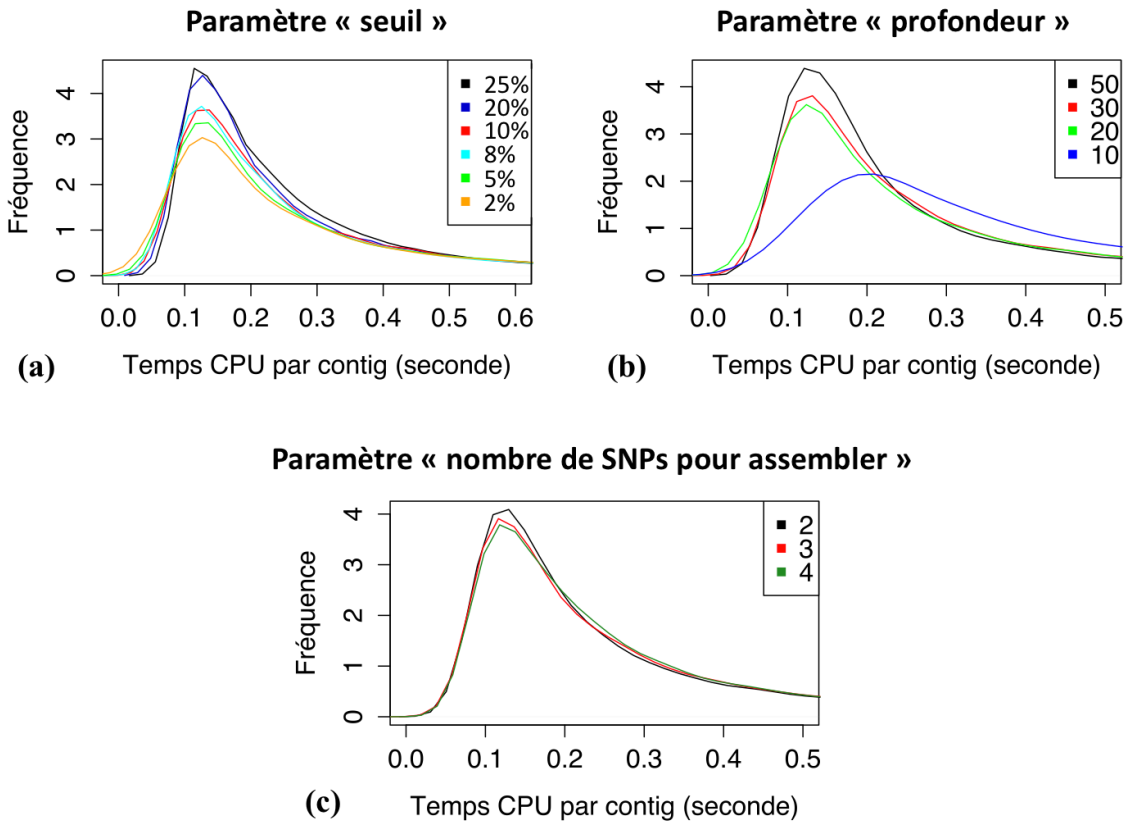


Figure 26 : Distribution des alignements en fonction du temps d'exécution pour chaque paramètre testé.

Afin de déterminer les paramètres qui influencent le plus les résultats, une Analyse en Composantes Principales (ACP) a été réalisée à l'aide de la librairie ade4 sous R (Dray et al. 2007). Cette analyse prend en compte 5 variables : les 3 paramètres testés, le temps de calcul CPU pour chaque contig et le nombre d'haplotypes détectés par alignement. Le plan factoriel a été réalisé en 3D, à l'aide la librairie rgl de R, afin de montrer la variabilité des données selon les 3 axes principaux de l'ACP.

La matrice des corrélations et un test de normalité ont été réalisés à l'aide d'un diagramme Quantile-Quantile (diagramme Q-Q). Les données ne suivant pas une loi Normale, la matrice des corrélations a été calculée grâce au coefficient de corrélation de Spearman. Nous pouvons observer une corrélation positive entre le nombre d'haplotypes et le temps de construction de ceux-ci pour chaque alignement (coefficient de corrélation de 0,71). Néanmoins le coefficient de corrélation n'étant pas assez élevé (inférieur à 0,9), nous avons conservé les deux variables lors de la réalisation de l'ACP.

Nous constatons d'après le cercle des corrélations que le temps et « nombre de SNPs pour assembler les haplotypes » sont liées sur un même axe ce qui est en adéquation avec la matrice de corrélation réalisée (Figure 27, a). D'après les contributions absolues et le cercle des corrélations, ces deux variables sont représentées par l'axe ayant le plus de variabilité (32,91%). Le plan factoriel, représentant les deux axes principaux de l'ACP, nous permet d'identifier trois nuages de points qui se distinguent par leurs valeurs du paramètre « nombre de SNP pour assembler les haplotypes » décrit par le deuxième axe (Figure 27, b). La représentation du nuage de points en 3D selon les axes 1, 2 et 3 nous permet de confirmer le regroupement des points par ce paramètre (Figure 27, c). Le graphique 3D nous permet également de visualiser le regroupement des contigs selon le paramètre « profondeur » (Figure 27, d).

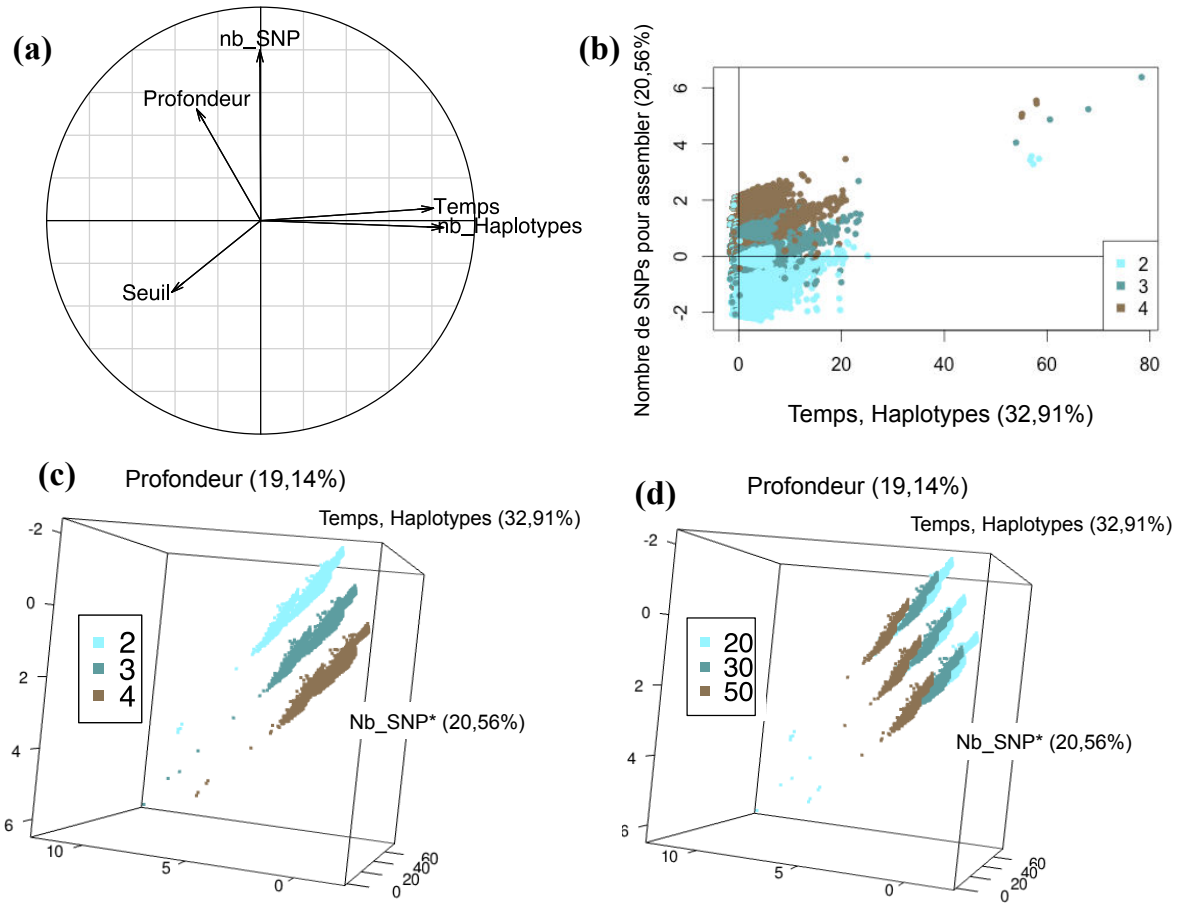


Figure 27 : Résultats de l'analyse en composante principale. (a) Cercle des corrélations entre les différentes variables. (b) Plan factoriel représentant les alignements selon les deux axes principaux en fonction du paramètre « nombre de SNPs pour assembler » réglé à 2, 3 ou 4. (c) Plan factoriel 3D représentant les alignements selon les axes 2 et 3 en fonction du paramètre « nombre de SNPs pour assembler ». (d) Plan factoriel 3D représentant les alignements selon les axes 2 et 3 en fonction du paramètre « profondeur » réglé à 20,30 ou 50 (le pourcentage de variance est indiqué sur les différents axes). *Nb_SNP correspond au paramètre « nombre de SNP pour assembler les haplotypes » (Giraud 2015).

v) Impact des paramètres sur la nature des duplicats détectés :

Pour étudier les haplotypes construits en modifiant les paramètres et déterminer les valeurs des paramètres pour lesquelles il est possible de détecter des allèles, des copies homéologues ou des copies paralogues, une étude phylogénétique a été réalisée pour un gène de *S. maritima* codant une « Pentatricopeptide repeat (PPR) superfamily protein ». Les haplotypes obtenus en modifiant le paramètre seuil (à 2, 5, 8, 10, 20 et 25%) ont été alignés. Une région de 217 bp présentant 81 sites variables a été sélectionnée pour réaliser une analyse phylogénétique à l'aide de la méthode du Maximum de Vraisemblance et du modèle de Kimura à deux paramètres (distribution Gamma). Un test de robustesse de 1 000

bootstraps a été réalisé et l'arbre a été enraciné à l'aide de plusieurs séquences de Poaceae identifiées par recherche d'homologie par BLASTn (Figure 28). Les séquences se distribuent dans deux clades A et B. Au sein du clade A, deux haplotypes (répartis dans deux sous-clades) ont été retrouvés pour l'ensemble des valeurs « seuils » testées (Figure 28, Clade A, copies 1 et 2). Ces deux copies A1 (hap_2) et A2 (hap_3) qui présentent 7 SNPs différents, pourraient correspondre à deux copies homéologues. Au sein du clade B, les haplotypes détectés ne sont retrouvés que lorsque le paramètre seuil est inférieur ou égal à 10%. Sachant que la diminution de la valeur du seuil permet de détecter des polymorphismes plus anciens, les copies A et B pourraient résulter d'une duplication ancienne. Plusieurs hypothèses pourraient être envisagées pour expliquer cette topologie : les trois clades B, A1 et A2 pourraient résulter de deux duplications génomiques successives (tétraploïdie et hexaploïdie) représentant ainsi trois groupes d'homéologues. Alternativement, le clade B pourrait résulter d'une duplication individuelle plus ancienne (paralogue). Le présent jeu de données ne permet pas de trancher entre ces hypothèses. De plus, s'agissant de données transcriptomiques, toutes les copies ne sont peut être pas exprimées.

En conclusion, il a été possible de déterminer les paramètres optimaux du logiciel « IlluHaplotyper » à appliquer sur des jeux de données Illumina. Les différentes analyses ont mis en évidence l'influence des paramètres « seuil » et « profondeur » sur le nombre de SNPs détectés, de reads utilisés pour construire les haplotypes, le nombre d'haplotypes construits et sur le temps de calcul CPU. Notre étude a également montré que le paramètre « nombre de SNPs pour assembler les reads entre eux » n'influençait pas les résultats, ce qui doit être dû à la taille des reads traités et à la longueur, de ce fait contrainte, des fenêtres glissantes utilisées pour l'assemblage de ces reads. Ces contraintes permettent également de réduire les assemblages de reads chimériques. Ces résultats ont été confirmés par le biais d'une analyse en composantes principales, ce paramètre n'a donc pas d'impact sur le nombre d'haplotypes construits et sur le temps de calcul du programme. Il est à présent possible d'appliquer le logiciel « IlluHaplotyper » sur l'ensemble des jeux de données transcriptomiques de Spartines disponibles au sein du laboratoire en utilisant les paramètres les plus adaptés à la recherche de copies homéologues. Les trois paramètres « seuil », « profondeur » et « nombre de SNPs pour assembler les reads » ont donc été réglés à 2%, 30

et 2 respectivement pour la suite des analyses. Ces valeurs se présentant comme les meilleurs paramètres pour l'analyse de nos jeux de données.

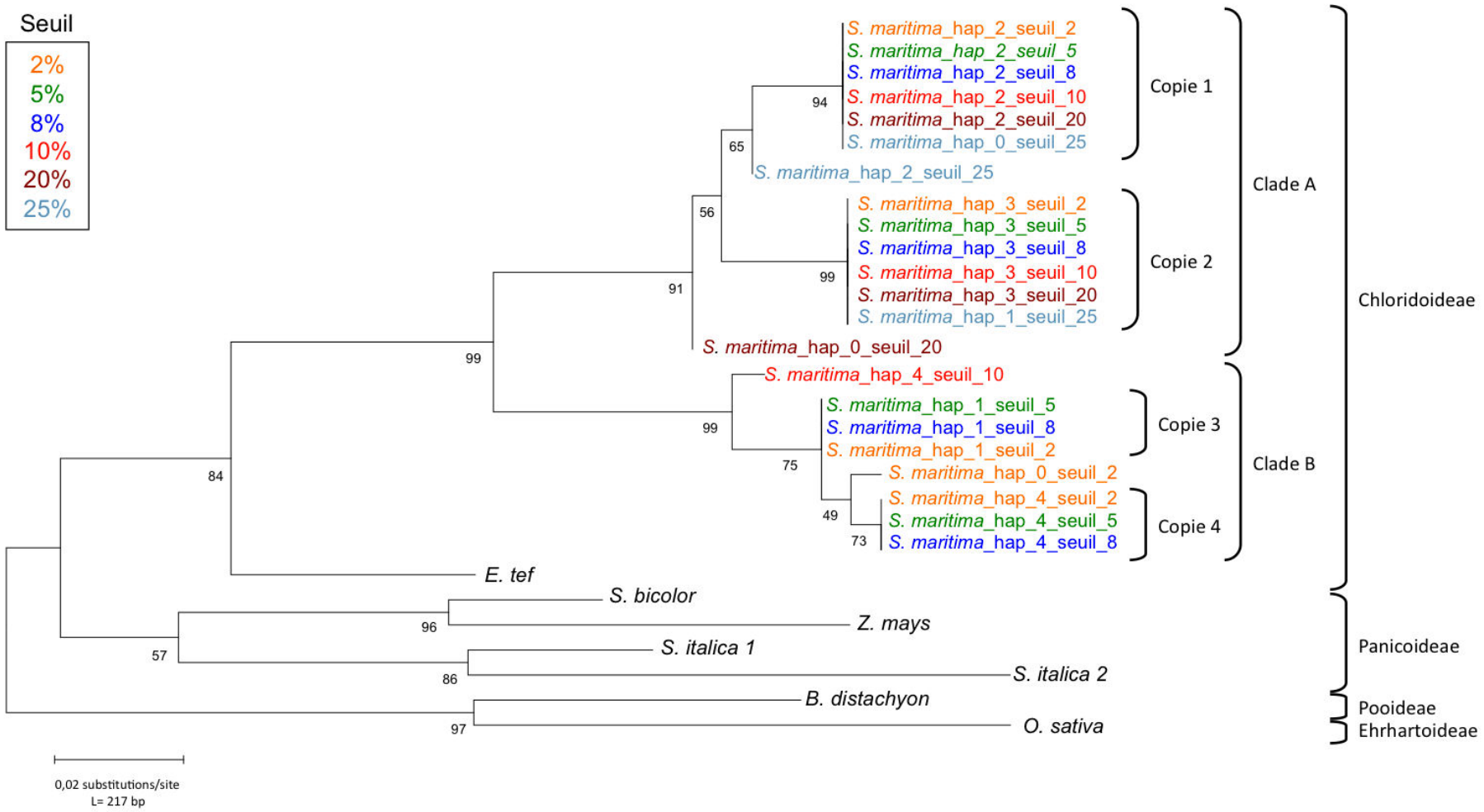


Figure 28 : Arbre phylogénétique réalisé à l'aide de la méthode du Maximum de Vraisemblance (modèle de Kimura à 2 paramètres, distribution Gamma) pour un gène de *S. maritima* codant une « Pentatricopeptide repeat (PPR) superfamily protein ». L'arbre a été enraciné avec des séquences de Panicoideae, Pooideae et Ehrhartoideae. Les accolades A et B indiquent un événement de duplication chez *S. maritima*. Les différentes couleurs des séquences correspondent à un jeu de données particulier obtenu en faisant varier le paramètre « seuil » du logiciel « IlluHaplotyper ». Les correspondances entre les couleurs et les valeurs de seuils utilisées sont indiquées à gauche. Les valeurs de bootstraps obtenu avec 1 000 réplicas sont indiquées en dessous des branches.

PARTIE B : Comparaison de polymorphismes et d'haplotypes détectés à partir de données Roche-454 et Illumina

Nous avons cherché à comparer les résultats de recherches de polymorphismes et d'haplotypes obtenus à l'aide des pipelines « PyroHaplotyper » et IlluHaplotyper » pour les mêmes références transcriptomiques.

I- Détection de sites polymorphes :

Nous avons recherché à l'aide du logiciel « IlluHaplotyper » (paramètres par défaut) les SNPs et haplotypes après alignement des reads Illumina sur les contigs Roche-454 présentés dans le Chapitre 4, Partie B et dont les alignements présentaient au moins un SNP (détecté à partir des données de pyroséquençage). Pour les deux espèces hexaploïdes *S. maritima* et *S. alterniflora*, nous avons détecté respectivement 3,95 et 3,83 SNPs pour 100 bp. Le nombre moyen de sites polymorphes pour les alignements des deux hybrides et de l'allopolyploïde est de 9,25 (pour *S. x townsendii*), 9,12 (pour *S. x neyrautii*) et de 9,52 (pour *S. anglica*) SNPs pour 100 bp (Figure 29, a). Le nombre de polymorphismes observés chez les deux espèces hexaploïdes est similaire (test de Student ; p-value > 0,05). Nous avons également observé un nombre similaire de SNPs chez les deux hybrides et l'allopolyploïde (test de Student ; p-value > 0,05). Le nombre de SNPs détectés chez les hybrides et *S. anglica* est plus important que le nombre de SNPs détectés chez les parents (test de Student ; p-value < 0,001). Le nombre de SNPs détectés à partir des données Illumina est plus important que le nombre de SNPs détectés à partir des données Roche-454, et ceci pour les cinq espèces étudiées. A partir des données Roche-454 nous avons détecté un nombre moyen de 2,66 (pour *S. maritima)*, 2,75 (pour *S. alterniflora*), 3,12 (pour *S. x townsendii*), 3,19 (pour *S. x neyrautii*) et de 2,74 (pour *S. anglica*) SNPs pour 100 bp (Chapitre 4, Partie B).

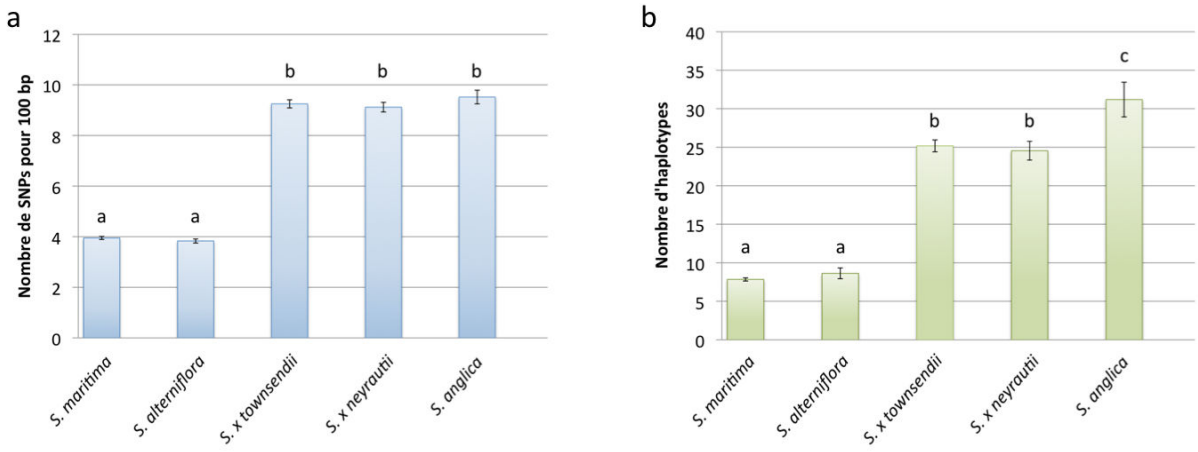


Figure 29 : (a) Nombre de SNPs pour 100 bp et (b) nombre moyen d'haplotypes pour les cinq espèces étudiées.

Nous avons ensuite comparé les différents sites polymorphes détectés au cours de cette étude. Les SNPs détectés à partir de l'assemblage des données Roche-454 (Chapitre 4, Partie B) ont été filtrés pour supprimer les blocs de SNPs correspondant à des groupes d'insertions/délétions en bloc pouvant apparaître dans les alignements de reads de longs fragments. Le nombre de SNPs détectés uniquement dans les données Roche-454 représente 6.83% à 29.37% de la totalité des SNPs, selon les espèces, et les SNPs détectés uniquement au sein des données Illumina représentent entre 53.55% et 78.59% de l'ensemble des SNPs (Figure 30). Finalement, le nombre de SNPs communs aux différents jeux de données varie entre 14.58 et 17.08% (Figure 30). Pour l'ensemble des espèces étudiées, la majorité des SNPs détectés ne se retrouvent que dans les données Illumina, les proportions de SNPs détectés dans les deux jeux de données (SNPs communs), celles détectées uniquement au sein des données Illumina et celles détectées uniquement au sein des données Roche-454 sont équivalentes d'une part entre les espèces hexaploïdes et d'autre part entre les hybrides et l'allopolyploïde (test exact de Fisher ; p-value > 0,05). Cependant, nous avons mis en évidence un nombre de SNPs Illumina et un nombre de SNPs Roche-454 plus important pour les espèces parentales hexaploïdes (test exact de Fisher ; p-value < 0,001).

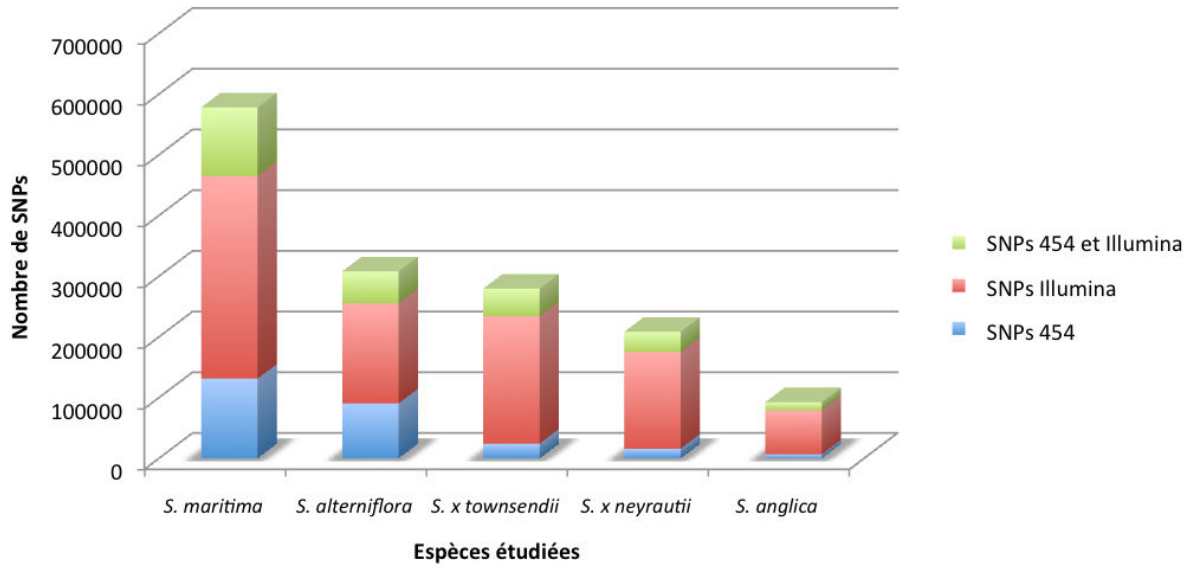


Figure 30 : Représentation du nombre total de SNPs pour chaque espèce étudiée. Les SNPs représentés en bleu correspondent au SNPs détectés uniquement au sein des données Roche-454, en rouge les SNPs détectés uniquement au sein des données Illumina. Les SNPs représentés en vert correspondent aux SNPs détectés à la fois au sein des données Roche-454 et Illumina.

II- Haplotypes détectés :

Consécutivement à la détection des SNPs, « IlluHaplotyper » reconstruit les différents haplotypes pour les 5 espèces étudiées (Figure 29, b). Les différents alignements des espèces hexaploïdes présentent en moyenne 7,84 (pour *S. maritima*) et 8,63 (pour *S. alterniflora*) haplotypes. Les alignements de séquences comptent en moyenne 25,18 (pour *S. x townsendii*), 24,55 (pour *S. x neyrautii*) et 31,19 (pour *S. anglica*) haplotypes. Nous observons que le nombre moyen d'haplotypes par alignement obtenu pour les deux espèces *S. x townsendii* et *S. x neyrautii* est similaire (test de Student ; p-value > 0,05). Les alignements des deux espèces hybrides et de l'allopolyploïde présentent un nombre d'haplotypes plus important par rapport aux alignements des deux parents *S. maritima* et *S. alterniflora* (test de Student ; p-value < 0,001). Le nombre moyen d'haplotypes détecté à partir des données Illumina est plus important que le nombre d'haplotypes détectés à partir des données Roche-454 qui comptent en moyenne 5,27 (pour *S. maritima*), 5,84 (pour *S. alterniflora*), 5,84 (pour *S. x townsendii*), 5,32 (pour *S. x neyrautii*) et 4,80 (pour *S. anglica*) haplotypes.

De la même manière que pour les alignements issus des données Roche-454, nous avons calculé le nombre d'haplotypes pour chaque alignement local pour chaque espèce.

Nous pouvons constater que le nombre d'alignements présentant entre 2 et 4 haplotypes d'une part et 5 ou 6 haplotypes d'autre part est présente dans des proportions différentes pour les cinq espèces étudiées (test exact de Fisher ; p-value < 0,001). Le nombre d'alignements présentant entre 2 et 6 haplotypes est similaire pour les deux parents et représente 45,81% (chez *S. maritima*) et 42,69% (chez *S. alterniflora*) des alignements (test de exact de Fisher ; p-value > 0,05). Comme attendu pour les deux espèces hybrides et l'espèce allopolyploïde, le nombre d'alignements présentant entre 2 et 6 copies ne représente que 6,12 à 12,13% des alignements. Au sein des ces espèces, la majorité des alignements vont présenter un nombre important d'haplotypes, ce qui est à mettre en relation avec le nombre moyen d'haplotypes détectés (entre 24,55 et 31,19). Ces résultats nous montrent que le programme reconstruit plus difficilement les haplotypes au sein de ces espèces, ce qui peut s'expliquer par le nombre de SNPs plus important, la présence de copies dupliquées peu divergentes (issues des deux parents hexaploïdes) et la volonté de ne pas créer d'haplotypes chimériques lors de l'assemblage des reads.

L'application du programme « IlluHaplotyper » sur ces jeux de données a permis d'identifier les SNPs communs aux données Roche-454. Le nombre de SNPs détectés à l'aide des données Illumina est plus important, comme attendu, puisque nous disposons d'un jeu de données plus important issu de cette technologie. De plus, les paramètres de détection de SNPs sont plus ou moins stringents selon la technologie utilisée et sont en adéquation avec différentes études (Oliphant et al. 2002; Udall et al. 2006; Tennessen et al. 2014). Néanmoins certains SNPs détectés uniquement avec la technologie Roche-454 pourraient correspondre à des faux positifs comme cela a pu être mis en évidence au sein des données d'ADN ribosomique de *S. maritima* (Chapitre 4, Partie A). Le nombre d'haplotypes détectés chez les espèces hybrides et allopolyploïde reste très important. Ces résultats indiquent qu'au sein d'alignements présentant un nombre trop important de polymorphismes les paramètres de stringence de détection d'haplotypes, pour éviter la création de contigs chimériques, conduisent à une augmentation du nombre d'haplotypes.

PARTIE C : Validation des haplotypes détectés par « IlluHaplotyper » sur des données transcriptomiques et génomiques.

I- Détection et validation de copies dupliquées au sein d'un jeu de données transcriptomiques.

i) Etude intégrative de l'expression de 13 gènes candidats en conditions naturelles.

Dans le cadre d'une étude d'expression de 13 gènes d'intérêts, préalablement identifiés comme étant différentiellement exprimés entre espèces (Chelaifa 2010) et dont les niveaux d'expression ont été étudiés par PCR quantitative (Q-PCR) chez les cinq espèces de Spartines dans des populations naturelles (Ferreira de Carvalho 2013) ; nous avons cherché à estimer le nombre des copies à l'aide de l'outil « IlluHaplotyper ». Le Tableau 11 présente la liste des gènes analysés et les jeux de données RNA-seq que nous avons utilisés pour chaque espèce. Ces travaux font l'objet d'un article « Changes in global gene expression in natural populations and identification of homoeologues in polyploid *Spartina* species » en préparation (Julie Ferreira de Carvalho, Julien Boutte, *et al. In prep*).

Tableau 11 : Fonction biologique et nombre de reads RNA-Seq Illumina alignés sur les 13 gènes étudiés, pour les 5 espèces de Spartines.

Fonction biologique :	*S. maritima*	*S. alterniflora*	*S. x townsendii*	*S. x neyrautii*	*S. anglica*
Xantine dehydrogenase 1	398	61	599	105	141
Metal tolerance protein A2	340	225	4 266	100	102
Transcriptional adapter ADA2a	1 357	462	1 147	394	660
WRKY24	1 132	361	566	187	318
Zinc finger protein STOP1 homolog	10 657	1 668	8 612	1 544	759
Putative GDP-mannose pyrophosphorylase	13	40	372	27	11
V-type proton ATPase catalytic subunit A	98	59	879	131	139
Cytochrome c oxidase subunit 6b-1	25	14	107	11	18
Cinnamoyl-CoA reductase-like protein	164	87	1 525	64	113
Transfactor, putative, expressed	897	279	653	265	449
Metallocarboxypeptidase inhibitor	10 657	25 655	47 436	17 387	18 705
Hexokinase 1	220	174	650	136	182
Similar to Transcription elongation factor SPT6	684	472	505	658	1 254

La technique employée (Q-PCR) a permis d'évaluer les niveaux d'expression globale pour chaque gène et de détecter pour chacun d'entre eux, à partir du logiciel « IlluHaplotyper », les différents haplotypes exprimés. Cette étude a permis d'analyser l'amplitude de la variation de l'expression des gènes entre individus de la même population, et entre populations d'une même espèce ou d'espèces différentes de Spartines (les deux parents, les hybrides et l'allopolyploïde) en conditions naturelles. Sera présentée ici ma contribution à cette étude, qui concerne plus particulièrement la détection automatisée des haplotypes au sein de ces gènes.

ii) Détection et validation de copies dupliquées à l'aide du programme « IlluHaplotyper ».

Les différentes copies dupliquées de ces 13 gènes d'intérêts pour les 5 espèces étudiées ont été recherchées à l'aide du logiciel « IlluHaplotyper ». Les résultats obtenus chez *S. maritima* pour le gène « Metal tolérance protein A2 » ont été comparés avec les résultats issus de clonage et séquençage obtenus à l'aide de la méthode Sanger. Le nombre maximum d'haplotypes (par fenêtre) détectés par alignement est présenté dans le Tableau 12. Pour les différents gènes étudiés, le nombre d'haplotypes construits chez les parents est similaire. Le nombre d'haplotypes détecté chez les hybrides et l'allopolyploïde est particulièrement important pour les différents alignements étudiés. Ces résultats sont en adéquation avec les résultats présentés dans ce chapitre qui montre une multiplication du nombre d'haplotypes lors de la reconstruction des haplotypes chez ces espèces.

Tableau 12 : Nombre maximum d'haplotypes détectés au sein de chaque alignement (et de chaque fenêtre). Les valeurs séparées par des virgules correspondent au nombre maximum pour chacune des fenêtres détectées.

Fonction biologique :	*S. maritima*	*S. alterniflora*	*S. x townsendii*	*S. x neyrautii*	*S. anglica*
Xantine dehydrogenase 1	7	1	3,5	1	3
Metal tolerance protein A2	6	3	25	1	5
Transcriptional adapter ADA2a	13	6, 7, 3	28	13	15
WRKY24	13,3	15	23	1	15
Zinc finger protein STOP1 homolog	18	17	48	24	18
Putative GDP-mannose pyrophosphorylase	1	1	9	1	1
V-type proton ATPase catalytic subunit A	3	1	13	1	1
Cytochrome c oxidase subunit 6b-1	1	1	10	1	1
Cinnamoyl-CoA reductase-like protein	1	2	16	1	7
Transfactor, putative, expressed	13	4	16	11	15
Metallocarboxypeptidase inhibitor	17	22	24, 8	16, 63	16, 131
Hexokinase 1	7	5	24	4	9
Similar to Transcription elongation factor SPT6	9	1	12	10	16

Nous avons comparé pour le gène « Metal tolérance protein A2 », les haplotypes construits à l'aide du logiciel « IlluHaplotyper » aux séquences obtenues par clonage (Figure 31). Six haplotypes de *S. maritima* et cinq haplotypes de *S. alterniflora* identifiés par « IlluHaplotyper » ont pu être comparés aux séquences clonées. Une seule copie détectée à l'aide des données Illumina n'est pas retrouvée dans les données de clonage de *S. maritima* comprenant des données génomiques et transcriptomiques. La Figure 31 présente un arbre phylogénétique construit à partir des séquences clonées. Trois de nos haplotypes sont similaires aux copies appartenant au clade A (dont un haplotype qui est 100% identique aux séquences du clade A). Deux haplotypes sont similaires aux séquences du groupe B. Sur les cinq haplotypes reconstruits à partir des données Illumina de *S. alterniflora*, trois ne sont pas retrouvées dans les données de clonage transcriptomiques, les deux autres sont similaires au groupe B. L'ensemble des séquences détectées à partir des données Illumina et similaire aux données de clonage sont identiques à 1 site près (qui pourrait correspondre à une erreur de séquençage). Ces résultats nous permettent de valider les copies du gène détectées à l'aide du logiciel « IlluHaplotyper » au sein de jeux de données RNA-seq Illumina.

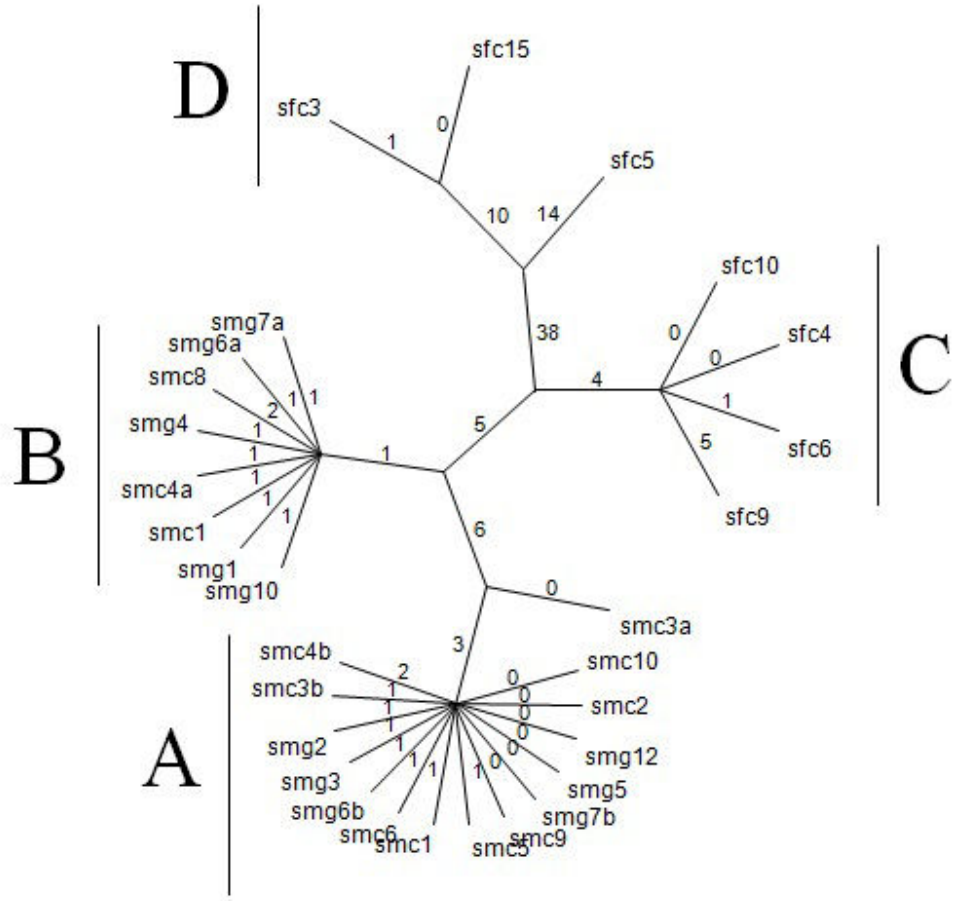


Figure 31 : Arbre phylogénétique (méthode de Maximum de Parcimonie) des séquences clonées correspondant au gène « Metal tolérance protein A2 ». Les clades A et B incluent des séquences génomiques (smg) et transcriptomiques (smc) de *S. maritima*. Les clades C et D incluent des séquences transcriptomiques de *S. alterniflora* (sfc). Les valeurs représentent le nombre de nucléotides différents entre les séquences (Bourdau 2012, Ferreira de Carvalho et al. in prep).

iii) Assignation de l'Origine des copies hybrides et allopolyploïdes.

Pour les différents alignements obtenus, nous avons assigné les copies dupliquées chez les hybrides et l'allopolyploïde à l'un des deux parents hexaploïdes selon l'approche présentée dans le Chapitre 3, Partie II.5. Nous pouvons observer pour les 3 espèces étudiées un nombre similaire d'haplotypes provenant respectivement de *S. maritima* et de *S. alterniflora* (test exact de Fisher ; p-value > 0,05). Nous assignons 221, 57 et 106 haplotypes respectivement de *S. x townsendii*, *S. x neyrautii* et *S. anglica* à l'espèce parentale *S. maritima*. Le nombre de copies assignées à *S. alterniflora* est respectivement de 148, 55 et 76 haplotypes pour les deux hybrides et l'allopolyploïde. Le nombre de copies non assignées représente 97 et 37 haplotypes pour les deux hybrides *S. x townsendii* et *S. x* neyrautii. 77 haplotypes présents chez l'allopolyploïde *S. anglica* ne sont pas assignés spécifiquement à

l'un des parents. On notera que pour les gènes étudiés ici, nous n'observons pas de mise sous silence de l'ensemble des copies de l'un des deux parents hexaploïdes. Nous avons étudié l'assignation des copies détectées chez l'hybride *S. x townsendii* pour le gène codant la « Transcriptional adapter ADA2a » (Figure 32). Pour une région de 129 bp, 7 haplotypes de *S .x townsendii* sur 10 ont été retrouvés et assignés à une copie parentale. Six de ces copies sont identiques à 100% aux copies parentales, le septième haplotype présente une différence d'un nucléotide par rapport à la copie parentale. L'une des copies détectée chez l'hybride a été identifiée chez les deux parents hexaploïdes.

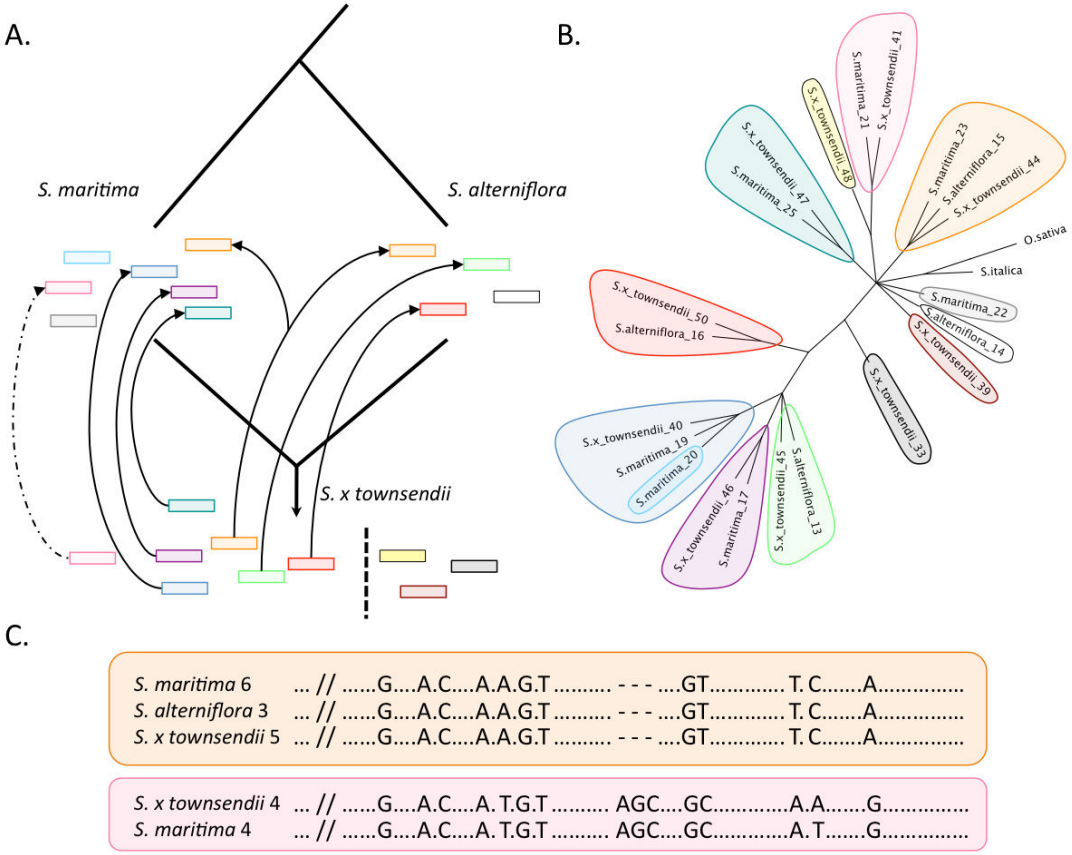


Figure 32 : (A.) Représentation schématique de l'assignation parentale des copies du gène « Transcriptional adapter ADA2a » de *S. x townsendii*. Les boites de couleurs représentent les différents haplotypes identifiés chez les trois espèces. Les haplotypes parentaux identiques sont de la même couleur (*e.g.* en orange). Les lignes indiquent les relations entre les haplotypes de l'hybride et les haplotypes parentaux présentant 100% de similarité. La ligne en pointillé indique que la copie de *S. maritima* est identique à la copie détectée chez *S. x townsendii* à un site près. (B.) Arbre phylogénétique (obtenu à l'aide de la méthode Neighbor Joining) montrant les relations entre les haplotypes. Les couleurs des haplotypes correspondent à celles représentées en A. (C.) Alignements des haplotypes correspondant aux copies oranges et roses. Les symboles '-', '.' et '//' au sein des séquences correspondent respectivement à un gap, un site invariable et une région invariable non représentée.

Comme souligné précédemment, dans notre système, les espèces parentales (*S. maritima* et *S. alterniflora*) sont déjà hexaploïdes et contiennent donc également des copies dupliquées à chaque locus. Dans le but d'explorer l'histoire des copies détectées au sein de ces parents, une phylogénie incluant les haplotypes détectés chez les deux espèces a été réalisée pour le gène « Hexokinase 1 ». Pour cela, la méthode du maximum de parcimonie (recherche heuristique) a été appliqué à l'aide du logiciel MEGA v5.2.1 (Tamura et al. 2011) sur un alignement de 275 bp et présentant 55 sites informatifs. L'analyse phylogénétique a été réalisée à l'aide d'outgroups appartenant à la sous-famille des Panicoideae (*Zea mays*: EU972171.1; *Setaria italica*: XM_004961391.1; *Sorghum bicolor*: XM_002440059.1). L'analyse de Bootstrap a été réalisée avec 1 000 réplicas (Figure 33). Nous pouvons observer que les séquences des spartines hexaploïdes se répartissent en deux clades (A et B) contenant chacun des haplotypes de *S. maritima* et *S. alterniflora*. Cette topologie suggère que les clades A et B résultent de la duplication de ce gène chez l'ancêtre commun (hexaploïde) de *S. maritima* et *S. alterniflora* et représenteraient probablement des homéologues. Pour chaque copie A et B nous avons pu distinguer plusieurs haplotypes par espèce. Chez *S. alterniflora,* tout comme chez *S. maritima* on note deux haplotypes différents par copie, avec un variant supplémentaire (différant d'une substitution) chez *S. maritima*.

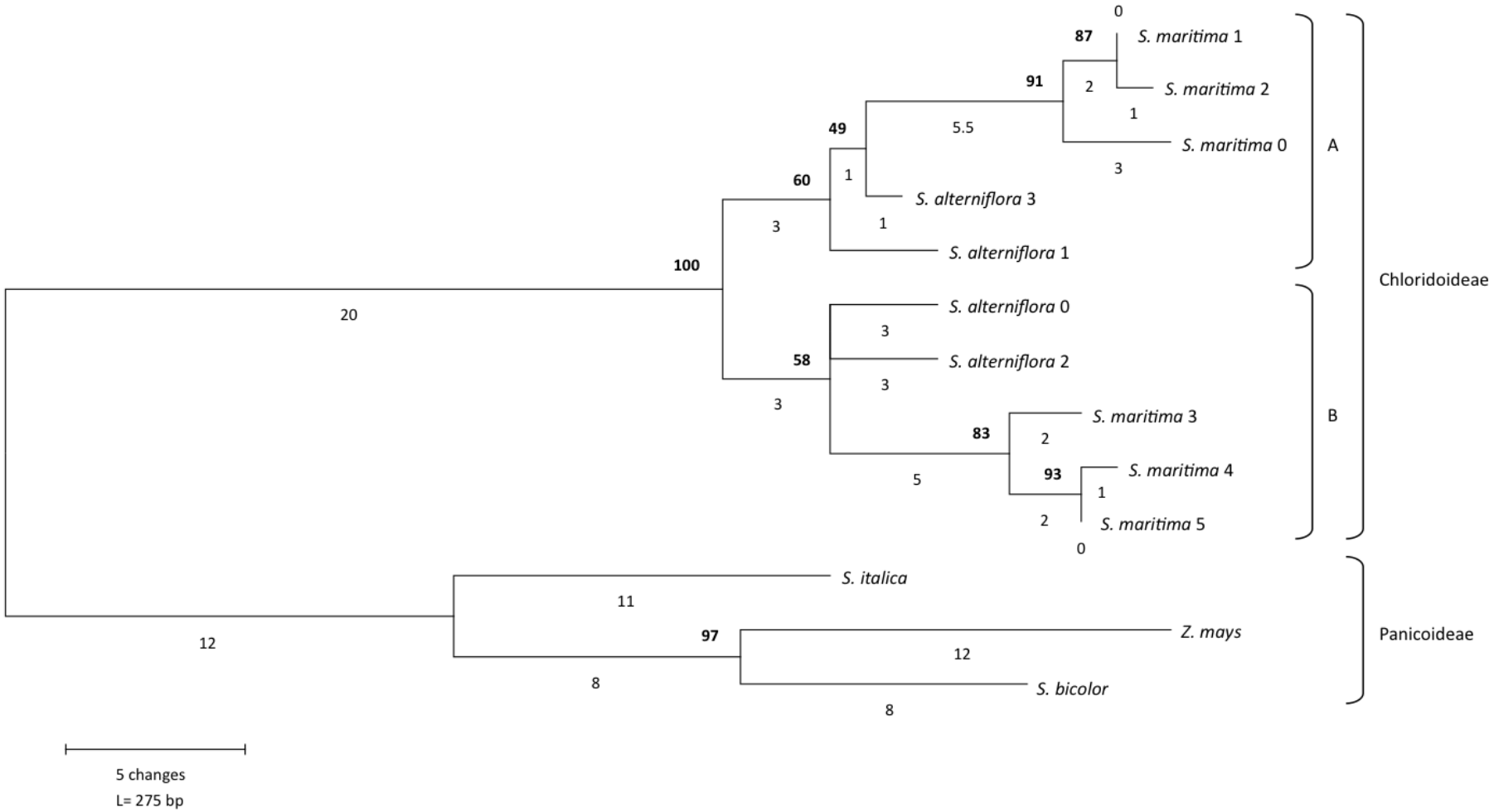


Figure 33 : Arbre phylogénétique obtenu à l'aide de la méthode du maximum de Parcimonie pour le gène « hexokinase 5 ». L'arbre a été enraciné avec des séquences de Panicoideae. Les accolades A et B indiquent un événement de duplication chez les espèces du genre *Spartina*. Le nombre de substitutions est indiqué en dessous des branches. Les valeurs de bootstraps obtenu avec 1 000 réplicas sont indiquées en gras au dessus des branches.

II- Détection et validation de copies dupliquées au sein de jeux de données génomiques : le cas du gène « *Waxy* ».

Afin de valider les haplotypes reconstruits à partir de données génomiques, nous avons choisi un gène en faible nombre de copies, le gène « *Waxy* » codant la GBSS I (« Granule Bound Sucrose Synthase I) initialement étudié dans le cadre de la thèse de Philippe Fortuné (2007). Les données obtenues ont été comparées aux résultats trouvés par Fortune et ses collaborateurs (2007) qui ont identifié les copies dupliquées du gène *Waxy* chez plusieurs espèces du genre *Spartina* à l'aide de clonage et de séquençage Sanger. Cette étude avait également permis d'analyser l'histoire de ce gène chez les Spartines polyploïdes et mis en évidence une rétention différentielle selon les espèces. Une et trois copies avaient été respectivement identifiées chez les espèces hexaploïdes *S. maritima* et *S. alterniflora* sur les 3 copies homéologues attendues (Fortune et al. 2007).

Nous avons donc recherché les copies dupliquées à l'aide du programme « IlluHaplotyper » (paramètres par défaut) en alignant des reads génomiques (Illumina) obtenus pour quatre espèces de Spartines (*S. bakeri* (2n=4x), *S. versicolor* (2n=4x), *S. maritima* (2n=6x) et *S. alterniflora* (2n=6x)) sur les séquences références d'une partie de l'exon 8 du gène « *Waxy* » (Fortune et al. 2007). Le nombre de reads alignés, de SNPs détectés et le nombre d'haplotypes construits sont indiqués dans le Tableau 13.

Tableau 13 : Nombre de reads totaux, alignés sur les séquences du gène *Waxy* des quatre espèces de Spartines. Le niveau de ploïdie de chaque espèce est indiqué ainsi que le nombre de SNPs, d'haplotypes et d'haplotypes maximum par bloc, détectés par le logiciel « IlluHaplotyper ».

Espèce :	Niveau de ploïdie:	Nombre de reads initiaux :	Nombre de reads alignés :	Nombre de SNPs :	Nombre total d'haplotypes :	Nombre maximum d'haplotypes :
S. maritima	6x	469 298 348	466	42	29	14
S. alterniflora	6x	354 079 558	185	14	7	3
S. bakeri	4x	323 705 910	638	72	40	17
S. versicolor	4x	333 349 850	886	67	28	13

Nous avons ensuite comparé les haplotypes détectés à l'aide du logiciel « IlluHaplotyper » entre les quatre espèces étudiées. Nous considérons deux haplotypes identiques s'ils présentent 100% d'identité sur la totalité de leur longueur ou s'ils présentent 100% d'identité sur leur région chevauchante et partagent au minimum 5 polymorphismes. Si un haplotype est retrouvé chez au moins deux espèces différentes (les jeux de données de chaque espèce étant indépendants), ce dernier est considéré comme un haplotype validé. Il a été ainsi possible de valider 13 haplotypes présents chez *S. maritima* et 4 haplotypes de *S. alterniflora*. Au sein des espèces tétraploïdes *S. bakeri* et *S. versicolor* nous avons validé respectivement 32 et 24 haplotypes préalablement détectés avec le programme « IlluHaplotyper » (Figure 34).

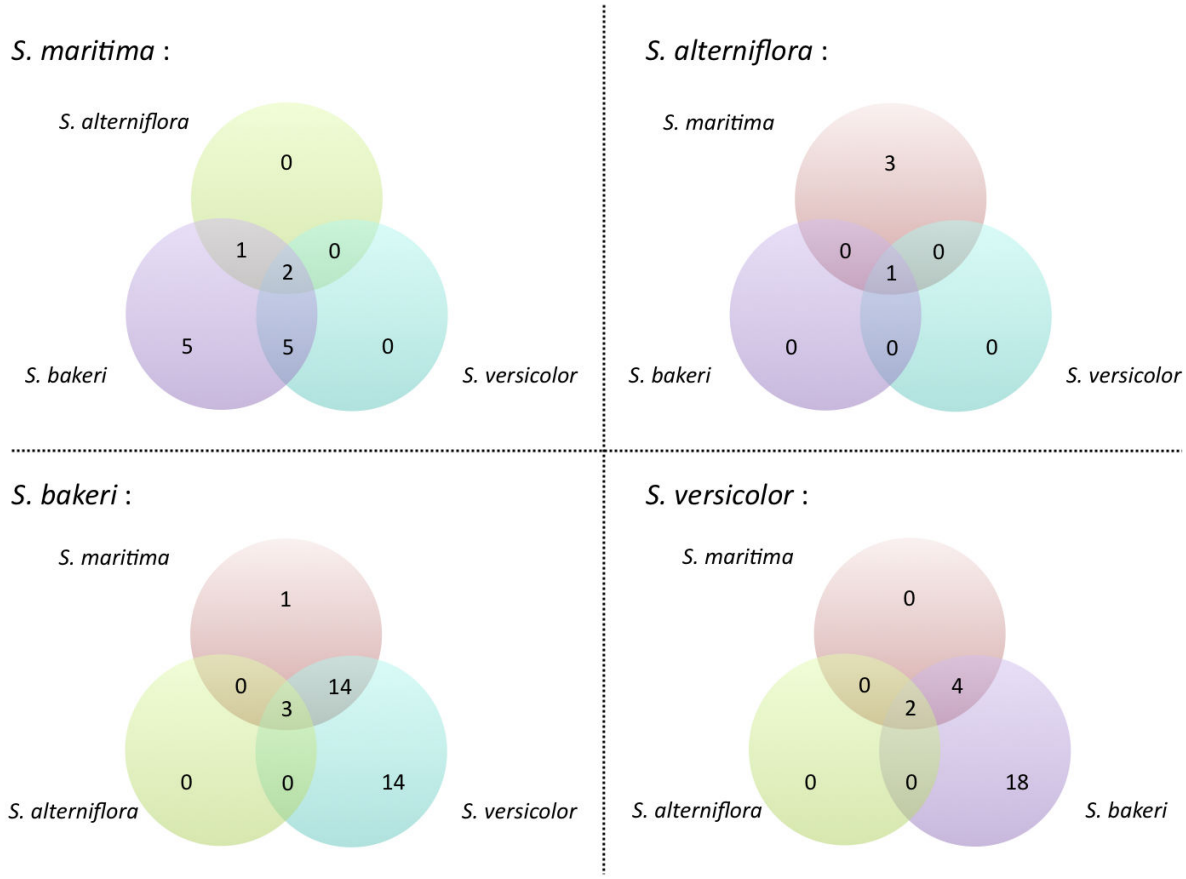


Figure 34 : Comparaison des haplotypes du gène *Waxy* détectés avec le programme « IlluHaplotyper » entre les quatre espèces de Spartines étudiées.

Dans un deuxième temps, nous avons comparé les haplotypes détectés avec la méthode de clonage et de séquençage par la méthode Sanger (Fortune et al. 2007) à ceux détectés avec notre programme. Pour cela nous avons sélectionné une région commune aux 2 jeux de données correspondant à un alignement de 103 bp (présentant 9 sites informatifs) et sur lequel une analyse phylogénétique a pu être réalisée (Figure 35). L'analyse phylogénétique a été effectuée à l'aide d'une analyse de maximum de Vraisemblance (ML), et du modèle d'évolution de Juke et Cantor. Un test de robustesse a été effectué à l'aide de 1 000 bootstraps.

Fortune *et al.* (2007) avaient mis en évidence une duplication ancienne du gène *Waxy*, ayant abouti à deux clades de séquences paralogues nommées A et B. Nous avons retrouvé au sein de notre jeu de données les deux copies des clades A et B mises en évidence par Fortune et ses collaborateurs (2007). Nous avons détecté 2 séquences de *S. bakeri* et 1 séquence de S*. versicolor* correspondant à la copie A. Deux groupes ont été identifiés au sein de la copie B. Au sein du groupe B1, se trouvent les 2 séquences de *S. versicolor* et les 5 séquences de *S. bakeri* détectées à l'aide des données Illumina ainsi que la copie B1 de *S. densiflora* détectée à l'aide de données de clonage. Les séquences détectées et identifiées comme les copies B2 et B3 par Fortune et ses collaborateurs (2007) sont localisées dans un deuxième groupe (B2-B3). Au sein de ce groupe se trouvent les 4 haplotypes de *S. maritima* et les 2 haplotypes de *S. alterniflora* détectés à l'aide de notre programme.

La Figure 36 montre la portion du gène *Waxy* alignée entre séquences clonées (Fortune et al. 2007) et haplotypes reconstruits (par « IlluHaplotyper »). Nous retrouvons dans nos haplotypes (qui sont ainsi validés) les copies clonées B3 de *S. maritima* et de *S. alterniflora* et la copie A2 de *S. bakeri*. Nous confirmons également d'autres haplotypes de *S. maritima* et de *S. alterniflora* identiques pour cette région à la copie clonée B2 *de S. argentinensis*. Il est à noter que Fortune *et al.* (2007) n'avait détecté qu'une seule copie (B3) chez *S. maritima* ce qui les avait amené à envisager l'hypothèse d'une perte de copie chez cette espèce. La profondeur du séquençage Illumina nous permet donc de détecter des copies qui pouvaient passer inaperçues dans les échantillons (forcément plus limités) de séquences clonées. La Figure 36 indique également que les séquences des haplotypes de *S. bakeri* et *S. versicolor* sont identiques à une copie clonée B1 de *S. densiflora*.

De plus, nous avons noté dans ces alignements une insertion-délétion (indiquée par une flèche en Figure 36), retrouvée dans les haplotypes et les séquences clonées (mais non prise en compte dans l'analyse phylogénétique ML présentée ci-dessous, Figure 35), qui est partagée par toutes les espèces analysées, et qui demanderait une analyse plus poussée (polymorphisme ancestral, recombinaisons homéologues).

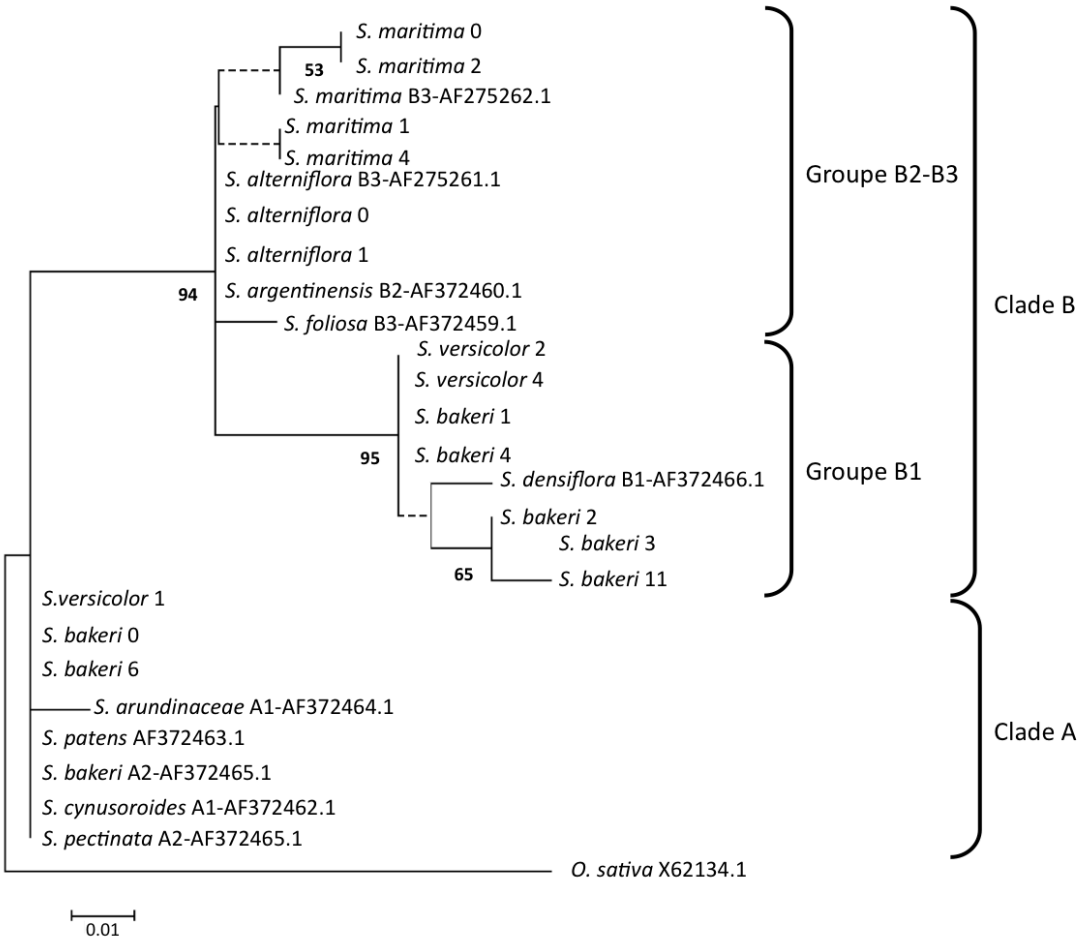


Figure 35 : Arbre phylogénétique obtenu à l'aide de la méthode du maximum de Vraisemblance pour un alignement de 103 bp du gène *Waxy*. L'arbre a été enraciné à l'aide de la séquence du riz (X62134.1). Les parenthèses à droite délimitent les deux copies (A et B) du gène *Waxy*. Les parenthèses à gauche indiquent les deux groupes distincts apparaissant dans le clade A. Les valeurs de bootstraps pour 1 000 réplicas sont indiquées en gras en dessous des branches. Seul les valeurs de bootstraps supérieures à 50% sont représentées. Les lignes en pointillées correspondent aux nœuds qui ne sont pas résolu à 50%. Les numéros d'accession des séquences publiées par Fortune et al. (2007) et utilisées pour cette analyse sont indiqués sur la figure. Les copies ne présentant pas de numéro d'accession correspondent aux haplotypes détectés avec le programme « IlluHaplotyper ».

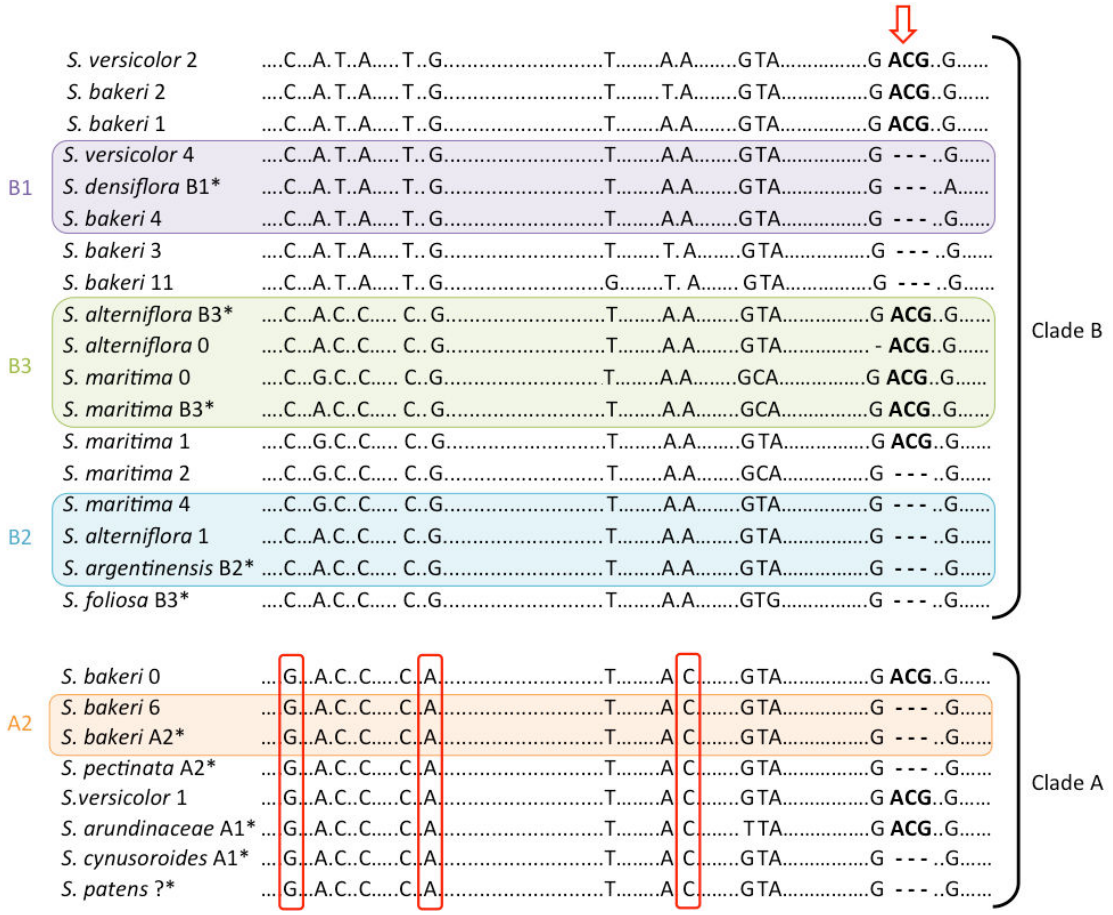


Figure 36 : Alignement nucléotidique d'une partie de l'exon 8 du gène *Waxy* utilisé pour l'analyse phylogénétique. Les astérisques (*) correspondent aux séquences clonées obtenues par Fortune *et al.* (2007). Les autres séquences correspondent aux haplotypes détectés à partir du logiciel « IlluHaplotyper ». Les séquences sont regroupées en deux clades correspondant aux clades A et B du gène *Waxy*. Les nucléotides entourés en rouge indiquent les sites polymorphes permettant de discriminer les deux clades. La flèche rouge signale une insertion-déletion de 3 bp qui permet de discriminer certaines copies.

PARTIE D : Construction de 5 nouveaux transcriptomes de référence chez les Spartines polyploïdes et identification de copies dupliquées à partir de données RNA-seq.

Le développement et la validation du pipeline « IlluHaplotyper » nous permettent à présent de détecter les différentes copies dans les jeux de données transcriptomiques Illumina de différentes espèces de Spartines. Pour cela, nous avons d'abord recherché parmi les outils d'assemblages existants, le logiciel le plus adapté à nos jeux de données. Nous avons ensuite co-assemblé les données de séquençage issues des technologies Roche-454 et Illumina pour obtenir de nouveaux transcriptomes de référence pour les cinq espèces : les parents hexaploïdes *S. maritima*, *S. alterniflora*, les deux hybrides *S. x townsendii*, *S. x neyrautii* et l'espèce alllododécaploïde *S. anglica*. Les données RNA-seq Illumina de chaque espèce ont été alignées sur chaque transcriptome de référence correspondant et la détection des copies a été réalisée à l'aide du pipeline « IlluHaplotyper ».

I- Quel Assembleur pour construire des transcriptomes de référence ?

Pour identifier le meilleur logiciel d'assemblage de données transcriptomiques « short reads » nous avons réalisé plusieurs analyses comparatives entre différents logiciels. Une analyse entre les logiciels Minia (Chikhi and Rizk 2012) et Trinity (Grabherr et al. 2011) qui se présentaient comme les deux meilleurs candidats pour l'assemblage de nos jeux de données a été réalisée en collaboration avec Erwan Scaon (doctorant à l'INRIA-IRISA de Rennes, équipe Genscale) et présentée en détail ci-dessous. Nous avons appliqué et comparé les résultats obtenus à partir des logiciels Minia et Trinity sur le jeu de données transcriptomiques de feuilles de *Spartina maritima* (Tableau 14).

Nous obtenons un total de 79 746 contigs à partir du logiciel Minia, d'une longueur moyenne de 200 bp. L'assemblage via le logiciel Trinity nous a permis d'obtenir 75 979 contigs d'une longueur moyenne de 390 bp. Pour obtenir le nombre moyen de reads par contig, nous avons aligné les reads sur les différents contigs à l'aide les logiciels Bowtie et Bowtie 2 (Langmead et al. 2009; Langmead and Salzberg 2012). A partir du logiciel Bowtie, le nombre total de reads mappés sur les contigs obtenus avec le logiciel Minia et le logiciel

Trinity représente respectivement 3,54% et 34,61% du nombre total de reads initiaux. Ces valeurs sont portées à 71,14% et 98,88% en utilisant le logiciel Bowtie 2 et des paramètres moins stringeants. Après mapping à l'aide du logiciel Bowtie, le nombre moyen de reads par contig est estimé respectivement à 17,59 reads et 123,23 reads pour les assemblages avec le logiciel Minia et Trinity. Ces valeurs sont plus importantes avec le logiciel Bowtie 2 : elles sont respectivement de 109,41 reads et 192,05 reads pour les assemblages obtenus avec les logiciels Minia et Trinity.

Tableau 14 : Nombre de contigs obtenus lors de l'assemblage à partir des logiciels Minia et Trinity, avec la longueur minimale, maximale et la médiane des contigs. Le temps de calcul et la mémoire nécessaire lors des assemblages sont également indiqués.

Logiciel :	Minia	Trinity
Temps d'exécution (en seconde) :	1426	72622
Mémoire vive utilisée au maximum :	2.4 Gb	184 Gb
Nombre de contigs obtenus :	79 746	75 979
Nombre total de bases :	16 006 047	29 701 312
Longueur moyenne des contigs (et médiane) en bp :	201 (158)	391 (2226)
Longueur minimale-maximale des contigs (en bp):	108 – 2 560	101 – 27 479
Longueur maximale (en bp) :	2560	27479

Les différents contigs obtenus ont été alignés par tblastx contre une base de données incluant 4 espèces de Poaceae (*Brachypodium distachyon*, *Oryza sativa*, *Zea mays* et *Sorghum bicolor*) avec une e-value réglée à 10^{-06}. Sur les 34 264 contigs issus de Minia, 30 366 ont un résultat d'alignement à plus de 70% de similitude contre la base de données. 33 788 des 37 585 contigs issus de l'assemblage via Trinity s'alignent à plus de 70% de similitude contre la base de données.

Les résultats obtenus à l'aide de ces deux logiciels d'assemblage nous permettent d'identifier l'outil **Trinity comme étant le logiciel le plus adapté à nos jeux de données**. Les différents assemblages *de novo* des données transcriptomiques ont donc été réalisés à l'aide de ce logiciel et les résultats obtenus sont présentés dans l'article ci dessous.

II- Construction de cinq transcriptomes de référence et détection d'haplotypes.

Nous avons utilisé de nouveaux jeux de données transcriptomiques issus de pyroséquençage et de séquençage par synthèse Illumina générés au Genoscope - Centre National de Séquençage (Evry, France) afin 1) d'enrichir les premiers transcriptomes de référence des espèces *S. maritima* et *S. alterniflora* (Ferreira de Carvalho et al. 2012) et 2) générer des transcriptomes de référence pour les espèces qui résultent de leur hybridation : *S. x townsendii*, *S. x neyrautii* et de l'espèce allopolyploïde *S. anglica*. Les différents jeux de données Roche-454 et Illumina ont été dans un premier temps assemblés indépendamment à l'aide des logiciels Newbler et Trinity. Les contigs obtenus ont ensuite été co-assemblés à l'aide du logiciel Newbler et de scripts développés au sein du laboratoire afin de supprimer les contigs redondants. Pour chaque transcriptome de référence, nous avons obtenu entre 44 158 et 65 099 contigs, dont 19 241 à 25 067 ont pu être annotés à l'aide d'une méthode similaire à celle développée par Ferreira de Carvalho et ses collaborateurs (2012) et du logiciel Pfam (Finn et al. 2014).

Nous avons ensuite reconstruit les haplotypes à l'aide d'« IlluHaplotyper » pour les 5 nouveux transcriptomes de référence de Spartines construits à l'aide de données Roche-454 et Illumina. A partir de ces haplotypes, l'origine des copies détectées au sein des hybrides et de l'allopolyploïde a été assignée à l'un des génomes parentaux (*S. maritima* ou *S. alterniflora*). Nous avons également identifié les différents événements de polyploïdisation récents survenus chez les Spartines en nous basant sur la distribution des divergences entre copies (estimées à partir des taux de substitutions synonymes ou Ks). Les différents « pics » de divergence (Ks) résultant de duplications génomiques ont été datés et comparés avec les estimations proposées par Rousseau-Gueutin *et al.* (2015) à partir de données de génomes chloroplastiques chez les spartines.

Ces différentes analyses sont présentées dans l'article ci-dessous.

Manuscript in preparation

Reference transcriptomes and detection of duplicated copies in hexaploid parents, hybrids and allododecaploid *Spartina* species (Poaceae)

Boutte J.[1], Ainouche M.[1] and Salmon A.[1]

[1]UMR CNRS 6553 Ecobio, University of Rennes 1, Bât 14A Campus Scientifique de Beaulieu, 35 042 Rennes Cedex (France)

Key-words: Paralogy, Homoeology, Non-model species, RNA sequencing, SNPs, *de novo* assembly, Duplication events, K_S

ABSTRACT:

In this study, we report the assembly and annotation of five reference transcriptomes for the European hexaploid Spartina species (*S. maritima*, *S. alterniflora* and their homoploid hybrids *S. x townsendii* and *S. x neyrautii*) and the allododecaploid invasive species *S. anglica*. These transcriptomes were constructed from various leaf and root cDNA libraries that were sequenced using both Roche-454 and Illumina technologies. As the repetitive nature of *Spartina* genomes and transcriptomes due to their high ploidy levels and their hybrid origin were challenging, and because no reference diploid genome is available for *Spartina*, we developed generic bioinformatics tools to 1) detect different haplotypes of each gene within each species and 2) assign a parental origin to haplotypes detected in the hexaploid hybrids and the neo-allopolyploid. The approach described here allows the detection of putative homeologs from sets of short reads. Synonymous substitution rate (Ks) comparisons between haplotypes from the hexaploid parental species revealed the presence of two Ks peaks (likely resulting from the tetraploid and hexaploid duplication events) which divergence times were estimated as 1.8-2.7 Myr and 8.7-9.6 Myr respectively. The procedure developed in this study can be applied for future differential gene expression or genomics experiments to study the fate of duplicated genes in the invasive allododecaploid *S. anglica*.

INTRODUCTION

Polyploidy (resulting from whole genome duplication) appears to be a major feature of eukaryote evolution (Otto and Whitton 2000; Van de Peer, Maere, and Meyer 2009; Mable, Alexandrou, and Taylor 2011). There are various ways by which polyploids may form in natural populations (reviewed in Ramsey and Schemske 1998; Ramsey and Schemske 2002; Tayalé and Parisod 2013). Two categories of polyploids may be distinguished, depending on the origin of the duplicated genomes: autopolyploids arise from genome duplication within species (and thus contain duplicated homologous chromosome sets) whereas allopolyploids result from the merger (by hybridization) and the duplication of two or more divergent (homoeologous) genomes. Several examples of polyploidization are reported in animals, as in amphibians where around 30 polyploid species have been identified (Gregory and Mable 2005) or in Teleostei fishes such as Salmonideae (Mable 2004). In plants polyploidy is a particularly prominent and recurrent process (Soltis et al. 2009) where genome duplication has contributed to speciation, phenotypic innovation and adaptation (Leitch and Leitch 2008). Polyploidy provides the raw genomic material for natural or artificial (*e.g.* domestication) selection (Wendel 2000). Most of our understanding of the consequences of polyploidy comes from relatively recent polyploids which include a large proportion of crops such as the allopolyploid wheats *Triticum turgidum* and *Triticum aestivum* (Feldman et al. 1997; Liu, Vega, and Feldman 1998; Ozkan, Levy, and Feldman 2001; Shaked et al. 2001; Kashkush, Feldman, and Levy 2002), cotton *Gossypium hirsutum* (Adams et al. 2003; Adams and Wendel 2005; Udall 2006; Hovav et al. 2008; Flagel et al. 2008; Flagel et al. 2009; Rapp, Udall, and Wendel 2009; Yoo, Szadkowski, and Wendel 2013), oilseed rape *Brassica napus* (Song et al. 1995; Osborn et al. 2003; Pires et al. 2004; Udall 2005; Albertin 2006; Lukens 2005; Gaeta et al. 2007; Szadkowski et al. 2010; Marmagne et al. 2010; Sarilar et al. 2013; Chalhoub et al. 2014), tobacco *Nicotiana tabaccum* (McCarthy et al. 2015) or *Coffea arabica* (Combes et al. 2012; Combes et al. 2013). Of particular interest are the nascent allopolyploid species which have been described in Asteraceae *(e.g. Tragopogon*, Malinska et al. (2011), *Senecio*, Abbott et al. (2008)), Brassicaceae (*e.g. Cardamine*, Marhold et al. (2009)), Phrymaceae (*e.g. Mimulus*, Vallejo-Marin (2012)) in Eudicots and Poaceae (*e.g. Spartina*, Ainouche, Baumel, and Salmon (2004)) in Monocots. These recently formed species can be compared to their actual parents and represent excellent model systems to understand the immediate consequences of hybridization and genome duplication in natural populations.

The rapidly accumulating genomic data has documented various, older genome duplication events (paleopolyploidy) in Eukaryotes (and most particularly plant) genomes (Blanc and Wolfe 2004; Van de Peer, Maere, and Meyer 2009; Jiao et al. 2011). Modern plant genomes then appear shaped by recurrent rounds of polyploidization and diploidization processes. Duplicated genes may undergo various evolutionary fates, including differential gene retention during the diploidization-fractionation

process (Langham et al. 2004), homoeologous recombination and gene conversion (Udall 2005; Nicolas et al. 2007; Salmon et al. 2009), or reprogramming of duplicated gene expression (Adams et al. 2003; Flagel et al. 2008; Yoo, Szadkowski, and Wendel 2013; Combes et al. 2013).

Distinguishing genes duplicated by polyploidy is then critical to undertsand the evolutionary history of plant species and to explore the short term and long-term evolution of duplicated genomes. In polyploids, allelic diversity (at orthologous loci) needs to be distinguished from homoeologous divergence (reflecting divergence between the parents and subsequent evolution after polyploid formation), and paralogs (resulting from individual gene duplications).

In recent years many progresses were accomplished toward identification of duplicated gene copies in polyploids, from EST (Expressed Sequence Tags, *e.g*. Udall 2006; Flagel et al. 2008) or from Next Generation Sequencing (NGS) such as in the cotton genus (Salmon et al. 2012; Page et al. 2013; Yoo, Szadkowski, and Wendel 2013), in oilseed rape (Higgins et al. 2012), *Coffea* (Combes et al. 2013), soybean (Ilut et al. 2012), *Tragopogon* (Buggs et al. 2012) or strawberry (Tennessen et al. 2014). The strategy developed in these studies is based on the preliminary identification of parental species-specific polymorphisms. The NGS dataset obtained for the polyploid is assembled using parameters adapted to optimize the recovery of paralogous and homoeologous copies. The constructed contigs are then compared with the diploid parental genomes using specific polymorphic sites (Flagel et al. 2008; Salmon et al. 2009; Ilut et al. 2012). Pipelines such as PolyCat (Page, Gingle, and Udall 2013), SNiPloid (Peralta et al. 2013) and HyLiTE (Duchemin et al. 2014) were designed to detect homeologs in allotetraploids, using their diploid parents as reference. These tools align diploid species reads (or sequenced ESTs) for detecting interspecific polymorphisms at homologous genomic regions. The detected polymorphisms are then considered as putative SNPs between homoeologous regions ("homeoSNPs") in the allotetraploid. The sequences from hybrid or allopolyploid species are then aligned or co-aligned to the parental homologous regions and the putative homeologs can be assigned to the corresponding parental genome according to the detected homeoSNPs. The POLiMAPS pipeline (Tennessen et al. 2014) associates homeolog-specific sites with genetic linkage maps, when a diploid genome reference is available. However when the diploid parents are unidentified or extinct, detection of homoeologous copies requires the development of adapted tools (Salmon and Ainouche 2015).

In this study, we aim at reconstructing reference transcriptomes from NGS datasets and detecting the various expected duplicated gene copies in the polyploid genus *Spartina* Schreb. (Poaceae, subfamily Chloridoideae), for two hexaploid parents, their two independently formed F1 hybrids and a neo-allododecaploid species. Genus *Spartina* is characterized by recurrent interspecific hybridization and genome duplication events that resulted in various ploidy levels ranging from tetraploid to dodecaploid, with a basic chromosome number x=10 (Ainouche et al. 2012). Hybridization and

polyploidy had major impacts on the genus diversification, and important ecological consequences in salt-marsh communities regarding the formation of invasive species (Ainouche et al. 2008; Strong and Ayres 2013). The history of the genus is now well-documented. *Spartina* represents a monophyletic lineage, embedded in the paraphyletic *Sporobolus* genus (Peterson et al. 2014). No diploid *Spartina* species are known among the 15 perennial species described by Mobberley (1956), which suggests that *Spartina* most likely emerged from an already polyploid common ancestor. Diploid species are reported in *Sporobolus* lineages which diverged from *Spartina* sometimes 14-20 MYA (Rousseau-Gueutin et al. 2015). *Spartina* has evolved in two lineages: a tetraploid clade and a hexaploid clade (Baumel et al. 2002a) which divergence was estimated as dating back to 6-10 mya from chloroplast genome sequences (Rousseau-Gueutin et al. 2015). Of particular interest are the hexaploid *Spartina alterniflora* Loisel. (2n=6x=62; East American coast origin) and *Spartina maritima* (Curtis) Fern. (2n=6x=60, European Atlantic coast origin) which have naturally hybridized in Europe following the introduction of the American species during the nineteenth-century. Two F1 hybrids were formed, *S. alterniflora* as female parent in both hybridization events: *Spartina x townsendii* (2n=6x=62) in Southampton water (England; Foucaud 1897) and *Spartina x neyrautii* (2n=6x=62; Marchant 1963) in Hendaye (Southwest France). Genome duplication of *S. x townsendii* (after 1890) resulted in a new allododecaploid species, *Spartina anglica* C.E. Hubbard, 2n= 120, 122, 124 (Marchant 1968). Expansion of this fertile and invasive species that rapidly colonized Western Europe and several continents (*e.g.* Australia and China) has important ecological consequences. *Spartina anglica* is now a classical example of recent allopolyploid speciation, and this system is an excellent model to explore the early evolutionary changes following hybridization and genome duplication in natural populations (Ainouche, Baumel, and Salmon 2004). No major genetic changes were detected in the hybrids and neoallopolyploid (Baumel et al. 2001; Baumel et al. 2002b, Parisod et al. 2009), but homogenization of parental rDNA homoeologous copies are being observed in populations of *S. anglica* (Dalibor et al. in press). Hybridization appears to have entailed significant epigenetic changes (Salmon, Ainouche, and Wendel 2005; Parisod et al. 2009). Using a single rice heterologous microarray, Chelaifa *et al.* (Chelaifa, Mahé and Ainouche 2010) have analyzed the differential expression between the hexaploid parental species (*S. maritima* and *S. alterniflora*). Non additive transcriptomic parental patterns were observed in the hybrids and allopolyploid (Chelaifa, Monnier, and Ainouche 2010), including maternal expression dominance (from *S. alterniflora*) and transgressive expression. However, only total gene expression levels were analyzed and the employed technology could not allow distinguishing the contribution of each homeolog. A first reference transcriptome was recently assembled for the parental hexaploid species using several leaf and root cDNA libraries and Roche-454 pyrosequencing (Ferreira de Carvalho et al. 2012). This led to the annotation of c.a. 17,000 genes.

The aim of the present work is (1) to extend transcriptome assembly and annotations by combining both 454 pyrosequencing and Illumina sequencing technologies in five polyploid species:

the hexaploid parents *S. maritima*, *S. alterniflora,* the F1 hybrids *S. x neyrautii* and *S. x townsendii*, and the allododecaploid *S. anglica* and (2) to detect duplicated gene copies in these highly redundant genomes, by developing a strategy aiming at reconstructing haplotypes with no diploid reference genome.

Material and Methods

Sampling, cDNA preparation and sequencing

This study focused on five *Spartina* species: the two hexaploid parents *S. maritima* and *S. alterniflora*, the F1 hybrids *S. x townsendii* and *S. x neyrautii* and the allododecaploid species *S. anglica*). *S. x townsendii* was collected in Hythe (Hampshire, England). Samples from *S. x neyrautii* were collected in Hendaye (Pyrénées Atlantiques, France) and *S. anglica* was sampled in Roscoff and l'Anse de Goulven (Finistère, France).

RNAs were extracted from leaves and roots, from plants grown in same conditions in the greenhouse as indicated in Ferreira de Carvalho *et al*. (2012).

Roche-454 data were sequenced at the Genoscope Platfom (Evry, France) and at the Environmental and Functional Genomics Platform of the University of Rennes 1 (Biogenouest, OSUR, France). Both normalized and non-normalized data were pyrosequenced to enhance the number of assembled contigs as previously published (Genbank accession: SRP015701 and SRP015702; (Ferreira de Carvalho et al. 2012). Roche-454 data of the hybrids and the allopolyploid were obtained using the same protocol as used by Ferreira de Carvalho *et al*. (2012).

Illumina libraries were prepared from cDNAs of the same samples as those used for the 454 pyrosequencing for each 5 species and Illumina (Hi-Seq 2000) sequencing and read-quality trimming (Phred score=20) were performed at the Genoscope Platform (Evry, France) .The number of cleaned reads obtained for each species is indicated in Table 1.

Strategy for assembling Roche-454 and Illumina reads

For each species we independently assembled Roche-454 and Illumina data using the most reliable approaches (Figure 1). Roche-454 reads were first assembled using the GS *de novo* assembler Software v.2.6, Roche (ml=80 bp; mi=90%; (Margulies et al. 2005). The Trinity algorithm (Grabherr et al. 2011) commonly recommended for Illumina RNA-seq assemblies (Chopra et al. 2014; Clarke et al. 2013; Liu et al. 2013) was used with the following parameters: k-mer size of 25 and minimum

contig length of 48. To avoid the formation of chimeric contigs, Roche-454 and Illumina contigs with a length higher than (or equal to) 100 bp were co-assembled using the Newbler software (ml=40 bp; mi=90%). The different contigs obtained after the co-assembly step and Roche-454 contigs and Illumina contigs with a length higher than (or equal to) 40 bp (which were not considered during the co-assembly step) were post-processed by deleting redundant contigs and self-blasted in order to maximize the length of overlapping contigs. Contigs overlapping on 50 bp or more with an identity percentage higher or equal to 90% were then assembled using custom python scripts. Redundancy of the contigs was checked again using a SELFBLAST (minimum length: 40 bp and minimum identity percent: 90%).

Functional annotation

Functional annotations were made following Ferreira de Carvalho *et al*. (2012) and using the Pfam software to detect annotated protein domains from alignments to protein families databases and using a profile Hidden Markov Model (HMM; Finn et al. 2014). All the contigs were analyzed using BLASTn and tBLASTx algorithms (*e*-value threshold of 10^{-5}; (Altschul et al. 1997) against a home-built CDS database including *Oryza sativa*, *Setaria italica*, *Brachypodium distachyon*, *Sorghum bicolor* (www.phytozome.net) and *Zea mays* (concatenation of two databases downloaded on www.phytozome.net and www.plantgdb.org websites). To obtain the homology-based functional annotation Best BLAST Hits (BBH) were selected. The Gene Ontology (GO) was analyzed using the BLAST2Go software (Conesa et al. 2005; Götz et al. 2008). GO annotations were performed using tBLASTx (*e*-value threshold of 10^{-5}) on the different assembled contigs against the *Arabidopsis thaliana* database (TAIR website, www.arabidopsis.org; e-value hit filter of 10^{-6} and a cutoff of 55 which corresponding to the maximum similarity). Pfam 27.0 database was used to enrich the protein domain annotations (6 reading frames tested by contig; PfamB option); Pfam results were filtered by significant hits and e-value lower or equal to 10^{-3} (Finn et al. 2014). Estimation of the number of exons and unigenes (transcripts from the same locus) in each *Spartina* species contigs was performed using BLASTn (≥70% of identity, ≥60pb of overlap) against the rice genome (GFF files downloaded from www.phytozome.net).

SNP detection and haplotype assembly using Illumina data

For each species, the Illumina reads dataset was mapped on the previously built reference contigs using Bowtie 2, v2.0 (Langmead and Salzberg 2012). The parameters used were "score-min: G, 52, 8" for the natural logarithmic function $f(x)= 52 + 8 * ln(x)$ where x correspond to the read length. Using these parameters, all reads (with a length of 80-120 bp) presenting at least 87.06% to 90.30% of

identity were mapped to the reference contig. The output file ".SAM" created by Bowtie 2 during the mapping step was converted to a ".PILEUP" format using the Samtools software suite (Li et al., 2009). We detected for each contig the different SNPs or Single Nucleotide Polymorphisms (minimum read depth=30; SNP detection threshold=2, corresponding to nucleotides that are not present more than 2/100 times per position), using custom python script. These parameters were chosen to remove potential sequencing errors in Illumina reads (below 0.1%) and to avoid the use of false positives SNPs in haplotype construction (Oliphant et al. 2002).

Within each alignment of homologous reads, the different haplotypes were assembled from the ".PILEUP" file and the previously detected SNPs. To construct these haplotypes, we first identified the different reads split in the ".PILEUP" file. The next step consisted in detecting the different haplotypes using each window with a minimum length of 240 nucleotides and containing at least 2 SNPs. Reads that were included in this window were used to detect and to assemble the different haplotypes using the same method as that developed by Boutte et al. (in press) for Roche-454 data (Figure 2). Pairwise comparisons of the reads previously assembled were then performed before creating a new haplotype, by assembling them if the two compared reads present the same SNPs and if no alternative assembly (creating another haplotype) was possible. This method has the advantage of not creating chimeric haplotypes (when two or more choices are possible, the program doesn't assemble reads) but creates many haplotypes (cascade phenomenon). To avoid this problem, we counted the maximum number of haplotypes by sliding windows (see Figure 2.D for description).

In order to explore phylogenetic relationships and the evolutionary history of the reconstructed haplotypes, Maximum Likelihood and Parsimony analyses were conducted using MEGA v5.2.1 (Tamura et al. 2011) on a Pentatricopeptide repeat (PPR) superfamily protein for *S. maritima* and *S. alterniflora*. Homologous sequences from grass sequenced genomes were included in this analysis, with representatives from Chloridoideae (*Eragrostis tef*), Panicoideae (*Zea mays*, *Sorghum bicolor*, and *Setaria italica*), Erhartoideae (*Oryza sativa*) and Pooideae (*Brachypodium distachyon*).

The Kimura two parameters plus Gamma (K2 + G) was selected for this analysis. Bootstrap analyses used 1 000 replicates for the dataset. Visual checking of alignments and SNPs were done using the Tablet software (Milne et al. 2009) for ".ACE" and ".SAM" alignment files and with the Jalview software (Waterhouse et al. 2009) for ".FASTA" files.

Parental haplotype assignation

Following detection of the haplotypes in each species, the parental origin of each haplotype (from *S. maritima* or *S. alterniflora*) was identified in the hybrids (*S. x townsendii*, *S. x neyrautii)* and the allopolyploid (*S. anglica*). The best homologous parental contig for each contig of the hybrid or

allopolyploid species were first identified using BLASTn (*e*-value threshold of 10^{-6}). The contigs of the two parents and the hybrid were then assembled using Newbler (ml=40 bp; mi=80%), before mapping the haplotypes of the three species on the new interspecific contig with Newbler (ml=40 bp; mi=10%). The parental haplotype presenting the maximum identity, the maximum common length and the maximum number of shared SNPs is associated to the hybrid haplotype. When both parental haplotypes are similar to the hybrid haplotype (or if the parental haplotypes are not found), hybrid haplotype was considered as "unassigned" (Figure 3).

K_A/K_s test and molecular dating of duplicate gene divergences

K_A/K_s ratios (Li, Wu, and Luo 1985) between homologous haplotypes of each alignment were calculated for the five species. A new python script was developed in order to (1) translate (using 6 reading frames) the homologous haplotypes from ".FASTA" files (created by the program of SNPs and haplotypes reconstruction) (2) select reading frame(s) with a minimum of stop codons (3) sort alignments by start position and length to select local alignment windows (with a length higher or equal to 120 bp (≥30 amino acids)) with a number of SNPs higher than (or equal to) two, without stop codon and no insertion/deletion polymorphism (4) select for each selected window, the best reading frame(s) and calculate the nucleotide and protein identities and (5) calculate the number of synonymous substitution per site (K_S) and the number of non-synonymous substitution per site (K_A), as estimated by Li *et al*. (1985) and by the Kimura two-parameter method (Kimura 1980). The numbers of transitions (A_i) and transversions (B_i) per *i*th site types are given by:

$$A_i = (1/2)\, ln\, (1/(1 - 2\, P_i - Q_i)) - (1/4)\, ln\, (1 - 2\, Q_i)$$

$$B_i = (1/2)\, ln\, (1/(1 - 2\, Q_i))$$

Where:

Proportion of type *i* transition rate: $P_i = S_i / L_i$

Proportion of type *i* transversion rate: $Q_i = V_i / L_i$

i = 0-fold, 2-fold, 4-fold

Allowing the calculation of K_A and K_S:

$$K_S = (L_2A_2 + L_4A_4 + L_4B_4) / (L_2/3 + L_4)$$

$$K_A = (L_0B_0 + L_2B_2 + L_0A_0) / ((2/3)\, L_2 + L_0)$$

The program outputs the length of each window, nucleotide and protein (amino acid) identities, the K_A, K_S and K_A/K_S ratios and other information for validating and/or filtering out the results. Frequency distributions of K_S values between pairs of haplotypes were performed using the R software (v. 2.13.0; (R Development Core Team 2011) to detect duplication events (Blanc and Wolfe 2004). We estimated the ages of the detected peaks using clock-like rates of synonymous substitution of 6.5 x 10^{-9} substitutions/synonymous site/year for Grasses (Gaut et al. 1996) for dating polyploïdization events.

Results

De novo Assemblies and Functional annotation

The number of contigs assembled from the five species ranged from 44 158 to 65 099 (Table 2). Using the *O. sativa* genome and its gene annotation as a reference, 35 039 and 32 734 exons were detected in the two parental species *S. maritima* and *S. alterniflora* respectively, and 40 365 and 34 792 in the two hybrids *S. x townsendii* and *S. x neyrautii*. In *S. anglica*, 35 062 were assembled. This search for exons from a reference annotated genome allowed identification of unigenes in the *Spartina* transcriptomes, ranging from 13 054 to 16 002 (which represents 26.61-32.62% of the expected number of unigenes). The number of Illumina contigs obtained using the Trinity assembler is more important in the F1 and the allopolyploid than in the parents (121 733, 110 455, 144 550 for the 3 hybrid species and 98 455, 76 010 for the parental species) while the number of Roche-454 contigs obtained with Newbler is more important in the parents due to a deeper sequencing and the presence of both normalized and non-normalized libraries (Table 1). The co-assembly (using Newbler) of Trinity (Illumina reads) and Newbler (Roche-454 reads) sub-assemblies formed 12 674, 9 232, 13 691, 8 768 and 11 201 Roche-454/Illumina hybrid contigs for *S. maritima*, *S. alterniflora*, *S. x townsendii*, *S. x neyrautii* and *S. anglica* respectively, with a proportion of non-co-assembled contigs ranging from 40.73% to 54.79% of the Roche-454 contigs and from 62.77% to 69.38% of the Illumina contigs. After comparisons of annotated contigs from the 5 transcriptomes, a total of 37 867 different annotated genes were obtained, 15 114 genes being redundant between species. The number of annotated contigs specific to each transcriptome was determined: 4 456, 3 760, 3 528, 6 751 and 4 168 contigs were found specific to *S. maritima, S. alterniflora*, *S. x townsendii*, *S. x neyrautii* and *S. anglica* respectively. For the 5 transcriptomes, the total sequence length is similar for 4 species (ranging from 27,548,352 bp to 29,841,745 bp) and less important for *S. alterniflora* (23,016,772 bp). The GC% and the N50 are

similar for the 5 species. The high values of the N50 are explained by the presence of Roche-454 contigs in the dataset (Table 2). For *S. maritima*, the length of the annotated contigs ranged from 58 to 10 313 bp (742 bp on average). For these contigs, the number of mapped reads (at 90% of identity) varied between 1 and 251 685 (643 reads per contig on average). The average coverage of each nucleotide within contigs is estimated to be 56.51x. The length of the unannotated contigs ranged from 40 to 4 701 bp (340 bp on average) and the number of mapped reads varied from 1 to 46 647 (146 reads per contig on average). For these contigs, the average read depth was estimated as 37.16x (Supplementary Figure 1). The assemblies of the four other species exhibit similar contig length, read depth, for both annotated and unannotated contigs to *S. maritima* and values are available in the Supplementary Table 1.

SNP detection and Haplotype construction

For the 2 parents *S. maritima* and *S. alterniflora*, we have detected a similar number of SNPs within processed alignments (3.854 and 3.850 SNPs for 100 bp respectively, Student's test p-value>0.05). The number of SNPs for 100 bp detected in the 2 hybrids and the allododecaploid *S. anglica* is more important than in the parents: 6.104 SNPs for 100 bp for S. *x townsendii* and a similar number of SNPs *in S. x neyrautii* and *S. anglica*, 5.323 and 5.327 respectively (Figure 4.a). After detecting haplotypes from the detected SNPs, the average number of haplotypes corresponding to the maximal number of haplotypes by region was calculated. At least two haplotypes were detected in 20 085 and 13 216 contigs of the parents *S. maritima* and *S. alterniflora*. For the hybrids *S. x townsendii*, *S. x neyrautii* and the allododecaploid *S. anglica*, 24 199, 16 776 and 18 839 contigs, respectively, exhibit at least 2 haplotypes (Table 2). A similar number of haplotypes was detected in the 2 parents (7.252 for *S. maritima* and 6.936 for *S. alterniflora*) while about twice more were detected in the other species, with 13.729, 11.631 and 11.722 haplotypes in *S. x townsendii*, *S. x neyrautii* and *S. anglica*, respectively (Figure 4.b). The number of SNPs for 100 bp and the maximal number of haplotypes by region is more important in *S. x townsendii* than in the parents, *S. x neyrautii* and the allopolyploid *S. anglica*. For each window where haplotypes reconstruction was possible, we have compared the number of windows presenting a similar number of haplotypes using a Fisher's exact test. For the two parental hexaploid species (*S. maritima* and *S. alterniflora*), 43.53% and 42.38% of the windows presented between 2 and 4 haplotypes. Within the two hexaploid hybrids and the allododecaploid species *S. anglica*, a lower percentage of windows presenting 2 to 4 haplotypes were detected (18.01% in *S. townsendii,* 21.80% in *S. x neyrautii* and 21.64% in *S. anglica*). However, the number of regions presenting between 5 to 12 haplotypes is similar for the two parents, the two F1 hybrids and the allopolyploid *S. anglica* (from 43.03% in *S. x townsendii* to 48.68% in *S. x neyrautii*). The number of regions presenting more than 12 haplotypes is more important in *S. x townsendii* and *S. x neyrautii* and

the allopolyploid (38.93%, 29.50% and 30.97% respectively) compared to the two hexaploid parents (12.16% and 10.51% for *S. maritima* and *S. alterniflora* respectively). The average length of the haplotypes reconstructed is more important than the initial length of the reads, the majority of haplotypes (≥76.44%) presents a length ranging from 150 and 450 bp for all 5 species (Figure 4.c).

Parental haplotype assignation

After haplotypes reconstruction in the 5 species studied, the haplotypes detected in the 3 hybrids (*S. x townsendii*, *S. x neyrautii* and *S. anglica*) were co-aligned together with their parents (*S. maritima* and *S. alterniflora*) to identify the parental origin of each haplotype and putative homeologs. For the F1 hybrid *S. x townsendii* we identified putative homoeologous copies for 7 293 contigs (10 693 windows totaling 266 820 local haplotypes); 135 298 and 108 548 haplotypes were assigned to *S. maritima* and *S. alterniflora* respectively. The number of unassigned haplotypes corresponds to haplotypes where the two parental copies are similar or where one parental copy is not present and correspond to 22 974 haplotypes in this hybrid. In the second hybrid *S. x neyrautii*, 97 516, 79 414 haplotypes were assigned to *S. maritima*, *S. alterniflora* and 18 154 were unassigned for 6 947 contigs (9 771 windows totaling 195 084 local haplotypes). In the allododecaploid *S. anglica:* 106 314, 87 884 haplotypes were assigned to *S. maritima*, *S. alterniflora* and 19 238 were unassigned for 7 153 contigs (10 159 windows totaling 213 436 local haplotypes).

For the 3 species, the number of haplotypes assigned to the parental species *S. maritima* is similar, ranging from 49.80% to 50.71% and represents the majority of the copy assigned. The number of haplotypes assigned to *S. alterniflora* is ranging from 40.69% to 41.18% and similar for the 3 species. The number of unassigned copies for the 3 dataset is less than 10% (similar for the 3 species, between 8.60% and 9.30%; Table 3).

K_A/K_S ratio test and molecular dating of duplicate genes

For each species, we calculated the K_A/K_S ratio between the different copies to evaluate the type of selective pressure that haplotypes have been subjected to. For the 5 species, 68.98% to 81.00% of the K_A/K_S ratios are included between 0 and 0.5 (indicating negative selective constraints). The number of K_A/K_S ratios included between 0.5 and 1 represent 13.52% to 22.08% of the comparisons. Only 5.48% to 10.08% of the ratios are greater than 1. These values are similar for the two parents and the hybrid *S. x townsendii* on the one hand and similar for *S. x neyrautii* and *S. anglica* on the other hand (Fisher's exact test, p-value > 0.05).

Frequency distributions of K_S values between pairs of haplotypes are presented in Figure 5. For the 2 parents, *S. maritima* and *S. alterniflora* and the two F1 hybrids a first peak (0.023 - 0.035) is observed. A second peak is present in the *S. maritima* K_S distribution (0.113 - 0.125). We estimated the age of these two peaks, most likely resulting from duplication events, using the calibrated molecular clock (for synonymous substitutions) in grasses (Gaut et al. 1996). The first peak (Ks 0.023 - 0.035) was estimated between 1.8 and 2.7 Mya. The second peak present in *S. maritima* data (Ks 0.113 - 0.125), resulting from an oldest duplication, was estimated between 8.7 and 9.6 Mya.

Phylogenetic analysis of the haplotypes detected in the PPR gene

Phylogeny of the different haplotypes (detected using the developed program) for the Pentatricopeptide repeat (PPR) superfamily protein gene is presented in Figure 6. All the 11 *Spartina* haplotypes form a monophyletic group with *Eragrostis tef* (Chloridoideae) as a sister lineage as expected from the organismal history. Only one PPR copy is found in the other grasses, exept in *Setaria italica* where two sister copies are encountered. These two copies most likely result from individual gene duplication in the diploid *S. italica*. The *Spartina* haplotypes are distributed in two clades (A and B) each containing sequences from both *S. alterniflora* and *S. maritima*. In clade A, the haplotypes from each hexaploid species *S. maritima* and *S. alterniflora* form two subclades containing respectively 3 and 2 haplotypes. The position of a third *S. alterniflora* haplotype is not resolved between these two subclades. In clade B, two subclades contain respectively two haplotypes of *S. alterniflora* and 2 of *S. maritima* and one *S. alterniflora* haplotype Figure 6).

According to the tree topology, clades A and B could be interpreted as homoeologous copies duplicated in the hexaploid ancestor of *S. maritima* and *S. alterniflora*. Divergence between the "maritima" and "alterniflora" subclades for each of these homeologs could reflect the divergence following speciation between *S. alterniflora* and *S. maritima*. The position of the *S. alterniflora*_0 haplotype which is branched within a "maritima" subclade is unexpected and needs further investigations.

Discussion

In this study, we report the assembly and annotation of five reference transcriptomes for the European hexaploid *Spartina* species (*S. maritima*, *S. alterniflora* and their homoploid hybrids *S. x townsendii* and *S. x neyrautii*) and the allododecaploid invasive species *S. anglica*. The use of a deep sequencing technology significantly enhanced the previously assembled and published reference transcriptomes

built for the hexaploid parental species (Ferreira de Carvalho et al. 2012) with 60 644 and 44 158 contigs against 25 239 and 14 317 for *S. maritima* and *S. alterniflora*, respectively and up to 30% more functionally annotated contigs. We also provide here the first reference transcriptomes of the two hybrids and the allododecaploid species *S. anglica*. As the repetitive nature of *Spartina* genomes and transcriptome due to their high ploidy levels and their hybrid origin was challenging, we developed generic bioinformatics tools to 1) detect different haplotypes of each gene within these species and 2) assign a parental origin to haplotypes detected in the hybrids and the allopolyploid. The approach described here allows the detection of putative homeologs from sets of short reads and can be applied for future differential gene expression or genomics experiments to study the fate of duplicated genes in the allododecaploid *S. anglica*

Spartina transcriptomics

Before the NGS revolution *Spartina* transcriptomic resources were restricted to few EST sequences available on NCBI databases (Baisakh, Subudhi, and Varadwaj 2008; Chelaifa, Mahé, and Ainouche 2010). Whole genome expression experiments were designed using heterologous rice microarrays to demonstrate differential expression in similar growing conditions of *S. maritima* and *S. alterniflora* (Chelaifa, Mahé and Ainouche 2010) and the relative effects of hybridization and genome duplication on non-additive expression in *S. x neyrautii*, *S. x townsendii* and *S. anglica* (Chelaifa, Monnier and Ainouche et al. 2010). Besides the non-species specific design of the array of this approach that was limiting the number of transcript detected, the measured signals were including all putative homeologs, disabling the study of each duplicated gene. NGS technologies were then first used to build reference transcriptomes of *Spartina pectinata*, a tetraploid species using Roche-454 data (Gedye et al. 2010) and for *S. maritima* and *S. alterniflora* (Ferreira de Carvalho *et al*. (Ferreira de Carvalho et al. 2012). In our study, we used a combination of Roche-454 and Illumina deep sequencing reads datasets to improve the reference transcriptomes of the two parents *S. maritima* and *S. alterniflora*, and to assemble the first reference transcriptomes for the two hybrids *S. x townsendii*, *S. x neyrautii* and the allododecaploid *S. anglica*. We chose to independently assemble the Roche-454 (with Newbler) and Illumina read datasets (with Trinity) before co-aligning them (with Newbler and custom scripts to enhance assemblies by self-blast). The Newbler software is commonly used for Roche-454 data (Margulies et al. 2005) and showed positive results on similar datasets (Ferreira de Carvalho et al. 2012). The choice of the Trinity software is based on the results of several studies (Clarke et al. 2013; Liu et al. 2013) and comparative tests on our dataset. The hybrid assembly strategy for Roche-454 and Illumina contigs showed good results in several studies (Barthelson et al. 2011; Jiang et al. 2011; Sirota-Madi et al. 2010) using assemblers such as Mira or Abyss. The length of Roche-454 contigs obtained in the first step of our assembly process motivated the choice of the

Newbler assembler. The large number of Illumina contigs obtained by the Trinity software (76 010 to 121 733 contigs) is explained by the presence of several similar copies (identity ≥ 90%) and were automatically re-assembled with Newbler in the co-assembly step. The parameters used for the different assemblies (90%) are consistent with the literature (Ferreira de Carvalho et al. 2012; Franchini, Van der Merwe, and Roodt-Wilding 2011; Liu et al. 2013) and fitted in order to get consensus sequences of all putative homeologs for each species. The functional annotations were made using a method similar to that used by Ferreira de Carvalho and collaborators (tblastx, Blast2Go approach; (Ferreira de Carvalho et al. 2012)) and using the Pfam software used in annotation pipeline such as TRAPID (Van Bel et al. 2013). The number of annotated contigs represents 36.50% to 43.57% of the total number of contig, the unannotated contigs have a lower average length, but also a lower average read depth and are reconstructed using a limited number of reads compared to annotated contigs (see Additional Figure 1 and Additional Table 1 for details); they also are shorter (40 to 200 bp). The annotated contigs of the two parents *S. maritima* and *S. alterniflora* have an average length of 741.75 and 761.11 bp respectively similar to contigs assembled by Ferreira de Carvalho and collaborators (2012) who reported an average length included between 415 and 759 (617 and 415 bp for *S. maritima* and 759 bp for *S. alterniflora*). The average length of the unannotated contigs corresponds to 339.56 and 336.00 bp for *S. maritima* and *S. alterniflora*, respectively (these results are similar for the other species). To validate the contig constructed and to detect the different SNPs and haplotypes, we have mapped (to 90% of identity) between 34.22% to 57,84% among different species. Contigs have an average read depth included to 25.12x to 90.52x. These values are similar to the study of Franchini, Van der Merwe, and Roodt-Wilding (2011) where 10 635 178 Illumina paired-end reads (42.10%) have been used (36.6x of average read depth) to construct the transcriptome of an abalone species.

Haplotype detection

Several studies have focused on the detection of different copies in polyploid genomes such as cotton, coffee, strawberry or even the paleopolyploid soybean genome. (Combes et al. 2013; Flagel et al. 2008; Ilut et al. 2012; Salmon et al. 2009; Tennessen et al. 2014). Nevertheless, the strategies developed in these studies can only be applied on species with known diploid parents. Detection of the different copies in *Spartina* hexaploid species using Roche-454 data was previously restricted to a few genes (Ferreira de Carvalho et al. 2012) and was recently automated for rDNA gene copies in *S. maritima* (Boutte et al, in press). Our study reports here the automated detection of haplotypes at a whole transcriptome scale enabling us to identify the parental origin of the hybrid and polyploid haplotypes.

In our study, the number of haplotypes detected in the investigated *Spartina* species is correlated with the number of SNPs detected and the number of copies expected. For the parents, we have detected around 7 haplotypes by windows and 12 to 14 for the hybrids and the allododecaploid species. These values are more important than the number of homoeologous copies expected (3 pairs for the hexaploids parents and for the hybrids and 6 pairs for the allopolyploid) suggesting co-alignements and detection of either paralogs or alleles. A previous study focusing on a few targeted genes demonstrated the detection of 4 haplotypes in the parental species with Roche-454 data that indicated the presence of two homoeologous copies (Ferreira de Carvalho et al. 2012). A study realized on Waxy genes using cloning and Sanger sequencing indicates the presence of one homoeologous copy in *S. maritima* and three copies in *S. alterniflora* (Fortune et al. 2007). Furthermore, the non-additive gene expression in polyploid species could lead to a number of detected copies less important (Yoo et al. 2014). The phylogeny of the different haplotypes for a Pentatricopeptide repeat (PPR) superfamily protein gene indicates the presence of two divergent homoeologous copies and additional alleles (2-3 per copy in *S. maritima* and 3 in *S. alterniflora*). Prevalence of reticulate evolution in *Spartina* and previous gene topologies (*e.g. Waxy* gene, Fortune et al. 2007) suggested an allopolyploid origin of the hexaploid ancestor to *S. maritima* and *S. alterniflora*; however, we cannot rule out a possible auto-allo hexaploid origin, which would result in divergent homeologs and additional related homologous alleles.

The number of haplotypes can be also explained by the difficulty to assemble the reconstructed haplotypes with Illumina data (cascade phenomenon) and the choice to not create chimeric haplotypes. The number of SNPs and haplotypes detected in the hybrid *S. x townsendii* is more important than values detected in the other hybrid and the allododecaploid. This information suggests the presence of more copies expressed in this hybrid compared to *S. x neyrautii* and *S. anglica*. Genome duplication in *S. anglica* could have reduced the number of copies expressed compared to *S. x townsendii* as a consequence of genome doubling. The parental haplotype assignation validates a majority of parental copies detected by the developed program using different datasets. The higher number of copies assigned to *S. maritima* for the two hybrids and the alllododecaploid species most likely results from the higher number of sequenced libraries (including normalized libraries) for this species. Another explanation would be that the number of *S. maritima* copies expressed in the hybrids and the allododecaploid is more important.

Frequency distributions of K_S values between pairs of haplotypes (Blanc and Wolfe 2004) exhibited two peaks, one common to four species (0.023 - 0.035) and a second one present in the hexaploid parent *S. maritima* (0.113 - 0.125). These peaks indicate a burst of the number of duplicated genes resulting from genome duplication events, and the presence of two divergent sets of duplicated copies and may be related to the tetraploidy and hexaploidy events in *Spartina*. The divergence between tetraploid and hexaploid clades on the one hand and between hexaploid species on the other were

recently estimated using chloroplast genomes of *S. maritima* (Rousseau-Gueutin et al. 2015). The divergence between the plastome of the hexaploid species *S. maritima* and *S. alterniflora* was estimated between 2 and 4 Myr and the divergence between the plastomes of the tetraploid and hexaploid species was estimated as 6-10 Myr. Using clock-like rates of synonymous substitution for nuclear genes indicated by Gaut et al. (1996) for Poaceae, we associated dating to the 2 peaks: 1.8-2.7 Myr and 8.7-9.6 Myr respectively. Haplotypes detected using our developed program could correspond to homoeologous copies of the two polyploidization events. Difference between dated events could be explained by the distinct evolution of chloroplast and nuclear genes (Wolfe, Li, and Sharp 1987). The oldest duplicated haplotypes detected here were estimated to 12-13 Myr, which is related to the mapping parameters selected for this study including reads between 88 and 100% of identity. No additional peak was observed in the recent allododecaploid species; this is due to the too large number of haplotypes present in the transcriptome of this species. Simulation of an allododecaploid species using parental haplotypes provided similar results (Additional Figure 2.A). The different peaks observed in the two hybrids *S. x townsendii* and *S. x neyrautii* is due to the presence of the haplotypes of the two parents. The hybrid simulation using parental mapping reads process confirmed these results (Additional Figure data 2.B). In fact, it was possible to observe 3 peaks, two common to *S. maritima* species and corresponding to duplication events and the third peak present in the hybrid species.

Conclusion

In conclusion, we have constructed five new reference *Spartina* transcriptomes using a specific approach that combine two NGS technologies: Roche-454 and Illumina. After transcriptomic assembly and annotation of the different contigs, SNP detection allowed reconstructing different haplotypes, which could correspond to paralogous, homoeologous and even allelic copies. We have detected around seven haplotypes for the hexaploid parents and around twelve to fourteen in the two hybrids and the allododecaploid that were assigned to a parental origin. K_s distributions peaks indicate two duplication events in the hexaploid *Spartina maritima*, dating of these events are similar with the literature and indicate the probable origin of the detected copies.

The *Spartina* reference transcriptomes constructed may provide useful informations to explore gene expression in the context of Spartina ecology, such as genes implicated in responses to abiotic stresses (salt and oxidative stress or to heavy metal stress for example), biotic interactions (Gray and Benham 1990) and in the context of allopolyploid speciation. Previous studies analyzed expression levels in the five *Spartina* species studied here using a rice heterologous array. Transcriptomic analyses were also realized in *S. alterniflora* to study salt stress and petroleum hydrocarbon response (Baisakh, Subudhi,

and Varadwaj 2008; RamanaRao et al. 2012). Transcriptomes constructed here offer the opportunity to study a large number of gene implicated in these pathways. It is now possible to study the different transcription levels of the detected copies using a specific experiment design; this opens a very interesting perspective to the study of the expression evolution of these duplicate copies in the context of the adaptive success of *S. anglica*.

References

Abbott, R. J., A. C. Brennan, J. K. James, D. G. Forbes, M. J. Hegarty, and S. J. Hiscock. **2008**. "Recent Hybrid Origin and Invasion of the British Isles by a Self-Incompatible Species, Oxford Ragwort (*Senecio squalidus* L., *Asteraceae*)." Biological Invasions **11** (5): 1145–58.

Adams, K. L., and J. F. Wendel. **2005**. "Polyploidy and Genome Evolution in Plants." Current Opinion in Plant Biology **8** (2): 135–41.

Adams, K. L., R. Cronn, R. Percifield, and J. F. Wendel. **2003**. "Genes Duplicated by Polyploidy Show Unequal Contributions to the Transcriptome and Organ-Specific Reciprocal Silencing." Proceedings of the National Academy of Sciences **100** (8): 4649–54.

Ainouche, M. L., A. Baumel, and A. Salmon. **2004**. "*Spartina anglica* CE Hubbard: A Natural Model System for Analysing Early Evolutionary Changes That Affect Allopolyploid Genomes." Biological Journal of the Linnean Society **82** (4): 475–84.

Ainouche, M L, H. Chelaifa, J. Ferreira de Carvalho, S. Bellot, A. K. Ainouche, and A. Salmon. **2012**. "Polyploid Evolution in *Spartina*." In Polyploidy and Genome Evolution: Dealing with Highly Redundant Hybrid Genomes, 225–43. Soltis, Pamela S.; Soltis, Douglas E. (eds) Polyploidy and Genome Evolution, Springer Berlin Heidelberg: Berlin, Heidelberg.

Ainouche, M. L., P. M. Fortune, A. Salmon, C. Parisod, M.-A. Grandbastien, K. Fukunaga, M. Ricou, and M.-T. Misset. **2008**. "Hybridization, Polyploidy and Invasion: Lessons from *Spartina* (Poaceae)." Biological Invasions **11** (5): 1159–73.

Albertin, W. **2006**. "Numerous and Rapid Nonstochastic Modifications of Gene Products in Newly Synthesized *Brassica napus* Allotetraploids." Genetics **173** (2): 1101–13.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. **1997**. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." Nucleic Acids Research **25** (17): 3389–3402.

Baisakh, N., P. K. Subudhi, and P. Varadwaj. **2008**. "Primary Responses to Salt Stress in a Halophyte, Smooth Cordgrass (*Spartina alterniflora* Loisel.)." Functional & Integrative Genomics **8** (3): 287–300.

Barthelson, R., A. J. McFarlin, S. D. Rounsley, and S. Young. **2011**. "Plantagora: Modeling Whole Genome Sequencing and Assembly of Plant Genomes." Edited by Matteo Pellegrini. PLoS ONE **6** (12): e28436.

Baumel, A, M. L. Ainouche, and J. E. Levasseur. **2001**. "Molecular Investigations in Populations of *Spartina anglica* CE Hubbard (Poaceae) Invading Coastal Brittany (France)." Molecular Ecology **10** (7): 1689–1701.

Baumel, A, M. L. Ainouche, R. J. Bayer, A. K. Ainouche, and M. T. Misset. **2002**. "Molecular Phylogeny of Hybridizing Species from the Genus *Spartina* Schreb. (Poaceae)." Molecular Phylogenetics and Evolution **22** (2): 303–14.

Baumel, A, M. L. Ainouche, R. Kalendar, and A. H. Schulman. **2002**. "Retrotransposons and Genomic Stability in Populations of the Young Allopolyploid Species *Spartina anglica* CE Hubbard (Poaceae)." Molecular Biology and Evolution **19** (8): 1218–27.

Blanc, G., and K. H. Wolfe. **2004**. "Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes." The Plant Cell Online **16** (7): 1667–78.

Buggs, R. J. A., S. Renny-Byfield, M. Chester, I. E. Jordon-Thaden, L. F. Viccini, S. Chamala, A. R. Leitch, et al. **2012**. "Next-Generation Sequencing and Genome Evolution in Allopolyploids." American Journal of Botany **99** (2): 372–82.

Chalhoub, B., F. Denoeud, S. Liu, I. A. P. Parkin, H. Tang, X. Wang, J. Chiquet, et al. **2014**. "Early Allopolyploid Evolution in the Post-Neolithic *Brassica napus* Oilseed Genome." Science **345** (6199): 950–53.

Chelaifa, H., F. Mahé, and M. L. Ainouche. **2010**. "Transcriptome Divergence between the Hexaploid Salt-Marsh Sister Species *Spartina maritima* and *Spartina alterniflora* (Poaceae)." Molecular Ecology **19** (10): 2050–63.

Chelaifa, H., A. Monnier, and M. L. Ainouche. **2010**. "Transcriptomic Changes Following Recent Natural Hybridization and Allopolyploidy in the Salt Marsh Species *Spartina x townsendii* and *Spartina anglica* (Poaceae)." New Phytologist **186** (1): 161–74.

Chelaifa, H., A. Monnier, and M. L. Ainouche. **2010**. "Transcriptomic Changes Following Recent Natural Hybridization and Allopolyploidy in the Salt Marsh Species *Spartina x townsendii* and *Spartina anglica* (Poaceae)." New Phytologist **186** (1): 161–74.

Chopra, R., G. Burow, A. Farmer, J. Mudge, C. E. Simpson, and M. D. Burow. **2014**. "Comparisons of *de novo* Transcriptome Assemblers in Diploid and Polyploid Species Using Peanut (*Arachis* Spp.) RNA-Seq Data." PloS One **9** (12): e115055.

Clarke, K., Y. Yang, R. Marsh, L. Xie, and K. K. Zhang. **2013**. "Comparative Analysis of *de novo* Transcriptome Assembly." Science China Life Sciences **56** (2): 156–62.

Combes, M.-C., A. Dereeper, D. Severac, B. Bertrand, and P. Lashermes. **2013**. "Contribution of Subgenomes to the Transcriptome and Their Intertwined Regulation in the Allopolyploid *Coffea arabica* Grown at Contrasted Temperatures." New Phytologist **200** (1): 251–60.

Combes, M.-C., A. Cenci, H. Baraille, B. Bertrand, and P. Lashermes. **2012**. "Homeologous Gene Expression in Response to Growing Temperature in a Recent Allopolyploid (*Coffea arabica* L.)." Journal of Heredity **103** (1): 36–46.

Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles. **2005**. "Blast2GO: A Universal Tool for Annotation, Visualization and Analysis in Functional Genomics Research." Bioinformatics **21** (18): 3674–76.

Duchemin, W., P.-Y. Dupont, M. A. Campbell, A. RD Ganley, and M. P. Cox. **2014**. "HyLiTE: Accurate and Flexible Analysis of Gene Expression in Hybrid and Allopolyploid Species." BMC Bioinformatics **16** (1).

Feldman, M., B. Liu, G. Segal, S. Abbo, A. A. Levy, and J. M. Vega. **1997**. "Rapid Elimination of Low-Copy DNA Sequences in Polyploid Wheat: A Possible Mechanism for Differentiation of Homoeologous Chromosomes." Genetics **147** (3): 1381–87.

Ferreira de Carvalho, J., J. Poulain, C. Da Silva, P. Wincker, S. Michon-Coudouel, A. Dheilly, D. Naquin, J. Boutte, A. Salmon, and M. L. Ainouche. **2012**. "Transcriptome *de novo* Assembly from next-Generation Sequencing and Comparative Analyses in the Hexaploid Salt Marsh Species *Spartina maritima* and *Spartina alterniflora* (Poaceae)." Heredity **110** (2): 181–93.

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, et al. **2014**. "Pfam: The Protein Families Database." Nucleic Acids Research **42** (D1): D222–30.

Flagel, L. E., L. Chen, B. Chaudhary, and J. F. Wendel. **2009**. "Coordinated and Fine-Scale Control of Homoeologous Gene Expression in Allotetraploid Cotton." Journal of Heredity **100** (4): 487–90.

Flagel, L., J. Udall, D. Nettleton, and J. Wendel. **2008**. "Duplicate Gene Expression in Allopolyploid Gossypium Reveals Two Temporally Distinct Phases of Expression Evolution." BMC Biology **6** (1): 16.

Fortune, P. M., K. A. Schierenbeck, A. K. Ainouche, J. Jacquemin, J. F. Wendel, and M. L. Ainouche. **2007**. "Evolutionary Dynamics of *Waxy* and the Origin of Hexaploid *Spartina* Species (Poaceae)." Molecular Phylogenetics and Evolution **43** (3): 1040–55.

Foucaud. **1897**. "Un *Spartina* Inédit." Ann Soc Sci Nat Char Inf **32**: 220–22.

Franchini, P., M. Van der Merwe, and R. Roodt-Wilding. **2011**. "Transcriptome Characterization of the South African Abalone *Haliotis midae* Using Sequencing-by-Synthesis." BMC Research Notes **4** (1): 59.

Gaeta, R. T., J. C. Pires, F. Iniguez-Luy, E. Leon, and T. C. Osborn. **2007**. "Genomic Changes in Resynthesized *Brassica napus* and Their Effect on Gene Expression and Phenotype." The Plant Cell Online **19** (11): 3403–17.

Gaut, B. S., B. R. Morton, B. C. McCaig, and M. T. Clegg. **1996**. "Substitution Rate Comparisons between Grasses and Palms: Synonymous Rate Differences at the Nuclear Gene Adh Parallel Rate Differences at the Plastid Gene rbcL." Proceedings of the National Academy of Sciences **93** (19): 10274–79.

Gedye, K., J. Gonzalez-Hernandez, Y. Ban, X. Ge, J. Thimmapuram, F. Sun, C. Wright, S. Ali, A. Boe, and V. Owens. **2010**. "Investigation of the Transcriptome of Prairie Cord Grass, a New Cellulosic Biomass Crop." The Plant Genome Journal **3** (2): 69.

Götz, S., J. M. García-Gómez, J. Terol, T. D. Williams, S. H. Nagaraj, M. J. Nueda, M. Robles, M. Talón, J. Dopazo, and A. Conesa. **2008**. "High-Throughput Functional Annotation and Data Mining with the Blast2GO Suite." Nucleic Acids Research **36** (10): 3420–35.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, and Q. Zeng. **2011**. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." Nature Biotechnology **29** (7): 644–52.

Gray, A. J., and P. E. M. Benham. **1990**. *Spartina anglica* - a Research Review. HMSO, London.

Gregory, T.R., and B. K. Mable. **2005**. "Polyploidy in Animals." In The Evolution of the Genome, 428–517. London: Elsevier Academic Press.

Higgins, J., A. Magusin, M. Trick, F. Fraser, and I. Bancroft. **2012**. "Use of mRNA-Seq to Discriminate Contributions to the Transcriptome from the Constituent Genomes of the Polyploid Crop Species *Brassica napus*." BMC Genomics **13** (1): 247.

Hovav, R., J. A. Udall, B. Chaudhary, R. Rapp, L. Flagel, and J. F. Wendel. **2008**. "Partitioned Expression of Duplicated Genes during Development and Evolution of a Single Cell in a Polyploid Plant." Proceedings of the National Academy of Sciences **105** (16): 6191–95.

Ilut, D. C., J. E. Coate, A. K. Luciano, T. G. Owens, G. D. May, A. Farmer, and J. J. Doyle. **2012**. "A Comparative Transcriptomic Study of an Allotetraploid and Its Diploid Progenitors Illustrates the Unique Advantages and Challenges of RNA-Seq in Plant Species." American Journal of Botany **99** (2): 383–96.

Jiang, Y., J. Lu, E. Peatman, H. Kucuktas, S. Liu, S. Wang, F. Sun, and Z. Liu. **2011**. "A Pilot Study for Channel Catfish Whole Genome Sequencing and *de novo* Assembly." BMC Genomics **12** (1): 629.

Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, *et al*. **2011**. "Ancestral Polyploidy in Seed Plants and Angiosperms." Nature **473** (7345): 97–100.

Kashkush, K., M. Feldman, and A. A. Levy. **2002**. "Transcriptional Activation of Retrotransposons Alters the Expression of Adjacent Genes in Wheat." Nature Genetics **33** (1): 102–6.

Kimura, M. **1980**. "A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences." Journal of Molecular Evolution **16** (2): 111–20.

Langham, R. J., J. Walsh, M. Dunn, C. Ko, S. A. Goff, and M. Freeling. **2004**. "Genomic Duplication, Fractionation and the Origin of Regulatory Novelty." Genetics **166** (2): 935–45.

Langmead, B., and S. L Salzberg. **2012**. "Fast Gapped-Read Alignment with Bowtie 2." Nature Methods **9** (4): 357–59.

Leitch A. R., Leitch I. J. **2008**. "Genomic Plasticity and the Diversity of Polyploid Plants." Science. **320**:481–483.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, 1000 Genome Project Data Processing Subgroup. **2009**. "The Sequence Alignment/Map Format and SAMtools." Bioinformatics **25** (16): 2078–79.

Liu, B., J. M. Vega, and M. Feldman. **1998**. "Rapid Genomic Changes in Newly Synthesized Amphiploids of *Triticum* and *Aegilops*. II. Changes in Low-Copy Coding DNA Sequences." Genome / National Research Council Canada **41** (4): 535–42.

Liu, S., W. Li, Y. Wu, C. Chen, and J. Lei. **2013**. "*De novo* Transcriptome Assembly in Chili Pepper (*Capsicum frutescens*) to Identify Genes Involved in the Biosynthesis of Capsaicinoids." Edited by Christian Schönbach. PLoS ONE **8** (1): e48156.

Li, W. H., C. I. Wu, and C. C. Luo. **1985**. "A New Method for Estimating Synonymous and Nonsynonymous Rates of Nucleotide Substitution Considering the Relative Likelihood of Nucleotide and Codon Changes." Molecular Biology and Evolution **2** (2): 150–74.

Lukens, L. N. **2005**. "Patterns of Sequence Loss and Cytosine Methylation within a Population of Newly Resynthesized *Brassica napus* Allopolyploids." Plant Physiology **140** (1): 336–48.

Mable, B. K. **2004**. "'Why Polyploidy Is Rarer in Animals than in Plants': Myths and Mechanisms." Biological Journal of the Linnean Society **82** (4): 453–66.

Mable, B. K., M. A. Alexandrou, and M. I. Taylor. **2011**. "Genome Duplication in Amphibians and Fish: An Extended Synthesis: Polyploidy in Amphibians and Fish." Journal of Zoology **284** (3): 151–82.

Malinska, H., J. A. Tate, E. Mavrodiev, R. Matyasek, K. Y. Lim, A. R. Leitch, D. E. Soltis, P. S. Soltis, and A. Kovarik. **2011**. "Ribosomal RNA Genes Evolution in Tragopogon: A Story of New and Old World Allotetraploids and the Synthetic Lines." Taxon **60** (2): 348.

Marchant, C. J. **1963**. "Corrected Chromosome Numbers for *Spartina x townsendii* and Its Parent Species." Nature **199**: 929.

Marchant, C. J. **1968**. "Evolution in *Spartina* (Gramineae). II. Chromosomes, Basic Relationships and the Problem of *Spartina x townsendii* Agg." Botanical Journal of the Linnean Society **60**: 381–409.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, and Z. Chen. **2005**. "Genome Sequencing in Microfabricated High-Density Picolitre Reactors." Nature **437** (7057): 376–80.

Marhold, K., H. Kudoh, J.-H. Pak, K. Watanabe, S. Spaniel, and J. Lihova. **2009**. "Cytotype Diversity and Genome Size Variation in Eastern Asian Polyploid Cardamine (Brassicaceae) Species." Annals of Botany **105** (2): 249–64.

Marmagne, A., P. Brabant, H. Thiellement, and K. Alix. **2010**. "Analysis of Gene Expression in Resynthesized *Brassica napus* Allotetraploids: Transcriptional Changes Do Not Explain Differential Protein Regulation." The New Phytologist **186** (1): 216–27.

McCarthy, E. W., S. E. J. Arnold, L. Chittka, S. C. Le Comber, R. Verity, S. Dodsworth, S. Knapp, *et al*. **2015**. "The Effect of Polyploidy and Hybridization on the Evolution of Floral Colour in *Nicotiana* (Solanaceae)." Annals of Botany **115** (7): 1117–31.

Milne, I., M. Bayer, L. Cardle, P. Shaw, G. Stephen, F. Wright, and D. Marshall. **2009**. "Tablet–next Generation Sequence Assembly Visualization." Bioinformatics **26** (3): 401–2.

Mobberley, D. G. **1956**. "Taxonomy and Distribution of the Genus *Spartina*." In , Iowa State Coll J Sci, **30**:471–574.

Nicolas, S. D., G. L. Mignon, F. Eber, O. Coriton, H. Monod, V. Clouet, V. Huteau, *et al*. **2007**. "Homeologous Recombination Plays a Major Role in Chromosome Rearrangements That Occur During Meiosis of *Brassica napus* Haploids." Genetics **175** (2): 487–503.

Oliphant, A., D. L. Barker, J. R. Stuelpnagel, and M. S. Chee. **2002**. "BeadArray Technology: Enabling an Accurate, Cost-Effective Approach to High-Throughput Genotyping." Biotechniques **32** (6): 56–58.

Osborn, T. C., J. C. Pires, J. A. Birchler, D. L. Auger, Z. J. Chen, H.-S. Lee, L. Comai, *et al*. **2003**. "Understanding Mechanisms of Novel Gene Expression in Polyploids." Trends in Genetics **19** (3): 141–47.

Otto, S. P., and J. Whitton. **2000**. "Polyploidy Incidence and Evolution." Annual Review of Genetics **34** (1): 401–37.

Ozkan, H., A. A. Levy, and M. Feldman. **2001**. "Allopolyploidy-Induced Rapid Genome Evolution in the Wheat (Aegilops–Triticum) Group." The Plant Cell **13** (8): 1735–47.

Page, J. T., A. R. Gingle, and J. A. Udall. **2013**. "PolyCat: A Resource for Genome Categorization of Sequencing Reads From Allopolyploid Organisms." G3; Genes|Genomes|Genetics **3** (3): 517–25.

Page, J. T., M. D. Huynh, Z. S. Liechty, K. Grupp, D. Stelly, A. M. Hulse, H. Ashrafi, A. Van Deynze, J. F. Wendel, and J. A. Udall. **2013**. "Insights into the Evolution of Cotton Diploids and Polyploids from Whole-Genome Re-Sequencing." G3; Genes|Genomes|Genetics **3** (10): 1809–18.

Parisod, C., A. Salmon, T. Zerjal, M. Tenaillon, M.-A. Grandbastien, and M. L. Ainouche. **2009**. "Rapid Structural and Epigenetic Reorganization near Transposable Elements in Hybrid and Allopolyploid Genomes in *Spartina*." New Phytologist **184** (4): 1003–15.

Peralta, M., M.-C. Combes, A. Cenci, P. Lashermes, and A. Dereeper. **2013**. "SNiPloid: A Utility to Exploit High-Throughput SNP Data Derived from RNA-Seq in Allopolyploid Species." International Journal of Plant Genomics **2013**: 1–6.

Peterson, P. M., K. Romaschenko, Y. Herrera Arrieta, and J. M. Saarela. **2014**. "A Molecular Phylogeny and New Subgeneric Classification of *Sporobolus* (Poaceae: Chloridoideae: Sporobolinae)." Taxon **63** (6): 1212–43.

Pires, J. C., J. Zhao, M. Schranz, E. J. Leon, P. A. Quijada, L. N. Lukens, and T. C. Osborn. **2004**. "Flowering Time Divergence and Genomic Rearrangements in Resynthesized *Brassica* Polyploids (Brassicaceae)." Biological Journal of the Linnean Society **82** (4): 675–88.

RamanaRao, M. V., D. Weindorf, G. Breitenbeck, and N. Baisakh. **2012**. "Differential Expression of the Transcripts of *Spartina alterniflora* Loisel (Smooth Cordgrass) Induced in Response to Petroleum Hydrocarbon." Molecular Biotechnology **51** (1): 18–26.

Ramsey, J., and D. W. Schemske. **1998**. "Pathways, Mechanisms, and Rates of Polyploid Formation in Flowering Plants." Annual Review of Ecology and Systematics, 467–501.

Ramsey, J., and D. W. Schemske. **2002**. "Neopolyploidy in Flowering Plants." Annual Review of Ecology and Systematics **33** (1): 589–639.

Rapp, R. A., J. A. Udall, and J. F. Wendel. **2009**. "Genomic Expression Dominance in Allopolyploids." BMC Biology **7** (1): 18.

R Development Core Team. **2011**. R: A Language and Environment for Statistical Computing. Vienna, Austria. http://www.R-project.org/.

Rousseau-Gueutin, M., S. Bellot, G.E. Martin, J. Boutte, H. Chelaifa, O. Lima, S. Michon-Coudouel, *et al*. **2015**. "The Chloroplast Genome of the Hexaploid *Spartina maritima* (Poaceae, Chloridoideae): Comparative Analyses and Molecular Dating." Molecular Phylogenetics and Evolution, July.

Salmon, A., L. Flagel, B. Ying, J. A. Udall, and J. F. Wendel. **2009**. "Homoeologous Nonreciprocal Recombination in Polyploid Cotton." New Phytologist **186** (1): 123–34.

Salmon, A., and M. L. Ainouche. **2015**. "Next Generation Sequencing and the Challenge of Deciphering Evolution of Recent and Highly Polyploid Genomes." In. Germany: Koeltz Scientific Books. www.iapt-taxon.org.

Salmon, A., M. L. Ainouche, and J. F. Wendel. **2005**. "Genetic and Epigenetic Consequences of Recent Hybridization and Polyploidy in *Spartina* (Poaceae)" Molecular Ecology **14** (4): 1163–75.

Salmon, A., J. A. Udall, J. A. Jeddeloh, and J. F. Wendel. **2012**. "Targeted Capture of Homoeologous Coding and Noncoding Sequence in Polyploid Cotton." G3; Genes|Genomes|Genetics **2** (8): 921–30.

Sarilar, V., P. M. Palacios, A. Rousselet, C. Ridel, M. Falque, F. Eber, A.-M. Chèvre, J. Joets, P. Brabant, and K. Alix. **2013**. "Allopolyploidy Has a Moderate Impact on Restructuring at Three Contrasting Transposable Element Insertion Sites in Resynthesized *Brassica napus* Allotetraploids." New Phytologist **198** (2): 593–604.

Shaked, H., K. Kashkush, H. Ozkan, M. Feldman, and A. A. Levy. **2001**. "Sequence Elimination and Cytosine Methylation Are Rapid and Reproducible Responses of the Genome to Wide Hybridization and Allopolyploidy in Wheat." The Plant Cell **13** (8): 1749–59.

Sirota-Madi, A., T. Olender, Y. Helman, C. Ingham, I. Brainis, D. Roth, E.t Hagi, L. Brodsky, D. Leshkowitz, and V. Galatenko, *et al*. **2010**. "Genome Sequence of the Pattern Forming Paenibacillus Vortex Bacterium Reveals Potential for Thriving in Complex Environments." BMC Genomics **11** (1): 710.

Soltis, D. E., V. A. Albert, J. Leebens-Mack, C. D. Bell, A. H. Paterson, C. Zheng, D. Sankoff, C. W. dePamphilis, P. K. Wall, and P. S. Soltis. **2009**. "Polyploidy and Angiosperm Diversification." American Journal of Botany **96** (1): 336–48.

Song, K., P. Lu, K. Tang, and T. C. Osborn. **1995**. "Rapid Genome Change in Synthetic Polyploids of *Brassica* and Its Implications for Polyploid Evolution." Proceedings of the National Academy of Sciences **92** (17): 7719–23.

Strong, D. R., and D. R. Ayres. **2013**. "Ecological and Evolutionary Misadventures of Spartina." Annual Review of Ecology, Evolution, and Systematics **44** (1): 389–410.

Szadkowski, E., F. Eber, V. Huteau, M. Lodé, C. Huneau, H. Belcram, O. Coriton, et al. **2010**. "The First Meiosis of Resynthesized *Brassica napus*, a Genome Blender." New Phytologist **186** (1): 102–12.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. **2011**. "MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods." Molecular Biology and Evolution **28** (10): 2731–39.

Tayalé, A., and C. Parisod. **2013**. "Natural Pathways to Polyploidy in Plants and Consequences for Genome Reorganization." Cytogenetic and Genome Research **140** (2-4): 79–96.

Tennessen, J. A., R. Govindarajulu, T.-L. Ashman, and A. Liston. **2014**. "Evolutionary Origins and Dynamics of Octoploid Strawberry Subgenomes Revealed by Dense Targeted Capture Linkage Maps." Genome Biology and Evolution **6** (12): 3295–3313.

Udall, J. A., P. A. Quijada, and T. C. Osborn **2005**. "Detection of Chromosomal Rearrangements Derived From Homeologous Recombination in Four Mapping Populations of *Brassica napus* L." Genetics **169** (2): 967–79.

Udall, J. A., J. M. Swanson, K. Haller, R. A. Rapp, M. E. Sparks, J. Hatfield, Y. Yu, *et al*. **2006**. "A Global Assembly of Cotton ESTs." Genome Research **16** (3): 441–50.

Vallejo-Marin, M. **2012**. "*Mimulus peregrinus* (Phrymaceae): A New British Allopolyploid Species." PhytoKeys **14** (0): 1–14. doi:10.3897/phytokeys.14.3305.

Van Bel, M., S. Proost, C. Van Neste, D. Deforce, Y. Van de Peer, and K. Vandepoele. **2013**. "TRAPID: An Efficient Online Tool for the Functional and Comparative Analysis of *de novo* RNA-Seq Transcriptomes." Genome Biology **14** (12): R134.

Van de Peer, Y., S. Maere, and A. Meyer. **2009**. "The Evolutionary Significance of Ancient Genome Duplications." Nature Reviews Genetics **10** (10): 725–32.

Waterhouse, A. M., J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton. **2009**. "Jalview Version 2--a Multiple Sequence Alignment Editor and Analysis Workbench." Bioinformatics **25** (9): 1189–91.

Wendel, J. F. **2000.** "Genome Evolution in Polyploids." Plant Molecular Biology **42** (1): 225–49.

Wolfe, K. H., W. H. Li, and P. M. Sharp. **1987**. "Rates of Nucleotide Substitution Vary Greatly among Plant Mitochondrial, Chloroplast, and Nuclear DNAs." Proceedings of the National Academy of Sciences of the United States of America **84** (24): 9054–58.

Yoo, M.-J., X. Liu, J. C. Pires, P. S. Soltis, and D. E. Soltis. **2014**. "Nonadditive Gene Expression in Polyploids." Annual Review of Genetics **48** (1): 485–517.

Yoo, M. J., E. Szadkowski, and J. F. Wendel. **2013**. "Homoeolog Expression Bias and Expression Level Dominance in Allopolyploid Cotton." Heredity **110** (2): 171–80.

Table 1: Sequencing statistics using Roche-454 and Illumina data and number of reads used for the analysis (Illumina cleaned reads length = 108 bp). *: Roche-454 data of *S. maritima* contain normalized and non-normalized cDNA libraries.

	454 reads			Illumina reads	
	Number of reads	Reads average length	Number of reads used in the assembly	Number of reads	Number of reads mapped on reference transcriptome
S. maritima *	984 006	463.24 ± 200.58 bp	755 309	76 985 267	28 837 359 (37.46%)
S. alterniflora	495 749	285.94 ± 160.69 bp	344 723	77 321 929	40 970 154 (52.99%)
S. x townsendii	322 773	261.40 ± 130.54 bp	193 619	71 358 554	41 277 405 (57.84%)
S. x neyrautii	367 577	241.46 ± 136.40 bp	206 750	65 483 843	22 411 036 (34.22%)
S. anglica	314 645	261.80 ± 143.96 bp	187 291	60 284 800	29 210 578 (48.45%)

Table 2: Summary of assemblies' steps and annotations of five *Spartina* species. *: Unigenes were detected using *O.sativa* genome only.

	Number of Contigs:							
	Reference transcriptome	**Annotated contigs**	Unigenes*	454 contigs	Illumina contigs	Number of contigs presenting two or more haplotypes	GC%	N50
S. maritima	**60 644**	**22 998**	13 771	25 239	98 455	20 085	40.79%	615 bp
S. alterniflora	**44 158**	**19 241**	13 054	17 062	76 010	13 216	41.24%	666 bp
S. x townsendii	**59 166**	**21 974**	16 002	9 042	121 733	24 199	42.53%	601 bp
S. x neyrautii	**65 099**	**25 067**	13 471	7 008	110 455	16 776	40.79%	519 bp
S. anglica	**57 920**	**21 143**	13 800	3 995	114 555	18 839	40.94%	563 bp

Table 3: Identification of the parental origin of the haplotypes in the hybrids, using 271,230 and 160,192 haplotypes of *S. maritima* and *S. alterniflora* respectively. Unassigned haplotypes correspond to those where the parental haplotypes are not found or to haplotypes where the two parental copies are similar.

		Parental haplotypes assignation:		
	Number of contigs (and windows) used for the assignation:	*S. maritima*	*S. alterniflora*	Unassigned haplotypes
S. x townsendii	7 293 (10 693)	135 298	108 548	22 974
S. x neyrautii	6 947 (9 771)	97 516	79 414	18 154
S. x anglica	7 153 (10 159)	106 314	87 884	19 238

Figure 1: Strategy developed for assemblies and haplotype detection. First, Roche-454 and Illumina reads were assembled separately using Newbler and Trinity, respectively. The contigs obtained were co-assembled using Newbler. The length of contigs was enhanced and redundant contigs using custom scripts. Reads Mapping were done with Bowtie 2 and the different contigs were annotated using 3 complementary methods: tblastx, Blast2Go and Pfam. Polymorphisms and haplotypes were detected and constructed using mapping data.

Figure 2: Description of the method and windows used to construct the different haplotypes. (A.) For each contig (dotted line), the developed program detects the different SNPs (grey boxes) using mapping data. (B.) For the windows created (1. and 2.), only reads (black lines) entirely included in the windows are selected. (C.) Detection of the different haplotypes in each window (thick black lines) using the method previously developed by Boutte et al. (in press). (D.) Example of "the maximum number of haplotypes by window" and "cascade phenomenon" leading to the detection of multiple haplotypes. Using reads mapped to a contig of *S. maritima* annotated as nucleotidyl transferase localized in the cytoplasm (GO annotations: 0009058, 0016740, 0016779, 0005737, 0005623), 11 SNPs are detected and a total of 11 haplotypes are constructed. If the maximum number of haplotypes for this gene is seven, the hypothetical minimum number of haplotypes for this gene should be 5 (haplotypes 7 and 9 might correspond to haplotypes 1,2,4 or 5). Because several choices are possible, the detected haplotypes are not assembled, illustratating the "cascade phenomenon".

Figure 3: Parental haplotype assignation process. (A.) For each read alignment, SNPs were detected and (B.) reads assembled to obtain the different haplotypes using the developed program. (C.) Hybrid haplotypes (detected following the procedure A. and B.) were aligned with parental haplotypes. Intra- and inter-specific polymorphisms were used to assign each hybrid haplotype to a specific parent.

Figure 4: Graphical representation of (a) the number of SNPs for 100 bp and (b) the maximum number of haplotypes by windows for each species studied (Student's test p-value>0.05 and p-value<0.001). Errors bars indicate confidence interval to 95%. (c) Distribution of the haplotypes length (in bp) reconstructed for the 5 species studied. Dotted vertical bar represent the length of the Illumina reads.

Figure 5: Ks distribution for the 2 parents *S. maritima*, *S. alterniflora,* and the 2 F1 hybrids *S. x townsendii* and *S. x neyrautii*. Dotted vertical bars represent the estimations of the 2 duplication events (0.113 - 0.125 and 0.113 - 0.125).

Figure 6: Phylogenetic analysis of the PPR gene (Pentatricopeptide repeat superfamily protein , GO annotations: 0003674, 0008150, 0005739) with Maximum Likelihood method (K2+G model). The numbers of substitutions indicated under the branches were obtained from a Maximum Parsimony analysis which generated the same tree topology. Bootstrap values obtained from 1 000 replicates are shown above the branches in bold. Stars indicate whole genome duplication events.

Additional Table 1: Statistics for the 5 species studied.

		Mean (Standard deviation)/Minimum-Maximum:		
		Length of contigs:	**Number of reads mapped:**	**Average read depth:**
S. maritima	Annotated	741.75 (620.91) / 58 – 10 313	643.36 (5005.43) / 1 – 251 685	56.51 (226.36) / 0.14 – 7 877.08
	Unannotated	339.56 (215.09) / 40 – 4 701	146.20 (826.68) / 1 – 46 647	37.16 (167.90) / 0.12 – 7 664.36
S. alterniflora	Annotated	761.11 (630.11) / 101 – 10 839	479.73 (3318.14) / 1 – 171 469	42.12 (200.31) / 0.08 – 7 843.19
	Unannotated	336.00 (188.14) / 41 – 3 444	110.47 (736.69) / 1 – 47 919	29.17 (160.81) / 0.13 – 7 760.78
S. x townsendii	Annotated	755.67 (578.83) / 92 – 9 423	912.40 (5 038.26) / 1 – 350 532	90.52 (314.00) / 0.69 – 7 871.55
	Unannotated	330.38 (170.46) / 40 – 3 473	173.18 (815.75) / 1 – 46 685	48.68 (208.80) / 0.42 – 7 135.0
S. x neyrautii	Annotated	664.86 (566.64) / 59 – 14 705	444.98 (4 393.17) / 3 – 253 813	40.10 (208.05) / 1.17- 7 560.28
	Unannotated	326.94 (172.01) / 40 - 4760	105.36 (1301.95) / 1 – 138 522	25.12 (175.81) / 0.29 – 7 815.79
S. anglica	Annotated	718.92 (587.83) / 74 – 8 833	643.62 (6 773.04) / 2 – 662 869	54.51 (257.76) / 0.49 – 7 792.45
	Unannotated	335.76 (180.99) / 40 – 3 965	121.37 (873.32) / 1 – 62 022	31.20 (190.89) / 0.28 – 7 901.54

Additional Figure 1: Statistics of the reference transcriptomes. Histograms show the distribution of the contigs length, the average read depth and the number of reads (mapped to around 90% of identity) of the annotated and unannotated *S. maritima* contigs. Distributions for the other species are similar to *S. maritima*.

Additional Figure 2: Ks distribution for the hybrid simulated using parental haplotypes mixing process and the allododecaploid *S. anglica* (A.) and for the hybrid simulated using parental read mapping process (B.). Dotted vertical bar represent the estimations of the 2 duplication events in the hexaploid species *S. maritima* (0.113 - 0.125 and 0.113 - 0.125; Figure 5). The number of contigs used for the simulation is 4 079 and 317 for the haplotypes mixing process and mapping read process respectively.

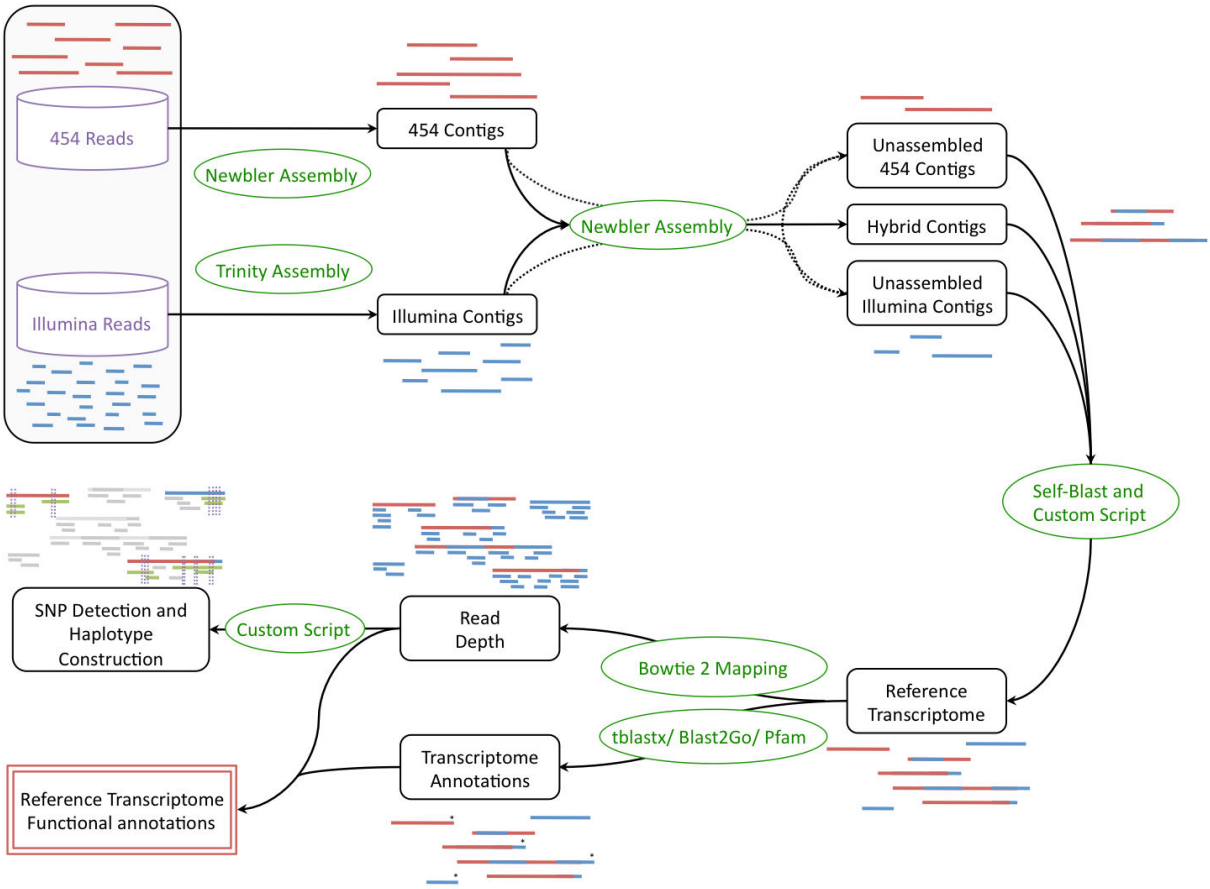


Figure 1

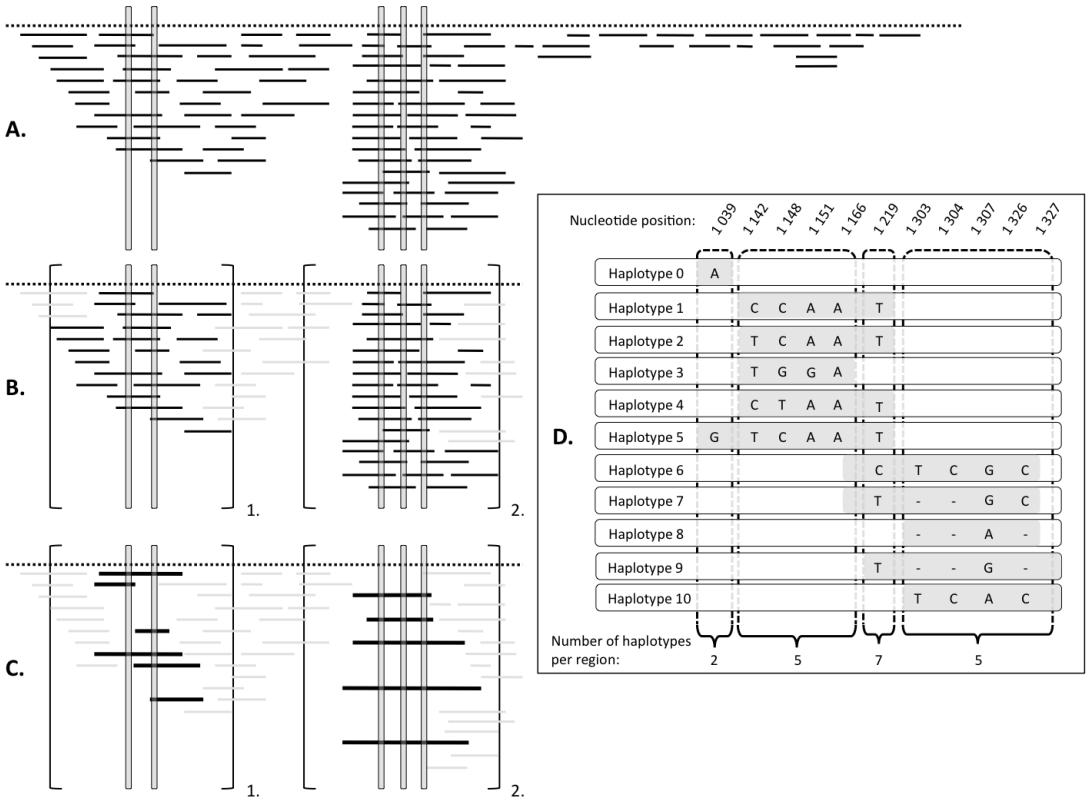


Figure 2

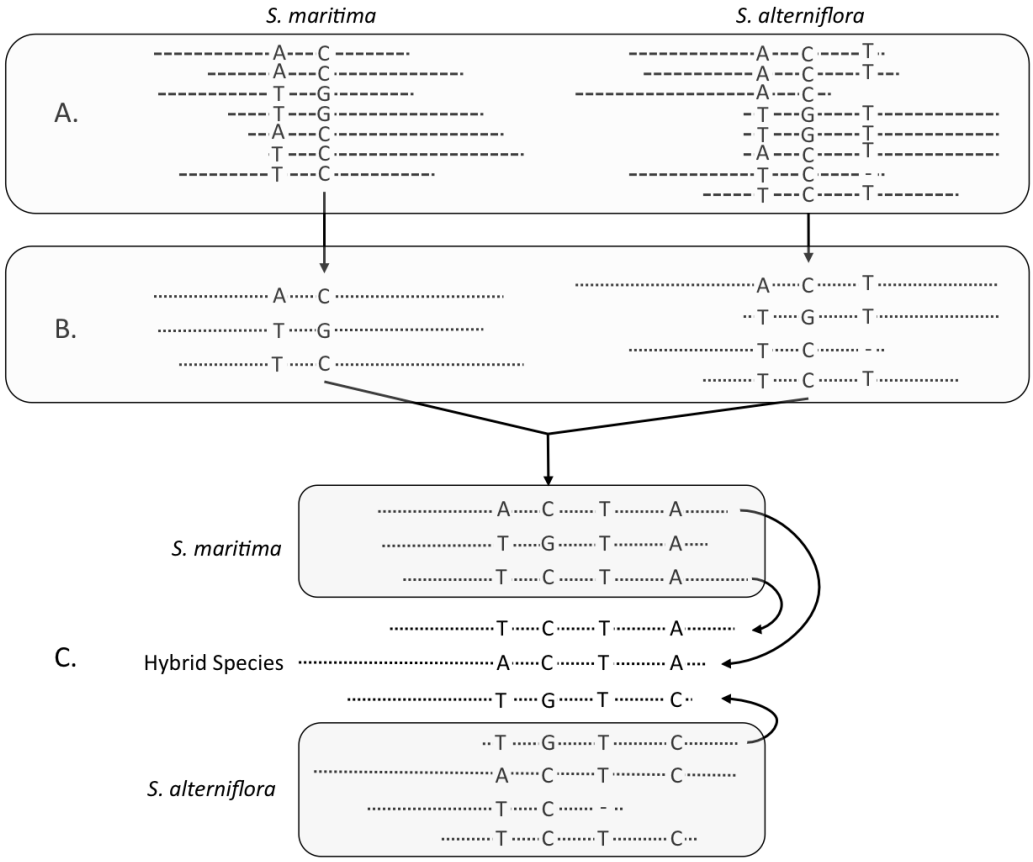


Figure 3

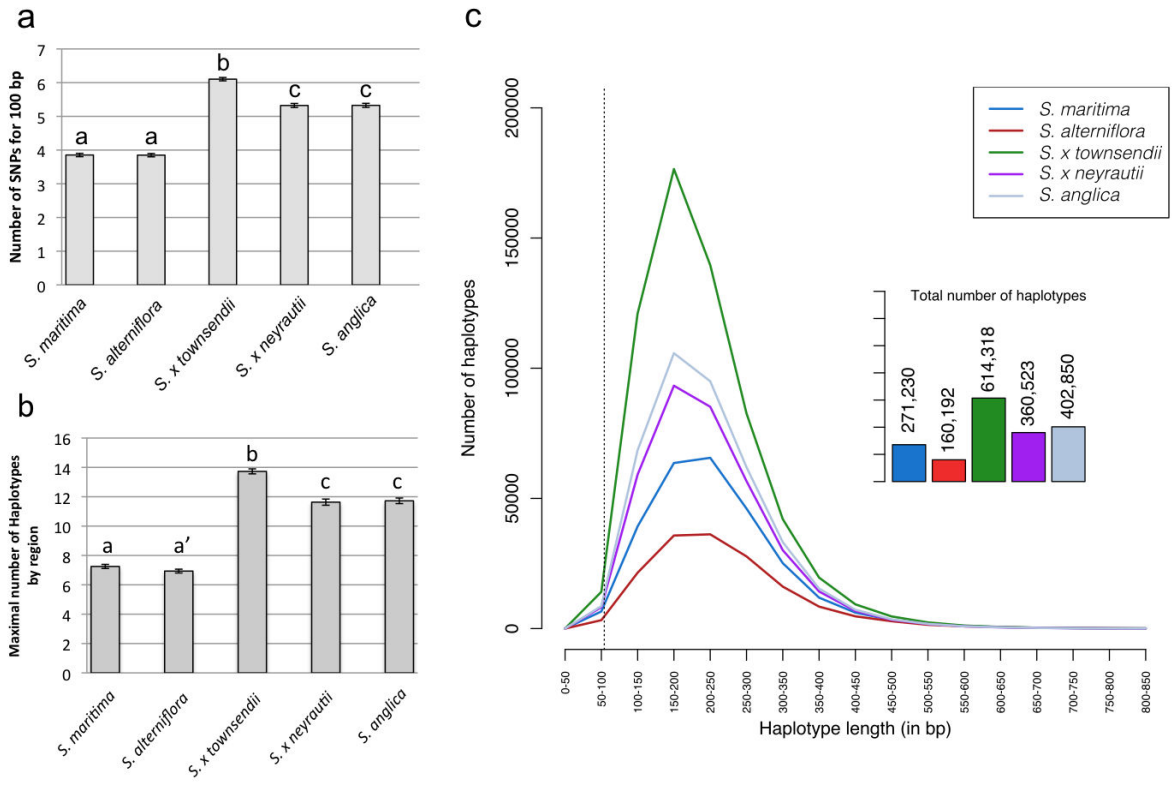


Figure 4

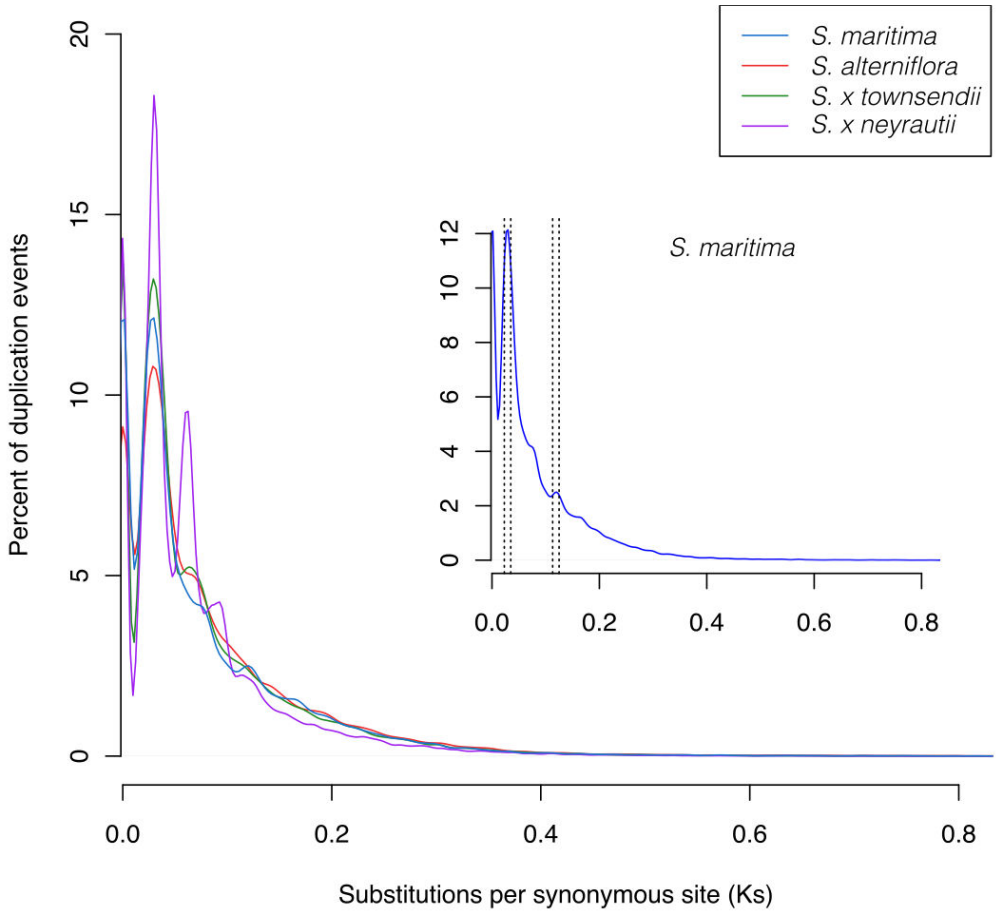


Figure 5

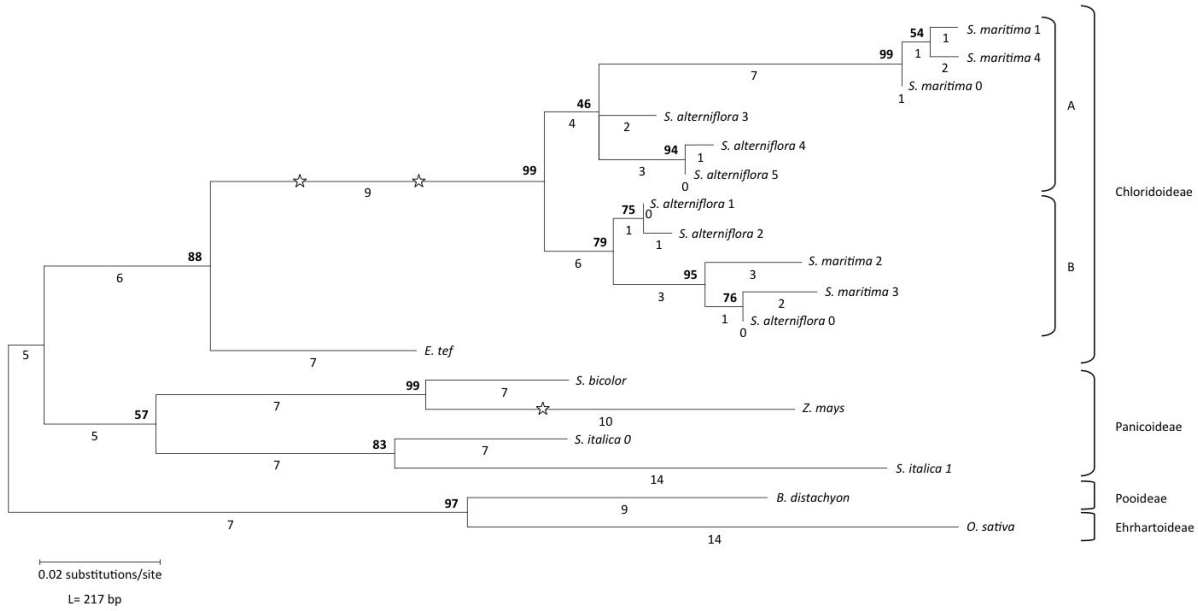
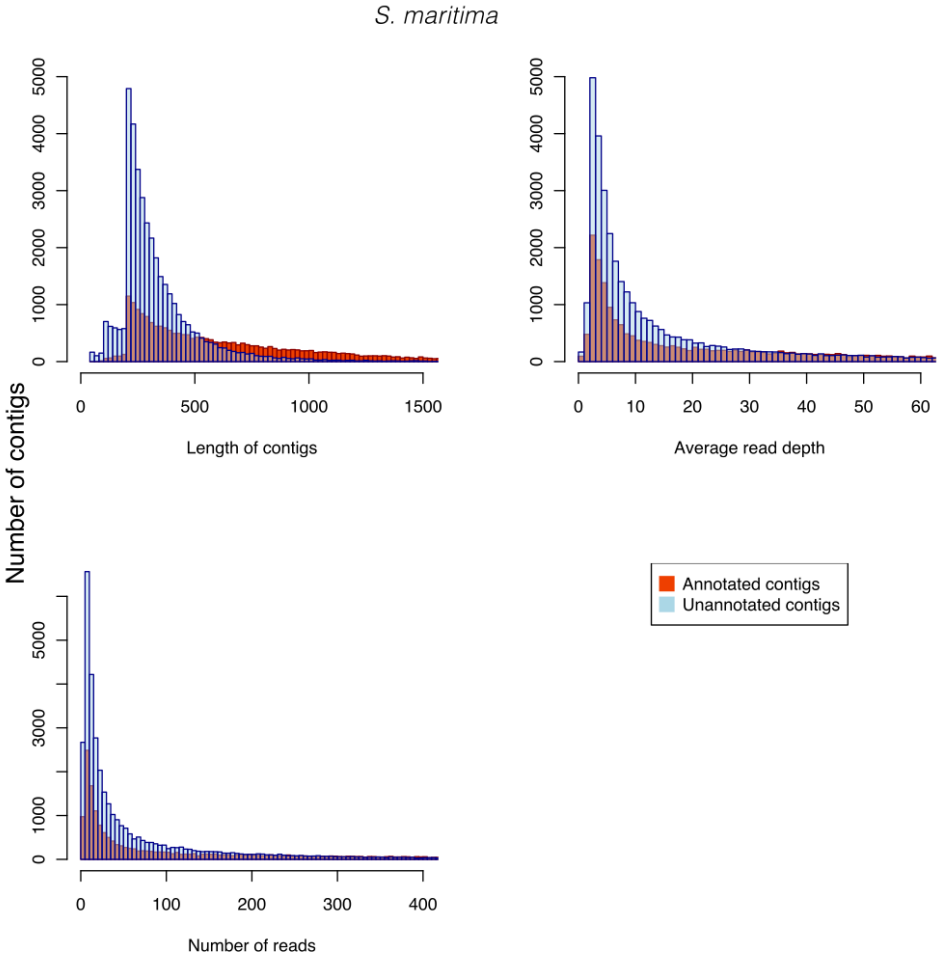
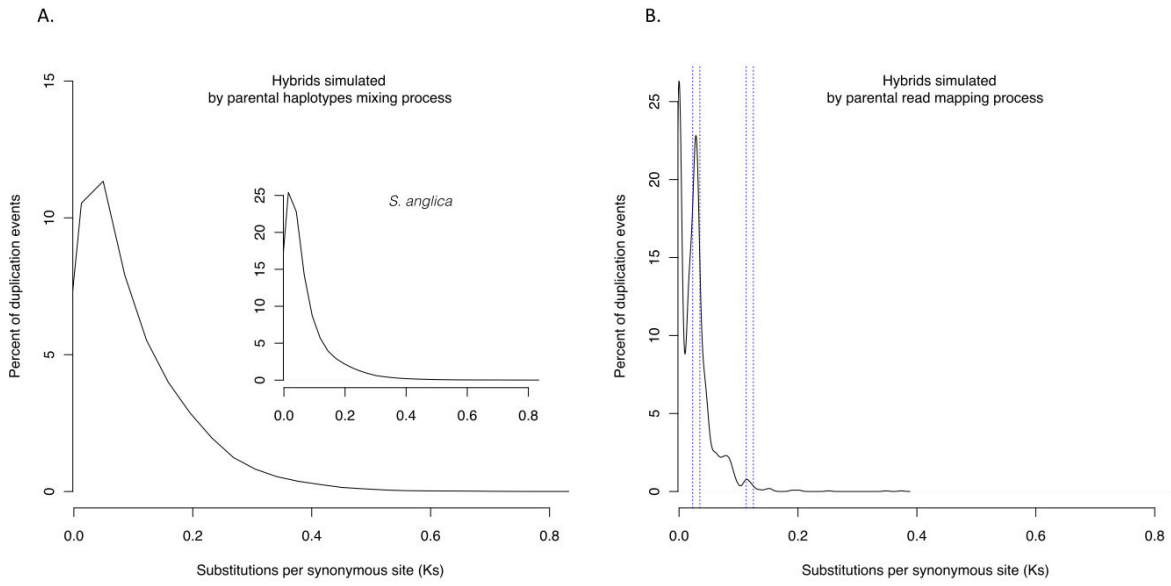


Figure 6



Supplemental Figure 1



Supplemental Figure 2

Partie E : Mise en place d'un « pipeline» d'analyses phylogénomiques pour explorer l'histoire des haplotypes détectés.

Aujourd'hui, les données de séquençage massif (NGS) offrent l'opportunité d'explorer les génomes et transcriptomes d'un plus grand nombre d'espèces. Ces données permettent notamment de revisiter les phylogénies traditionnelles basées sur un nombre limité de gènes nucléaires ou cytoplasmiques. De telles études commencent à se développer, par exemple, Bond *et al.* (2014) ont ainsi analysé deux supermatrices de 327 et 128 gènes orthologues chez 33 familles d'araignées (Araneae), Stephens *et al.* (2015) ont exploré les relations entre plantes carnivores du genre *Sarracenia* (à l'aide de plus de 200 gènes) ou Song *et al.* (2012) qui ont réalisé une phylogénie de 33 espèces de Mammifères Euthériens à l'aide de 447 gènes nucléaires.

Ces analyses à grande échelle nécessitent le développement d'une méthodologie appropriée. Les démarches à suivre dépendent largement de la particularité des jeux de données. Dans cette partie, nous présenterons une méthode adaptée à l'étude de génomes complexes, dans le but de réaliser des analyses phylogénomiques à partir d'un nombre de séquences très variables par gène issues de données NGS et d'explorer l'histoire évolutive des haplotypes détectés dans les génomes polyploïdes.

Nous comparons ici les haplotypes reconstruits à l'aide du logiciel « IlluHaplotyper » sur les transcriptomes des deux espèces hexaploïdes *S. maritima* et *S. alterniflora*. Ces analyses permettront, à terme, de comprendre l'histoire des génomes du clade des spartines hexaploïdes. Dans un premier temps, nous avons détecté dans chaque jeu de données les régions homologues aux deux espèces. Une sélection d'espèces a été choisie comme outgroup parmi les génomes séquencés d'angiospermes : *Amborella* (lignée basale des angiospermes), *Arabidopsis* (Eudicots), chez les Monocots *Phoenix, Musa* ainsi que plusieurs représentant de la famille des Poaceae (Tableau 15). Les régions homologues à celles identifiées chez les spartines ont été recherchées chez les différents outgroups. Comme attendu, un nombre plus important de copies homologues ont été détectées chez les espèces les plus proches phylogénétiquement.

Pour chaque région homologue où au moins un outgroup a pu être représenté, nous avons recherché les différents haplotypes de Spartines construits par le logiciel « IlluHaplotyper », ces séquences ont ensuite été alignées via le logiciel Mafft (Katoh and Toh 2010). Les zones présentant un mauvais alignement (en début et fin d'alignement) et les séquences redondantes ont été supprimées. Seuls les alignements présentant un pourcentage d'identité multiple supérieur ou égal à 50% (ou supérieur ou égal à 40% et contenant un nombre moyen de gap inférieur à 30%) ont été conservés. Nous avons ainsi obtenu un total de 7 295 matrices.

Le nombre de séquences provenant des outgroups est présenté dans le Tableau 15. Les alignements obtenus ont une longueur moyenne de 588,84 bp (σ=467,24, $IC_{95\%}$= 578,12 - 599,57) et sont constitués de 5,08 séquences en moyenne (σ=4,60, $IC_{95\%}$= 4,98 - 5,19).

Tableau 15 : Nombre de séquences disponibles pour les 10 outgroups utilisées au cours des différentes étapes de recherche de régions homologues avec les Spartines et des analyses phylogénomiques réalisées. La dernière colonne correspond au nombre d'alignements étant ré-enracinés à l'aide des différents outgroups.

Espèce :	Nombre de séquences			Nombre d'alignements où l'espèce est utilisée pour l'enracinement :
	dans la base de données :	utilisées dans les alignements :	utilisées dans les phylogénies :	
Amborella trichopoda	26 846	68	68	61
Arabidopsis thaliana	33 410	71	71	35
Phoenix dactylifera	28 889	246	246	141
Musa acuminata	36 549	335	333	82
Oryza sativa	49 061	1 628	1 609	1 226
Brachypodium distachyon	31 029	1 644	1 622	438
Setaria italica	40 599	2 598	2 462	737
Zea mays	48 162	2 622	2 604	155
Sorghum bicolor	39 441	2 206	2 174	81
Eragrostis tef	70 860	4 903	4 463	486

Une analyse phylogénomique à l'aide de la méthode du maximum de Vraisemblance (modèle GTR-Gamma) a ensuite été réalisée sur les 7 295 matrices via le logiciel RAxML (v 7.2.8; Stamatakis 2014) ce qui a permis d'obtenir un total de 3 442 arbres. Le choix de ce logiciel est motivé par sa capacité à traiter automatiquement un nombre important de matrices (en ligne de commandes ; Stamatakis 2014) et son utilisation courante pour les analyses phylogénomiques : pour la sélection du modèle évolutif et/ou l'application de la méthode du maximum de Vraisemblance (*e.g.* McKain et al. 2012; Song et al. 2012; Dunn, Howison, and Zapata 2013; Cronin et al. 2014; Grover et al. 2015). L'approche développée au cours de ce travail est similaire à l'outil « Agalma » qui permet, à partir de reads Illumina d'obtenir de manière automatique différentes matrices pour réaliser plusieurs analyses phylogénomiques (Dunn, Howison, and Zapata 2013).

Les 3 442 arbres obtenus ont été ré-enracinés à l'aide du logiciel TreeCollapseCL 4 (Hodcroft 2015) et un consensus de ces arbres a été créé en conservant uniquement les nœuds ayant un bootstrap supérieur ou égal à 50%. Le nombre de matrices éliminées (sur les 7 295 matrices initiales) du processus de phylogénie peut s'expliquer par le fait que nous n'avons pas trié au préalable les matrices : les alignements présentant un faible nombre de sites polymorphes et/ou une longueur insuffisante n'ont en effet pas été traités par le logiciel RAxML. Pour les 3 442 arbres obtenus, 63,5% des alignements ont une longueur comprise entre 150 et 600 bp (Figure 37). Les arbres sont composés en moyenne 5,27 séquences (σ=3,96, $IC_{95\%}$= 5,13 – 5,40) et 2 453 arbres ne présentent qu'une seule séquence de *S. maritima* et une seule séquence de *S. alterniflora* (Figure 37).

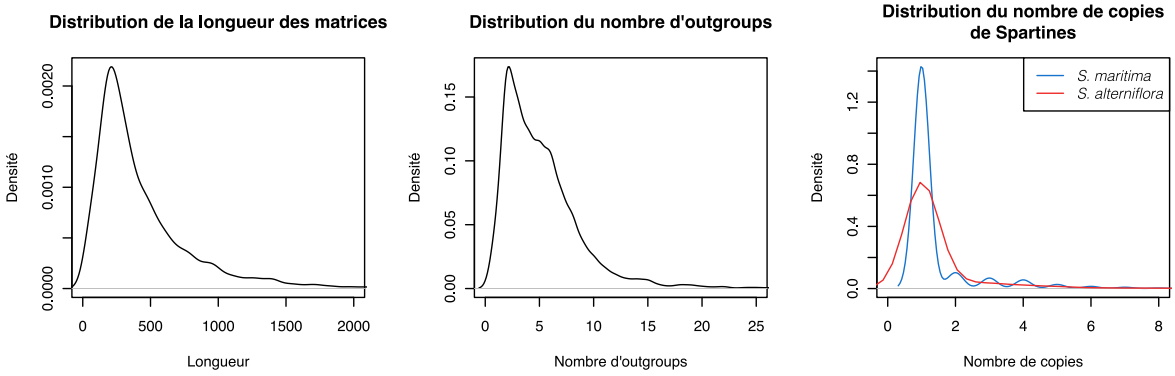


Figure 37 : Représentation de la longueur des 3 442 matrices, du nombre d'outgroups et de séquences de Spartines présentent au sein des alignements.

Les longueurs de ces matrices sont le plus souvent dues à la longueur des haplotypes construits à l'aide du logiciel « IlluHaplotyper ». Il serait intéressant de regarder la profondeur de séquençage des régions ne présentant qu'une seule copie de *S. maritima* et de *S. alterniflora* (pour 2 453 gènes) afin de déterminer si elles correspondent à des régions conservées chez les spartines ou à des régions pour lesquelles il n'a pas été possible de détecter les différentes copies potentielles. Si la profondeur de séquençage est importante, il serait intéressant de regarder si ces régions correspondent à des domaines protéiques particuliers et d'étudier leur divergence avec les autres espèces de Poaceae.

Les arbres obtenus présentent des topologies différentes en fonction du nombre de séquences, d'outgroups et de la résolution de l'arbre. En filtrant les arbres obtenus sur la base du nombre de séquences, nous avons examiné les différentes topologies obtenues. A titre illustratif, nous présentons deux arbres construits à partir de matrices de 181 bp (gène codant une « membrane-anchored ubiquitin-fold protein », Figure 38) et 205 bp (gène codant une « Profilin », Figure 39). Ces deux jeux de données ont également fait l'objet d'une analyse phylogénétique (Maximum de Vraisemblance) à l'aide du logiciel MEGA, un logiciel couramment utilisé pour des analyses phylogénétiques (Yang and Rannala 2012), dans le but de comparer les résultats avec ceux obtenus à l'aide de RAxML.

Nous constatons pour ces deux gènes la présence de deux clades (A et B) dans lesquels les deux espèces hexaploïdes de Spartines sont représentées. Pour le gène codant une « membrane-anchored ubiquitin-fold protein » (Figure 38), les haplotypes de Spartines forment un groupe frère avec *Eragrostis tef* (Chloridoideae). Nous retrouvons à l'extérieur de ces groupes les séquences des Panicoideae (*Setaria italica*, *Sorghum bicolor* et *Zea mays*) puis *Oryza sativa* (outgroup), en accord avec l'histoire connue des Poacées. Au sein du clade A, nous observons la présence de six haplotypes : quatre de *S. alterniflora* groupés en deux sous-clades et deux de *S. maritima* formant un troisième sous-clade. Ces trois sous-clades (formés par des séquences de la même espèce) forment une polytomie. Le clade B contient cinq copies de Spartines (trois de *S. alterniflora* et deux de *S. maritima*) entre lesquelles les relations ne sont pas résolues. Les topologies des arbres obtenus par le logiciel RAxML (Figure 38a) et MEGA (Figure 38b) sont similaires. Les clades A et B qui contiennent chacun des séquences de *S. maritima* et *S. alterniflora* résulteraient d'une duplication intervenue chez leur ancêtre commun hexaploïde et pourraient donc représenter des groupes

homéologues (duplication génomique) ou paralogues (duplication génique antérieure à la spéciation entre *S. maritima* et *S. alterniflora*). Les topologies mal résolues au sein des clades A et B ne permettent pas de statuer sur l'origine des différents haplotypes de *S. maritima* et *S. alterniflora*.

Pour le gène codant la « Profilin » (Figure 39), les deux analyses phylogénétiques (réalisées à l'aide de RAxML et MEGA), fournissent les même topologies, avec les haplotypes de Spartines distribués dans deux clades. Le premier clade (A) contient deux copies sœurs de *S. maritima* et deux copies de S*. alterniflora* regroupés sur une polytomie. Le clade B est composé d'une copie de *S. alterniflora* et de deux copies sœurs de *S. maritima*. Ces deux clades résultant d'une duplication chez l'ancêtre de ces deux espèces, pourraient correspondre à deux copies homéologues comme dans le cas du gène précédent. Dans le clade A, nous serions alors en présence de paires d'allèles pour chaque espèce. Le clade B correspondrait à un second jeu de copies homéologues, avec la présence d'une paire d'allèles de *S. maritima* et d'un allèle chez *S. alterniflora*.

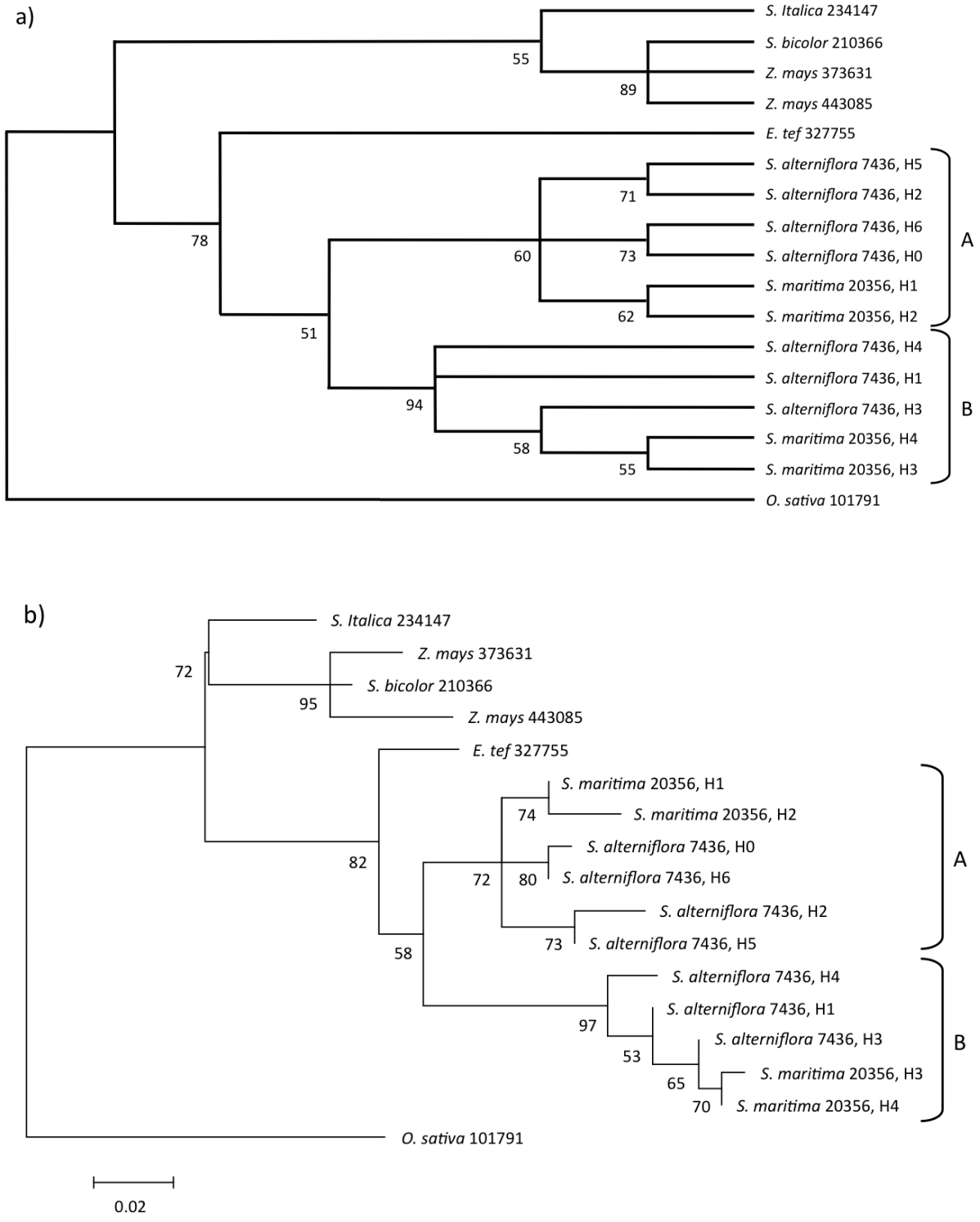


Figure 38 : Arbres phylogénétiques réalisés à l'aide de la méthode du Maximum de Vraisemblance pour un gène codant une « membrane-anchored ubiquitin-fold protein ». La longueur de la matrice est de 181 bp. Les arbres ont été enracinés avec une séquence d'*Oryza sativa*. Les valeurs de bootstraps obtenues avec 500 réplicas sont indiquées en dessous des branches. L'arbre a) a été obtenu à partir du logiciel RAxML (modèle d'évolution GTR-GAMMA) et l'arbre b) à partir du logiciel MEGA (modèle d'évolution de Kimura à 2 paramètres et distribution Gamma).

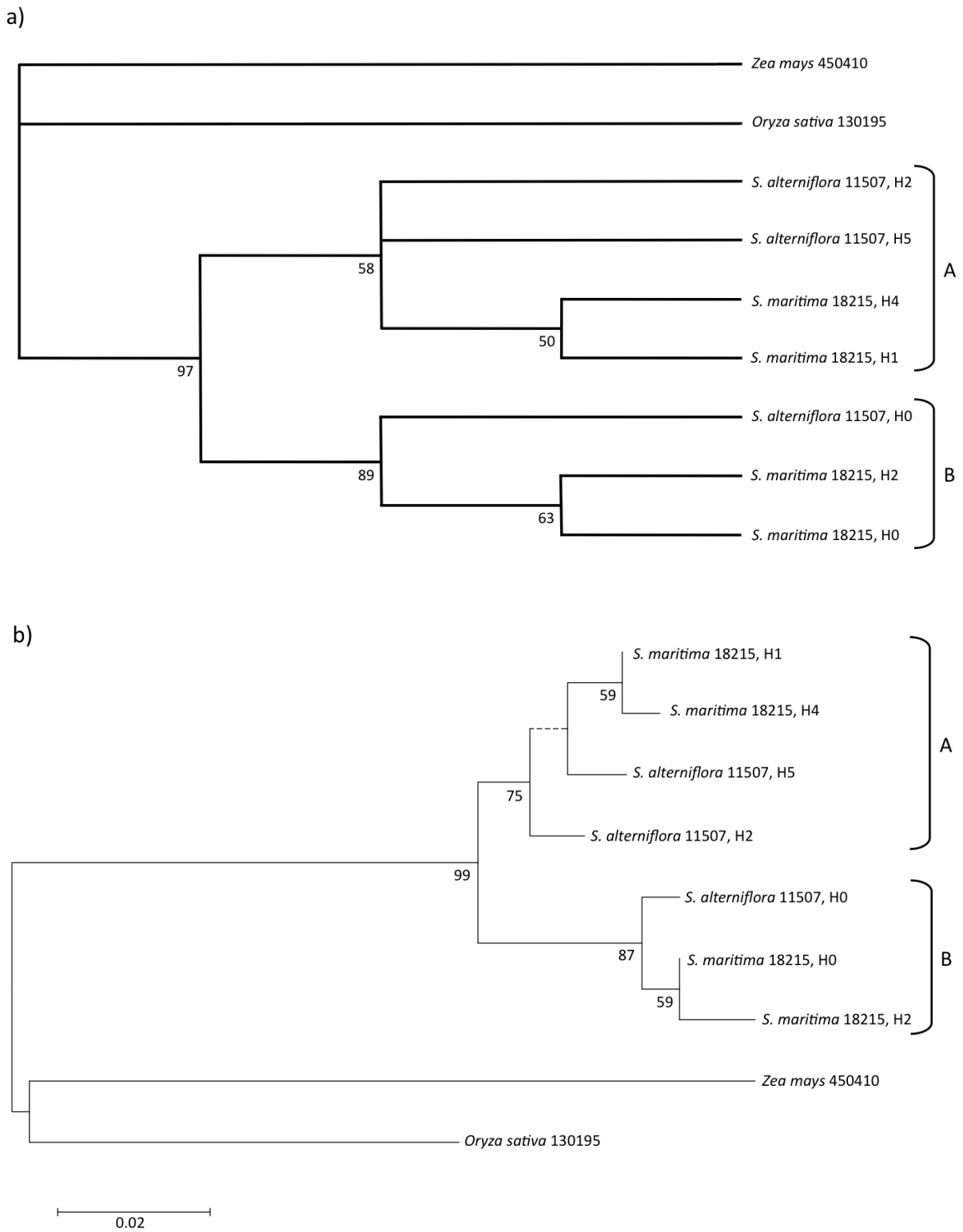


Figure 39 : Arbres phylogénétiques réalisés à l'aide de la méthode du Maximum de Vraisemblance pour un gène codant une « Profilin ». La longueur de la matrice est de 205 bp. Les arbres ont été enracinés avec les séquences du riz et du maïs. Les valeurs de bootstraps obtenues avec 500 réplicas sont indiquées en dessous des branches. L'arbre a) a été obtenu à partir du logiciel RAxML (modèle d'évolution GTR-GAMMA) et l'arbre b) à partir du logiciel MEGA (modèle d'évolution de Jukes et Cantor). La ligne en pointillée correspond à un nœud qui n'est pas résolu à 50%.

Il sera intéressant de voir si les topologies similaires entre les deux gènes présentés ci-dessus se retrouvent plus fréquemment sur l'ensemble des matrices analysées. Toutefois, les analyses de congruences entre les différents arbres ne peuvent être réalisées en l'état actuel du jeu de données. Pour chaque analyse (chaque matrice), le nombre de séquences regroupant les outgroups et les copies détectées diffère. Il est donc impossible de traiter globalement ces alignements en créant des supermatrices comme cela a été dans les travaux précédemment mentionnés (Bond et al. 2014, Stephens et al. 2015).

Nos jeux de données étant particulièrement complexes et pouvant présenter un nombre important d'haplotypes, nous ne pouvons pas développer un programme présentant une approche cas par cas. En effet, pour un alignement présentant *n* séquences, nous pouvons avoir :

$$N_{\mathrm{R}} = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$ arbres résolus (avec n≥2).

Dans le cas où nous serions en présence des 6 copies de chaque espèce hexaploïde de Spartines et de trois outgroups (soit 15 séquences), un total de 213 458 046 676 875 topologies d'arbre résolus est possible.

Différentes investigations peuvent être envisagées qui dépassent le temps imparti à nos travaux de thèse, et nous proposons ci-dessous quelques pistes : pour comparer nos arbres nous pourrions envisager de développer une approche de clustering. Les clusters pourraient être analysés à l'aide de logiciels tel que STAR (Liu et al. 2009), PHYLDOG (Boussau et al. 2013) ou ASTRAL (Mirarab et al. 2014) par exemple qui permettent d'obtenir un arbre consensus à partir de plusieurs centaines d'arbres de gènes.

En complément aux analyses basées sur le maximum de Vraisemblance, il serait également intéressant de réaliser des analyses phylogénomiques basées sur des méthodes Bayésiennes. Ceci pourrait être réalisé avec un logiciel tel que BUCKy (Larget et al. 2010), qui a déjà été appliqué sur des jeux de données génomiques importants comme des données RADseq pour étudier l'inférence phylogénétique et les introgressions chez les plantes (Eaton

and Ree 2013) ou pour re-évaluer la phylogénie des cotons allopolyploïdes à partir de 52 gènes (Grover et al. 2015). Il serait également possible de réaliser des analyses phylogénomiques à partir des données Roche-454 étudiées au cours de cette thèse (Chapitre 4). En effet, une analyse récente réalisée à partir des données Roche-454 a permis de réaliser la phylogénie du genre *Daucus* incluant la carotte cultivée (Arbizu et al. 2014). Pour cela 94 gènes ont été concaténés en 2 lots, en séparant les gènes présentant un allèle des gènes présentant deux allèles. Les matrices obtenues ont été analysées à l'aide des logiciel PAUP pour l'analyse du maximum de Parcimonie, RAxML pour les analyses de maximum de Vraisemblance et BUCKy et BEAST pour les analyses Bayésiennes. Au cours de leur étude, Arbizu et ses collaborateurs (2014) ont dû nettoyer les différentes séquences issues de la technologie Roche-454 en corrigeant principalement les régions homopolymèriques. Les haplotypes que nous avons détectés à l'aide du logiciel « PyroHaplotyper » sont déjà corrigés et pourraient directement être analysés par les logiciels RAxML et BUCKy par exemple. La longueur des haplotypes issues des données Roche-454 étant plus importante, il serait plus aisé d'obtenir des alignements orthologues plus informatifs pour les analyses phylogénomiques.

Les approches de phylogénie qui pourraient ainsi être menées (sur des données transcriptomiques mais aussi des données génomiques) devraient apporter un éclairage nouveau sur l'histoire ancienne de la lignée des spartines, et l'évolution des séquences au cours des évènements successifs de polyploïdisation.

Discussion :

Dans ce deuxième volet de nos travaux, nous avons développé le logiciel « IlluHaplotyper » adapté aux alignements de séquences de type Illumina dans la perspective de détecter les polymorphismes intragénomiques et de reconstruire les haplotypes chez les spartines polyploïdes. Cette démarche a été réalisée en suivant différentes étapes qui sont discutées ci dessous.

A la recherche des paramètres optimaux du programme « Illuhaplotyper » :

Nous avons dans un premier temps fait varier les différents paramètres de ce programme afin de déterminer leurs valeurs optimales, ceci dans le but de détecter le plus finement possible les copies dupliquées tout en évitant la détection de faux positifs. Pour cela nous avons utilisé un sous-ensemble du transcriptome de référence de *S. maritima* assemblé à partir de données Roche-454 et Illumina (18 861 contigs).

Une première **analyse des paramètres de mapping** (effectuée sur plusieurs logiciels) nous a montrée qu'en deçà de 90%, la diminution du pourcentage d'identité de mapping n'entrainait pas une augmentation significative du nombre de reads alignés. Nous avons par ailleurs montré que le logiciel Bowtie 2 nous permettait d'aligner plus de reads sur le transcriptome de référence que les autres logiciels testés (Bowtie, BWA, Novoalign et SOAP2). Ces résultats sont en adéquation avec les études menées sur les différents outils de mapping qui ont montré que le meilleur logiciel dépendait des jeux de données à analyser (Hatem et al. 2013; Schbath et al. 2012). De ce fait, nous avons étudié les répercussions des paramètres du programme « IlluHaplotyper » sur un seul sous ensemble de reads alignés à 90% d'identité avec le logiciel Bowtie 2.

La variation du **paramètre « seuil » de détection des SNPs** entraine une modification du nombre d'haplotypes construit par alignement. L'augmentation de ce paramètre entraine une diminution du nombre de SNPs détectés et donc du nombre d'haplotypes reconstruits. Avec un seuil réglé à 10%, 6 haplotypes sont détectés en moyenne ce qui correspond au nombre maximal d'allèles attendus chez l'espèce hexaploïde *S. maritima*. Si le seuil est réglé

au dessus de 10%, le risque de supprimer des SNPs valides est trop important, notamment si on veut détecter des homéologues divergents. Rousseau-Gueutin *et al.* (2015) ont estimé le temps de divergence maximum entre les plastomes des espèces tétraploïdes et hexaploïdes comme pouvant remonter jusqu'à 6 à 10 MA. L'analyse phylogénétique du gène de *S. maritima* codant une « Pentatricopeptide repeat (PPR) superfamily protein » nous a montré qu'une des copies du gène (Clade B, Figure 28) ne pouvait être détectée qu'avec des seuils inférieurs ou égaux à 10%.

La modification du **paramètre « profondeur » qui définit la profondeur minimale en reads pour détecter des polymorphismes**, nous montre qu'il existe des valeurs seuils critiques. La modification de ce paramètre à 10 reads entraine un temps d'exécution trop important ce qui peut représenter un inconvénient majeur lors de la manipulation de données NGS. A l'inverse, avec une profondeur en reads de 50 au minimum, le nombre de reads utilisés diminue fortement. Cette valeur est donc trop importante pour permettre une utilisation optimale des données Illumina. Cette valeur devrait donc être fixée à 30 reads pour obtenir des résultats optimaux.

Le **paramètre «nombre de SNPs pour assembler » les reads entre eux** ne fait pas varier les résultats. En effet, quelle que soit la valeur testée (2, 3 ou 4 SNPs pour assembler), le logiciel « IlluHaplotyper » détecte en moyenne 11 SNPs et 6 haplotypes par alignement de séquences homologues. De plus, ces résultats sont confirmés par une analyse contig par contig où l'on peut observer que la totalité des alignements étudiés ont une valeur de SNPs détectés et utilisés identique quelle que soit la valeur du paramètre testé. En comparant les différents nombres d'haplotypes détectés par le logiciel, nous pouvons observer que 88% des contigs ont un nombre d'haplotypes construits égal quelle que soit la valeur de ce paramètre. Ce paramètre n'est donc pas un facteur limitant pour la détection des haplotypes et n'influence pas le temps de calcul.

L'analyse en composantes principales nous a permis d'identifier que **le temps CPU d'exécution du programme est fortement corrélé au nombre d'haplotypes construits**. La représentation du plan factoriel selon les deux axes principaux montre que le temps CPU et le nombre d'haplotypes construits ne varient pas en fonction du paramètre « nombre de SNPs pour assembler ».

L'étude de la sensibilité et des répercussions directes des outils et paramètres sur les jeux de données nous a permis de sélectionner les valeurs des paramètres du logiciel « IlluHaplotyper » à appliquer sur les jeux de données génomiques et transcriptomiques étudiés.

« IlluHaplotyper » versus « PyroHaplotyper » :

Au cours des travaux présentés dans ce chapitre, il a été possible de détecter les sites polymorphes et de reconstruire les différentes copies dupliquées de gènes à partir d'alignements de reads de « courts fragments » de type Illumina. Nous détectons un nombre de SNPs plus important chez les deux espèces hybrides et l'allopolyploïde. Les données obtenues à partir des données Illumina sont cohérentes avec les résultats obtenus à partir des jeux de données Roche-454 et indiquent une même tendance chez les espèces étudiées. Comme attendu, les hybrides et l'allododécaploide présentent au sein de leurs génomes un polymorphisme intragénomique plus important dû à la présence des deux génomes parentaux divergents.

Pour les 5 espèces étudiées, le nombre de SNPs détectés au sein des données Illumina est plus important que le nombre de SNPs détectés au sein des données Roche-454. En effet, sur les 1 467 703 SNPs détectés sur l'ensemble des jeux de données des cinq espèces, 941 055 SNPs soit 64,12% sont uniquement présents dans les données Illumina. Ces résultats s'expliquent par la différence de profondeur entre les jeux de données Roche-454 et Illumina et par le réglage du paramètre « seuil » des deux logiciels. Les reads 454 présentant un nombre d'erreur de séquençage plus important que les reads Illumina, le paramètre seuil du logiciel « PyroHaplotyper » a été réglé à 20% tandis que le paramètre seuil du logiciel « IlluHaplotyper » a été réglé à 2%. Ces valeurs ont été fixées en fonction de plusieurs études similaires et des tests réalisés sur nos jeux de données. Le choix de régler le seuil de détection de SNPs à 20% pour les données Roche-454 est en adéquation avec des méthodes usuellement utilisées pour l'analyse de génomes polyploïdes. Pour l'analyse de

données d'EST (« Expressed Sequence Tag ») du coton tétraploïde, Udall et ses collaborateurs (2006) ont choisi de fixer ce paramètre à 25%. Tennessen et ses collaborateurs (2014) ont choisi d'ajuster ce paramètre à 12,5% (soit une fréquence de 1/8) pour détecter les variants présents au sein des fraisiers octoploïdes. Notre paramètre est moins stringent que le paramètre utilisé pour la détection de SNPs homéologues (fixé à 40%) chez le coton tétraploïde (Page et al. 2013a; Page et al. 2013b). Le choix de fixer le paramètre de détection de SNPs à 2% pour le logiciel « IlluHaplotyper » nous permet de supprimer les erreurs de séquençages de l'ordre de 1% présents dans les jeux de données Illumina (Oliphant et al. 2002). Néanmoins, la faible valeur de ce paramètre nous permet également de détecter certains SNPs potentiellement liés à des copies paralogues (para-SNPs) ou d'anciens homéologues (homéo-SNPs). L'étude des taux de mutations synonymes (K_S) nous montre la présence de 2 lots de copies dupliquées (potentiellement homéologues, de 1,8-2,7 MA et 8,7-9,6 MA, Figure 5 du papier « Reference transcriptomes and detection of duplicated copies in hexaploid parents, hybrids and allododecaploid *Spartina* species (Poaceae) » présenté ci dessus). La valeur du paramètre « seuil de détection » est donc difficile à fixer et dépend de l'histoire et du nombre de copies de chaque gène étudié. Le nombre de SNPs détectés uniquement à partir des données Roche-454 peut s'expliquer par le fait que certains de ces SNPs apparaissent dans les données Illumina, mais la profondeur Illumina reste trop faible pour valider ces derniers (profondeur réglé à 30 reads par défaut). De plus, nos résultats sont en accord avec plusieurs études qui ont montré qu'il était nécessaire d'associer des jeux de données de différentes technologies pour détecter les polymorphismes et contourner les erreurs de séquençage propres à chaque technologie (Schulz et al. 2014). Notre étude menée sur l'ADN ribosomique de *S. maritima* (Chapitre 4, Partie A) a permis de valider la majorité des polymorphismes détectés avec des données de pyroséquençage. Une minorité des polymorphismes détectés dans des régions homopolymères et qui ne sont pas validés par les données du logiciel « IlluHaplotyper » pourraient cependant correspondre à des faux-positifs comme cela a été montré chez l'ADNr de *S. maritima* ou 4 SNPs détectés avec les données Roche-454 et présents dans des régions homopolymères n'ont pas pu être validés (Boutte *et al.* in press).

Enfin, Le programme « IlluHaplotyper » permet de reconstruire des haplotypes à partir de données de type Illumina et les résultats obtenus ont pu être validés par des jeux

de données indépendants. Néanmoins, le nombre d'haplotypes détectés chez les hybrides et l'allopolyploïde reste très important. Ces résultats indiquent que le logiciel n'assemble pas les reads issus de copies dupliquées au sein d'alignements présentant un nombre trop important de polymorphismes. Une approche intéressante permettant de contourner ce problème serait d'aligner les reads des espèces hybrides et allopolyploïdes sur les génomes parentaux et de détecter indépendamment les haplotypes. Les jeux de données pourraient ensuite être poolés et un re-assemblage des copies identiques (présentes chez les deux parents) pourrait être réalisé. Néanmoins, une telle approche nécessite d'avoir les deux copies parentales pour chaque gène étudié. De plus, les copies parentales doivent être assez divergentes pour séparer les jeux de données.

Validation du logiciel « IlluHaplotyper » :

L'analyse des données génomiques du gène *Waxy* a permis de détecter différentes copies qui ont pu être validées et comparées. Treize haplotypes détectés au sein de *S. maritima* ont été retrouvé chez les trois autres espèces étudiées. Pour la deuxième espèce hexaploïde *S. alterniflora*, quatre haplotypes ont été retrouvés chez les autres espèces. Chez les deux espèces tétraploïdes, respectivement 32 et 24 haplotypes ont été retrouvés chez les autres espèces. Il est intéressant de noter que la majorité des haplotypes que nous détectons avec le logiciel « IlluHaplotyper » est validée par les données de clonage (Fortune et al. 2007). Nous détectons également des copies qui n'avait pas pu être mises en évidence dans les études précédentes (comme la copie B2 de *S. maritima*). Ceci confirme la contribution majeure que pourra apporter le séquençage en masse dans les phylogénies de polyploïdes (Tennessen et al. 2014 ; Brassac et al. 2015).

Les logiciels « PyroHaplotyper » et « IlluHaplotyper » développés ici n'utilisent pas de génomes diploïdes de référence pour identifier les différentes copies présentent chez des espèces polyploïdes contrairement à d'autres logiciels disponibles (Duchemin et al. 2014; Page et al. 2014). De plus, ces logiciels ont pour avantage de détecter eux-mêmes les

polymorphismes présents chez les polyploïdes contrairement aux logiciels PolyCat, PolyDog et SNiPloid pour lesquels il est nécessaire de fournir une banque de SNPs détectés chez les génomes parentaux (Page, Gingle, and Udall 2013; Peralta et al. 2013; Page and Udall 2015). Il est à présent possible d'étudier en détail les conséquences de la polyploïdie sur les génomes polyploïdes ne présentant pas de références diploïdes. Un développement intéressant de nos résultats concerne la perspective d'analyser les niveaux d'expression de chaque copie homéologue dans le génome des spartines polyploïdes. Peu d'études ont été réalisées de ce point de vue chez les polyploïdes si on excepte le coton (Udall et al. 2006; Flagel et al. 2008; Flagel et al. 2009; Flagel, Wendel, and Udall 2012; Yoo, Szadkowski, and Wendel 2013), *Arabidopsis* (Chang et al. 2010), le blé (Akhunova et al. 2010), le caféier (Combes et al. 2012; Combes et al. 2013) et le soja (Ilut et al. 2012), chez lesquels les espèces parentales sont diploïdes et non pas comme dans notre cas des espèces hexaploïdes.

Détection des copies dupliquées dans 5 nouveaux transcriptomes de Spartines hexaploïdes et allododécaploïde assemblés à partir de données Roche-454 et Illumina.

Cinq nouveaux transcriptomes construits au cours de cette thèse permettent d'enrichir les données de séquençage déjà disponibles pour les espèces du genre *Spartina*. Une première analyse de transcriptome avait été réalisée chez une espèce tétraploïde, *Spartina pectinata*, à l'aide de données Roche-454 (Gedye et al. 2010). Pour les espèces hexaploïdes, les quelques séquences disponibles sur le NCBI (Baisakh, Subudhi, and Varadwaj 2008; Chelaifa, Mahé, and Ainouche 2010) ont été nettement enrichies par les travaux de Ferreira de Carvalho *et al.* (2012) qui ont reconstruit les deux premiers transcriptomes de référence des espèces hexaploïdes *S. maritima* et *S. alterniflora* et annoté 16 753 gènes (à partir de 38 478 contigs). Au cours de ce travail, nous avons obtenu un nombre de contigs compris entre 44 158 et 65 099. L'assemblage de nos données transcriptomiques obtenues à partir des données Roche-454 est en accord avec les résultats obtenus avec des données similaires (Ferreira de Carvalho et al. 2012). Le choix d'utiliser le logiciel Trinity pour l'assemblage des données Illumina est motivé par plusieurs études

(Clarke et al. 2013; Liu et al. 2013) et des tests réalisés sur nos jeux de données, notamment par l'étude comparative des logiciels Trinity et Minia. Nous avons annoté 36.50% à 43.57% des contigs reconstruits à l'aide d'une recherche d'homologie de séquences (contre des bases de données d'angiospermes annotées) similaire à celles utilisées par Ferreira de Carvalho et ses collaborateurs (tblastx, Blast2Go; Ferreira de Carvalho et al. 2012) et à l'aide du logiciel Pfam qui a été utilisé dans des pipelines d'annotation tel que TRAPID (Van Bel et al. 2013). Le nombre de contigs non annotés peut s'expliquer par le choix d'avoir conservé les contigs d'une longueur comprise entre 40 et 200 bp qui sont habituellement supprimés des transcriptomes construits. Ce choix a été motivé par le fait que plusieurs contigs inférieurs à 100 bp ont pu être fonctionnellement annotés. Les longueurs des contigs annotés construits au cours de cette étude sont similaires à celles des contigs annotés par Ferreira de Carvalho (2012).

Après un mapping peu stringent à l'aide du logiciel Bowtie 2, nous avons détecté un nombre de SNPs similaires au sein des espèces parentales *S. maritima* et *S. alterniflora*. Le nombre de SNPs détectés au sein des hybrides et de l'allopolyploïde est plus important, comme attendu puisqu'ils combinent les génomes de deux espèces divergentes (*S. maritima* et *S. alterniflora*). Le nombre d'haplotypes reconstruits est corrélé au nombre de SNPs détectés : nous avons trouvé en moyenne 7 haplotypes au maximum par fenêtre d'alignement chez les espèces parentales et entre 12 et 14 haplotypes au maximum chez les hybrides et *S. anglica*. Ces valeurs sont plus importantes que le nombre de copies homéologues attendues au sein de ces espèces. Si nous attendions jusqu'à 3 et 6 paires de copies selon les espèces, la nature du jeu de données (transcriptomique) peut expliquer le nombre de copies moins important détecté dans le cas de la non-expression de certaines copies. Pour plusieurs gènes, nous avons montré qu'un maximum de quatre haplotypes sont détectés chez les parents hexaploïdes à l'aide de données Roche-454, ce qui indique la présence de deux paires de copies homéologues exprimées (Ferreira de Carvalho et al. 2012). Pour certains loci, la déviation du nombre de copies observées par rapport à l'attendu pourrait aussi résulter de pertes de copies (Buggs et al. 2012) : l'étude menée sur le gène *Waxy* à l'aide de clonage a permis de montrer la présence d'une copie homéologue chez *S. maritima* et trois copies homéologues chez *S. alterniflora* (Fortune et al. 2007). Bien que nos analyses de données NGS de type Illumina nous aient permis de mette en évidence la

présence d'une seconde copie chez *S. maritima*, il est aujourd'hui montré que le génome de l'espèce hexaploïde *S. maritima* a subit des phénomènes de diploïdisations, notamment pour l'unité 45S de l'ADN ribosomique (Boutte et al. in press, Chapitre 4). Il est intéressant de noter qu'une étude en cours au sein du laboratoire, portant sur des données de BACs de Spartines montre la présence de deux homéologues (et non trois copies comme attendu) dans deux régions génomiques différentes de *S. maritima* (Charron et al. in prep).

L'analyse des taux de substitutions synonymes entre paires d'haplotypes selon la démarche proposée par Blanc et Wolfe (2004) a permis de mettre en évidence la présence de 2 « pics » de duplications au sein des jeux de données de l'espèce hexaploïde *S. maritima*, résultant des deux évènements de polyploïdisation qui ont eu lieu dans cette lignée au cours des 10 derniers millions d'années (Ainouche et al. 2012; Rousseau-Gueutin et al. 2015). En nous basant sur l'horloge moléculaire classiquement considérée chez les Poaceae (Gaut et al. 1996) nous avons estimé le temps de divergence correspondant à ces deux « pics » respectivement à 1,8-2,7 MA et 8,7-9,6 MA. Ces valeurs se rapprochent de celles estimées entre les plastomes des espèces hexaploïdes *S. maritima* et *S. alterniflora* (moins de 2 à 4 MA) et entre les génomes chloroplastiques des hexaploïdes et tétraploïdes (moins de 6 à 10 MA). Il est à noter que contrairement à ce qui est parfois abusivement indiqué dans la littérature, les « pics » de Ks observés suite à la polyploïdisation **n'indiquent pas forcément l'âge de la formation du polyploïde** mais plutôt l'âge de la divergence entre les homéologues, remontant à **la divergence des ancêtres diploïdes** (Doyle et al. 2010). Les données disponibles à ce jour concernant la position des spartines au sein des Chloridoideae (Peterson et al. 2014b) suggèrent que les diploïdes actuels les plus proches des spartines appartiennent à des représentants du genre *Sporobolus* ayant divergé du clade des spartines il y a 12 à 20 MA (estimé à partir d'ADN chloroplastique, Rousseau-Gueutin et al. 2015). Les datations moléculaires sont de plus dépendantes de différents paramètres (pressions évolutives agissants sur les gènes comparés, taux d'évolution variables entre génomes nucléaires et chloroplastiques (Wolfe, Li, and Sharp 1987)).

Au cours de cette thèse, nous avons initié des approches phylogénomiques sur des données NGS obtenues chez les spartines polyploïdes. Le développement de ces approches pourra être appliqué sur gènes présents en faible nombre de copies chez les Poaceae, tels que ceux décrits dans le Chapitre 3 (*e.g. Waxy*, *AGT1*, *Bio2*, *PHYa*, *PHYC*, *DMC*-1) et dont la

profondeur de séquençage pourra être augmentée à l'aide de la méthode de capture de séquence en utilisant des sondes exoniques (Grover, Salmon, and Wendel 2012, Salmon et al. 2012). Pour ces gènes, nous pourrons analyser les espèces à différent niveaux de ploïdie : espèces tétraploïdes (*S. versicolor*, *S. pectinata*, *S. bakeri, S. argentinensis* et *S. gracilis*), hexaploïdes (*S. maritima*, *S. alterniflora*, *S. x townsendii*, *S. x neyrautii* et *S. foliosa*), heptaploïde (*S. densiflora*) et dodécaploïde (*S. anglica*) ainsi que des outgroups proches (*e.g*. *Sporobolus*) ce qui nous permettra d'identifier plus aisément l'origine des copies dupliquées et l'histoire ancienne de ce genre polyploïde.

Chapitre 6 :

Conclusion générale et perspectives.

CHAPITRE 6 : Conclusion générale et perspectives.

Au cours de ce travail, nous avons cherché à détecter les différentes copies dupliquées (homéologues et/ou paralogues) présentes au sein de génomes hautement polyploïdes à partir de données de séquençage à haut débit (NGS). Les différents outils développés au cours de cette thèse ont été appliqués et validés sur différentes ressources génomiques et transcriptomiques (issues de données Roche-454 et Illumina) de Spartines.

Dans un premier volet de cette thèse, nous nous sommes intéressé au développement d'outils permettant la détection de polymorphismes et de copies dupliquées à partir de données de Pyroséquençage Roche-454 (ou de diverses technologies générant de longs fragments). L'application de ce programme sur la région hautement dupliquée de l'ADN ribosomique 45S de *S. maritima* a permis de détecter de nombreux haplotypes dans les régions non codantes de ce gène et plus particulièrement dans les régions IGS et ETS. Nous avons mis en évidence une homogénéisation intragénomique des régions codantes (18S, 5.8S et 25S) et des ITS. Une hétérogénéité des régions contenant l'espaceur intergénique (IGS) et l'espaceur externe transcrit (ETS) a également été observée. Les résultats obtenus à l'aide de cet outil « PyroHaplotyper » ont été validés à l'aide de données de clonage et (re)-séquençage obtenu par la méthode Sanger et de données de séquençage par synthèse Illumina. La présence d'un seul locus de l'ADNr 45S a été validé en cytogénétique par la méthode FISH indiquant que cette région a donc subi un phénomène de diploïdisation au sein de cette espèce hexaploïde. Nous avons également identifié les différents sites polymorphes et les copies dupliquées au sein d'un sous jeu de données transcriptomiques de cinq espèces de Spartines polyploïdes obtenues à l'aide de Pyroséquençage.

Dans un deuxième temps, nous avons construit cinq nouveaux transcriptomes de référence pour deux espèces hexaploïdes *S. maritima* et *S. alterniflora*, leurs deux hybrides (*S. x townsendii* et *S. x neyrautii*) et l'allopolyploïde *S. anglica* dérivant de la duplication du génome de *S. x townsendii*. Ces nouveaux transcriptomes viennent enrichir les deux premiers transcriptomes de *S. martima* et *S. alterniflora*, à l'aide de données issues de

séquençage Illumina. Nous avons également développé « IlluHaplotyper », un programme permettant de détecter les sites polymorphes et de reconstruire les haplotypes à partir de données de lectures courtes de type Illumina. L'application de cet outil sur les cinq transcriptomes de référence a permis d'identifier les différentes copies dupliquées et d'évaluer la divergence (Ks) entre ces copies. Les dates de divergence correspondant aux « pics » de Ks (résultant de duplications génomiques) ont été estimées.

Enfin, nous avons initié une approche phylogénomique adaptée à nos ressources permettant d'identifier les différentes régions orthologues de plusieurs espèces de Poaceae, à partir des haplotypes des espèces hexaploïdes *S. maritima* et *S. alterniflora*, préalablement construits à l'aide du logiciel « IlluHaplotyper ». Il a été ainsi possible de réaliser 3 442 analyses phylogénomiques à l'aide du logiciel RAxML. L'application de ce pipeline sur des jeux de données génomiques de capture de séquences devrait permettre à terme, d'élucider l'histoire des différents clades de Spartines polyploïdes.

Le défi biologique

Le genre *Spartina*, dans lequel se trouve un exemple classique de spéciation récente par allopolyploïdie (*Spartina anglica*) se présente comme un excellent modèle pour étudier les conséquences de la polyploïde à court terme ; de plus, ce genre présente un intérêt incontestable en écologie évolutive, par son rôle « d'ingénieur d'écosystèmes » sur les marais salés côtiers, ses capacités à stabiliser les estrans, et sa tolérance à la pollution (accumulation de Cuivre, de Cadmium et de Plomb) et en phytoremédiation (Ainouche, Baumel, and Salmon 2004; Ainouche et al. 2008; Strong and Ayres 2013). Toutefois, si le genre *Spartina* représente l'un des rares cas connus de spéciation récente permettant de comparer les espèces parentales (*S. maritima* et *S. alterniflora*), des hybrides formés indépendamment (*S. x townsendii* et *S. x neyrautii*) et de l'espèce allododécaploïde (*S. anglica*, issue du doublement génomique de l'hybride *S. x townsendii*), l'étude de ce genre représente un défi majeur, notamment pour des approches de génomique (Salmon and Ainouche 2015). En effet, au sein du genre *Spartina*, marqué par des évènements récurrents

de duplications de génomes au cours de son histoire, aucune espèce diploïde n'est recensée à ce jour. Ainsi, pour étudier les espèces du genre *Spartina*, il a été nécessaire de développer des approches adaptées à des génomes complexes, pour lesquels aucune référence diploïde n'est disponible. De plus, bien qu'appartenant à la famille des Poaceae (regroupant des espèces modèles ou à intérêt agronomique séquencées tel que le riz (Ouyang et al. 2007), le sorgho (Paterson et al. 2009) ou le maïs (Schnable et al. 2009)), les Spartines appartiennent à la sous famille des Chloridoideae, qui reste très peu étudiée. Récemment, le séquençage du génome d'*Eragrostis tef* (2n=4x=40 ; Cannarozzi et al. 2014), une espèce tétraploïde à intérêt agronomique de la sous famille des Chloridoideae, a permis de comparer nos jeux de données à un génome plus proche de la même sous famille. Néanmoins, les analyses comparatives que nous avons menées sur des données génomiques et transcriptomiques du riz, du sorgho, du tef et des Spartines ont montré d'importantes différences entre le tef et les autres espèces étudiées. Ces informations qui peuvent indiquer d'importants réarrangements au niveau du génome *d'Eragrostis tef* ou des erreurs survenues lors de l'assemblage des données de cette première version (« draft ») du génome nous ont obligés à écarter cette espèce de plusieurs analyses.

Le défi méthodologique

La recherche de copies homéologues au sein de génomes polyploïdes se présente comme une question essentielle pour retracer l'histoire évolutive des espèces ou identifier le devenir des copies dupliquées. De nombreuses études ont développé des approches similaires pour identifier ces copies, au sein de nombreuses espèces polyploïdes à intérêt scientifique et/ou agronomique majeur, comme le coton (Udall 2006; Flagel et al. 2008; Salmon et al. 2009; Flagel, Wendel, and Udall 2012), le caféier (Combes et al. 2012; Combes et al. 2013; Combes et al. 2015), le soja (Ilut et al. 2012), le blé (Akhunova et al. 2010), *Arabidopsis thaliana* (Chang et al. 2010) ou plus récemment le fraisier (Tennessen et al. 2014). Plusieurs outils permettant la détection de copies homéologues ont ainsi été développés dans ce but : SNiPLoid (Peralta et al. 2013), PolyCat (Page, Gingle, and Udall 2013), BamBam (Page et al. 2014), PolyDog (Page and Udall 2015) ou HyLiTE (Duchemin et

al. 2014). Pour l'ensemble de ces approches, les SNPs spécifiques à chacune des espèces diploïdes sont identifiés pour discriminer les différentes copies chez l'espèce allopolyploïde. Chez les Spartines hautement polyploïdes, aucune espèce diploïde n'est référencée à ce jour ; il a donc été nécessaire de développer une approche spécifique permettant de discriminer les différentes copies dupliquées (homéologues et/ou paralogues). C'est dans cette optique que nous avons développé les programmes « PyroHaplotyper » et « IlluHaplotyper » présentés dans les chapitres précédents. Il a été nécessaire d'adapter notre méthode en fonction de la nature des jeux de données de séquençage. Les algorithmes ainsi que les différentes options de ces programmes ont donc été adaptés en fonction des reads générés par les séquenceurs (« courts fragments » pour la technologie Illumina ou « longs fragments » pour la technologie Roche-454) et en fonction des erreurs de séquençage propres à chaque technologie. L'étude de l'impact des paramètres sur les résultats obtenus nous a permis d'identifier les paramètres optimaux à appliquer sur nos jeux de données. Un résultat intéressant obtenu lors de notre étude sur ces paramètres concerne le paramètre « seuil » du logiciel « IlluHaplotyper » (Chapitre 5) qui permet d'identifier les SNPs et d'éliminer les erreurs de séquençage ainsi que les faux-positifs. En effet, nous détectons un nombre d'haplotypes (de copies) variables plus ou moins divergents en fonction de la valeur de ce paramètre. Il serait intéressant d'analyser plus en détails l'impact de ce paramètre sur des gènes bien connus et d'identifier pour chaque alignement considéré, une éventuelle valeur clé nous permettant de différencier les copies homéologues des copies paralogues, sans avoir à les discriminer à l'aide de méthodes phylogénétiques coûteuses en temps de calcul pour ces jeux de données à très haut-débit. Les programmes développés au cours de cette thèse (Chapitre 4 et 5), présentent un intérêt majeur pour l'étude des espèces polyploïdes pour lesquelles aucune référence diploïde n'est connue. Ces outils peuvent également être appliqués sur des jeux de données d'espèces diploïdes pour la recherche d'allèles dans le cadre d'études en génétique des populations. C'est dans cet optique que le logiciel « PyroHaplotyper » a été appliqué sur des données de Nématodes à kyste (en collaboration avec Cécile Gracianne et Eric Petit (UMR ESE, INRA, Rennes)) et de Gorilles (en collaboration avec Alice Baudouin et Pascaline Le Gouard (UMR-CNRS Ecobio, Université de Rennes 1)). Pour rendre ces outils disponibles à la communauté scientifique et permettre leur application sur divers jeux de données, le logiciel « PyroHaplotyper » est en cours d'intégration sous l'environnement Galaxy de la Plateforme

de BioGenouest (Namour 2015 ; en collaboration avec Yvan Le Bras (INRIA/IRISA, Genouest)). Une perspective à cours terme sera d'intégrer « IlluHaplotyper », le second outil de détection d'haplotypes, sous Galaxy. Ce travail pourra être réalisé au sein du laboratoire, en collaboration avec la plateforme GenOuest de Rennes. De plus, cet outil, ne dépend d'aucun programme/logiciel sous licence propriétaire (comme le logiciel Newbler pour « PyroHaplotyper »), ce qui devrait permettre une intégration aisée et rapide.

L'évolution de l'ADN ribosomique chez les Spartines

Au cours de cette thèse nous nous sommes intéressé à l'ADN ribosomique de *S. maritima*. Ce choix a été motivé par le fait que les gènes codant l'ARN ribosomique (ARNr) forment une famille multigénique qui est très souvent utilisée en phylogénie moléculaire. Leur analyse est facilitée par la possibilité de définir des amorces « universelles » dans les régions codantes ce qui permet d'analyser la région contenant les deux espaceurs internes transcrits (ou ITS) connues pour être des régions variables entre espèces proches (White, Bruns, and Taylor 1990). Ces séquences sont connues pour être soumisses à un processus d'homogénéisation par conversion génique aboutissant à une évolution concertée, ce qui permet d'éviter les biais liés à la paralogie (Baldwin et al. 1995) et de réaliser des analyses phylogénétiques interspécifiques. Dans le cadre de l'étude d'espèce d'origine hybride et/ou polyploïde, il est nécessaire de combiner l'analyse phylogénétique obtenue avec les ITS avec d'autres analyses de gènes nucléaires en copie unique et/ou chloroplastiques. En effet, l'évolution différentielle des copies parentales ainsi que les taux variables d'évolution concertée représentent une importante limite pour les gènes ribosomiques (Álvarez and Wendel 2003). Au cours de notre étude, nous avons mis en évidence une homogénéisation intragénomique des régions codantes et des ITS, comme attendu. Nous avons également montré l'hétérogénéité des régions contenant l'espaceur intergénique et l'espaceur externe transcrit. Le nombre de copies détectées et la faible présence de certaines de ces copies nous montre cependant un processus d'homogénéisaton en cours dans ces régions. L'analyse de ces données a mis en évidence plusieurs éléments intéressants (la taille importante de l'ETS en 5' ; le faible pourcentage en GC dans les ITS 1 et 2 ; le nombre de

NORs détectés (1 contre 3 attendu)) qui montrent une évolution particulière de l'ADN ribosomique chez cette espèce. Néanmoins plusieurs résultats montrent que la majorité de ces éléments ne sont pas propre à *S. maritima*. Une analyse préliminaire des ITS de plusieurs espèces de Poaceae et en particulier de Chloridoideae que nous avons réalisé à partir des bases de données a permis de mettre en évidence que le faible taux en GC au sein des ITS serait commun aux espèces appartenant aux tribus des Eragrostideae, Zoysieae et Cynodonteae. Le pourcentage en GC chez les Tritaphideae et Centropodieae étant plus important et pouvant contenir 10 à 20% de GC en plus au sein de ces régions. La perte de loci qui est un phénomène courant chez les polyploïdes (*e.g.* le fraisier octoploïde *Fragaria iturupensis*; (Liu and Davis 2011) ou le colza *Brassica napus* (Snowdon, Köhler, and Köhler 1997)) et que nous avons détectée chez *S. maritima* serait un phénomène commun aux différentes espèces de Spartines. Une étude menée au sein du laboratoire a également mis en évidence la perte d'une paire loci chez l'espèce tétraploïde *Spartina versicolor* (Malika Ainouche et Olivier Coriton, données non publiées). La perte d'une paire de loci serait peut être commune à l'ensemble des espèces du genre *Spartina* ce qui indiquerait que ce processus aurait eu lieu après l'évènement de tétraploïdisation des Spartines. Une perspective intéressante serait d'analyser plus en détail l'ADN ribosomique des différentes espèces de Spartines et d'espèces proches, ce qui permettrait de mieux comprendre l'histoire évolutive de ce genre.

Les NGS, des outils pour explorer les génomes des Spartines

Les avancées technologiques et scientifiques permettent de mieux appréhender et comprendre les mécanismes des génomes non-modèles et particulièrement complexes dans le genre *Spartina*. Un premier transcriptome de Spartine tétraploïde (*S. pectinata*) a été obtenu à l'aide de données 454 (Gedye et al. 2010). Au cours de notre étude nous avons enrichi à l'aide de données Illumina les premiers transcriptomes de Spartines hexaploïdes construits pour les deux parents *S. maritima* et *S. alterniflora* (Ferreira de Carvalho et al. 2012). Nous avons également construits *de novo* trois nouveaux transcriptomes de référence pour les espèces hybrides (*S. x neyrautii* et *S. x townsendii*) et l'espèce

allododécaploïde *S. anglica* à partir de données Roche-454 et Illumina. Ces cinq transcriptomes construits à l'aide d'une approche combinant directement les données Roche-454 et Illumina, nous avons pu obtenir entre 44 158 et 65 099 contigs (dont 19 241 à 25 067 contigs annotés). Pour les différents transcrits construits, nous avons détecté les différentes copies dupliquées, au sein des données Roche-454 (Chapitre 4) et Illumina (Chapitre 5). A partir des données Roche-454, nous détectons comme attendu un nombre moyen de SNPs (pour 100 bp) plus important chez les espèces hybrides *S. x neyrautii* et *S. x townsendii*, ces deux espèces combinant les génomes des deux parents divergents. Nous obtenons des résultats similaires à l'aide des données Illumina, le nombre de moyen de SNPs détectés chez les deux hybrides et l'allopolyploïde (entre 5,32 et 6,10 SNPs pour 100 bp) est plus important que le nombre de SNPs détectés chez les parents (3,85 SNPs pour 100bp). Au cours de notre étude, nous avons observé que le nombre d'haplotypes détectés à partir des données Roche-454 est similaire pour les cinq espèces étudiées (entre 4,80 et 5,84 haplotypes par alignement de séquences), ce qui n'est pas le cas avec les données Illumina, où nous avons détecté un nombre d'haplotypes transcrits plus important chez les hybrides et l'espèce allopolyploïde. En effet, en moyenne 7 haplotypes ont été détectés au maximum par fenêtre chez les espèces parentales et entre 12 et 14 haplotypes au maximum chez les hybrides et *S. anglica*. Ces résultats peuvent indiquer la présence de copies homéologues (et leurs allèles) et paralogues. Cette divergence de résultats, malgré des paramètres d'alignement identiques (bien que réalisé à l'aide de différents outils de mapping : Newbler et Bowtie 2) est essentiellement due à la taille des reads séquencés qui joue un rôle majeur dans la construction des haplotypes. En effet, le logiciel « IlluHaplotyper » est ainsi contraint dans l'assemblage des reads des espèces hybrides et de l'allododécaploïde, ce qui est en relation avec les paramètres de nos programmes et notre volonté de limiter au maximum la création d'haplotypes chimériques.

Les technologies de séquençage à haut débit nous permettent, via les outils développés au cours de cette thèse de mieux appréhender les génomes complexes des Spartines. Nos résultats comparatifs avec les données génomiques (issues de clonage) du gène *Waxy* (Fortune et al. 2007) ont, par exemple, permis de montrer l'importance et l'intérêt de ces approches. Nous avons ainsi mis en évidence la présence de nouvelles copies

au sein de *S. maritima*. Il a également été possible d'identifier jusqu'à six haplotypes du gène *Waxy* chez l'espèce tétraploïde *S. bakeri*.

Les copies dupliquées au sein de génomes polyploïdes peuvent évoluer de différentes façons (Adams and Wendel 2005). Suite à la potentielle relâche des contraintes sélectives sur l'une des copies, cette dernière pourra évoluer librement en accumulant des mutations délétères (pseudogénéisation) ou en étant éliminée, ce qui participe au phénomène de diploïdisation des génomes polyploïdes. D'un point de vue fonctionnel, l'expression des copies dupliquées peut également tendre vers une sous-fonctionnalisation ou néo-fonctionnalisation (Ohno 1970; Wendel 2000; Lynch and Force 2000; Adams et al. 2003). Par exemple, plusieurs études ont montré chez des espèces allopolyploïdes, dans le cas de la rétention des copies parentales, la mise sous silence de la copie de l'un des parents (Comai et al. 2000, Adams et al. 2003, Adams 2004, Wang et al. 2004, Buggs et al. 2010). Les génomes polyploïdes sont de ce fait particulièrement remaniés et complexes. La détection des copies dupliquées et l'identification de leur origine phylogénétique sont donc essentielles pour étudier la rétention ou la perte des copies et ainsi étudier leur devenir.

Les résultats obtenus et les outils développés au cours de cette thèse ont des applications directes pour étudier notamment l'évolution d'une fonction d'intérêt écologique majeur. En effet, certaines Spartines sont productrices de DMSP (Dimethylsulfoniopropionate) qui est le précurseur du DMS (Dimethylsulfide). Ce composé volatile est impliqué dans le cycle du souffre, dans les précipitations acides ainsi que dans la régulation du climat (Yoch 2002). Le travail de thèse d'Hélène Rousseau (en cours au sein du laboratoire) se concentre sur l'étude de l'émergence de cette nouvelle fonction chez certaines espèces de Spartines. Pour identifier les gènes (putatifs) responsables de cette fonction, les différents transcriptomes de référence des Spartines construit au cours de cette thèse ont été utilisés (Hélène Rousseau *et al.*, in prep). A partir des données de capture de séquence, il sera possible d'étudier plus en détail les gènes sélectionnés, d'identifier les différentes copies dupliquées et de les comparer entre espèces productrices et non-productrices de DMSP.

A partir des outils développés au cours de cette thèse, il est également possible d'analyser les niveaux d'expression de chaque copie homéologue dans le génome des

Spartines polyploïdes. Peu d'études ont été effectuées de ce point de vue chez les polyploïdes si on excepte le coton (Udall et al. 2006; Flagel et al. 2009; Flagel, Wendel, and Udall 2012; Flagel, Wendel, and Udall 2012), *Arabidopsis* (Chang et al. 2010), le soja (Ilut et al. 2012), le caféier (Combes et al. 2012; Combes et al. 2013) et le blé (Akhunova et al. 2010) chez lesquels les espèces parentales sont diploïdes et non hexaploïdes comme dans notre cas. C'est dans cette optique que des jeux de données transcriptomiques ont été séquencés au sein du laboratoire dans le cadre des travaux de Mathieu Rousseau-Gueutin (UMR-CNRS Ecobio, Université de Rennes 1 / UMR IGEPP, INRA, Le Rheu). Les travaux menés se concentrent sur 2 espèces hexaploïdes (*S. maritima* et *S. alterniflora*), les deux hybrides *S. x townsendii* et *S. x neyrautii*, l'espèce allododécaploïde *S. anglica* et deux espèces tétraploïdes (*S. bakeri* et *S. versicolor*). Des réplicas techniques et biologiques (7 lanes, 1 123 977 025 reads) ont ainsi été effectués dans le but de mesurer les niveaux d'expression des gènes entre les différentes espèces, mais également pour mesurer les niveaux d'expression (au niveau intraspécifique et interspécifique) des différentes copies dupliquées présentent au sein de chaque espèce cultivée dans les mêmes conditions.

D'autres approches, cherchant à identifier les potentielles régulations épigénétiques de l'expression des génomes pourront également bénéficier de ces programmes développés ; pour l'identification par exemple, de sites Uracile converti en Thymine après traitement au bisulfite et des haplotypes affectés par des changements de méthylation.

Vers la compréhension de l'origine des Spartines

Le développement des technologies de séquençage permet d'explorer plus en détail l'histoire du genre *Spartina*. A partir du génome chloroplastique de *S. maritima* (obtenu à l'aide de données Roche-454 et Illumina), Rousseau-Gueutin et ses collaborateurs (2015) ont comparé différentes régions chloroplastiques chez les spartines et daté la divergence entre le clade tétraploïde et hexaploïde (estimée entre 6 et 10 millions d'années). La divergence entre des espèces hexaploïdes *S. maritima* et *S. alterniflora* a été estimée entre 2 et 4 millions d'années. Ces résultats sont en accord avec les divergences estimées au cours de

cette thèse à partir des données transcriptomiques de Spartines entre haplotypes (datation des « pics » de Ks). L'analyse du gène *Waxy* avait permis de suggérer une origine allopolyploïde des Spartines hexaploïdes (Fortune et al. 2007). Néanmoins, les résultats sur l'hétérogénéité des séquences chez les Spartines (obtenus à partir de données transcriptomiques et génomiques), montrent dans la plupart des cas la présence de deux copies homéologues (Ferreira de Carvalho et al. 2012; Charron et al. in prep; Chapitre 5) ce qui soulève plusieurs hypothèses : une origine auto-allohexaploïde (ou allo-autohexaploïde) des Spartines hexaploïdes, des processus de diploïdisation ou de la conversion génique entre les copies (Wendel 2000). A partir des analyses effectuées au cours de ce travail (assemblage de jeux de données de Spartines), nous avons contribué à désigné des sondes pour une sélection de 17 gènes d'intérêt utilisés en phylogénie moléculaire (dont le gène *Waxy*), qui seront analysés par la méthode de capture de séquence (Grover, Salmon, and Wendel 2012; Salmon et al. 2012). Ceci nous permettra d'obtenir une profondeur de séquençage génomique très importante. Il sera alors possible, à l'aide des outils développés au cours de cette thèse, de détecter la totalité des haplotypes présents au sein du génome des Spartines et d'explorer leur histoire. Pour cela plusieurs espèces de Spartines (tetra, hexa et dodecaploïdes) seront analysées ainsi qu'un outgroup proche des Spartines (*Sporobolus heterolepis*) dans le but d'établir l'origine évolutive des copies détectées. Il sera également possible de préciser l'histoire du gène *Waxy* à l'aide de données NGS par capture de séquences.

Au cours de cette thèse nous avons répondu aux défis méthodologiques majeurs que représente l'application des NGS chez des espèces polyploïdes complexes afin d'apporter des éléments de connaissances sur l'évolution de différents types de séquences du génome des spartines, ce qui permettra, à terme, d'améliorer la compréhension de leur origine. Le développement constant des NGS offre de nouvelles possibilités ; par exemple des technologies tel que Nanopore (Wang, Yang, and Wang 2015) ou SMRT (Eid et al. 2009) qui génèrent des fragments de plusieurs milliers de paires de bases permettront de reconstruire des haplotypes pleines longueurs. La reconstruction et l'étude des génomes complexes de certaines espèces seront facilitées par ces technologies et l'utilisation d'outils bioinformatiques adaptés.

Bibliographie :

A

Abbott, R. J., Brennan A. C., James J. K., Forbes D. G., Hegarty M. J., and Hiscock S. J. **2008**. "Recent Hybrid Origin and Invasion of the British Isles by a Self-Incompatible Species, Oxford Ragwort (*Senecio Squalidus L.*, Asteraceae)." Biological Invasions **11** (5): 1145–58.

Abegunde, T. **2010**. "Comparison of DNA Sequence Assembly Algorithms Using Mixed Data Sources." Saskatoon, Saskatchewan Canada: University of Saskatchewan.

Abrouk, M., Murat F., Pont C., Messing J., Jackson S., Faraut T., Tannier E., Plomion C., Cooke R., and Feuillet C. **2010**. "Palaeogenomics of Plants: Synteny-Based Modelling of Extinct Ancestors." Trends in Plant Science **15** (9): 479–87.

Adams, K. L., Cronn, R., Percifield R., and Wendel, J. F. **2003**. "Genes Duplicated by Polyploidy Show Unequal Contributions to the Transcriptome and Organ-Specific Reciprocal Silencing." Proceedings of the National Academy of Sciences **100** (8): 4649–54.

Adams, K. L. **2004**. "Organ-Specific Silencing of Duplicated Genes in a Newly Synthesized Cotton Allotetraploid." Genetics **168** (4): 2217–26.

Adams, K. L., and Wendel. J. F. **2005**. "Polyploidy and Genome Evolution in Plants." Current Opinion in Plant Biology **8** (2): 135–41.

Ainouche, M. L., Baumel, A., Salmon, A., and Yannic, G. **2003**. "Hybridization, Polyploidy and Speciation in *Spartina* (Poaceae)." New Phytologist **161** (1): 165–72.

Ainouche, M. L., Baumel A., and Salmon, A. **2004**. "*Spartina Anglica* CE Hubbard: A Natural Model System for Analysing Early Evolutionary Changes That Affect Allopolyploid Genomes." Biological Journal of the Linnean Society **82** (4): 475–84.

Ainouche, M. L., Fortune, P. M., Salmon, A., Parisod, C., Grandbastien,M.-A., Fukunaga, K., Ricou, M. and Misset, M.-T. **2008**. "Hybridization, Polyploidy and Invasion: Lessons from *Spartina* (Poaceae)." Biological Invasions **11** (5): 1159–73.

Ainouche, M. L., Baumel, A., Bayer,R., Fukunaga, K., Cariou, T., and Misset, M. T. **2010**. "Speciation, Genetic and Genomic Evolution in Spartina." **1**:15–21. San Fransisco, CA, USA: Ayres, D.R., Kerr D.W., Ericson S.D. and Olofson P.R.

Ainouche, M L, Chelaifa H., Ferreira de Carvalho J., Bellot S., Ainouche, A K., and Salmon, A. **2012**. "Polyploid Evolution in Spartina: Dealing with Highly Redundant Hybrid Genomes." In Polyploidy and Genome Evolution: Dealing with Highly Redundant Hybrid Genomes, 225–43. Soltis, Pamela S.; Soltis, Douglas E. (eds) Polyploidy and Genome Evolution, Springer Berlin Heidelberg: Berlin, Heidelberg.

Ainouche, M. L., and Wendel, J. F. **2014**. "Polyploid Speciation and Genome Evolution: Lessons from Recent Allopolyploids." In Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life, edited by Pierre Pontarotti, 87–113. Cham: Springer International Publishing.

Akama, S., Shimizu-Inatsugi, R., Shimizu, K. K., and Sese, J. **2014**. "Genome-Wide Quantification of Homeolog Expression Ratio Revealed Nonstochastic Gene Regulation in Synthetic Allopolyploid Arabidopsis." Nucleic Acids Research **42** (6): e46–e46.

Akhunova, A. R., Matniyazov, R. T., Liang, H. and Akhunov, E. D. **2010**. “Homoeolog-Specific Transcriptional Bias in Allopolyploid Wheat.” BMC Genomics **11** (1): 505.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. **1990**. “Basic Local Alignment Search Tool.” Journal of Molecular Biology 215 (3): 403–10.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z.,Miller W., and Lipman, D. J. **1997**. “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs.” Nucleic Acids Research **25** (17): 3389–3402.

Álvarez, I., and Wendel, J. F. **2003**. “Ribosomal ITS Sequences and Plant Phylogenetic Inference.” Molecular Phylogenetics and Evolution **29** (3): 417–34.

An, S. Q., Gu, B. H., Zhou C. F., Wang, Z. S., Deng, Z. F., Zhi, Y. B., Li, H. L., Chen, L., Yu, D. H. , and Liu, Y. H. **2007**. “*Spartina* Invasion in China: Implications for Invasive Species Management and Future Research.” Weed Research **47** (3): 183–91.

Arabidopsis Genome Initiative. **2000**. “Analysis of the Genome Sequence of the Flowering Plant *Arabidopsis thaliana*.” Nature **408** (6814): 796–815.

Arbizu, C., Ruess, H., Senalik, D., Simon, P. W., and Spooner, D. M. **2014**. “Phylogenomics of the Carrot Genus (*Daucus*, *Apiaceae*).” American Journal of Botany **101** (10): 1666–85.

Arnold, M. L. **2008**. Evolution through Genetic Exchange. Repr. Oxford: Oxford Univ. Press.

Ayres, D. R., Grotkopp, E., Zaremba, K., Sloop, C. M., Blum, M. J., Bailey, J. P., Anttila, C. K., and Strong. D. R. **2008**. “Hybridization between Invasive *Spartina Densiflora* (Poaceae) and Native *S. Foliosa* in San Francisco Bay, California, USA.” American Journal of Botany **95** (6): 713–19.

B

Badaeva, E. D., Friebe, B., and Gill, B. S. **1996**. “Genome Differentiation in *Aegilops*. 2. Physical Mapping of 5S and 18S-26S Ribosomal RNA Gene Families in Diploid Species.” Genome / National Research Council Canada **39** (6): 1150–58.

Baisakh, N., Subudhi, P. K., and Varadwaj, P. **2008**. “Primary Responses to Salt Stress in a Halophyte, Smooth Cordgrass (*Spartina Alterniflora* Loisel.).” Functional & Integrative Genomics **8** (3): 287–300.

Baldwin, B. G., Sanderson, M. J., Porter, J. M., Wojciechowski, M. F., Campbell, C. S. and Donoghue, M. J. **1995**. “The Its Region of Nuclear Ribosomal DNA: A Valuable Source of Evidence on Angiosperm Phylogeny.” Annals of the Missouri Botanical Garden **82** (2): 247.

Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X., and Song, Y. **2011**. “Evaluation of next-Generation Sequencing Software in Mapping and Assembly.” Journal of Human Genetics **56** (6): 406–14.

Barakat, A., DiLoreto, D., S., Zhang, Y., Smith, C., Baier, K., Powell, W. A., Wheeler, N., Sederoff, R., and Carlson, J. E. **2009**. “Comparison of the Transcriptomes of American

Chestnut (*Castanea Dentata*) and Chinese Chestnut (*Castanea Mollissima*) in Response to the Chestnut Blight Infection." BMC Plant Biology **9** (1): 51.

Bardil, A., Dantas de Almeida, J., Combes, M.-C., Lashermes, P., and Bertrand, B. **2011**. "Genomic Expression Dominance in the Natural Allopolyploid *Coffea Arabica* is Massively Affected by Growth Temperature." New Phytologist **192** (3): 760–74.

Barker, R. F., Harberd, N. P., Jarvis, M. G., and Flavell., R. B. **1988**. "Structure and Evolution of the Intergenic Region in a Ribosomal DNA Repeat Unit of Wheat." Journal of Molecular Biology **201** (1): 1–17.

Barthelson, R., McFarlin, A. J., Rounsley, S. D., and Young, S. **2011**. "Plantagora: Modeling Whole Genome Sequencing and Assembly of Plant Genomes." Edited by Matteo Pellegrini. PLoS ONE **6** (12): e28436.

Baumel, Alex, Ainouche, M. L., and Levasseur, J. E. **2001**. "Molecular Investigations in Populations of *Spartina anglica* CE Hubbard (Poaceae) Invading Coastal Brittany (France)." Molecular Ecology **10** (7): 1689–1701.

Baumel, A., Ainouche, M. L., Bayer, R. J., Ainouche, A. K., and Misset, M. T. **2002**. "Molecular Phylogeny of Hybridizing Species from the Genus *Spartina* Schreb. (Poaceae)." Molecular Phylogenetics and Evolution **22** (2): 303–14.

Baumel, A., Ainouche, M. L., Misset, M. T., Gourret, J-P., and Bayer, R. J. **2003**. "Genetic Evidence for Hybridization Between the Native *Spartina maritima* and the Introduced *Spartina alterniflora* (Poaceae) in South-West France: *Spartina x neyrautii* Re-Examined." Plant Systematics and Evolution **237** (1-2): 87–97.

Bedada, G., Westerbergh, A., Müller, T., Galkin, E., Bdolach, E., Moshelion, M., Fridman, E., and Schmid, K., J. **2014**. "Transcriptome Sequencing of Two Wild Barley (*Hordeum Spontaneum L.*) Ecotypes Differentially Adapted to Drought Stress Reveals Ecotype-Specific Transcripts." BMC Genomics **15** (1): 995.

Berger, B., Peng, J., and Singh, M. **2013**. "Computational Solutions for Omics Data." Nature Reviews Genetics **14** (5): 333–46.

Berlin, K., Koren, S., Chin, C., Drake, J. P., Landolin, J. M., and Phillippy. A. M. **2015**. "Assembling Large Genomes with Single-Molecule Sequencing and Locality-Sensitive Hashing." Nature Biotechnology **33** (6): 623–30.

Bikard, D., Patel, D., Le Mette, C., Giorgi, V., Camilleri, C., Bennett, M. J., and Loudet, O. **2009**. "Divergent Evolution of Duplicate Genes Leads to Genetic Incompatibilities Within *A. thaliana*." Science **323** (5914): 623–26.

Blanc, G., and Wolfe, K. H. **2004**. "Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes." The Plant Cell Online **16** (7): 1667–78.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. **2000**. "Extensive Duplication and Reshuffling in the *Arabidopsis* Genome." The Plant Cell **12** (7): 1093–1101.

Blum, M. J., Jun Bando, K., Katz, M., and Strong. D. R. **2007**. "Geographic Structure, Genetic Diversity and Source Tracking of *Spartina alterniflora*." Journal of Biogeography **34** (12): 2055–69.

Boland, E. J., Pillai, A., Odom, M. W., and Jagadeeswaran, P. **1994**. "Automation of the Maxam-Gilbert Chemical Sequencing Reactions." BioTechniques **16** (6): 1088–92, 1094–95.

Bond, J. E., Garrison, N. L., Hamilton, C. A., Godwin, R. L., Hedin, M. and Agnarsson, I. **2014**. "Phylogenomics Resolves a Spider Backbone Phylogeny and Rejects a Prevailing Paradigm for Orb Web Evolution." Current Biology **24** (15): 1765–71.

Bortolus, A. **2006**. "The Austral Cordgrass *Spartina densiflora* Brong.: Its Taxonomy, Biogeography and Natural History." Journal of Biogeography **33** (1): 158–68.

Bouchenak-Khelladi, Y., Salamin, N., Savolainen, V., Forest, F., Van der Bank, M., Chase, M. W., and Hodkinson, T. R. **2008**. "Large Multi-Gene Phylogenetic Trees of the Grasses (Poaceae): Progress towards Complete Tribal and Generic Level Sampling." Molecular Phylogenetics and Evolution **47** (2): 488–505.

Bourdau, P. **2012**. Polyploïdie et Variation de l'Expression Génique dans les Populations Naturelles de Spartines Envahissant les Marais Salés. Université de Rennes 1: Rapport de stage de Master 1 EFCE (Ecologie Fonctionnelle, Comportementale et Evolutive)

Boussau, B., Szollosi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. **2013**. "Genome-Scale Coestimation of Species and Gene Trees." Genome Research **23** (2): 323–30.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S. *et al.* **2013**. "Assemblathon 2: Evaluating *de novo* Methods of Genome Assembly in Three Vertebrate Species." GigaScience 2 (1): 10.

Brassac, J., and Blattner, F. R. **2015**. "Species Level Phylogeny and Polyploid Relationships in *Hordeum* (Poaceae) Inferred by Next-Generation Sequencing and In-Silico Cloning of Multiple Nuclear Loci." Systematic Biology, syv035.

Buggs, R. J. A., Elliott, N. M., Zhang, L., Koh, J., Viccini, L. F., Soltis, D. E., and Soltis, P. S. **2010**. "Tissue-Specific Silencing of Homoeologs in Natural Populations of the Recent Allopolyploid *Tragopogon mirus*." New Phytologist **186** (1): 175–83.

Buggs, R. J. A., Chamala, S., Wu, W., Tate, J. A., Schnable, P. S., Soltis, D. E., Soltis, P. S., and Barbazuk, W. B. **2012**. "Rapid, Repeated, and Clustered Loss of Duplicate Genes in Allopolyploid Plant Populations of Independent Origin." Current Biology **22** (3): 248–52.

C

Caboche, S., Audebert, C., Lemoine, Y., and Hot, D. **2014**. "Comparison of Mapping Algorithms Used in High-Throughput Sequencing: Application to Ion Torrent Data." BMC Genomics **15** (1): 264.

Cannarozzi, G., Plaza-Wüthrich, S., Esfeld, K., Larti, S., Wilson, Y., Girma, D., De Castro, E., *et al.* **2014**. "Genome and Transcriptome Sequencing Identifies Breeding Targets in the Orphan Crop Tef (*Eragrostis Tef*)." BMC Genomics **15** (1): 581.

Castillo, J. M., Ayres, D. R., Leira-Doce, P., Bailey, J., Blum, M., Strong, D. R., Luque, T., and Figueroa, E. **2010**. "The Production of Hybrids with High Ecological Amplitude between

Exotic *Spartina densiflora* and Native *S. maritima* in the Iberian Peninsula: New *Spartina* Hybrid in Iberian Peninsula." Diversity and Distributions **16** (4): 547–58.

Cavé-Radet, A. 2015. "Etude de la Tolérance des *Spartines* Polyploïdes aux Hydrocarbures Aromatiques Polycycliques (HAPs)." Université de Rennes 1: Rapport de stage de Master 2 STS, Spécialité BioVIGPA.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., Chiquet, J., *et al.* **2014**. "Early Allopolyploid Evolution in the Post-Neolithic *Brassica napus* Oilseed Genome." Science **345** (6199): 950–53.

Chang, P. L., Dilkes, B. P., McMahon, M., Comai, L, and Nuzhdin, S. V. **2010**. "Homoeolog-Specific Retention and Use in Allotetraploid *Arabidopsis suecica* Depends on Parent of Origin and Network Partners." Genome Biol **11** (12): R125.

Chapman, B. A., Bowers, J. E., Feltus, F. A., and Paterson, A. H. **2006**. "Buffering of Crucial Functions by Paleologous Duplicated Genes May Contribute Cyclicality to Angiosperm Genome Duplication." Proceedings of the National Academy of Sciences of the United States of America **103** (8): 2730–35.

Chaw, S., Chang, C., Chen, H., and Li, W. **2004**. "Dating the Monocot-Dicot Divergence and the Origin of Core Eudicots Using Whole Chloroplast Genomes." Journal of Molecular Evolution **58** (4): 424–41.

Check Hayden, E. **2012**. "Nanopore Genome Sequencer Makes Its Debut." Nature, February.

Check Hayden, E. **2014a**. "Data from Pocket-Sized Genome Sequencer Unveiled." Nature, February.

Check Hayden, E. **2014b**. "Technology: The $1,000 Genome." Nature **507** (7492): 294–95.

Chelaifa, H., Monnier, A., and Ainouche., M. L. **2010**. "Transcriptomic Changes Following Recent Natural Hybridization and Allopolyploidy in the Salt Marsh Species *Spartina x townsendii* and *Spartina anglica* (Poaceae)." New Phytologist **186** (1): 161–74.

Chelaifa, H., Mahé, F., and Ainouche, M. L. **2010**. "Transcriptome Divergence between the Hexaploid Salt-Marsh Sister Species *Spartina maritima* and *Spartina alterniflora* (Poaceae)" Molecular Ecology **19** (10): 2050–63

Chelaifa, H. **2010**. "Spéciation Allopolyploïde et Dynamique Fonctionnelle du Génome chez les Spartines." Rennes: Université de Rennes 1.

Chen, Y., and Huang, X. **2009**. "DNA Sequencing by Denaturation: Principle and Thermodynamic Simulations." Analytical Biochemistry **384** (1): 170–79.

Chen, Z. J. **2007**. "Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids." Annual Review of Plant Biology **58** (1): 377–406.

Chevreux, B., Wetter, T., and Suhai, S. **1999**. "Genome Sequence Assembly Using Trace Signals and Additional Sequence Information." In German Conference on Bioinformatics, 45–56.

Chikhi, R., and Rizk, G. **2012**. "Space-Efficient and Exact de Bruijn Graph Representation Based on a Bloom Filter." In Algorithms in Bioinformatics, edited by Ben Raphael and Jijun Tang, **7534**:236–48. Berlin, Heidelberg: Springer Berlin Heidelberg.

Chopra, R., Burow, G., Farmer, A., Mudge, J., Simpson, C. E., and Burow, M. D. **2014**. "Comparisons of *de novo* Transcriptome Assemblers in Diploid and Polyploid Species Using Peanut (*Arachis* Spp.) RNA-Seq Data." PloS One **9** (12): e115055.

Christin, P., Spriggs, E., Osborne, C. P., Stromberg, C. A. E., Salamin, N., and Edwards,E. J. **2014**. "Molecular Dating, Evolutionary Rates, and the Age of the Grasses." Systematic Biology **63** (2): 153–65.

Chu, H., Hsiao, W. W. L., Chen, J., Yeh, T., Tsai, M., Lin, H., Liu, Y., *et al*. **2013**. "EBARDenovo: Highly Accurate *de novo* Assembly of RNA-Seq with Efficient Chimera-Detection." Bioinformatics **29** (8): 1004–10.

Clarke, K., Yang, Y., Marsh, R., Xie, L., and Zhang, K. K. **2013**. "Comparative Analysis of *de novo* Transcriptome Assembly." Science China Life Sciences **56** (2): 156–62.

Clayton, W.D., Harman, K.T., and Williamson, H. **2008**. "The Online World Grass Flora." http://www.kew.org/data/grasses-db.html.

Clevenger, J. P., and Ozias-Akins,P. . **2015**. "SWEEP: A Tool for Filtering High-Quality SNPs in Polyploid Crops." G3; Genes|Genomes|Genetics **5** (9): 1797–1803.

Clevenger, J. P., Chavarro, C., Pearl, S. A., Ozias-Akins, P., and Jackson S. A. **2015**. "Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations." Molecular Plant **8** (6): 831–46.

Columbus, J. T., Cerros-Tlatilpa, R., Kinney, M. S., Siqueiros-Delgado, M. E., Bell, H. L., Griffith, M. P. and Refulio-Rodriguez, N. F. **2007**. "Phylogenetics of Chloridoideae (Gramineae): A Preliminary Study Based on Nuclear Ribosomal Internal Transcribed Spacer and Chloroplast trnL–F Sequences." Aliso: A Journal of Systematic and Evolutionary Botany **23** (1): 565–79.

Comai, L. **2000**. "Genetic and Epigenetic Interactions in Allopolyploid Plants." Plant Molecular Biology **43** (2-3): 387–99.

Comai, L. **2005**. "The Advantages and Disadvantages of Being Polyploid." Nature Reviews Genetics **6** (11): 836–46.

Combes, M.-C., Cenci, A., Baraille, H., Bertrand, B., and Lashermes. P., **2012**. "Homeologous Gene Expression in Response to Growing Temperature in a Recent Allopolyploid (*Coffea arabica L.*)." Journal of Heredity **103** (1): 36–46.

Combes, M.-C., Dereeper, A., Severac, D., Bertrand, B., and Lashermes, P. **2013**. "Contribution of Subgenomes to the Transcriptome and Their Intertwined Regulation in the Allopolyploid *Coffea arabica* Grown at Contrasted Temperatures." New Phytologist **200** (1): 251–60.

Combes, M.-C., Hueber, Y., Dereeper, A., Rialle, S., Herrera, J.-C., and Lashermes, P. **2015**. "Regulatory Divergence between Parental Alleles Determines Gene Expression Patterns in Hybrids." Genome Biology and Evolution **7** (4): 1110–21.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. **2005**. "Blast2GO: A Universal Tool for Annotation, Visualization and Analysis in Functional Genomics Research." Bioinformatics **21** (18): 3674–76.

Cosson, E., and de Maisonneuve, D. **1867**. "Introduction à La Flore d'Algérie. Phanérogamie. Groupe Des Glumacées (seu Descriptio Glumacearum in Algeria Nascentium).

Exploration Scientifique de l'Algérie, Publiée Par Ordre Du Gouvernement." Sciences Naturelles. Botanique. Imprimerie Impériale. Paris.

Coste, H. **1906**. "Flore de La France. T.III, Librairie Des Sciences Naturelles, Paris, 807p."

Cottet, M., de Montaudouin, X., Blanchet, H., and Lebleu. P. **2007**. "*Spartina anglica* Eradication Experiment and in Situ Monitoring Assess Structuring Strength of Habitat Complexity on Marine Macrofauna at High Tidal Level." Estuarine, Coastal and Shelf Science **71** (3-4): 629–40.

Crawford, N. G., Parham, J. F., Sellas, A. B., Faircloth, B. C., Glenn, T. C., Papenfuss, T. J., Henderson, J. B., Hansen, M. H., and Simison. W. B. **2015**. "A Phylogenomic Analysis of Turtles." Molecular Phylogenetics and Evolution **83** (February): 250–57.

Cronin, M. A., Rincon, G., Meredith, R. W., MacNeil, M. D., Islas-Trejo, A., Canovas, A., and Medrano, J. F. **2014**. "Molecular Phylogeny and SNP Variation of Polar Bears (*Ursus maritimus*), Brown Bears (*U. arctos*), and Black Bears (*U. americanus*) Derived from Genome Sequences." Journal of Heredity **105** (3): 312–23.

D

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., Noel, B., *et al.* **2012**. "The Banana (*Musa acuminata*) Genome and the Evolution of Monocotyledonous Plants." Nature **488** (7410): 213–17.

Daehler, C. C., and Strong, D. R. **1996**. "Status, Prediction and Prevention of Introduced Cordgrass *Spartina* Spp. Invasions in Pacific Estuaries, USA." Biological Conservation **78** (1-2): 51–58.

Daveau, J. **1897**. "La Flore Littorale Du Portugal." Boletim Da Sociedade Broteriana **14**: 4–54.

Delsuc, F., and Douzery, E. J. P. **2004**. "Les Méthodes Probabilistes en Phylogénie moléculaire:(2) L'approche Bayésienne." Biosystema **22**: 75–86.

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C., *et al.* **2014**. "The Coffee Genome Provides Insight into the Convergent Evolution of Caffeine Biosynthesis." Science **345** (6201): 1181–84.

Di Ventra, M. **2013**. "Fast DNA Sequencing by Electrical Means Inches Closer." Nanotechnology 24 (34): 342501.

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. **2007**. "SHARCGS, a Fast and Highly Accurate Short-Read Assembly Algorithm for *de novo* Genomic Sequencing." Genome Research **17** (11): 1697–1706.

Doyle, J. J., Doyle, J. L., Rauscher, J. T., and Brown, A. H. D. **2003**. "Diploid and Polyploid Reticulate Evolution throughout the History of the Perennial Soybeans (*Glycine* Subgenus *Glycine*): Research Review." New Phytologist **161** (1): 121–32.

Doyle, J. J., Flagel, L. E., Paterson, A. H., Rapp, R. A., Soltis, D. E., Soltis, P., S., and Wendel., J. F. **2008**. "Evolutionary Genetics of Genome Merger and Doubling in Plants." Annual Review of Genetics **42** (1): 443–61.

Doyle, J. J., and Egan A. N. **2010**. "Dating the Origins of Polyploidy Events." New Phytologist **186** (1): 73–85.

Dray, S. and Dufour, A.B. **2007**. "The ade4 package: implementing the duality diagram for ecologists." Journal of Statistical Software **22**(4): 1-20

Drummond, A.J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Heled, J., Kearse, M., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T., Wilson, A., **2010**. Geneious v5.1, http://www.geneious.com.

Duan, J., Xia, C., Zhao, G., Jia, J., and Kong, X. **2012**. "Optimizing *de novo* Common Wheat Transcriptome Assembly Using Short-Read RNA-Seq Data." BMC Genomics **13** (1): 392.

Duchemin, W., Dupont, P.-Y., Campbell, M. A., Ganley, A. R. D., and Cox, M. P. **2014**. "HyLiTE: Accurate and Flexible Analysis of Gene Expression in Hybrid and Allopolyploid Species." BMC Bioinformatics **16** (1).

Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. **2014**. "Recent Progress and Challenges in Population Genetics of Polyploid Organisms: An Overview of Current State-of-the-Art Molecular and Statistical Tools." Molecular Ecology **23** (1): 40–69.

Dunn, C. W., Howison, M., and Zapata, F. **2013**. "Agalma: An Automated Phylogenomics Workflow." BMC Bioinformatics **14** (1): 330.

E

Eaton, D. A. R., and Ree, R. H. **2013**. "Inferring Phylogeny and Introgression Using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae)." Systematic Biology **62** (5): 689–706.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., *et al.* **2009**. "Real-Time DNA Sequencing from Single Polymerase Molecules." Science **323** (5910): 133–38.

El-Metwally, S, Ouda, O. M., and Helmy, M. **2014**. Next Generation Sequencing Technologies and Challenges in Sequence Assembly. Springer Briefs in Systems Biology. New York, NY: Springer.

El-Twab, M. H. A. **2007**. "Physical Mapping of the 45S rDNA on the Chromosomes of *Triticum turgidum* and *T. aestivum* using Fluorescence *in Situ* Hybridization for chromosome ancestors." Arab J. Biotechnol., **10**: 69-80.

Étienne, J., and Millot, F. **1998**. Biochimie Génétique, Biologie Moléculaire. Paris; Milan; Barcelone: Masson.

F

Fabre, ME. **1849**. "Description d'une Nouvelle Espèce de *Spartina*, Abondante sur une Portion du Littoral Méditerranéen." Annales Des Sciences Naturelles. Botanique (Paris) **3**: 122–25.

Feliner, G., and Rosselló, J. A. **2012**. “Concerted Evolution of Multigene Families and Homoeologous Recombination.” In Plant Genome Diversity Volume 1, edited by Jonathan F. Wendel, Johann Greilhuber, Jaroslav Dolezel, and Ilia J. Leitch, 171–93. Springer Vienna.

Felsenstein, J. **1978**. “Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading.” Systematic Zoology **27** (4): 401.

Felsenstein, J. **1985**. “Confidence Limits on Phylogenies: An Approach Using the Bootstrap.” Evolution **39** (4): 783–91.

Ferreira de Carvalho, J., Poulain, J., Da Silva, C., Wincker, P., Michon-Coudouel, S., Dheilly, A., Naquin, D., Boutte, J., Salmon, A., and Ainouche. M. L. **2012**. “Transcriptome *de novo* Assembly from next-Generation Sequencing and Comparative Analyses in the Hexaploid Salt Marsh Species *Spartina maritima* and *Spartina alterniflora* (Poaceae).” Heredity **110** (2): 181–93.

Ferreira de Carvalho, J., Chelaifa, H., Boutte, J., Poulain, J., Couloux, A., Wincker, P., Bellec, A., *et al.* **2013**. “Exploring the Genome of the Salt-Marsh *Spartina maritima* (Poaceae, Chloridoideae) through BAC End Sequence Analysis.” Plant Molecular Biology **83** (6): 591–606.

Ferreira de Carvalho, J. **2013**. “Evolution du Génome des *Spartines* Polyploïdes Envahissant les Marais Salés : Apport des Nouvelles Techniques de Séquençage Haut-Débit.” Rennes: Université de Rennes 1.

Ferris, C., King, R. A., and Gray, A. J. **1997**. “Molecular Evidence for the Maternal Parentage in the Hybrid Origin of *Spartina anglica* C.E. Hubbard.” Molecular Ecology **6** (2): 185–87.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., *et al.* **2014**. “Pfam: The Protein Families Database.” Nucleic Acids Research **42** (D1): D222–30.

Flagel, L. E., and Wendel. J. F., **2009**. “Evolutionary Rate Variation, Genomic Dominance and Duplicate Gene Expression Evolution during Allotetraploid Cotton Speciation.” New Phytologist **186** (1): 184–93.

Flagel, L. E, Udall, J., Nettleton, D., and Wendel, J. F. **2008**. “Duplicate Gene Expression in Allopolyploid *Gossypium* Reveals Two Temporally Distinct Phases of Expression Evolution.” BMC Biology **6** (1): 16.

Flagel, L. E., Chen, L., Chaudhary, B., and Wendel, J. F. **2009**. “Coordinated and Fine-Scale Control of Homoeologous Gene Expression in Allotetraploid Cotton.” Journal of Heredity **100** (4): 487–90.

Flagel, L. E., J. F. Wendel, and J. A. Udall. **2012**. “Duplicate Gene Evolution, Homoeologous Recombination, and Transcriptome Characterization in Allopolyploid Cotton.” BMC Genomics 13 (1): 302.

Flicek, P., and Birney, E. **2009**. “Sense from Sequence Reads: Methods for Alignment and Assembly.” Nature Methods **6** (11s): S6–12.

Fonseca, N. A., Rung, J., Brazma, A., and Marioni, J. C. **2012**. “Tools for Mapping High-Throughput Sequencing Data.” Bioinformatics **28** (24): 3169–77.

Fortune, P. M., Schierenbeck, K. A., Ainouche, A. K., Jacquemin, J., Wendel, J. F., and Ainouche,M. L. **2007**. "Evolutionary Dynamics of Waxy and the Origin of Hexaploid *Spartina* Species (Poaceae)." Molecular Phylogenetics and Evolution **43** (3): 1040–55.

Fortune, P. M., Schierenbeck, K., Ayres, D., Bortolus, A., Catrice, O., Brown, S., and Ainouche, M. L. **2008**. "The enigmatic invasive *Spartina densiflora* : a history of hybridizations in a polyploidy context." Molecular Ecology **17** (19): 4304–16.

Fortuné, P. **2007**. "Phylogénie et Dynamique des Gènes Dupliqués Chez les Plantes Polyploïdes : Evolution dans les Genres *Bromus* L. et *Spartina* Schreb. (Poaceae)." Rennes: Université de Rennes 1.

G

Gabaldón, T., and Koonin, E. V. **2013**. "Functional and Evolutionary Implications of Gene Orthology." Nature Reviews Genetics **14** (5): 360–66.

Gaeta, R. T., Pires, J. C., Iniguez-Luy, F., Leon, E., and Osborn, T. C. **2007**. "Genomic Changes in Resynthesized *Brassica napus* and their Effect on Gene Expression and Phenotype." The Plant Cell Online **19** (11): 3403–17.

Gaeta, R., T., and Pires, J., C. **2010**. "Homoeologous Recombination in Allopolyploids: The Polyploid Ratchet: Research Review." New Phytologist **186** (1): 18–28.

Gallardo, M. H., Bickham, J. W., Honeycutt, R. L., Ojeda, R. A., and Köhler, N. **1999**. "Discovery of Tetraploidy in a Mammal." Nature **401** (6751): 341–341.

Gallardo, M. H., Kausel, G., Jiménez, A., Bacquet, C., González, C., Figueroa, J., Köhler, N., and Ojeda,R. **2004**. "Whole-Genome Duplications in South American Desert Rodents (Octodontidae)." Biological Journal of the Linnean Society **82** (4): 443–51.

Garg, R., Patel, R. K., Tyagi, A. K., and Jain, M. **2011**. "*De novo* Assembly of Chickpea Transcriptome Using Short Reads for Gene Discovery and Marker Identification." DNA Research **18** (1): 53–63.

Gaut, B. S., Morton, B. R., McCaig, B. C., and Clegg, M. T. **1996**. "Substitution Rate Comparisons between Grasses and Palms: Synonymous Rate Differences at the Nuclear Gene Adh Parallel Rate Differences at the Plastid Gene rbcL." Proceedings of the National Academy of Sciences **93** (19): 10274–79.

Gedye, K., Gonzalez-Hernandez, J., Ban, Y., Ge, X., Thimmapuram, J., Sun, F., Wright C., Ali, S., Boe, A., and Owens, V. **2010**. "Investigation of the Transcriptome of Prairie Cord Grass, a New Cellulosic Biomass Crop." The Plant Genome Journal **3** (2): 69.

Gharib, K. **2012**. "Diversité Génétique et Origine des Populations de *Spartina alterniflora* Introduites en Europe." Université de Rennes 1: Rapport de stage de Master 2 STS, Spécialité BioVIGPA.

Giraud, D. **2015**. "Détection de Copies Dupliquées Chez les Spartines Polyploïdes à Partir de Données NGS et de l'Outil de Détection d'Haplotypes « IlluHaplotyper »." Université de Rennes 1: Rapport de stage de Master 1 BIG (Bio-Informatique et Génomique).

Goldberg, S. M. D., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., Halpern, A., *et al.* **2006**. "A Sanger/pyrosequencing Hybrid Approach for the Generation of High-Quality Draft Assemblies of Marine Microbial Genomes." Proceedings of the National Academy of Sciences **103** (30): 11240–45.

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M., and McCombie, W. R. **2015**. "Oxford Nanopore Sequencing and *de novo* Assembly of a Eukaryotic Genome." BMC Genomics **16**:327.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talón, M., Dopazo, J., and Conesa, A. **2008**. "High-Throughput Functional Annotation and Data Mining with the Blast2GO Suite." Nucleic Acids Research **36** (10): 3420–35.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., and Zeng, Q. **2011**. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." Nature Biotechnology **29** (7): 644–52.

Grass Phylogeny Working Group II. **2012**. "New Grass Phylogeny Resolves Deep Evolutionary Relationships and Discovers C4 Origins." New Phytologist **193** (2): 304–12.

Gray, A.J., Marshall, D.F., and RaybouldA.F. **1991**. "A Century of Evolution in *Spartina anglica*." In Advances in Ecological Research, **21**:1–62. Elsevier.

Gregory, T. R., and Mable, B. K. **2005**. "Polyploidy in Animals." In the Evolution of the Genome, 428–517. London: Elsevier Academic Press.

Griffin, P. C., Robin, C., and Hoffmann, A. A. **2011**. "A next-Generation Sequencing Method for Overcoming the Multiple Gene Copy Problem in Polyploid Phylogenetics, Applied to Poa Grasses." BMC Biology **9** (1): 19.

Grover, C. E., Salmon, A., and Wendel, J. F. **2012**. "Targeted Sequence Capture as a Powerful Tool for Evolutionary analysis1." American Journal of Botany **99** (2): 312–19.

Grover, C. E., Gallagher, J. P., Szadkowski, E. P., Yoo, M. J., Flagel, L. E., and Wendel, J. F. **2012**. "Homoeolog Expression Bias and Expression Level Dominance in Allopolyploids." New Phytologist **196** (4): 966–71.

Grover, C. E., Gallagher, J. P., Jareczek, J. J., Page, J. T., Udall, J. A., Gore, M. A., and Wendel. J. F. **2015**. "Re-Evaluating the Phylogeny of Allopolyploid *Gossypium* L." Molecular Phylogenetics and Evolution ***92*** (November): 45–52.

H

Ha, M., Kim, E.-D., and Chen. Z. J. **2009**. "Duplicate Genes Increase Expression Diversity in Closely Related Species and Allopolyploids." Proceedings of the National Academy of Sciences **106** (7): 2295–2300.

Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M. , *et al.* **2008**. "Single-Molecule DNA Sequencing of a Viral Genome." Science **320** (5872): 106–9.

Hatem, A., Bozdağ, D., Toland, A. E., and Çatalyürek. Ü. V. **2013**. "Benchmarking Short Sequence Mapping Tools." BMC Bioinformatics **14** (1): 184.

Hernandez, D., Francois, P., Farinelli, L., Osteras, M., and Schrenzel. J., **2008**. "*De novo* Bacterial Genome Sequencing: Millions of Very Short Reads Assembled on a Desktop Computer." Genome Research **18** (5): 802–9.

Higgins, J., Magusin, A., Trick, M., Fraser, F., and Bancroft, I. **2012**. "Use of mRNA-Seq to Discriminate Contributions to the Transcriptome from the Constituent Genomes of the Polyploid Crop Species *Brassica napus*." BMC Genomics **13** (1): 247.

Hilu, KW, and Alice, L. A. **2001**. "A Phylogeny of *Chloridoideae* (Poaceae) Based on matK Sequences." Systematic Botany **26**: 386–405.

Hittinger, C. T., and Carroll, S. B. **2007**. "Gene Duplication and the Adaptive Evolution of a Classic Genetic Switch." Nature **449** (7163): 677–81.

Hodcroft, E. 2015. TreeCollapseCL 4. University of Edinburgh. http;//emmahodcroft.com/TreeColapseCL.html.

Hossain, M., Azimi, N., and Skiena, S. **2009**. "Crystallizing Short-Read Assemblies around Seeds." BMC Bioinformatics **10** (Suppl 1): S16.

Hounsome, G. 2013. "*Spartina patens* in West Sussex, v.c.13," BSBI News, **132**: 66–67.

Huang, X., and Yang, S.-P. **2005**. "Generating a Genome Assembly with PCAP." In Current Protocols in Bioinformatics, edited by Andreas D. Baxevanis, Daniel B. Davison, Roderic D.M. Page, Gregory A. Petsko, Lincoln D. Stein, and Gary D. Stormo. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Huson, D. H., and Scornavacca, C. **2012**. "Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks." Systematic Biology **61** (6): 1061–67.

I

Idury, R. M., and Waterman, M. S. **1995**. "A New Algorithm for DNA Sequence Assembly." Journal of Computational Biology: A Journal of Computational Molecular Cell Biology **2** (2): 291–306.

Ilut, D. C., Coate, J. E., Luciano, A. K., Owens, T. G., May, G. D., Farmer, A., and Doyle, J. J. **2012**. "A Comparative Transcriptomic Study of an Allotetraploid and Its Diploid Progenitors Illustrates the Unique Advantages and Challenges of RNA-Seq in Plant Species." American Journal of Botany **99** (2): 383–96.

IUCN. **2000**. "World's Worst Invasive Alien Species." The World Conservation Union.

J

Jacobs, S.W. **1987**. “Systematics of the Chloridoid Grasses.” In Grass Systematics and Evolution, T. R. Soderstrom, K. W. Hilu, C. S. Campbell, and M. E. Barkworth, 277–86. Washington, D.C., USA.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., *et al.* **2007**. “The Grapevine Genome Sequence Suggests Ancestral Hexaploidization in Major Angiosperm Phyla.” Nature **449** (7161): 463–67.

Jeanmonod, D., and Burdet, H. M. **1989**. “Notes et Contributions à La Flore de Corse IV. Candollea” **44**: 337–401.

Jeck, W. R., Reinhardt, J. A., Baltrus, D. A., Hickenbotham, M. T., Magrini, V., Mardis, E. R., Dangl, J. L., and Jones, C. D. **2007**. “Extending Assembly of Short DNA Sequences to Handle Error.” Bioinformatics **23** (21): 2942–44.

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., Appels, R., *et al.* **2013**. “*Aegilops tauschii* Draft Genome Sequence Reveals a Gene Repertoire for Wheat Adaptation.” Nature **496** (7443): 91–95.

Jiang, Y., Lu, J., Peatman, E., Kucuktas, H., Liu, S., Wang, S., Sun, F., and Liu, Z. **2011**. “A Pilot Study for Channel Catfish Whole Genome Sequencing and *de novo* Assembly.” BMC Genomics **12** (1): 629.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., *et al.* **2011**. “Ancestral Polyploidy in Seed Plants and Angiosperms.” Nature **473** (7345): 97–100.

Jukes, T.H., and Cantor, C.R. **1969**. “Evolution of Protein Molecules. In Mammalian Protein Metabolism (Vol. 3).” ed., Academic Press. Munro, H.N.

K

Kasahara, M., and Morishita, S. **2006**. Large-Scale Genome Sequence Processing. London : Singapore ; Hackensack, NJ: Imperial College Press ; Distributed by World Scientific.

Katoh, K, and Toh, H. **2010**. “Parallelization of the MAFFT Multiple Sequence Alignment Program.” Bioinformatics **26** (15): 1899–1900.

Kent, W. J. **2002**. “BLAT—the BLAST-like Alignment Tool.” Genome Research **12** (4): 656–64.

Kerpedjiev, P., Frellsen, J., Lindgreen, S., and Krogh, A. **2014**. “Adaptable Probabilistic Mapping of Short Reads Using Position Specific Scoring Matrices.” BMC Bioinformatics **15** (1): 100.

Kim, C., Tang, H., and Paterson, A. H. **2009**. “Duplication and Divergence of Grass Genomes: Integrating the Chloridoids.” Tropical Plant Biology **2** (1): 51–62.

Kim, S., Rayburn, A. L. and Lee, D. K. **2010**. “Genome Size and Chromosome Analyses in Prairie Cordgrass.” Crop Science **50** (6): 2277.

Kimura, M. **1980**. “A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences.” Journal of Molecular Evolution **16** (2): 111–20.

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., Weinstock, G. M., Wilson, R. K., and Ding, L. **2009**. “VarScan: Variant Detection in Massively Parallel Sequencing of Individual and Pooled Samples.” Bioinformatics **25** (17): 2283–85.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. **2012**. “VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing.” Genome Research **22** (3): 568–76.

Koren, S., and Phillippy, A. M. **2015**. “One Chromosome, One Contig: Complete Microbial Genomes from Long-Read Sequencing and Assembly.” Current Opinion in Microbiology **23** (February): 110–20.

Korlach, J., Bjornson, K. P., Chaudhuri, B. P., Cicero, R. L., Flusberg, B. A., Gray, J. J., Holden, D., Saxena,R., Wegener, J., and Turner, S. W. **2010**. “Real-Time DNA Sequencing from Single Polymerase Molecules.” In Methods in Enzymology, **472**:431–55. Elsevier.

Kovarik, A. **2005**. “Rapid Concerted Evolution of Nuclear Ribosomal DNA in Two Tragopogon Allopolyploids of Recent and Recurrent Origin.” Genetics **169** (2): 931–44.

Kovarik, A, Dadejova, M., Lim, Y. K., Chase, M. W., Clarkson, J. J., Knapp, S., and Leitch, A. R. **2008**. “Evolution of rDNA in Nicotiana Allopolyploids: A Potential Link between rDNA Homogenization and Epigenetics.” Annals of Botany **101** (6): 815–23.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. **2009**. “Circos: An Information Aesthetic for Comparative Genomics.” Genome Research **19** (9): 1639–45.

Kumar, S., Banks, T. W., and Cloutier, S. **2012**. “SNP Discovery through Next-Generation Sequencing and Its Applications.” International Journal of Plant Genomics **2012**: 1–15.

L

Lagercrantz, U., and Lydiate, D.J. **1996**. “Comparative Genome Mapping in *Brassica*.” Genetics **144** (4): 1903–10.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., *et al.* **2001**. “Initial Sequencing and Analysis of the Human Genome.” Nature **409** (6822): 860–921.

Langham, R. J., Walsh, J., Dunn, M., Ko, C., Goff, S. A., and Freeling, M. **2004**. “Genomic Duplication, Fractionation and the Origin of Regulatory Novelty.” Genetics **166** (2): 935–45.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. **2009**. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” Genome Biol **10** (3): R25.

Langmead, B., and Salzberg, S. L. **2012**. “Fast Gapped-Read Alignment with Bowtie 2.” Nature Methods **9** (4): 357–59.

Larget, B. R., Kotha, S. K., Dewey, C. N., and Ane, C. **2010**. “BUCKy: Gene Tree/species Tree Reconciliation with Bayesian Concordance Analysis.” Bioinformatics **26** (22): 2910–11.

Laszlo, A. H., Derrington, I. M., Ross, B. C., Brinkerhoff, H., Adey, A., Nova, I. C., Craig, J. M., *et al.* **2014**. “Decoding Long Nanopore Sequencing Reads of Natural DNA.” Nature Biotechnology **32** (8): 829–33.

Lee, R. W. **2003**. “Physiological Adaptations of the Invasive Cordgrass *Spartina* a*nglica* to Reducing Sediments: Rhizome Metabolic Gas Fluxes and Enhanced O 2 and H 2 S Transport.” Marine Biology **143** (1): 9–15.

Leitch, I. J., and Bennett, M. D. **2004**. “Genome Downsizing in Polyploid Plants.” Biological Journal of the Linnean Society **82** (4): 651–63.

Lewis, W. H. **1980**. “Polyploidy in Angiosperms: Dicotyledons.” In Polyploidy, edited by W. H. Lewis, **13**:241–68. Basic Life Sciences. Springer US.

Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., Li, Q., *et al.* **2014**. “Genome Sequence of the Cultivated Cotton *Gossypium* Arboreum.” Nature Genetics **46** (6): 567–72.

Li, H., and Durbin, R. **2009**. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” Bioinformatics **25** (14): 1754–60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. **2009**. “The Sequence Alignment/Map Format and SAMtools.” Bioinformatics **25** (16): 2078–79.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. **2009**. “SOAP2: An Improved Ultrafast Tool for Short Read Alignment.” Bioinformatics **25** (15): 1966–67.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., *et al.* **2010**. “*De novo* Assembly of Human Genomes with Massively Parallel Short Read Sequencing.” Genome Research **20** (2): 265–72.

Li, R., Li, Y., Kristiansen, K., and Wang, J. **2008**. “SOAP: Short Oligonucleotide Alignment Program.” Bioinformatics **24** (5): 713–14.

Li, W. H., Wu, C. I., and Luo, C. C. **1985**. “A New Method for Estimating Synonymous and Nonsynonymous Rates of Nucleotide Substitution Considering the Relative Likelihood of Nucleotide and Codon Changes.” Molecular Biology and Evolution **2** (2): 150–74.

Lim, K. Y., Matyášek, R., Lichtenstein, C. P., and Leitch., A. R. **2000**. “Molecular Cytogenetic Analyses and Phylogenetic Studies in the *Nicotiana* Section Tomentosae.” Chromosoma **109** (4): 245–58.

Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C. J., *et al.* **2011**. “Comparative Studies of *de novo* Assembly Tools for next-Generation Sequencing Technologies.” Bioinformatics **27** (15): 2031–37.

Liu, B., and Davis, T. M. **2011**. “Conservation and Loss of Ribosomal RNA Gene Sites in Diploid and Polyploid *Fragaria* (Rosaceae).” BMC Plant Biology **11** (1): 157.

Liu, C.-M., Wong, T., Wu, E., Luo, R., Yiu, S.-M., Li, Y., Wang, B., *et al.* **2012**. "SOAP3: Ultra-Fast GPU-Based Parallel Alignment Tool for Short Reads." Bioinformatics **28** (6): 878–79.

Liu, L., Yu, L., Pearl, D. K., and Edwards, S. V. . **2009**. "Estimating Species Phylogenies Using Coalescence Times among Sequences." Systematic Biology **58** (5): 468–77.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. **2012**. "Comparison of Next-Generation Sequencing Systems." Journal of Biomedicine and Biotechnology **2012**: 1–11.

Liu, S., Li, W., Wu, Y., Chen, C., and Lei, J. **2013**. "*De novo* Transcriptome Assembly in Chili Pepper (*Capsicum frutescens*) to Identify Genes Involved in the Biosynthesis of Capsaicinoids." Edited by Christian Schönbach. PLoS ONE **8** (1): e48156.

Lloyd, A. H., Ranoux, M., Vautrin, S., Glover, N., Fourment, J., Charif, D., Choulet, F., *et al.* **2014**. "Meiotic Gene Evolution: Can You Teach a New Dog New Tricks?" Molecular Biology and Evolution **31** (7): 1724–27.

Luchetta, P. **2005**. Évolution Moléculaire: Cours et Questions de Révision. Paris: Dunod.

Luo, R., Wong, T., Zhu, J., Liu, C.-M., Zhu, X., Wu, E., Lee, L.-K., *et al.* **2013**. "SOAP3-Dp: Fast, Accurate and Sensitive GPU-Based Short Read Aligner." Edited by Frederick C. C. Leung. PLoS ONE **8** (5): e65632.

Lynch, M., and Force, A. **2000**. "The Probability of Duplicate Gene Preservation by Subfunctionalization." Genetics **154** (1): 459–73.

M

Ma, L., Chen, C., Liu, X., Jiao, Y., Su, N., Li, L., Wang, X., *et al.* **2005**. "A Microarray Analysis of the Rice Transcriptome and Its Comparison to *Arabidopsis*." Genome Research **15** (9): 1274–83.

Ma, P.-F., Zhang, Y.-X., Zeng, C.-X., Guo, Z.-H., and Li, D.-Z., **2014**. "Chloroplast Phylogenomic Analyses Resolve Deep-Level Relationships of an Intractable Bamboo Tribe *Arundinarieae* (Poaceae)." Systematic Biology **63** (6): 933–50.

Mable, B. K. **2004**. "'Why Polyploidy is Rarer in Animals than in Plants': Myths and Mechanisms." Biological Journal of the Linnean Society **82** (4): 453–66.

Mable, B. K., Alexandrou, M. A., and Taylor. M. I., **2011**. "Genome Duplication in Amphibians and Fish: An Extended Synthesis: Polyploidy in Amphibians and Fish." Journal of Zoology **284** (3): 151–82.

Madlung, A. **2002**. "Remodeling of DNA Methylation and Phenotypic and Transcriptional Changes in Synthetic Arabidopsis Allotetraploids." Plant Physiology **129** (2): 733–46.

Malinska, H., Tate, J. A., Mavrodiev, E., Matyasek, R., Lim, K. Y., Leitch, A. R., Soltis, D. E., Soltis, P. S., and Kovarik. A., **2011**. "Ribosomal RNA Genes Evolution in *Tragopogon*: A Story of New and Old World Allotetraploids and the Synthetic Lines." Taxon **60** (2): 348.

Mallet, J. **2007**. “Hybrid Speciation.” Nature **446** (7133): 279–83.

Marchant, C. J. **1968a**. “Evolution in *Spartina* (Graminae). III Species Chromosome Numbers and Their Taxonomic Significance.” Botanical Journal of the Linnean Society **60**: 411–17.

Marchant, C. J. **1968b**. “Evolution in *Spartina* (Gramineae). II. Chromosomes, Basic Relationships and the Problem of *Spartina x townsendii* Agg.” Botanical Journal of the Linnean Society **60**: 381–409.

Mares, M. A., Braun, J. K., Barquez, R. M., and Díaz, M. M . **2000**. “Two New Genera and Species of Halophytic Desert Mammals from Isolated Salt Flats in Argentina.” Occasional Papers of the Museum of Texas Technical University **203**: 1–27.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., and Chen, Z. **2005**. “Genome Sequencing in Microfabricated High-Density Picolitre Reactors.” Nature **437** (7057): 376–80.

Marhold, K., Kudoh, H., Pak, J.-H., Watanabe, K., Spaniel, S., and Lihova J. **2009**. “Cytotype Diversity and Genome Size Variation in Eastern Asian Polyploid *Cardamine* (Brassicaceae) Species.” Annals of Botany **105** (2): 249–64.

Maricle, B. R., and Lee R. W. **2002**. “Aerenchyma Development and Oxygen Transport in the Estuarine Cordgrasses *Spartina alterniflora* and *S. anglica*.” Aquatic Botany **74** (2): 109–20.

Mathews, S., Tsai, R. C., and Kellogg, E. A. **2000**. “Phylogenetic Structure in the Grass Family (Poaceae): Evidence from the Nuclear Gene Phytochrome B.” American Journal of Botany **87** (1): 96–107.

Maxam, A. M., and Gilbert, W. **1977**. “A New Method for Sequencing DNA.” Proceedings of the National Academy of Sciences of the United States of America **74** (2): 560–64.

McKain, M. R., Wickett, N., Zhang, Y., Ayyampalayam, S., McCombie, W. R., Chase, M. WPires, ., J. C., dePamphilis, C. W. and Leebens-Mack J . **2012**. “Phylogenomic Analysis of Transcriptome Data Elucidates Co-Occurrence of a Paleopolyploid Event and the Origin of Bimodal Karyotypes in *Agavoideae* (Asparagaceae).” American Journal of Botany **99** (2): 397–406.

McLysaght, A., Hokamp, K., and Wolfe, K. H. **2002**. “Extensive Genomic Duplication during Early Chordate Evolution.” Nature Genetics **31** (2): 200–204.

Metzker, M. L. **2009**. “Sequencing Technologies—the next Generation.” Nature Reviews Genetics **11** (1): 31–46.

Michael, T. P., and Jackson, S. **2013**. “The First 50 Plant Genomes.” The Plant Genome **6** (2).

Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. **2008**. “Aggressive Assembly of Pyrosequencing Reads with Mates.” Bioinformatics **24** (24): 2818–24.

Miller, J. R., Koren, S., and SuttonG. **2010**. “Assembly Algorithms for next-Generation Sequencing Data.” Genomics **95** (6): 315–27.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., and Marshall, D. **2009**. "Tablet--next Generation Sequence Assembly Visualization." Bioinformatics **26** (3): 401–2.

Mirarab, SReaz, ., R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. **2014**. "ASTRAL: Genome-Scale Coalescent-Based Species Tree Estimation." Bioinformatics **30** (17): i541–48.

Mobberley, D. G. **1956**. "Taxonomy and Distribution of the Genus *Spartina*." In , Iowa State Coll J Sci, **30**:471–574.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A, *et al.* **2000**. "A Whole-Genome Assembly of Drosophila." Science (New York, N.Y.) **287** (5461): 2196–2204.

N

Namour, Y. **2015**. "Intégration sous Galaxy de «PyroHaplotyper», Logiciel de Détection d'Haplotypes à Partir de Données NGS (Next-Generation Sequencing)." Université de Rennes 1: Rapport de stage de Master 1 BIG (Bio-Informatique et Génomique).

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): 443–53.

Nicolas, S. D., Mignon, G. L., Eber, F., Coriton, O., Monod, H., Clouet, V., Huteau, V. *et al.* **2007**. "Homeologous Recombination Plays a Major Role in Chromosome Rearrangements that Occur During Meiosis of *Brassica napus* Haploids." Genetics *175* (2): 487–503.

Novaes, E., Drost, D. R., Farmerie, W. G., Pappas, G. J., Grattapaglia, D., Sederoff, R. R., and Kirst, M. **2008**. "High-Throughput Gene and SNP Discovery in Eucalyptus Grandis, an Uncharacterized Genome." BMC Genomics **9** (1): 312.

Novák, P., Neumann, P., and Macas, J. **2010**. "Graph-Based Clustering and Characterization of Repetitive Sequences in next-Generation Sequencing Data." BMC Bioinformatics **11** (1): 378.

O

Ohno, S. **1970**. "Evolution by Gene Duplication. Soukup, S." Soukup, S. W.Springer-Verlag, New York. **9**: 250–51.

Ohno, S. **1974**. "Evolution by Gene Duplication." Soukup, S. W. Springer-Verlag, New York.

Oliphant, A., Barker, D. L., Stuelpnagel, J. R., and Chee, M. S. **2002**. "BeadArray Technology: Enabling an Accurate, Cost-Effective Approach to High-Throughput Genotyping." Biotechniques **32** (6): 56–58.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., *et al.* **2007**. "The TIGR Rice Genome Annotation Resource: Improvements and New Features." Nucleic Acids Research **35** (Database issue): D883–87.

P

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., and Trajanoski, Z. **2013**. "A Survey of Tools for Variant Analysis of next-Generation Genome Sequencing Data." Briefings in Bioinformatics **15** (2): 256–78.

Page, J. T., Gingle, A. R., and Udall, J. A. **2013**. "PolyCat: A Resource for Genome Categorization of Sequencing Reads From Allopolyploid Organisms." G3; Genes|Genomes|Genetics **3** (3): 517–25.

Page, J. T., Huynh, M. D., Liechty, Z. S., Grupp, K., Stelly, D., Hulse, A. M., Ashrafi, H., Van Deynze, A., Wendel, J. F., and Udall, J. A. **2013**. "Insights into the Evolution of Cotton Diploids and Polyploids from Whole-Genome Re-Sequencing." G3; Genes|Genomes|Genetics **3** (10): 1809–18.

Page, J. T., Liechty, Z. S., Huynh, M. D., and Udall, J. A. **2014**. "BamBam: Genome Sequence Analysis Tools for Biologists." BMC Research Notes **7** (1): 829.

Page, J. T., and Udall, J.A. **2015**. "Methods for Mapping and Categorization of DNA Sequence Reads from Allopolyploid Organisms." BMC Genetics **16** (Suppl 2): S4.

Panahiazar, M., Sheth, A. P., Ranabahu, A., Vos, R. A., and Leebens-Mack, J. **2013**. "Advancing Data Reuse in Phyloinformatics Using an Ontology-Driven Semantic Web Approach." BMC Medical Genomics **6** (Suppl 3): S5.

Pareek, C. S., Smoczynski, R., and Tretyn, A. **2011**. "Sequencing Technologies and Genome Sequencing." Journal of Applied Genetics **52** (4): 413–35.

Parisod, C., Salmon, A., Zerjal, T., Tenaillon, M., Grandbastien, M.-A., and Ainouche, M. L. **2009**. "Rapid Structural and Epigenetic Reorganization near Transposable Elements in Hybrid and Allopolyploid Genomes in Spartina." New Phytologist **184** (4): 1003–15.

Parlatore, F. **1850**. "Flora Italiana. I. Tipografia Le Monnier. Firenze."

Paterson, A. H., Bowers, J. E. , and Chapman, B. A. **2004**. "Ancient Polyploidization Predating Divergence of the Cereals, and its Consequences for Comparative Genomics." Proceedings of the National Academy of Sciences of the United States of America **101** (26): 9903–8.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., *et al.* **2009**. "The *Sorghum bicolor* Genome and the Diversification of Grasses." Nature **457** (7229): 551–56.

Paterson, A. H., Wendel, J. F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., Llewellyn, D., *et al.* **2012**. "Repeated Polyploidization of *Gossypium* Genomes and the Evolution of Spinnable Cotton Fibres." Nature **492** (7429): 423–27.

Paterson, A. H., Wang, X., Li, J., and Tang, H. **2012**. "Ancient and Recent Polyploidy in Monocots." In Polyploidy and Genome Evolution, edited by Pamela S. Soltis and Douglas E. Soltis, 93–108. Berlin, Heidelberg: Springer Berlin Heidelberg.

Peralta, M., Combes, M.-C., Cenci, A., Lashermes, P., and Dereeper, A. **2013**. "SNiPloid: A Utility to Exploit High-Throughput SNP Data Derived from RNA-Seq in Allopolyploid Species." International Journal of Plant Genomics **2013**: 1–6.

Perkins, E. J, Streever, W. J., Davis, E., and Fredrickson, H. E. **2002**. "Development of Amplified Fragment Length Polymorphism Markers for *Spartina alterniflora*." Aquatic Botany **74** (1): 85–95.

Peterson, P. M., Romaschenko, K., and Johnson, G. **2010a**. "A Phylogeny and Classification of the Muhlenbergiinae (Poaceae: *Chloridoideae*: *Cynodonteae*) Based on Plastid and Nuclear DNA Sequences." American Journal of Botany **97** (9): 1532–54.

Peterson, P. M., Romaschenko, K., Johnson, G. **2010b**. "A Classification of the *Chloridoideae* (Poaceae) Based on Multi-Gene Phylogenetic Trees." Molecular Phylogenetics and Evolution **55** (2): 580–98.

Peterson, P. M., Romaschenko, K., Herrera Arrieta, Y., and Saarela, J. M. **2014a**. "(2332) Proposal to Conserve the Name *Sporobolus* against *Spartina*, *Crypsis*, *Ponceletia*, and *Heleochloa* (Poaceae: *Chloridoideae*: *Sporobolinae*)." Taxon **63** (6): 1373–74.

Peterson, P. M., Romaschenko, K., Herrera Arrieta, Y., and Saarela, J. M. **2014b**. "A Molecular Phylogeny and New Subgeneric Classification of Sporobolus (Poaceae: *Chloridoideae*: *Sporobolinae*)." Taxon **63** (6): 1212–43.

Pevzner, P. A. **1989**. "1-Tuple DNA Sequencing: Computer Analysis." Journal of Biomolecular Structure & Dynamics **7** (1): 63–73.

Pfeifer, M., Kugler, K. G., Sandve, S. R., Zhan, B., Rudi, H., Hvidsten, T. R., International Wheat Genome Sequencing Consortium, Mayer, K. F. X., and Olsen, O.-A. **2014**. "Genome Interplay in the Grain Transcriptome of Hexaploid Bread Wheat." Science **345** (6194): 1250091–1250091.

Plomion, C., Aury, J.-M., Amselem, J., Alaeitabar, T., Barbe, V., Belser, C., Bergès, H., *et al.* **2015**. "Decoding the Oak Genome: Public Release of Sequence Data, Assembly, Annotation and Publication Strategies." Molecular Ecology Resources, may 2015.

Poole, R., Barker, G., Wilson, I. D., Coghill, J. A., and Edwards, K. J. **2007**. "Measuring Global Gene Expression in Polyploidy; a Cautionary Note from Allohexaploid Wheat." Functional & Integrative Genomics **7** (3): 207–19.

Prasad, V., Strömberg, C., Alimohammadian, H., and Sahni, A. **2005**. "Dinosaur Coprolites and the Early Evolution of Grasses and Grazers." Science **310** (5751): 1177–80.

Prieto, J., Cires, E., Sánchez Corominas, T., and Vázquez, V. **2011**. "Systematics and Management of Natural Resources: The Case of Spartina Species on European Shores." Biologia **66** (6).

Q

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. **2012**. "A Tale of Three next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers." BMC Genomics **13** (1): 341.

R

R Development Core Team. **2011**. R: A Language and Environment for Statistical Computing. Vienna, Austria. http://www.R-project.org/.

Raicu, P. **1976**. "Les complexes polyploïdes chez les végétaux." Bulletin de la Société Botanique de France **123** (5-6): 249–60.

Ramsey, J., and Schemske, D. W. **1998**. "Pathways, Mechanisms, and Rates of Polyploid Formation in Flowering Plants." Annual Review of Ecology and Systematics, 467–501.

Raybould, A. F., Gray, A. J., Lawrence, M. J., and Marshall, D. F. **1991**. "The Evolution of *Spartina anglica* C.E. Hubbard (Gramineae): Origin and Genetic Variability." Biological Journal of the Linnean Society **43** (2): 111–26.

Reddy, T. B. K., Thomas, A. D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J. Mallajosyula, J., Pagani, I., Lobos, E. A., and Kyrpides, N. C. **2015**. "The Genomes OnLine Database (GOLD) v.5: A Metadata Management System Based on a Four Level (meta) genome Project Classification." Nucleic Acids Research **43** (D1): D1099–1106.

Reinhardt, J. A., Baltrus, D. A., Nishimura, M. T., Jeck, W. R., Jones, C. D., and Dangl. J. L., **2008**. "*De novo* Assembly Using Low-Coverage Short Read Sequence Data from the Rice Pathogen Pseudomonas Syringae Pv. Oryzae." Genome Research **19** (2): 294–305.

Ren, X., Liu, T., Dong, J., Sun, L., Yang, J., Zhu, Y., and Jin, Q. **2012**. "Evaluating de Bruijn Graph Assemblers on 454 Transcriptomic Data." Edited by Michael Watson. PLoS ONE **7** (12): e51188.

Renny-Byfield, S., and Wendel, J. F. **2014**. "Doubling down on Genomes: Polyploidy and Crop Plants." American Journal of Botany **101** (10): 1711–25.

Renny-Byfield, S., Ainouche, M. L., Leitch, I. J., Lim, K. Y., Le Comber, S. C., and Leitch, A. R. **2010**. "Flow Cytometry and GISH Reveal Mixed Ploidy Populations and *Spartina* Nonaploids with Genomes of *S. alterniflora* and *S. maritima* Origin." Annals of Botany **105** (4): 527–33.

Richardson, B. A., Page, J. T., Bajgain, P., Sanderson, S. C., and Udall. J. A., **2012**. "Deep Sequencing of Amplicons Reveals Widespread Intraspecific Hybridization and Multiple Origins of Polyploidy in Big Sagebrush (*Artemisia tridentata*; Asteraceae)." American Journal of Botany **99** (12): 1962–75.

Robasky, K., Lewis, N. E., and Church, G. M. **2014**. "The Role of Replicates for Error Mitigation in next-Generation Sequencing." Nature Reviews Genetics **15** (1): 56–62.

Roberts, P. D., and Pullin, A. S. **2008**. “The Effectiveness of Management Interventions for the Control of Spartina Species: A Systematic Review and Meta-Analysis.” Aquatic Conservation: Marine and Freshwater Ecosystems **18** (5): 592–618.

Rodríguez, F., Oliver, J. F., Marín, A., and Medina, J. R. **1990**. The general stochastic model of nucleotide substitution. Journal of Theoretical Biology **142**: 485–501.

Ronquist, F., and Huelsenbeck, J. P. **2003**. “MrBayes 3: Bayesian Phylogenetic Inference under Mixed Models.” Bioinformatics (Oxford, England) **19** (12): 1572–74.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, Mel., Leamon, J. H. *et al.* **2011**. “An Integrated Semiconductor Device Enabling Non-Optical Genome Sequencing.” Nature **475** (7356): 348–52.

Rousseau-Gueutin, M., Bellot, S., Martin, G.E., Boutte, J., Chelaifa, H., Lima, O., Michon-Coudouel, S. *et al.* **2015**. “The Chloroplast Genome of the Hexaploid *Spartina maritima* (Poaceae, *Chloridoideae*): Comparative Analyses and Molecular Dating.” Molecular Phylogenetics and Evolution, **93**:5-16

S

Saitou, N., and Nei, M. **1987**. “The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees.” Molecular Biology and Evolution **4** (4): 406–25.

Salmon, A., Ainouche, M. L. and Wendel, J. F. **2005**. “Genetic and Epigenetic Consequences of Recent Hybridization and Polyploidy in *Spartina* (Poaceae)” Molecular Ecology **14** (4): 1163–75.

Salmon, A., Flagel, L., Ying, B., Udall, J. A., and Wendel, J. F. **2009**. “Homoeologous Nonreciprocal Recombination in Polyploid Cotton.” New Phytologist **186** (1): 123–34.

Salmon, A., Udall, J. A., Jeddeloh, J. A., and Wendel, J. F. **2012**. “Targeted Capture of Homoeologous Coding and Noncoding Sequence in Polyploid Cotton.” G3; Genes|Genomes|Genetics **2** (8): 921–30.

Salmon, A., and Ainouche M. L. **2015**. “Next Generation Sequencing and the Challenge of Deciphering Evolution of Recent and Highly Polyploid Genomes.” In . Germany: Koeltz Scientific Books. www.iapt-taxon.org.

Salse, J., Bolot, S., Throude, M., Jouffe, V., Piegu, B., Quraishi, U. M., Calcagno, T., Cooke, R., Delseny, M., and Feuillet, C. **2008**. “Identification and Characterization of Shared Duplications between Rice and Wheat Provide New Insight into Grass Genome Evolution.” The Plant Cell Online **20** (1): 11–24.

Sanchéz-Gullón, E. **2001**. “*Spartina versicolor* (Poaceae): Novedad Agrotológica Para Andalucía.” Acta Liotanica Malacitana **26**: 279–80.

Sanger, F., and Coulson, A. R. **1975**. “A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase.” Journal of Molecular Biology **94** (3): 441–48.

Sanger, F., Nicklen, S., and Coulson, A. R. **1977**. "DNA Sequencing with Chain-Terminating Inhibitors." Proceedings of the National Academy of Sciences of the United States of America **74** (12): 5463–67.

Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M., and Smith, M. **1978**. "The Nucleotide Sequence of Bacteriophage ϕX174." Journal of Molecular Biology **125** (2): 225–46.

Sanmiguel, P., and Bennetzen, J. L. **1998**. "Evidence That a Recent Increase in Maize Genome Size Was Caused by the Massive Amplification of Intergene Retrotransposons." Annals of Botany **82** (suppl 1): 37–44.

Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., and Gibrat, J.-F. **2012**, "Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis." Journal of Computational Biology **19** (6): 796–813.

Schnable, J. C., Springer, N. M., and Freeling, M. **2011**. "Differentiation of the Maize Subgenomes by Genome Dominance and Both Ancient and Ongoing Gene Loss." Proceedings of the National Academy of Sciences **108** (10): 4069–74.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., *et al.* **2009**. "The B73 Maize Genome: Complexity, Diversity, and Dynamics." Science **326** (5956): 1112–15.

Schranz, M. E., and Osborn T. C. **2000**. "Novel Flowering Time Variation in the Resynthesized Polyploid *Brassica napus*." The Journal of Heredity **91** (3): 242–46.

Schulz, M. H., Weese, D., Holtgrewe, M., Dimitrova, V., Niu, S., Reinert, K., and Richard, H. **2014**. "Fiona: A Parallel and Automatic Strategy for Read Error Correction." Bioinformatics **30** (17): i356–63.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J.M., and Birol, I. **2009**. "ABySS: A Parallel Assembler for Short Read Sequence Data." Genome Research **19** (6): 1117–23.

Sirota-Madi, A., Olender, T., Helman, Y., Ingham, C., Brainis, I., Roth, D., Hagi, E., *et al.* **2010**. "Genome Sequence of the Pattern Forming Paenibacillus Vortex Bacterium Reveals Potential for Thriving in Complex Environments." BMC Genomics **11** (1): 710.

Sleep, J. A., Schreiber, A. W., and Baumann, U. **2013**. "Sequencing Error Correction without a Reference Genome." BMC Bioinformatics **14** (1): 367.

Smith, T. F., and Waterman, M. S. **1981**. "Identification of Common Molecular Subsequences." Journal of Molecular Biology **147** (1): 195–97.

Snowdon, R. J., Köhler, W. , and Köhler, A. **1997**. "Chromosomal Localization and Characterization of rDNA Loci in the *Brassica* A and C Genomes." Genome **40** (4): 582–87.

Song, S., Liu, L., Edwards, S. V., and Wu, S. **2012**. "Resolving Conflict in Eutherian Mammal Phylogeny Using Phylogenomics and the Multispecies Coalescent Model." Proceedings of the National Academy of Sciences **109** (37): 14942–47.

Soreng, R. J., Peterson, P. M., Romaschenko, K., Davidse, G., Zuloaga, F. O., Judziewicz, E. J., Filgueiras, T. S., Davis, J. I., and Morrone, O. **2015**. "A Worldwide Phylogenetic

Classification of the Poaceae (Gramineae): Phylogenetic Classification of the Grasses." Journal of Systematics and Evolution **53** (2): 117–37.

Stamatakis, A. **2014**. "RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies." Bioinformatics **30** (9): 1312–13.

Stebbins, G. L. **1950**. "Variation and Evolution in Plants." Columbia University Press, New York.

Stebbins, G. L. **1971**. "Chromosomal Evolution in Higher Plants." London: Edward Arnold.

Stephens, J. D., Rogers, W. L., Heyduk, K., Cruse-Sanders, J. M., Determann, R. O., Glenn, T. C., and Malmberg, R. L. **2015**. "Resolving Phylogenetic Relationships of the Recently Radiated Carnivorous Plant Genus Sarracenia Using Target Enrichment." Molecular Phylogenetics and Evolution **85** (April): 76–87.

Strong, D. R., and Ayres D. A. **2013**. "Ecological and Evolutionary Misadventures of Spartina." Annual Review of Ecology, Evolution, and Systematics **44** (1): 389–410.

Stupar, R. M., Bhaskar, P. B., Yandell, B. S., Rensink, W. A., Hart, A. L., Ouyang, S., Veilleux, R. E., *et al.* **2007**. "Phenotypic and Transcriptomic Changes Associated With Potato Autopolyploidization." Genetics **176** (4): 2055–67.

Szadkowski, E., Eber, F., Huteau, V., Lodé, M., Huneau, C., Belcram, H., Coriton, O., *et al.* **2010**. "The First Meiosis of Resynthesized *Brassica napus*, a Genome Blender." New Phytologist **186** (1): 102–12.

T

Tagu, D., and Moussard, C. **2006**. Techniques for Molecular Biology. Enfield, NH: Science Publishers.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. **2011**. "MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods." Molecular Biology and Evolution **28** (10): 2731–39.

Tang, H., Bowers, J. E. , Wang, X., and Paterson, A. H. . **2010**. "Angiosperm Genome Comparisons Reveal Early Polyploidy in the Monocot Lineage." Proceedings of the National Academy of Sciences **107** (1): 472–77.

te Beest, M., Le Roux, J. J., Richardson, D. M., Brysting, A. K., Suda, J., Kubesova, M., and Pysek, P. **2012**. "The More the Better? The Role of Polyploidy in Facilitating Plant Invasions." Annals of Botany **109** (1): 19–45.

Tennessen, J. A., Govindarajulu, R., Ashman, T.-L. and Liston, A. **2014**. "Evolutionary Origins and Dynamics of Octoploid Strawberry Subgenomes Revealed by Dense Targeted Capture Linkage Maps." Genome Biology and Evolution **6** (12): 3295–3313.

The International Wheat Genome Sequencing Consortium (IWGSC), Mayer, K. F. X., Rogers, J., Dole el, J., Pozniak, C., Eversole, K., Feuillet, C. *et al.* **2014**. "A Chromosome-Based

Draft Sequence of the Hexaploid Bread Wheat (*Triticum aestivum*) Genome." Science **345** (6194): 1251788–1251788.

Thompson, J. D., McNeilly, T. and Gray, A. J. **1991**. "Population Variation in *Spartina anglica* C. E. Hubbard. I. Evidence from a Common Garden Experiment." New Phytologist **117** (1): 115–28.

Toledo-Silva, Guilherme, Cardoso-Silva, C. B., Jank, L., and Souza, A. P. **2013**. "*De novo* Transcriptome Assembly for the Tropical Grass Panicum Maximum Jacq." Edited by Zhanjiang Liu. PLoS ONE **8** (7): e70781.

Toloczyki, C., and Feix, G. **1986**. "Occurrence of 9 Homologous Repeat Units in the External Spacer Region of a Nuclear Maize rRNA Gene Unit." Nucleic Acids Research **14** (12): 4969–86.

U

Udall, J. A., Swanson, J. M., Nettleton, D., Percifield, R. J., and Wendel. J. F. **2006**. "A Novel Approach for Characterizing Expression Levels of Genes Duplicated by Polyploidy." Genetics **173** (3): 1823–27.

Udall, J. A., Swanson, J. M., Haller, K., Rapp, R. A., Sparks, M. E., Hatfield, J., Yu, Y., *et al.* **2006**. "A Global Assembly of Cotton ESTs." Genome Research **16** (3): 441–50.

Utomo, H. S., Wenefrida, I., Materne, M. D., and Harrison. S. A. **2009**. "Genetic Diversity and Population Genetic Structure of Saltmarsh *Spartina alterniflora* from Four Coastal Louisiana Basins." Aquatic Botany **90** (1): 30–36.

V

Vallejo-Marin, M. **2012**. "*Mimulus peregrinus* (Phrymaceae): A New British Allopolyploid Species." PhytoKeys **14** (0): 1–14.

Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van de Peer, Y., and Vandepoele, K. **2013**. "TRAPID: An Efficient Online Tool for the Functional and Comparative Analysis of de Novo RNA-Seq Transcriptomes." Genome Biology **14** (12): R134.

Van de Peer, Y., Maere, S., and Meyer, A. **2009**. "The Evolutionary Significance of Ancient Genome Duplications." Nature Reviews Genetics **10** (10): 725–32.

Van Den Borre, A., and Watson L. **1997**. "On the Classification of the Chloridoideae (Poaceae)." Australian Systematic Botany **10**: 491–531.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., *et al.* **2013**. "From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline: The Genome Analysis Toolkit Best Practices Pipeline." In Current Protocols in Bioinformatics **11**(1110)1–33.

Voelkerding, K. V., Dames, S. A., and Durtschi, J. D. **2009**. "Next-Generation Sequencing: From Basic Research to Diagnostics." Clinical Chemistry **55** (4): 641–58.

Volkov, R., Kostishin, S., Ehrendorfer, F., and Schweizer, D. **1996**. "Molecular Organization and Evolution of the External Transcribed rDNA Spacer Region in Two Diploid Relatives of *Nicotiana tabacum* (Solanaceae)." Plant Systematics and Evolution **201** (1-4): 117–29.

W

Wall, P. K., Leebens-Mack, J., Chanderbali, A. S., Barakat, A., Wolcott, E., Liang, H., Landherr, L., *et al.* **2009**. "Comparison of next Generation Sequencing Technologies for Transcriptome Characterization." BMC Genomics **10** (1): 347.

Walter, M., and Briggs, D. **1969**. Les Plantes: Variations et Evolution. Hachette. L'Univers Des Connaissances.

Wang, J. **2004**. "Stochastic and Epigenetic Changes of Gene Expression in Arabidopsis Polyploids." Genetics **167** (4): 1961–73.

Wang, Y., Yang, Q., and Wang, Z. **2015**. "The Evolution of Nanopore Sequencing." Frontiers in Genetics 5 (January).

Warren, R. L., Sutton, G. G., Jones, S. J. M., and Holt, R. A. **2007**. "Assembling Millions of Short DNA Sequences Using SSAKE." Bioinformatics **23** (4): 500–501.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. **2009**. "Jalview Version 2--a Multiple Sequence Alignment Editor and Analysis Workbench." Bioinformatics **25** (9): 1189–91.

Watson, J. D., and Crick, F. H. **1953**. "Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid." Nature **171** (4356): 737–38.

Watson, L., and Dallwitz, M.J. **1992**. "Grass Genera of the World: Descriptions, Illustrations, Identification, and Information Retrieval; Including Synonyms, Morphology, Anatomy, Physiology, Phytochemistry, Cytology, Classification, Pathogens, World and Local Distribution, and References."

Wendel, J. F. **2000**. "Genome Evolution in Polyploids." Plant Molecular Biology **42** (1): 225–49.

Wendel, J. F., Schnabel, A., and Seelanan, T. **1995**. "Bidirectional Interlocus Concerted Evolution Following Allopolyploid Speciation in Cotton (Gossypium)." Proceedings of the National Academy of Sciences **92** (1): 280–84.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., *et al.* **2008**. "The Complete Genome of an Individual by Massively Parallel DNA Sequencing." Nature **452** (7189): 872–76.

White, T.J., Bruns, T., and Taylor, J. **1990**. "Amplification and Direct Sequencing of Fungal Ribo- Somal RNA Genes for Phylogenetics. In 'PCR Protocols: A Guide to Methods and

Ap- Plications.'" In , Academic Press, 315–22. San Diego, CA: M. Innis, D. Gelfand, J. Sninsky, and T. White eds.

Wolfe, K. H., Li, W. H., and Sharp, P. M. **1987**. "Rates of Nucleotide Substitution Vary Greatly among Plant Mitochondrial, Chloroplast, and Nuclear DNAs." Proceedings of the National Academy of Sciences of the United States of America **84** (24): 9054–58.

Wolfe, K. H. **2001**. "Yesterday's Polyploids and the Mystery of Diploidization." Nature Reviews. Genetics **2** (5): 333–41.

Y

Yang, Z., and Rannala, B. **2012**. "Molecular Phylogenetics: Principles and Practice." Nature Reviews Genetics **13** (5): 303–14.

Yannic, G., Baumel, A., and Ainouche, M. L. **2004**. "Uniformity of the Nuclear and Chloroplast Genomes of *Spartina maritima* (Poaceae), a Salt-Marsh Species in Decline along the Western European Coast." Heredity **93** (2): 182–88.

Yoch, D. C. **2002**. "Dimethylsulfoniopropionate: Its Sources, Role in the Marine Food Web, and Biological Degradation to Dimethylsulfide." Applied and Environmental Microbiology **68** (12): 5804–15.

Yoo, M. Szadkowski, J., E., and Wendel, J. F., **2013**. "Homoeolog Expression Bias and Expression Level Dominance in Allopolyploid Cotton." Heredity **110** (2): 171–80.

You, N., Murillo, G., Su, X., Zeng, X., Xu, J., Ning, K., Zhang, S., Zhu, J., and Cui, X. **2012**. "SNP Calling Using Genotype Model Selection on High-Throughput Sequencing Data." Bioinformatics **28** (5): 643–50.

Young, N. D., Debellé, F., Oldroyd, G. E. D., Geurts, R., Cannon, S. B., Udvardi, M. K., Benedito, V. A., *et al.* **2011**. "The Medicago Genome Provides Insight into the Evolution of Rhizobial Symbioses." Nature, November. **480**(7378):520-4.

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., *et al.* **2005**. "The Genomes of Oryza Sativa: A History of Duplications." PLoS Biology **3** (2): e38.

Yu, X, and Sun, S. **2013**. "Comparing a Few SNP Calling Algorithms Using Low-Coverage Sequencing Data." BMC Bioinformatics **14** (1): 274.

Z

Zerbino, D. R., and Birney, E. **2008**. "Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs." Genome Research **18** (5): 821–29.

Zhang, J. **2006**. "Parallel Adaptive Origins of Digestive RNases in Asian and African Leaf Monkeys." Nature Genetics **38** (7): 819–23.

Annexes :

ANNEXES :

Les articles pubiés au cours de cette thèse et dont je suis co-auteur sont présentés dans cette partie.

Annexe 1 :

Ferreira de Carvalho, J., Poulain, J., Da Silva, C., Wincker, P., Michon-Coudouel, S., Dheilly, A., Naquin, D., Boutte, J., Salmon, A., and Ainouche. M. L. **2012**. "Transcriptome *de novo* Assembly from next-Generation Sequencing and Comparative Analyses in the Hexaploid Salt Marsh Species *Spartina maritima* and *Spartina alterniflora* (Poaceae)." Heredity **110** (2): 181–93.

Annexe 2 :

Ferreira de Carvalho, J., Chelaifa, H., Boutte, J., Poulain, J., Couloux, A., Wincker, P., Bellec, A., *et al.* **2013**. "Exploring the Genome of the Salt-Marsh *Spartina maritima* (Poaceae, Chloridoideae) through BAC End Sequence Analysis." Plant Molecular Biology **83** (6): 591–606.

Annexe 3 :

Rousseau-Gueutin, M., Bellot, S., Martin, G.E., Boutte, J., Chelaifa, H., Lima, O., Michon-Coudouel, S. *et al.* **2015**. "The Chloroplast Genome of the Hexaploid *Spartina maritima* (Poaceae, *Chloridoideae*): Comparative Analyses and Molecular Dating." Molecular Phylogenetics and Evolution, **93**:5-16

Heredity

Heredity (2012), 1–13

www.nature.com/hdy

ORIGINAL ARTICLE

Transcriptome *de novo* assembly from next-generation sequencing and comparative analyses in the hexaploid salt marsh species *Spartina maritima* and *Spartina alterniflora* (Poaceae)

J Ferreira de Carvalho[1], J Poulain[2], C Da Silva[2], P Wincker[2], S Michon-Coudouel[3], A Dheilly[3], D Naquin[3], J Boutte[1], A Salmon[1] and M Ainouche[1]

Spartina species have a critical ecological role in salt marshes and represent an excellent system to investigate recurrent polyploid speciation. Using the 454 GS-FLX pyrosequencer, we assembled and annotated the first reference transcriptome (from roots and leaves) for two related hexaploid *Spartina* species that hybridize in Western Europe, the East American invasive *Spartina alterniflora* and the Euro-African *S. maritima*. The *de novo* read assembly generated 38 478 consensus sequences and 99% found an annotation using Poaceae databases, representing a total of 16 753 non-redundant genes. *Spartina* expressed sequence tags were mapped onto the *Sorghum bicolor* genome, where they were distributed among the subtelomeric arms of the 10 *S. bicolor* chromosomes, with high gene density correlation. Normalization of the complementary DNA library improved the number of annotated genes. Ecologically relevant genes were identified among GO biological function categories in salt and heavy metal stress response, C4 photosynthesis and in lignin and cellulose metabolism. Expression of some of these genes had been found to be altered by hybridization and genome duplication in a previous microarray-based study in *Spartina*. As these species are hexaploid, up to three duplicated homoeologs may be expected per locus. When analyzing sequence polymorphism at four different loci in *S. maritima* and *S. alterniflora*, we found up to four haplotypes per locus, suggesting the presence of two expressed homoeologous sequences with one or two allelic variants each. This reference transcriptome will allow analysis of specific *Spartina* genes of ecological or evolutionary interest, estimation of homoeologous gene expression variation using RNA-seq and further gene expression evolution analyses in natural populations.
Heredity advance online publication, 14 November 2012; doi:10.1038/hdy.2012.76

Keywords: transcriptome assembly; polyploidy; invasive species; *Spartina*; chloridoideae

INTRODUCTION

The recent advent of next-generation sequencing (NGS) technologies has opened unique avenues to address ecological and evolutionary questions involving non-model biological systems for which there are limited genomic resources (Hudson, 2008; Ekblom and Galindo, 2010). This is particularly relevant for complex and redundant genomes of polyploid species, which represent a major fraction of eukaryotic lineages (Otto, 2007). Although sequencing performance is rapidly improving in read depth, technologies generating long-sequence fragments such as 454 Roche pyrosequencing have proven particularly useful in *de novo* sequencing and development of new resources for non-model species, without an available reference genome (Wheat, 2008). High-throughput transcriptome sequencing allows assembly of reference transcriptomes that may be used for various purposes in evolutionary ecology, such as functionally important gene annotation or discovery (for example, Alagna *et al.*, 2009; Barakat *et al.*, 2009; Sun *et al.*, 2010; Logacheva *et al.*, 2011), molecular marker (for example, microsatellite, single-nucleotide polymorphism (SNP)) detection (Barbazuk *et al.*, 2007; Novaes *et al.*, 2008; Bundock *et al.*, 2009) or gene expression variation (Buggs *et al.*, 2010; Swarbreck *et al.*, 2011; Ilut *et al.*, 2012; Yoo *et al.*, 2012). As polyploidy is a recurrent process, many lineages exhibit superimposed traces of genome duplication. Large-scale sequencing and deep read coverage offer a unique opportunity to explore the redundant genome and transcriptome of polyploids, even when diploid progenitors are unidentified or extinct, which makes identification of duplicated homoeologous gene copies particularly challenging.

Recurrent polyploidy is particularly well illustrated in the genus *Spartina* (Poaceae), where all extant species are polyploids (reviewed in Ainouche *et al.* (2012)). The grass genus *Spartina* belongs to the Chloridoideae subfamily, a genomically poorly explored Poaceae lineage, contrasting with well-investigated crops, such as rice, sorghum, maize or wheat that belong to other grass subfamilies. Divergence between *Spartina* and these grass models is currently estimated to be 35–40 million years ago (MYA) with Panicoideae (including *Sorghum* and maize) and at least 50 MYA with Ehrhartoideae (including rice) (Christin *et al.*, 2008). *Spartina* is

[1]UMR CNRS 6553 Ecobio, University of Rennes 1, Rennes Cedex, France; [2]Genoscope, 2 rue Gaston Crémieux, Evry, France and [3]Environmental Genomics Platform (Biogenouest), Rennes Cedex, France
Correspondence: Professor ML Ainouche, UMR CNRS 6553 Ecobio, University of Rennes 1, Bât 14A Campus Scientifique de Beaulieu, 35 042 Rennes Cedex, France.
E-mail: malika.ainouche@univ-rennes1.fr
Received 31 May 2012; revised 10 September 2012; accepted 1 October 2012

composed of 13–15 perennial species (Mobberley, 1956), colonizing coastal or inland salt marshes. The basic chromosome numbers in *Spartina* is x = 10, as in most Chloridoideae (Marchant, 1968). *Spartina* species exhibit various ploidy levels ranging from tetraploid to dodecaploid (Ainouche *et al.*, 2004a). Two closely related hexaploid species, *Spartina maritima* (Curt.) Fern., and *S. alterniflora* Lois., are derived from a common hexaploid ancestor (Baumel *et al.*, 2002a; Fortune *et al.*, 2007); although divergence time has not been definitively ascertained, analysis of chloroplast DNA divergence suggests that they diverged less than 3 MYA. They have a critical ecological role in coastal salt marshes at the interface of land and sea, and represent classical models involved in reticulate evolution and recent polyploid speciation (Ainouche *et al.*, 2004a, b; Ainouche *et al.*, 2009). They thus make a good model in evolutionary ecology to investigate the consequences of polyploidy at different evolutionary time scales in natural populations, and to explore the adaptive processes accompanying hybridization, polyploid species formation and expansion.

As for most *Spartina* species, *S. alterniflora* is native to the New World, where it is distributed from Canada to southern Argentina along the North and South American Atlantic coast (Mobberley, 1956), whereas *S. maritima* is distributed along the western European and African Atlantic coasts. Divergence between the two species across the Atlantic Ocean was accompanied by ecological and phenotypic differentiation. *Spartina alterniflora* has a larger distribution and displays invasive abilities in most regions where it was introduced: in California (Ayres *et al.*, 2004; Civille *et al.*, 2005), in China (Li *et al.*, 2009) and in western Europe (Campos *et al.*, 2004; Ainouche *et al.*, 2009; Querné *et al.*, 2011). In contrast, *S. maritima* populations are regressing. The recession of *S. maritima* in its northern range limit (southern England and Brittany) is interpreted as a consequence of climatic changes and anthropogenic habitat disturbance (Raybould *et al.*, 1991), but may also be related to the biological and morphological differences between these two species. *Spartina alterniflora* exhibits strong rhizomes facilitating lateral expansion and sediment accretion, and thus has an important role in the salt marsh dynamics where it is considered as an ecosystem engineer, whereas *S. maritima* is a non-rhizomatous, genetically depauperate species (Yannic *et al.*, 2004) with very low seed production (Marchant and Goodman, 1969; Castellanos *et al.*, 1994; Castillo *et al.*, 2008). *Spartina maritima* and *S. alterniflora* also exhibit chromosome number differences, as the former has a regular hexaploid number ($2n = 6x = 60$) whereas the latter presents aneuploidy ($2n = 62$), and genome size differences ($2C = 3.8$ pg for *S. maritima* and $2C = 4.3$ pg for *S. alterniflora*, Fortune *et al.*, 2008). Less than 5% nucleotide divergence was encountered at 10 putative orthologous-coding loci between the two species, but consistent gene expression differences (13% of the examined genes) were detected using heterologous rice microarrays (Chelaifa *et al.*, 2010a). Genes involved in cellular growth were found highly expressed in *S. alterniflora* and downregulated in *S. maritima*, whereas stress-related genes were highly expressed in *S. maritima* (Chelaifa *et al.*, 2010a).

Spartina alterniflora and *S. maritima* are involved in one of the textbook examples of recent allopolyploid speciation (reviewed in Ainouche *et al.*, 2004b; Ainouche *et al.*, 2009). *Spartina alterniflora* was accidentally introduced during the 19th century in Europe, where it hybridized with the native *S. maritima.* In England, hybridization (with *S. alternifora* as maternal genome donor, Ferris *et al.*, 1997; Baumel *et al.*, 2001) resulted in *Spartina* × *townsendii,* a perennial sterile hybrid first recorded around 1870 (Groves and Groves, 1880), that gave rise by chromosome doubling to a fertile, vigorous and highly invasive allo-dodecapolyploid species, *Spartina anglica,* which has now been introduced on several continents. An independent hybridization event between *S. maritima* and *S. alterniflora* occurred also in southwest France with *S. alterniflora* as the maternal parent (Baumel *et al.*, 2003), contributing to the formation of another sterile F1 hybrid, *Spartina* × *neyrautii*.

Recent studies have been aimed at examining the evolutionary fate of the homoeologous parental genomes from *S. maritima* and *S. alterniflora* in the neo-allododecaploid species to understand the genomic determinants of the ecological success of the invasive neopolyploid (Baumel *et al.*, 2002b; Ainouche *et al.*, 2004a; Salmon *et al.*, 2005; Parisod *et al.*, 2009). These studies have revealed that epigenetic reprogramming (for example, DNA methylation, Salmon *et al.*, 2005; Parisod *et al.*, 2009) and evolution of gene expression (Chelaifa *et al.*, 2010b) represent important components of the speciation process in polyploid *Spartina* species, and are most likely having a critical role in the ecology of the species. However, the previously employed technology for transcriptome analysis (heterologous hybridization on rice microarrays) had several limitations (for example, only a fraction of the genes that hybridized on the array could be analyzed, only global gene expression variation could be evaluated, with no possible distinction of the copies duplicated by polyploidy). Developments in sequencing technology mean that there is now the potential to develop more advanced genomic resources in this important model system for understanding the ecological and evolutionary consequences of hybridization and polyploidy.

When analyzing species such as these where genomic resources are lacking, constitution of a reference transcriptome represents a first critical step to explore the genic compartment. In polyploids, assembled contigs from sequence reads represent consensus sequences among potentially different alleles at strictly orthologous loci, more or less divergent homoeologues (parental orthologues duplicated by polyploidy), or recent paralogues (resulting from individual gene duplication); thus necessitating a more complicated analytical strategy than for diploids. The goal of this study is to build a reference 'consensus' transcriptome in the hexaploid parental species *S. alterniflora* and *S. maritima* using NGS technology, which will allow annotation and identification of specific *Spartina* genes, including genes of ecological or evolutionary (that is, genes whose expression is altered following speciation) interest. The strategy was then to (i) choose the appropriate high-throughput sequencing that generates long reads facilitating *de novo* assemblies in the absence of a reference genome (that is, the GS-FLX Roche 454 technology) and (ii) to sequence as many diverse transcripts as possible (to annotate a maximum of genes), by using different types of complementary DNA (cDNA) libraries (normalized and non-normalized) from different tissues (leaves, roots) and from different (natural or controlled) environmental conditions. Sequence heterogeneity at putative homologous loci (within 'consensus' contigs) is discussed in the context of the (hexaploid) redundant genomes of *S. maritima* and *S. alterniflora*. Beyond the *Spartina* model, the procedure presented here may be applicable to any polyploid system for which no reference genome is available and whose parental species (that is, homoeologous copies) are unknown.

MATERIALS AND METHODS

Plant material

Samples from *S. alterniflora* were collected in Landerneau (Finistère, France). *Spartina maritima* was collected at two sites from the French Atlantic coast: Pointe du Verdon (Morbihan) and Noirmoutier (Vendée). Several individuals were collected at each site, and plants were transplanted in the greenhouse (University of Rennes 1).

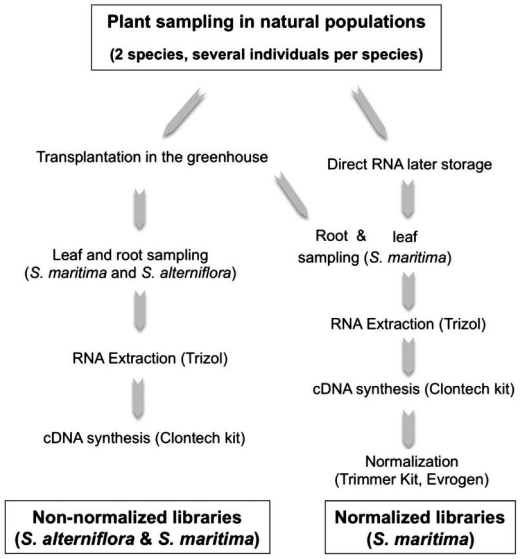


Figure 1 Sampling strategy and construction of the normalized (*S. maritima*) and non-normalized (*S. maritima* and *S. alterniflora*) cDNA libraries.

To maximize detection and annotation of various expressed *Spartina* genes, RNA extraction was performed on different organs (leaves and roots) from plants sampled either from wild populations and so grown in variable natural conditions (normalized cDNA libraries) or transplanted in a common greenhouse environment (non-normalized cDNA libraries) (Figure 1). Non-normalized libraries usually offer an overview of the most transcribed genes, whereas normalization facilitates the assessment of rare transcripts by decreasing the prevalence of abundant transcripts. For practical reasons, the normalized library could be done only on one species (the European native *S. maritima*), which was chosen because a larger population sampling was available as part of an ongoing project in our laboratory, involving genome sequencing of this species.

Non-normalized cDNA libraries for both *S. maritima* (from Pointe du Verdon) and *S. alterniflora* (from Landerneau) were created from plants grown in the same conditions in the greenhouse (30 cm^3 daily watered pots containing a mixture of soil, fertilizer and sand) under a day temperature of 20 °C and night temperature of 14 °C. After 21 days of acclimatization, 1–2 g of young leaves and roots per plant were collected separately from three different individuals (from the same population), frozen in liquid nitrogen and stored at −80 °C until RNA extraction.

A normalized library (for *S. maritima*) was created using leaves from eight individuals collected in the population from Noirmoutier and sampled along a tidal gradient to capture subtleties in gene expression under varying environmental conditions. Two additional *S. maritima* individuals collected from Pointe du Verdon and transplanted in the greenhouse were also included in the normalized library. Five young leaves were selected for each individual plant, and stored in RNAlater solution (Ambion Inc., Austin, TX, USA) at −20 °C until RNA extraction. For practical reasons, the root normalized library was performed from the same plants used for the non-normalized library that were transplanted in the greenhouse. Roots were carefully washed in distilled water, and then young roots were cut and collected in liquid nitrogen.

For each sample, total RNA was extracted from frozen leaves and roots with Trizol reagent (Sigma-Aldrich Inc., St. Louis, MO, USA) using three cycles of precipitation with isopropanol (Sigma-Aldrich), according to a procedure previously described for *Spartina* (Chelaifa *et al.*, 2010a, b). All RNA samples were quantified using a Nanodrop Spectrophotometer ND 1000 (Nanodrop Technologies, Thermo Fisher Scientific Inc. Waltham, MA, USA) and the RNA quality (absence of degradation and DNA contamination) was checked on an Agilent 2100 Bioanalyzer (DNA 7500 Chip, Agilent Technologies, Santa Clara, CA, USA). After processing, RNA was stored at −80 °C.

cDNA preparation

cDNA synthesis was performed with 1 µg of total RNA using the SMARTer cDNA Synthesis Kit (Clontech, Mountain View, CA, USA), following the protocol recommended by manufacturers. Briefly, first-strand cDNA synthesis was primed with a modified oligo(dT) primer (the 3′SMART CDS Primer II A). When SMARTScribe RT reaches the 5′-end of the mRNA, the enzyme adds a few additional nucleotides to the 3′-end of the cDNA. After a second-strand cDNA synthesis reaction, double-stranded cDNAs were amplified (21 cycles with primer 5′ PCR Primer II A). This procedure yielded about 2–6 µg of cDNAs that were purified using the Qiaquick PCR Purification Kit (Qiagen, Hilden, Germany). An equimolar mix of samples was constituted for each organ and each species to reach 10 µg of total cDNA and stored at −20 °C until sequencing.

Normalization of *S. maritima* cDNA

A total of 1 µg of cDNAs from each organ (leaves and roots) of *S. maritima* was separately normalized as following: 4 µl 4 × hybridization buffer were added and the samples denatured at 95 °C for 5 min and then allowed to anneal at 68 °C for 5 h. The following preheated reagents from the Trimmer kit (Evrogen, Moscow, Russia) were added to the hybridization reaction at 68 °C: 3.5 µl milliQ water, 1 µl 5 × DNAse buffer, 1 µl double-strand nuclease (DSN) enzyme. After incubation at 68 °C for 25 min, the DSN enzyme was inactivated by adding 10 µl of DSN stop solution and heating at 68 °C for 5 min. The normalized cDNA samples were diluted by adding 40.5 µl milliQ water and used for two PCR amplifications. The first PCR (50 µl) contained 1 µl diluted cDNA, 5 µl 10 × Advantage 2 PCR buffer (Clontech), 1 µl 50 × dNTPs mix, 1.5 µl PCR primer M1 10 µM (Evrogen), 1 µl 50 × Advantage 2 Polymerase mix (Clontech) and was amplified as following: initial denaturation at 95 °C for 1 min, followed by 18 cycles (95 °C for 15 s, 66 °C for 20 s, 72 °C for 3 min). The second PCR reaction (100 µl) was performed using 2 µl of diluted normalized cDNA, 1 µl of 10 × Advantage 2 PCR Buffer (Clontech), 2 µl 50 × dNTP mix, 4 µl PCR Primer M2 10 µM (Evrogen), 2 µl 50 × Advantage 2 Polymerase mix (Clontech) and was amplified following an initial denaturation at 95 °C for 1 min, then 12 cycles (95 °C for 15 s, 64 °C for 20 s, 72 °C for 3 min), and a final extension step (64 °C for 15 s and 72 °C for 3 min). The normalized double-stranded cDNAs were checked on an agarose gel and on an Agilent 2100 bioanalyzer DNA chip (DNA 7500 chip), quantified with a ND 1000 Spectrophotometer (Nanodrop Technologies Inc., Wilmington, DE, USA), and stored at −20 °C.

Sequencing, cleaning and assembly

The four non-normalized cDNA libraries (roots and leaves from *S. maritima* and *S. alterniflora*) were sheared by nebulization and sequenced at the Genoscope Platform (Evry). A total of 500 ng of cDNAs were sequenced for each library in two runs on a 454 GS XLR70 Titanium Genomic Sequencer (Roche Inc., Basel, Switzerland). The tissues (leaves and roots) were distinctly distributed on two half regions of the sequencing plate.

Sequencing of the normalized *S. maritima* cDNA libraries was performed at the Environmental Genomic Platform of the University of Rennes 1. A total of 500 ng of each normalized cDNA library from *S. maritima* leaves and roots were nebulized and sequenced separately in two half-plates on a 454 GS XLR70 Titanium Genomic Sequencer (Roche Inc.).

The 454 sequence primers (Roche Inc.) and low-quality sequences were removed during signal processing. GS Assembler version 2.3 (Roche, Inc.) was employed to assemble reads into contigs; this program was already successfully used for assembly in transcriptome analyses (Bellin *et al.* (2009) in *Vitis vinifera*; Gedye *et al.* (2010) in *S. pectinata*; Sun *et al.* (2010) in *Panax ginseng*).

Different assemblies were performed for each separate library or for combined data sets per species, tissue and normalization type. Finally, a global assembly of all the obtained reads provided the reference transcriptome for both hexaploids.

As hexaploid *Spartina* species are expected to potentially express up to six allelic transcripts per locus (resulting from three duplicated pairs of homoeologous genes), the assembly strategy aimed at assembling potentially homologous reads (orthologues and homoeologues) with relatively low stringency to construct consensus contigs constituting the 'hexaploid reference transcriptome' that will be used for identification and annotation of *Spartina*

genes. In this perspective, effects of different minimum match percentages (90, 95, 96 and 97%) on the assembly process were explored. Analyses presented in this paper are based on *de novo* assemblies executed with 90% of minimum match on at least 100 bp and GS Assembler version 2.3 (Roche, Inc.) default parameters for cDNA. This low minimum match percentage (90%) was chosen to maximize assembly of reads corresponding to putative orthologous and homoeologous transcripts, although we cannot rule out assembling weakly divergent paralogs. Useful information (such as the number of reads used in the assembly, the number of contigs and singletons, mean length and read depth) was extracted from assembly files. Read depth is estimated by GS Assembler as the total number of included bases from all the obtained 454 sequence reads aligned to generate the consensus contig sequence, divided over the contig length. To test validity of the assembly, we aligned 10 contigs against homologous expressed sequence tags (ESTs), which were sequenced using the Sanger method (Chelaifa *et al.*, 2010a) and sequence identities were calculated.

Contigs from *S. maritima* and *S. alterniflora* were then mapped to the *Sorghum bicolor* genome, the closest related species to *Spartina* that has a fully sequenced and annotated genome (Paterson *et al.*, 2009), to compare the distribution and density of the identified *Spartina* homologous genes across the different *Sorghum* chromosomes. The *Sorghum bicolor* gene annotation was retrieved from the Sbicolor_79_gene.gff3 annotation file available at http://genome.jgi-psf.org/Sorbi1/ and gene density was estimated from the proportion of annotated genes per 100 kb intervals. Colinearity between *Spartina* and *Sorghum* has not been investigated previously, but conservation of gene colinearity is expected according to what is known from related lineages (for example, finger millet, Chloridoideae and rice, Ehrhartoideae, Srinivasachary *et al.* (2007)) in the grass family. The BLASTn algorithm was used with a *P*-value of 10^{-5} and Best BLAST Hit (corresponding to the highest *e*-value and bit score) parsed for each query sequence. The proportion of *Spartina* homologs was calculated by 100 kb windows (delimitated from *Sorghum*) and the results were represented using the Circos v.0.55 software (Krzywinski *et al.*, 2009). To evaluate the genome-wide representation of the assembled contigs on the *Sorghum* genome, Pearson's correlations and linear regressions were calculated between gene densities (number of genes per 100 kb window) in *Sorghum* and corresponding homologs in the investigated *Spartina* species. Both statistics were calculated for all 10 *Sorghum* chromosomes and by individual chromosomes using the R software (R Development Core Team, 2011).

Annotation

BLASTn and tBLASTx (Altschul *et al.*, 1990) analyses of contigs and singletons were conducted against two nucleotide databases: *Oryza sativa* ESTs database (http://rapdb.dna.affrc.go.jp), and a home-built regularly updated Poaceae database, including ESTs from *Oryza sativa*, *Zea mays*, *Brachypodium distachyon* and *Sorghum bicolor* (www.gramene.org). All BLAST searches were performed with an *e*-value of 10^{-5}. Best BLAST Hit from all BLAST results were parsed for a homology-based functional annotation.

GO annotations using BLAST2Go (Conesa *et al.*, 2005; Götz *et al.*, 2008) were performed using tBLASTx (*e*-value 10^{-6}) on assembled contigs against the *Arabidopsis thaliana* database from the TAIR website (www.arabidopsis.org) (with GO IDs and term assigned), with an annotation *e*-value hit filter of 10^{-6} and a cutoff of 55 (maximum similarity).

The annotated *Spartina* transcriptome was examined to identify genes of potential ecological interest (for example, genes involved in salt stress response, oxidative stress, heavy metal tolerance or growth). Genes whose expression was previously found altered following hybridization and genome duplication from a rice microarray-based study on these species (Chelaifa *et al.*, 2010a) were investigated. The corresponding accession numbers of the rice oligos spotted on Agilent microarrays (44 K Agilent G2519F) employed in that study were used to retrieve putative homologs in our *Spartina* reference transcriptome using BLASTn (*e*-value 10^{-5}).

Sequence heterogeneity at homologous gene copies

As both *Spartina* species studied here are hexaploid, sequence read heterogeneity is expected in the assembled contigs, resulting from both genome duplication and allelic variation within homoeologues (heterozygosity at orthologous loci). In this study, we chose the 454 technology because it generates long read sequences to facilitate *de novo* assembly, but this sequencing method offers less read depth than alternative technologies generating short reads to capture all the allelic variants that may be transcribed at each locus. As a preliminary evaluation of sequence heterogeneity among assembled reads obtained with the 454 pyrosequencing technique, we have selected contigs with relatively good coverage (at least 50 reads) that were present in both *S. maritima* and *S. alterniflora* data sets.

We looked at polymorphisms within contigs by mapping the corresponding reads (using Genome Assembler v 2.5.3, Roche) to a subset of selected homologous contigs between the two species. We then scanned the resulting alignments for SNPs using the Ace.py program from the biopython package (http://biopython.org/). Rare SNPs or SNPs detected within homopolymeric regions were removed from the analysis to avoid putative false-positive SNPs. We then assembled reads presenting 100% similarity (using at least one shared SNP) to maximize the consensus sequence length. This consensus sequence was then considered as a haplotype, representing a particular copy in the corresponding contig.

RESULTS

De novo assemblies and contig annotation

Spartina maritima. Sequencing of the non-normalized and normalized cDNA libraries from roots and leaves resulted in 425 274 reads (average length 314 ± 147.3 bp) and 558 732 reads (average length 203 ± 102.8 bp), respectively. Data are available in Genbank under accession references SRP015701 and SRP015702 for *S. maritima* and *S. alterniflora*, respectively.

Assemblies and annotations were first performed separately on the sequences obtained from the non-normalized and normalized cDNA libraries for each tissue, respectively, then on the pooled reads from both normalized and non-normalized libraries. A total of nine different assemblies (as presented in Table 1) were performed using individual (by tissue and normalization) or combined data sets, allowing the comparison of annotated contigs by tissue and evaluation of the normalization process efficiency.

After trimming the adapter sequences and removing sequences shorter than 50 bases, 405 386 and 359 159 reads remained for the *S. maritima* non-normalized library and *S. maritima* normalized library, respectively. Assembly of the trimmed reads resulted in 12 309 contigs for the non-normalized library and 17 182 contigs for the normalized library. The mean contig length was 617 bp (s.d. = 540.3, range = 50–8036) and 415 bp (s.d. = 246.9, range = 50–2252) for the non-normalized and normalized libraries, respectively.

Separate assemblies for roots and leaves were also processed for each library, as well as global assembly of all the reads from *S. maritima* to get a global gene annotation for this species. Unequal read numbers were obtained for leaf and root cDNA sequencing in both the normalized and non-normalized libraries. In the non-normalized cDNA library, the read number in leaves was twice that of roots. In the root normalized library, read number was three times larger than the number obtained in the non-normalized library (Table 1). Equivalent number of contigs were assembled for leaves (5866) and roots (5910) in the non-normalized cDNA library, but many more contigs were assembled for roots (13 315) than for leaves (3654) in the normalized library. When pooling all reads from *S. maritima* (normalized and non-normalized for both organs), 25 239 contigs were assembled. Separate assemblies of roots and leaves resulted in 19 069 and 10 098 contigs, respectively.

Functional annotation was performed by sequence comparisons with public databases. The different *S. maritima* data sets (from non-normalized and normalized cDNAs in each tissue) were first compared with the *Oryza sativa* EST database, then to a larger database including four sequenced Poaceae genomes. As expected, the

Table 1 Summary of assemblies and annotations of the *Spartina maritima* and *Spartina alterniflora* complementary DNA libraries

Analysis	*Assemblies*			*Annotations*	
	Number of reads used in the assembly	*Number of contigs*	*Number of singletons*	*tBLASTX Oryza sativa*	*tBLASTX Poaceae*
S. maritima (non-normalized)					
Leaves	273 659	5866	63 064	5237 (3143)	5505 (3825)
Roots	131 727	5910	43 945	5275 (3029)	5551 (3824)
Leaves and roots	405 386	12 309	83 878	11 118 (5705)	11 718 (7290)
S. maritima (normalized)					
Leaves	95 045	3654	43 993	1797 (1589)	2069 (1938)
Roots	264 114	13 315	74 418	9948 (7193)	10 550 (9115)
Leaves and roots	359 159	17 182	89 765	10 805 (8195)	12 518 (10 629)
S. maritima total (non-normalized + normalized)					
Leaves	371 111	10 098	79 436	8517 (4821)	9002 (6100)
Roots	398 991	19 069	84 789	17 409 (8485)	18 162 (11 149)
Total (all organs and all cDNAs)	755 309	25 239	114 857	16 137 (9958)	17 307 (13 786)
S. alterniflora					
Leaves	140 733	3217	30 480	2995 (1806)	3127 (2169)
Roots	203 990	11 155	43 904	10 805 (5 281)	11 201 (6811)
Leaves and roots	344 723	14 317	58 298	13 919 (6430)	14 123 (8370)
S. maritima and *S. alterniflora*					
Leaves	511 844	13 824	93 274	12 246 (5999)	12 910 (7773)
Roots	602 981	29 187	102 638	28 164 (10 268)	29 054 (14 135)
Total	1 114 825	38 478	153 409	36 549 (11 776)	38 089 (16 753)

In brackets are numbers of non-redundant gene annotations.

use of this homemade Poaceae database improved the number of annotated genes (Table 1). In the non-normalized library, 5705 different genes were annotated with the *O. sativa* database and 7290 with the Poaceae database. In the normalized library, 8195 were annotated with the *O. sativa* database and 10 629 with the Poaceae database. The normalization of the cDNA library significantly increased the number of annotated genes, as among these 10 629 annotated genes, 3620 were common to both libraries and 6642 genes were specific to the normalized data set (Figure 2a).

The Poaceae database allowed annotation of 6100 different genes for *S. maritima* leaves and 11 149 genes for roots (Table 1). Among these, 2938 genes were found in both root and leaf transcriptomes, (Figure 2b). When pooling all the read data sets (both tissues and both normalization types), 13 786 genes were annotated in total for *S. maritima* with the Poaceae database (Table 1).

Spartina alterniflora. Sequencing of the *S. alterniflora* non-normalized cDNA library from roots and leaves resulted in 495 749 reads, with an average length of 285 ± 160.6 bp. After trimming, 344 723 reads were used for the assembly, which resulted in 14 137 contigs (Table 1). The *S. alterniflora* contigs have an average length of 759 bp (s.d. = 637.1, range = 50–12 334) and a mean read depth of 14.3. Separate assemblies of roots and leaves were processed as for *S. maritima* and resulted in 3217 contigs for leaves and 11 155 contigs for roots. More reads and more contigs were obtained for roots than for leaves, as observed in *S. maritima* (Table 1).

Functional annotation of the *S. alterniflora* contigs using the *Oryza* and Poaceae databases, respectively, resulted in 1806 and 2169 different genes annotated in leaves. For roots, 5281 (*Oryza* database) and 6811 (Poaceae database) genes were annotated. When pooling root and leaf data sets, 6430 genes were annotated when using the *Oryza* database, and 8370 genes were annotated in total for *S. alterniflora*, when using the Poaceae database (Table 1).

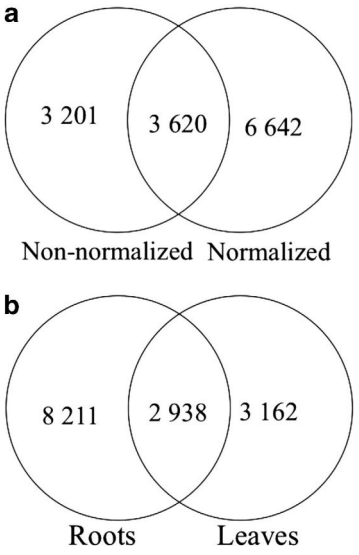


Figure 2 Common annotated contigs (using the Poaceae database) of *S. maritima* (**a**) between non-normalized and normalized cDNA libraries (leaves + roots) (**b**) between roots and leaves.

Spartina leaf and root transcriptomes

To maximize the number of contigs and annotated genes per tissue, *S. maritima* and *S. alterniflora* reads were pooled, which resulted in

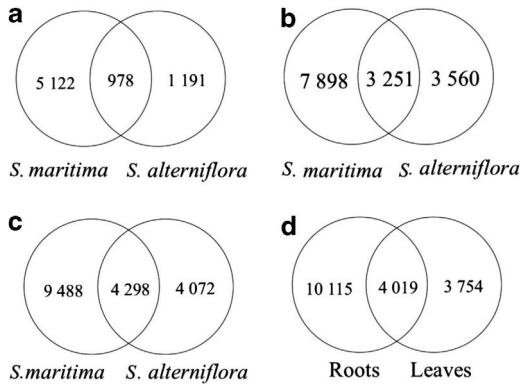


Figure 3 Common annotated contigs (using the Poaceae database) between *S. maritima* and *S. alterniflora.* (**a**) Comparison between *S. maritima* and *S. alterniflora* leaves. (**b**) Comparison between *S. maritima* and *S. alterniflora* roots. (**c**) Comparison between *S. maritima* and *S. alterniflora* (combined data from leaves and roots). (**d**) Comparison between roots and leaves (combined data from both species).

13 824 and 29 187 assembled contigs for leaves and roots, respectively (Table 1). When using the Poaceae database for functional annotation, 7773 and 14 135 different genes were annotated for leaves and roots, respectively. Among these, 4019 (22.5%) genes were common to root and leaf *Spartina* transcriptomes (Figure 3d).

When examining leaf and root transcriptomes between species, 978 and 3251 annotated genes were found common to *S. maritima* and *S. alterniflora* for leaves and roots, respectively (Figures 3a and b). Overall, *S. maritima* and *S. alterniflora* share 4298 expressed genes (pooled leaf and root data sets) with 9488 genes annotated only in *S. maritima* and 4072 genes only in *S. alterniflora* (Figure 3c). The total data set (both species and organs) resulted in 38 478 contigs and 16 753 annotated *Spartina* genes (Table 1), which represent the first reference transcriptome for the hexaploid *Spartina* species.

Distribution of the contigs on the Sorghum genome. The number of homologous genes sequenced in *Spartina* hexaploid species was about half the number found in *Sorghum bicolor* per 100 kb sliding window. Mapping of the *Spartina* contigs to the *Sorghum* genome revealed similar relative gene densities for both *Spartina* EST libraries among the 10 chromosomes (Figure 4b, Supplementary Figure 1). High correlation between *Sorghum* gene densities along chromosomes and the number of homologous *Spartina* genes in a 100-kb *Sorghum* window were encountered for most chromosomes. A relatively lower correlation was found for chromosomes 5 and 8 (Supplementary Figure 1), which could suggest more extensive rearrangements during evolution of these taxa. Furthermore, we observed that *Spartina* gene densities were higher in the corresponding subtelomeric *Sorghum* chromosome positions than in pericentromeric ones, as expected from gene distributions in *Sorghum* (Paterson *et al.*, 2009).

Most-represented genes in the normalized and non-normalized Spartina data sets. The 20 most-represented transcripts (according to read depth) in the non-normalized libraries appear very similar in *S. alterniflora* and *S. maritima* (Supplementary Table 1). In both leaves and roots, they are mainly involved in respiratory pathways (for example, cytochrome *c* oxidase, ATP synthase), and in RNA and ribosomal protein synthesis. In roots, NADH-ubiquinone oxidoreductase and acylCoA-binding protein were also well-represented. Genes involved in stress responses were observed mainly in root transcripts. Among the most represented are the metallothionein and zinc finger (A20 and AN1) domains involved in metal binding and control of oxidative stress. A transcription elongation factor (EF) was also well represented in the root transcriptome of *S. maritima* and *S. alterniflora* (Supplementary Table 1); this gene is involved in protein elongation during translation (Andersen *et al.*, 2003) and is also found highly represented in the roots of other grass species (for example, in *Zea mais,* Poroyko *et al.* (2005) or *Avena barbata,* Swarbreck *et al.* (2011)). The chaperone protein *DnaJ* gene was also encountered in the root transcriptome of *S. maritima*. This gene is induced by heat shock and prevents apoptosis (Gotoh *et al.*, 2004). In addition, in *S. alterniflora*, two contigs annotated with a pathogenesis-related Bet V family protein were highly represented. This gene can be induced by different pathogens, such as viruses, bacteria and fungi (Liu and Ekramoddoullah, 2006).

The most abundant sequences annotated from the normalized cDNA data set in *S. maritima* belong to a larger set of gene categories compared with those encountered in the non-normalized data sets for both *S. alterniflora* and *S. maritima*. In leaves, all of the important functions are represented: we encountered genes involved in flowering control (tetratricopeptide repeat protein 1), in cell wall structure (glycine-rich protein), in the C4 assimilation process (phosphoenol-pyruvate carboxykinase, carbonic anhydrase) and in fatty acid metabolism (Acyl coA-binding protein). The *thioredoxin* gene has a critical role in redox regulation in the apoplast, which regulates cell division (Tian *et al.*, 2009), cell differentiation (Takeda *et al.*, 2003), pollen germination (Ge *et al.*, 2011) and stress responses (Song *et al.*, 2011).

In the normalized root cDNA data set, apart from three highly represented contigs annotated as ribosomal genes, all others were genes and proteins involved in primary metabolism, such as cell transport (ADP-ribosylation factor, ranBP1 domain-containing protein), cell organization (mps 1 binder kinase activator-like 1A, steroid-binding protein, FYVE zinc finger domain-containing protein), plant growth (peptidase T1 family, tetratricopeptide repeat protein 1) and stress response (calreticulin precursor protein, phosphatase 2C, *cytosolic ascorbate peroxidase* gene, peroxiredoxin).

GO (Gene ontology) annotation and biological process analyses

Functional annotation. Using the *A. thaliana* protein database of the TAIR website, GO functions could be assigned to *Spartina* transcripts. Among the various biological processes, cellular (5865) and metabolic (5660) processes, as well as biological regulations (2125) were most highly represented (Figure 5). Important functions were also identified, such as response to stimulus, protein localization and transport and developmental process. Similarly, cell and organelle were most represented between the cellular component and binding and catalytic activities among the various molecular functions (Figure 5).

Identification of ecologically relevant genes. Annotated *Spartina* genes with potential ecological relevance are listed in Supplementary Table 2, with the corresponding number of putative homologous regions identified in the *Sorghum* genome. Transcription factors, such as zinc finger proteins, anti-oxidants (for example, gdp-mannose pyrophosphorylase) and osmolyte synthetic transporters were identified. Heat shock proteins, such as zeaxanthin epoxidase, a precursor of abscisic acid (ABA), which is involved in response to abiotic stress (including salt and heavy metal tolerance), were also encountered.

Among the known genes of the lignin biosynthetic pathway (Humphreys and Chapple, 2002), we were able to identify the *cinnamoyl-CoA reductase* and *cinnamyl alcohol dehydrogenase*

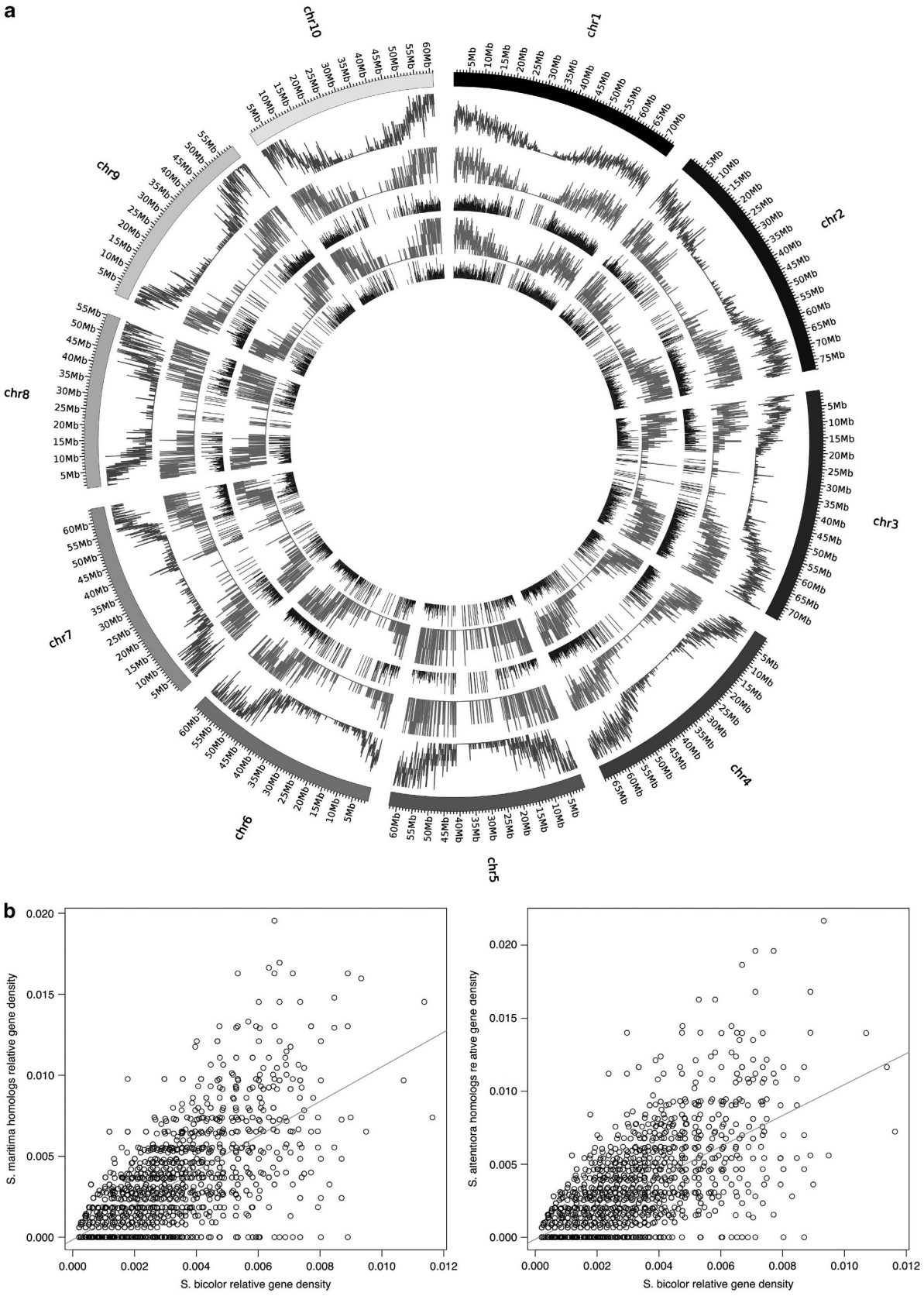


Figure 4 (**a**) *Spartina* contigs mapped to the *Sorghum* genome. The 10 individual chromosomes are shown in the outer circle. Relative gene densities on each chromosome are displayed successively inward as following: (i) gene density in *Sorghum bicolor*, (ii) gene density in *Spartina maritima*, (iii) *Spartina maritima* gene density relative to *Sorghum* gene density, (iv) gene density in *Spartina alterniflora*, (v) *Spartina alterniflora* gene density relative to *Sorghum* gene density (by 100 kb region). (**b**) Correlations between *Sorghum* density and *S. maritima* and *S. alterniflora* homologous gene densities by 100 kb region (P-value $< 2.2\,e^{-16}$).

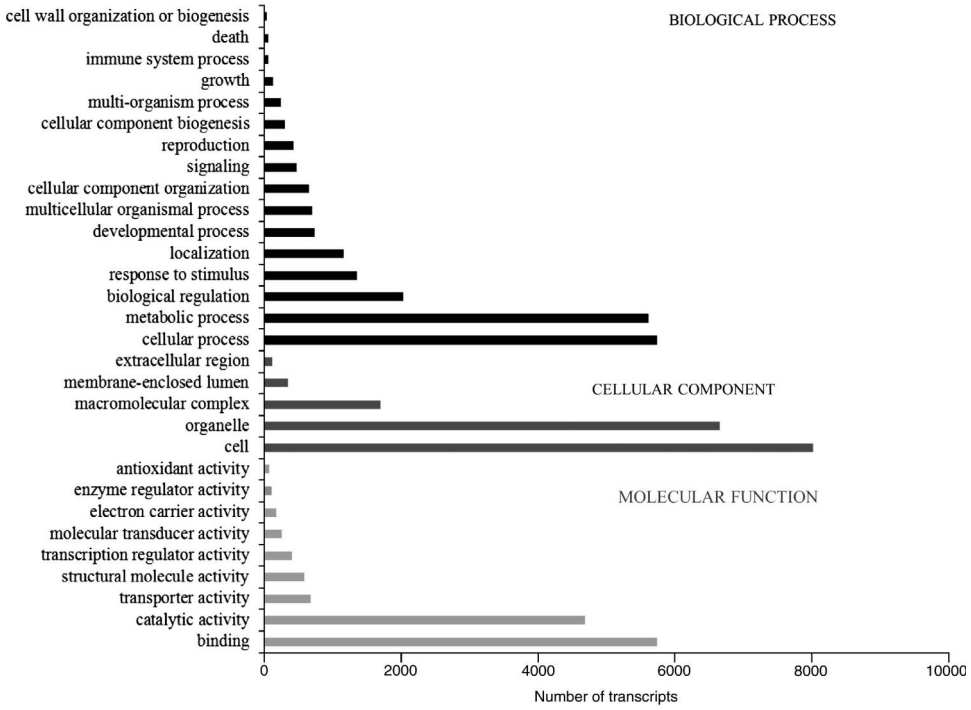


Figure 5 Functional classification of the leaf transcriptome of *S. maritima* and *S. alterniflora*. GO annotations were used for classification for GO cellular component, GO molecular function and GO biological process.

genes. Gene families associated with the production of cellulose, such as cellulose synthases (*CesA*) and glycosyl transferases, were also identified in the reference *Spartina* transcriptome (Supplementary Table 2).

Identification of genes whose expression is altered following speciation in Spartina. When searching for the differentially expressed genes between the parental species (*S. maritima* and *S. alterniflora*) and between the parents and their hybrid (*Spartina* × *townsendii*) or allopolyploid (*S. anglica*) derivatives detected using rice microarrays by Chelaifa *et al.* (2010a, b), we found 409 *Spartina* contigs exhibiting similarities to rice sequences (Supplementary Table 3). A BLAST2Go analysis was performed on these 409 sequences, of which 271 were found to have different functional annotation. Sequences whose expression is altered following speciation according to Chelaifa *et al.* (2010a,b) such as transcription factors, retrotransposons, peptide transport system genes, glutathione transferases, peroxidases and cytochrome *c* oxidase were parsed to provide a sequence database. This database now constitutes a reference for future studies regarding genomic and transcriptomic consequences of polyploidy speciation in *Spartina*.

Polymorphism analysis at homologous genes

Because of the polyploid nature of these highly redundant genomes, up to three duplicated homoeologs may be encountered at each locus, leading to sequence heterogeneity among reads. Contigs from four genes (phosphoenol-pyruvate carboxykinase, HECT domain-containing protein, homeobox domain-containing protein and a heat shock protein) were analyzed in detail to identify homologous sequences, polymorphic sites and putative haplotypes. In these contigs, three to four haplotypes could be distinguished within individuals when comparing the homologous sequences for each gene (Table 2). The polymorphism analysis is illustrated in Table 3, for a 200-bp region of the HECT domain-containing protein gene. In this window, seven haplotypes (over the two species) were aligned. Six polymorphic sites were detected in each species, including four polymorphic sites shared between *S. maritima* and *S. alterniflora*, and two species-specific polymorphic sites. The shared polymorphisms allow distinction of two divergent haplotypes (where all six polymorphic sites are different) present in both hexaploids, and one (in *S. maritima*) or two (in *S. alterniflora*) additional less divergent variants (one or two nucleotide difference). Although the number of polymorphic sites defining haplotypes is variable among the other analyzed contigs, we observed the same pattern distinguishing two divergent haplotypes and one or two less divergent variants within individuals (Table 2).

DISCUSSION

We have explored the transcriptome of two related *Spartina* species (*S. maritima* and *S. alterniflora*) using 454 sequencing technology. Before this study, only a limited number of *Spartina* ESTs were deposited in the NCBI EST database. If we exclude a recent transcriptome analysis in the tetraploid *Spartina pectinata* that generated 556 198 ESTs (Gedye *et al.*, 2010), a few hundred sequences only were available for *S. maritima* (Chelaifa *et al.*, 2010a) and *S. alterniflora* (Baisakh *et al.*, 2008). Our work represents the first effort to analyze the transcriptome of the hexaploid *Spartina* species, resulting in a reference transcriptome of more than 16 700 annotated genes from leaves and roots.

***De novo* transcriptome assembly using 454 sequencing technology**

Compared with other NGS technologies, the Roche platform offers long read lengths that facilitate assembly and annotation (Morozova *et al.*, 2009) and for this reason it is the most widely used technology for *de novo* EST sequencing (Sun *et al.*, 2010). In total, 25 239 (normalized and non-normalized libraries) and 14 317 contigs were assembled for *S. maritima* and *S. alterniflora*, respectively,

Table 2 Nucleotide polymorphisms detected among reads within four annotated contigs from *S. maritima* and *S. alterniflora*

Gene annotation	*Contig length*		*Number of reads*		*Number of polymorphisms*		*Number of haplotypes*	
	S. maritima	*S. alterniflora*	*S. maritima*	*S. alterniflora*	*S. maritima*	*S. alterniflora*	*S. maritima*	*S. alterniflora*
Phosphoenol-pyruvate Carboxykinase (LOC_Os03g15050.4|13103.m01762|cDNA)	632	470	132	50	8	3	4	3
HECT domain-containing protein, expressed (LOC_Os12g24080.1|13112.m02448|cDNA)	4294	3961	127	85	113	103	4	4
Homeobox domain-containing protein, expressed—Transcription factor MEIS1 and related HOX domain proteins (LOC_Os12g43950.4|13112.m08878|cDNA)	3031	2777	185	129	6	4	4	3
Heat shock protein, putative, expressed (LOC_Os06g50300.1|13106.m05403|cDNA)	3019	3391	67	263	42	5	4	4

representing 65.1% and 57.8% of the reads, the remaining of the reads left as singletons. Using a similar technology and assembly software, Gedye *et al.* (2010) assembled 65% of the reads into contigs for *S. pectinata*. The contig lengths found for both species are comparable to the length range reported in similar studies on other species (for example 299 bp in *Oryza longistaminata*, Yang *et al.* (2010); 394 bp in *S. pectinata*, Gedye *et al.* (2010); 526 bp in *Panax quinquefolius*, Sun *et al.* (2010)). From this data set, 17 307 contigs were annotated for *S. maritima* and 14 123 contigs were annotated for *S. alterniflora* (38 089 total annotated contigs for both species) corresponding to 16 753 different genes. These results are situated in the range of reported studies in non-model species (69.8% in ginseng; 72.6% in *S. pectinata*; 82% in amaranth; 85.5% in *Cicer*). Functional annotation could be assigned to 68.6% of the *S. maritima* contigs and 98.6% of the *S. alterniflora* contigs. Nonetheless, a large number of unique reads (singletons) were found, that is, 15% for our data set compared with other studies using the same assembler: 13% in *S. pectinata* (Gedye *et al.*, 2010); 10–25% in *Mytilus galloprovincialis* (Craft *et al.*, 2010); 8.8% in *Palomero* maize (Vega-Arreguin *et al.*, 2009) and 7% in *Amaranthus* and Ginseng (Sun *et al.*, 2010; Délano-Frier *et al.,* 2011). This could result from various causes such as the presence of rare transcripts from lowly expressed genes. The 454 sequencing technology also has some limitations resulting mainly from sequencing errors associated with homopolymers (Margulies *et al.*, 2005; Moore *et al.*, 2006; Wicker *et al.*, 2006), A/T bias (Moore *et al.*, 2006; Wicker *et al.*, 2006) or random nucleotide misincorporation (Huse *et al.*, 2007; Holt and Jones, 2008). The error rate for 454 sequencing is higher than the rate usually observed with Sanger sequencing (0.04 and 0.01%, respectively (Ewing and Green, 1998; Margulies *et al.*, 2005; Moore *et al.*, 2006)). Nevertheless, the error rate drops significantly to 0.4 bp errors per 10 kb after assembly (Margulies *et al.*, 2005; Moore *et al.*, 2006). We checked the quality of our sequence assemblies from 454 sequencing by comparing 10 assembled contigs to their putative homologs in *S. maritima* ESTs sequenced with the Sanger method (Chelaifa *et al.*, 2010a, b). The identity between the sequences was found very high (99.5%), which validates the procedures employed.

As there is no reference genome for *Spartina*, we used information from several EST and protein databases for gene annotation, a procedure successfully employed for other non-model species (for example, Barakat *et al.*, 2009; Gedye *et al.*, 2010; Franssen *et al.*, 2011; Garg *et al.*, 2011). In *de novo* sequencing projects transcriptome coverage efficiency has been evaluated by comparing the number of unique genes to the nearest transcriptome available (Parchman *et al.*, 2010). We compared our data to the nearest sequenced grass genomes: *Oryza sativa* (51 258 protein-coding transcripts, Yu *et al.*, 2005 and the Rice Genome Annotation project, http://rice.plantbiology.msu.edu//) and *Sorghum bicolor* (36 338 protein-coding transcripts, Paterson *et al.*, 2009). Using combined cDNA libraries, we identified 16 753 putative (non-redundant) genes by homology searches, which represent more than half of the genes found in fully sequenced related plant genomes. Interestingly, these genes appear distributed among the different *Sorghum* chromosomes, particularly in high gene density subtelomeric regions. Global gene colinearity is known to be well conserved among grass genomes (Feuillet and Keller, 2002; Srinivasachary *et al.*, 2007) and the comparison here between hexaploid *Spartina* and *Sorghum bicolor* validates the utilization of *Sorghum* as a comparative model, as first observed in Gedye *et al.* (2010) for *S. pectinata*. The percentage of contigs without a BLAST hit in our study is quite low (1.01%), with 389 contigs that did not match any putative homolog in the Poaceae database. This fraction varies among other studies fluctuating from 14.5% in *Cicer* (Garg *et al.*, 2011) to 30.2% in *Panax* (Sun *et al.*, 2010), for instance. These sequences without homology hit can be attributed to technical biases, such as low-quality data, inaccurate assembly, assembly parameters and contamination by genomic DNA. The causes can

Table 3 Single-nucleotide polymorphisms among assembled reads of the gene coding the HECT domain-containing protein in *Spartina maritima* and *Spartina alterniflora*

HECT domain-containing protein, expressed

S. alterniflora contig 03059 (length = 3961, reads = 85)						
Nucleotide position	1034	1085	1100	1119	1130	1167
Haplotype 1	C	T	C	A	A	T
Haplotype 2	T	T	T	A	A	C
Haplotype 3	T	C	T	A	A	C
Haplotype 4	C	T	C	G	C	T
S. maritima contig 02799 (length = 4294, reads = 127)						
Nucleotide position	1344	1352	1371	1382	1401	1419
Haplotype 1	C	C	G	C	C	T
Haplotype 2	T	C	G	C	C	T
Haplotype 3	T	T	A	A	A	C

Analysis of a 200-bp window, including two species-specific polymorphic sites (positions 1304, 1085 in *S. alterniflora* and positions 1344 and 1401 in *S. maritima*) and four polymorphic sites shared between the two species. These shared polymorphic sites are vertically aligned in the table.

also be biological: some cDNAs are non-coding, lineage-specific or highly variable (Logacheva *et al.*, 2011). Specific *Spartina* (or Chloridoideae) sequences also might be too divergent from the grass model species used.

In this study, among the 13 786 genes annotated in *S. maritima*, 6642 were retrieved in the normalized library, 3201 genes in the non-normalized and only 3620 genes overlapping both libraries, which indicates that normalization significantly improved the number of annotated genes. The normalization reduces oversampling of abundant transcripts and maximizes the potential to sequence less abundant transcripts (Zhulidov *et al.*, 2004). RNA-Seq studies on Zebra finch and rice have reported a higher efficiency in gene discovery using normalized cDNA libraries compared with non-normalized libraries (Yang *et al.*, 2010; Ekblom *et al.*, 2012). In contrast, Hale *et al.* (2009) demonstrated that normalization has a limited influence on increasing sequenced gene number. Ekblom *et al.* (2012) suggest that differences in technologies used and sequencing efforts can affect the outcome of the comparison between normalized and non-normalized libraries. In our present study, the normalized library was constructed from plants grown under natural conditions along a tidal gradient, which might also have increased the number of transcripts annotated. The transcriptome size, unknown in most non-model species may also affect the coverage and the sequencing effort. Therefore, it can affect indirectly the efficiency of normalization: normalized libraries show less efficiency when the non-normalized library already covers the whole transcriptome. This suggests that the combination of both normalized and non-normalized libraries is essential for gene discovery in non-model species, particularly in species exhibiting redundant genomes such as hexaploid *Spartina*.

Functional aspects: biology and ecology of *Spartina*

The 16 753 *Spartina* unigenes annotated in this study represent an important resource to explore genes involved in functions of ecological and adaptive interest. The genus *Spartina* exhibits a C4-type photosynthesis, which evolved in the Chloridoideae between 25 and 32 MYA (Christin *et al.*, 2008), and which uses the ATP-dependent phosphoenolpyruvate carboxykinase (PCK) as decarboxylating enzyme (Christin *et al.*, 2009). C4 metabolism confers high plant productivity under warm, arid and saline conditions, although *Spartina* species (and most particularly the hexaploids) colonize temperate regions (Long *et al.*, 1975). In the study conducted by Christin *et al.* (2009), one PCK-sequence-type was found in *S. maritima*, whereas two sequence types were found in *S. anglica*, one being sister to the *maritima*-type sequence and the other one most likely originating from the other parent of *S. anglica* (*S. alterniflora*, which was not analyzed by these authors). When analyzing an 830-bp partial PCK-coding region in *S. maritima* and *S. alterniflora,* Chelaifa *et al.* (2010a) found high nucleotide identity (99.7%) between *S. maritima* and *S. alterniflora*. In our study, a fragment of the *PCK* gene was found well represented in the leaf transcriptome of both *S. maritima* (623 bp) and *S. alterniflora* (470 bp), which is less than 25% of the total CDS length of *O. sativa* being 2820 bp but provides an indication of levels of heterogeneity. SNPs examined in this region revealed the presence of up to two haplotypes for each species. The identity between the two most divergent haplotypes of S. *maritima* was 98.5%, whereas the two less divergent sequences exhibited 99.4% identity. Our results then indicate that at least two different, putative homoeologous PCK sequences are expressed in the leaves of the hexaploid *S. maritima* and *S. alterniflora* species.

S. alterniflora and *S. maritima* are low-marsh species that have developed particular adaptation to tolerate several hours of immersion under seawater at high tide (Adams and Bate, 1995; Daehler and Strong, 1996). Survival of low-marsh *Spartina* species in anoxic sediments is facilitated by their ability to develop aerenchyma systems (studied particularly in *S. alterniflora*) that supply the submerged plants with atmospheric oxygen and efficiently transport oxygen to the roots (Maricle and Lee, 2002). High salinity can be damaging by salt toxicity and dehydration caused by low water potential. Thus, plants living in saline, high-light environments are adapted to minimize water loss to prevent dehydration, and have developed particular adaptive anatomical features with this regard (Maricle *et al.*, 2007). Salt marsh *Spartina* species have thick leaves with pronounced ridges on the adaxial side. They are adapted to controlling water loss by having stomata on the adaxial side and by having large leaf ridges that fit together as the leaf rolls during water stress (Maricle *et al.*, 2009). To prevent salt toxicity, *Spartina* have large vacuoles for salt storage (Munns and Tester, 2008) and salt-secreting glands to excrete inorganic ions (Zhu, 2001). Phenotypic adaptations are well documented but little is known about genes involved in these responses. The first *Spartina* transcriptome analyses under salt stress were performed in *S. alterniflora* using cDNA amplified fragment length polymorphism (Baisakh *et al.*, 2006) and EST analyses (Baisakh *et al.*, 2008); these analyses identified various transcripts involved in ion transport and compartmentalization, osmolyte production, cell division, metabolism and protein synthesis, as well as previously unknown genes induced by salt stress. Although our transcriptome analysis of *S. maritima* and *S. alterniflora* was not performed under salt stress, we retrieved 937 (4642 contigs) of the 1266 ESTs Baisakh *et al.* (2008) and Subudhi and Baisakh (2011) generated. Using *A. thaliana* as a functional reference transcriptome, we were also able to annotate 130 genes (305 contigs) involved in salt stress response. These genes include transcription factors, heat shock protein and cytochrome *c* oxidase that have been found to respond to salt and oxidative stress by balancing ion concentrations in *Spartina* (Maricle *et al.*, 2006).

We also annotated 71 genes (190 contigs) involved in heavy metal tolerance. *Spartina* species are of high interest regarding their ecological role in polluted coastal environment, where they exhibit particular tolerance to oil spill and where they are considered for phytoremediation purposes (Maricle and Lee, 2002; Martinez-Dominguez *et al.,* 2008; Mateo-Naranjo *et al.,* 2008). Ramana Rao *et al.* (2011) found 28 differentially expressed genes following experimental petroleum hydrocarbon exposure in *S. alterniflora*. We retrieved in our data set 8 of these genes (52 contigs).

Genes involved in stress response or in developmental and cellular growth were found to be differentially expressed in controlled conditions in the two species, the former being overexpressed in *S. maritima*, whereas the latter were overexpressed in *S. alterniflora* (Chelaifa *et al.*, 2010a). Most of these genes have also been found to be predominantly affected following hybridization between these species (that is, in *Spartina* × *townsendii*) and subsequent genome duplication in *S. anglica* (Chelaifa *et al.*, 2010b). Here, we identified 409 contigs corresponding to 271 different genes matching the putative homologous rice probes described in Chelaifa *et al.* (2010b). Our *Spartina* sequence data set may provide useful information to target genes of ecological and evolutionary interest (that is, whose expression is affected by divergent and reticulate speciation). Specific primers may be now designed to explore gene expression evolution in natural conditions and under various ecological situations.

Sequence polymorphism at homologous loci in hexaploid *Spartina*

The contigs assembled from the 454 reads in each of these hexaploid species actually represent a consensus sequence among strictly homologous (that is, orthologous) sequences but may also include homoeologous sequences (generated by polyploidy). Within homoelogues (at strictly orthologous loci), levels of heterozygosity have been poorly investigated in *S. maritima,* although this species is well known for its predominant clonal propagation and weak inter-individual genetic variation (Yannic *et al.*, 2004). *Spartina alterniflora* has a mixed, predominantly outcrossing mating system (Travis *et al.*, 2004); thus, more allelic variation within homoeologues might be expected than for *S. maritima*. Reads were assembled using a 90% identity threshold, to avoid potential comparisons involving divergent paralogs, but homoeologous sequences are expected to exhibit more similarity at each locus, and thus will most likely be aligned in the same contig. Homology assessment requires sequence comparison examined in a phylogenetic context. Such an analysis was performed for *Spartina* for the granule-bound starch synthase I (*Waxy*) gene (Fortune *et al.*, 2007). Molecular cloning, sequencing, and phylogenetic analyses allowed detection of paralogous, homoeologous and orthologous copies. In *S. alterniflora*, three homoeologous waxy copies were detected, exhibiting substitution rates ranging from 0.0218 to 0.0479. When analyzing sequence polymorphism among the assembled reads at four putative homologous loci between *S. maritima* and *S. alterniflora*, we found at each of these loci four different haplotypes that include two divergent sequences and two other less divergent variants. These results suggest the presence of two expressed homoeologous sequences with, respectively, two allelic variants; complementary phylogenetic analyses involving tetraploid *Spartina* species and outgroups will help to elucidate the evolutionary origin of these different sequences. As *S. maritima* and *S. alterniflora* are hexaploid, up to three duplicated homoeologs may be expected per locus. The fact that only two homoeologs were encountered in the analyzed transcripts might result from either homoeologous silencing as observed in the various cases of subfunctionalization reported in allopolyploids (reviewed in Osborn *et al.*, 2003; Adams and Wendel, 2005; Doyle *et al.*, 2008), from physical loss of the duplicated copies that may occur more or less rapidly following polyploid speciation (for example, Gaeta *et al.*, 2007; Tate *et al.*, 2009; Koh *et al.*, 2010) or from homoeologous recombination (Cifuentes *et al.*, 2010; Salmon *et al.*, 2010; Gaeta and Pires, 2010). For the *Waxy* gene mentioned above, Fortune *et al.* (2007) found a variable number of retained copies per homologous locus. Two paralogs (A and B) were identified in the genus *Spartina*, only one B copy was found in *S. maritima,* whereas three distinct B copies were encountered in *S. alterniflora.* The A copy was apparently lost in these two species but is maintained in the hexaploid *S. foliosa,* which is sister species to *S. alterniflora.*

CONCLUSIONS

NGS technologies open new opportunities to screen large sets of genes and their evolution in polyploid species (Buggs *et al.,* 2012). This first reference transcriptome, coupled with ongoing studies in our laboratory, involving deeper coverage from (Illumina INC., San Diego, CA, USA) RNA-Seq, and high-throughput genomic DNA sequencing, will facilitate a more accurate estimate of the level of duplicated homoeologous gene retention and relative expression in the hexaploid *Spartina* species and their hybrid and allopolyploid derivatives, in controlled and natural conditions. The analysis of the retained gene copies will also shed light into the origin of the hexaploid lineage and improve our understanding of the deepest *Spartina* history.

DATA ARCHIVING

Data have been deposited at Genbank (Sequence Read Archive SRA) under accession references SRP015701 and SRP015702 for *Spartina maritima* and *Spartina alterniflora*, respectively.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work is being developed in the frame of the International Associated Laboratory 'Ecological Genomics of Polyploidy' supported by CNRS (INEE, UMR CNRS 6553 Ecobio), University of Rennes 1 and Iowa State University (Ames, USA). Sequencing was supported by Genoscope funds (GENOSPART Project). This work benefited from BioGenouest (Environmental and Functional Genomics) and Genouest (Bioinformatics) Plateform facilities. J Ferreira de Carvalho benefited from a PhD grant (ARED EVOSPART) from the Regional Council of Brittany. B Mable, JF Wendel and one anonymous reviewer are thanked for their helpful comments on an earliest version of this manuscript.

Adams JB, Bate GC (1995). Ecological implications of tolerance of salinity and inundation by *Spartina maritima*. *Aquat Bot* **52**: 183–191.

Adams KL, Wendel JF (2005). Novel patterns of gene expression in polyploid plants. *Trends Genet* **21**: 539–543.

Ainouche ML, Baumel A, Salmon A (2004a). *Spartina anglica* C. E. Hubbard: a natural model system for analysing early evolutionary changes that affect allopolyploid genomes. *Biol J Linn Soc Lond* **82**: 475–484.

Ainouche ML, Baumel A, Salmon A, Yannic G (2004b). Hybridization, polyploidy and speciation in *Spartina* (Poaceae). *New Phytol* **161**: 165–172.

Ainouche ML, Chelaifa H, Ferreira de Carvalho J, Bellot S, Ainouche AK, Salmon A (2012). Polyploid evolution in *Spartina:* dealing with highly redundant genomes. In: Soltis PS, Soltis DE (eds). *Polyploidy and Genome Evolution.* Springer: Berlin, Heidelberg. pp 225–244.

Ainouche ML, Fortune PM, Salmon A, Parisod C, Grandbastien M-A, Fukunaga K *et al.* (2009). Hybridization, polyploidy and invasion: lessons from *Spartina* (Poaceae). *Biol Invasions* **11**: 1159–1173.

Alagna F, D'Agostino N, Torchia L, Servili M, Rao R, Pietrella M *et al.* (2009). Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics* **10**: 399.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Andersen GR, Nissen P, Nyborg J (2003). Elongation factors in protein biosynthesis. *Trends Biochem Sci* **28**: 434–441.

Ayres DR, Smith DL, Zaremba K, Klohr S, Strong DR (2004). Spread of exotic cordgrasses and hybrids (*Spartina sp.*) in the tidal marshes of San Francisco Bay, California, USA. *Biol Invasions* **6**: 221–231.

Baisakh N, Subudhi PK, Parami NP (2006). cDNA-AFLP analysis reveals differential gene expression in response to salt stress in a halophyte *Spartina alterniflora* Loisel. *Plant Sci* **170**: 1141–1149.

Baisakh N, Subudhi PK, Varadwaj P (2008). Primary responses to salt stress in a halophyte, smooth cordgrass (*Spartina alterniflora* Loisel.). *Funct Integr Genomics* **8**: 287–300.

Barakat A, DiLoreto D, Zhang Y, Smith C, Baier K, Powell W *et al.* (2009). Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. *BMC Plant Biol* **9**: 51.

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007). SNP discovery via 454 transcriptome sequencing. *Plant J* **51**: 910–918.

Baumel A, Ainouche M, Kalendar R, Schulman AH (2002a). Retrotransposons and genomic stability in populations of the young allopolyploid species *Spartina anglica* CE Hubbard (Poaceae). *Mol Biol Evol* **19**: 1218–1227.

Baumel A, Ainouche ML, Bayer RJ, Ainouche AK, Misset MT (2002b). Molecular phylogeny of hybridizing species from the genus *Spartina* Schreb. (Poaceae). *Mol Phylogenet Evol* **22**: 303–314.

Baumel A, Ainouche ML, Levasseur JE (2001). Molecular investigations in populations of *Spartina anglica* C.E. Hubbard (Poaceae) invading coastal Brittany (France). *Mol Ecol* **10**: 1689–1701.

Baumel A, Ainouche ML, Misset MT, Gourret JP, Bayer RJ (2003). Genetic evidence for hybridization between the native *Spartina maritima* and the introduced *Spartina alterniflora* (Poaceae) in South-West France: *Spartina* x *neyrautii* re-examined. *Plant Syst Evol* **237**: 87–97.

Bellin D, Ferrarini A, Chimento A, Kaiser O, Levenkova N, Bouffard P *et al.* (2009). Combining next-generation pyrosequencing with microarray for large scale expression analysis in non-model species. *BMC genomics* **10**: 555.

Buggs RJA, Chamala S, Wu W, Gao L, May GD, Schnable PS *et al.* (2010). Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Mol Ecol* **19**: 132–146.

Buggs RJA, Renny-Byfield S, Chester M, Jordon-Thaden IE, Viccini LF, Chamala S *et al.* (2012). Next-generation sequencing and genome evolution in allopolyploids. *Am J Bot* **99**: 372–382.

Bundock PC, Eliott FG, Ablett G, Benson AD, Casu RE, Aitken KS *et al.* (2009). Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant Biotechnol J* **7**: 347–354.

Campos JA, Herrera M, Biurrun I, Loidi J (2004). The role of alien plants in the natural coastal vegetation in central-northern Spain. *Biodivers Conserv* **13**: 2275–2293.

Castellanos E, Figueroa M, Davy A (1994). Nucleation and facilitation in salt-marsh succession—interactions between *Spartina maritima* and *Arthrocnemum perenne*. *J Ecol* **82**: 239–248.

Castillo JM, Leira-Doce P, Rubio-Casal AE, Figueroa E (2008). Spatial and temporal variations in aboveground and belowground biomass of *Spartina maritima* (small cordgrass) in created and natural marshes. *Estuar Coast Shelf Sci* **56**: 2037–2042.

Chelaifa H, Mahé F, Ainouche M (2010a). Transcriptome divergence between the hexaploid salt-marsh sister species *Spartina maritima* and *Spartina alterniflora* (Poaceae). *Mol Ecol* **19**: 2050–2063.

Chelaifa H, Monnier A, Ainouche M (2010b). Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species *Spartina* × *townsendii* and *Spartina anglica* (Poaceae). *New Phytol* **186**: 161–174.

Christin P-A, Petitpierre B, Salamin N, Büchi L, Besnard G (2009). Evolution of C4 phosphoenolpyruvate carboxykinase in grasses, from genotype to phenotype. *Mol Biol Evol* **26**: 357–365.

Christin PA, Besnard G, Samaritani E, Duvall MR, Hodkinson TR, Savolainen V *et al.* (2008). Oligocene CO2 decline promoted C4 photosynthesis in grasses. *Curr Biol* **18**: 37–43.

Cifuentes M, Grandont L, Moore G, Chèvre AM, Jenczewski E (2010). Genetic regulation of meiosis in polyploid species: new insights into an old question. *New Phytol* **186**: 29–36.

Civille JC, Sayce K, Smith SD, Strong DR (2005). Reconstructing a century of *Spartina alterniflora* invasion with historical records and contemporary remote sensing. *Ecoscience* **12**: 330–338.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**: 3674–3676.

Craft JA, Gilbert JA, Temperton B, Dempsey KE, Ashelford K, Tiwari B *et al.* (2010). Pyrosequencing of *Mytilus galloprovincialis* cDNAs: tissue-specific expression patterns. *PLoS One* **5**: e8875.

Daehler CC, Strong DR (1996). Status, prediction and prevention of introduced cordgrass *Spartina spp.* invasions in Pacific estuaries, USA. *Biol Conserv* **78**: 51–58.

Délano-Frier J, Aviles-Arnaut H, Casarrubias-Castillo K, Casique-Arroyo G, Castrillon-Arbelaez P, Herrera-Estrella L *et al.* (2011). Transcriptomic analysis of grain amaranth (*Amaranthus hypochondriacus*) using 454 pyrosequencing: comparison with *A. tuberculatus*, expression profiling in stems and in response to biotic and abiotic stress. *BMC genomics* **12**: 363.

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS *et al.* (2008). Evolutionary genetics of genome merger and doubling in plants. *Annu Rev Genet* **42**: 443–461.

Ekblom R, Galindo J (2010). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**: 1–15.

Ekblom R, Slate J, Horsburgh GJ, Birkhead T, Burke T (2012). Comparison between normalised and unnormalised 454-sequencing libraries for small-scale RNA-Seq studies. *Comp Funct Genomics* **2012**: 1–8.

Ewing B, Green P (1998). Base-calling of automated sequencer traces using Phred II error probabilities. *Genome Res* **8**: 186–194.

Ferris C, King RA, Gray AJ (1997). Molecular evidence for the maternal parentage in the hybrid origin of *Spartina anglica*. *Mol Ecol* **6**: 185–187.

Feuillet C, Keller B (2002). Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution. *AoB Plants* **89**: 3–10.

Fortune PM, Schierenbeck K, Ayres D, Bortolus A, Catrice O, Brown S *et al.* (2008). The enigmatic invasive *Spartina densiflora*: A history of hybridizations in a polyploidy context. *Mol Ecol* **17**: 4304–4316.

Fortune PM, Schierenbeck KA, Ainouche AK, Jacquemin J, Wendel JF, Ainouche ML (2007). Evolutionary dynamics of *Waxy* and the origin of hexaploid *Spartina* species (Poaceae). *Mol Phylogenet Evol* **43**: 1040–1055.

Franssen S, Shrestha R, Brautigam A, Bornberg-Bauer E, Weber A (2011). Comprehensive transcriptome analysis of the highly complex *Pisum sativum* genome using next generation sequencing. *BMC Genomics* **12**: 227.

Gaeta RT, Pires JC (2010). Homoeologous recombination in allopolyploids: the polyploid ratchet. *New Phytol* **186**: 18–28.

Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC (2007). Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. *Plant Cell* **19**: 3403–3417.

Garg R, Patel RK, Tyagi AK, Jain M (2011). *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res* **18**: 53.

Ge W, Song Y, Zhang C, Zhang Y, Burlingame AL, Guo Y (2011). Proteomic analyses of apoplastic proteins from germinating *Arabidopsis thaliana* pollen. *Biochim Biophys Acta* **1814**: 1964–1973.

Gedye K, Gonzalez-Hernandez J, Ban Y, Ge X, Thimmapuram J, Sun F *et al.* (2010). Investigation of the transcriptome of prairie cord grass, a new cellulosic biomass crop. *Plant Genome* **3**: 69.

Gotoh T, Terada K, Oyadomari S, Mori M (2004). Hsp70-DnaJ chaperone pair prevents nitric oxide- and CHOP-induced apoptosis by inhibiting translocation of Bax to mitochondria. *Cell Death Differ* **11**: 390–402.

Groves H, Groves J (1880). *Spartina townsendii* Nobis. *Rep Bot Soc Exch Club Bri Id* **1**: 37.

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ *et al.* (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* **36**: 3420–3435.

Hale MC, McCormick CR, Jackson JR, DeWoody JA (2009). Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery. *BMC Genomics* **10**: 203.

Holt RA, Jones SJM (2008). The new paradigm of flow cell sequencing. *Genome Res* **18**: 839–846.

Hudson ME (2008). Sequencing breakthroughs for genomic ecology and evolutionary biology. *Mol Ecol Resour* **8**: 3–17.

Humphreys JM, Chapple C (2002). Rewriting the lignin roadmap. *Curr Opin Plant Biol* **5**: 224–229.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch D (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8**: R143.

Ilut DC, Coate JE, Luciano AK, Owens TG, May GD, Farmer A *et al.* (2012). A comparative transcriptomic study of an allotetraploid and its diploid progenitors illustrates the unique advantages and challenges of RNA-seq in plant species. *Am J Bot* **99**: 383–396.

Koh J, Soltis P, Soltis D (2010). Homeolog loss and expression changes in natural populations of the recently and repeatedly formed allotetraploid *Tragopogon mirus* (Asteraceae). *BMC Genomics* **11**: 97.

Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D *et al.* (2009). Circos: an information aesthetic for comparative genomics. *Genome Res* **19**: 1639–1645.

Li B, Liao C-h, Zhang X-d, Chen H-l, Wang Q, Chen Z-y *et al.* (2009). *Spartina alterniflora* invasions in the Yangtze River estuary, China: an overview of current status and ecosystem effects. *Ecol Eng* **35**: 511–520.

Liu J-J, Ekramoddoullah AKM (2006). The family 10 of plant pathogenesis-related proteins: their structure, regulation, and function in response to biotic and abiotic stresses. *Physiol Mol Plant Pathol* **68**: 3–13.

Logacheva M, Kasianov A, Vinogradov D, Samigullin T, Gelfand M, Makeev V *et al.* (2011). *De novo* sequencing and characterization of floral transcriptome in two species of buckwheat (*Fagopyrum*). *BMC Genomics* **12**: 30.

Long SP, Incoll LD, Woolhouse HW (1975). C4 photosynthesis in plants from cool temperate regions, with particular reference to *Spartina* × *townsendii*. *Nature* **257**: 622–624.

Marchant C, Goodman P (1969). *Spartina maritima* (Curtis) Fernald. *J Ecol* **57**: 287–291.

Marchant CJ (1968). Evolution in *Spartina* (Gramineae). II. Chromosomes, basic relationships and the problem of *Spartina* × *townsendii*. *Biol J Linn Soc Lond* **60**: 381–409.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.

Maricle BR, Cobos DR, Campbell CS (2007). Biophysical and morphological leaf adaptations to drought and salinity in salt marsh grasses. *Environ Exp Bot* **60**: 458–467.

Maricle BR, Crosier JJ, Bussiere BC, Lee RW (2006). Respiratory enzyme activities correlate with anoxia tolerance in salt marsh grasses. *J Exp Mar Bio Ecol* **337**: 30–37.

Maricle BR, Koteyeva NK, Voznesenskaya EV, Thomasson JR, Edwards GE (2009). Diversity in leaf anatomy, and stomatal distribution and conductance, between salt marsh and freshwater species in the C4 genus *Spartina* (Poaceae). *New Phytol* **184**: 216–233.

Maricle BR, Lee RW (2002). Aerenchyma development and oxygen transport in the estuarine cordgrasses *Spartina alterniflora* and *S. anglica*. *Aquat Bot* **74**: 109–120.

Martinez-Dominguez D, Heras MA de las, Navarro F, Torronteras R, Cordoba F (2008). Efficiency of antioxidant response in *Spartina densiflora*: an adaptative success in a polluted environment. *Environ Exp Bot* **62**: 69–77.

Mateos-Naranjo E, Redondo-Gomez S, Cambrolle J, Luque T, Figueroa ME (2008). Growth and photosynthetic responses to zinc stress of an invasive cordgrass, *Spartina densiflora*. *Plant Biol* **10**: 754–762.

Mobberley DG (1956). Taxonomy and distribution of the genus *Spartina*. *Iowa State Coll J Sci* **30**: 471–574.

Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM *et al.* (2006). Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol* **6**: 17.

Morozova O, Hirst M, Marra MA (2009). Applications of new sequencing technologies for transcriptome analysis. *Annu Rev Genomics Hum Genet* **10**: 135–151.

Munns R, Tester M (2008). Mechanisms of salinity tolerance. *Annu Rev Plant Biol* **59**: 651–681.

Novaes E, Drost D, Farmerie W, Pappas G, Grattapaglia D, Sederoff R *et al.* (2008). High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**: 312.

Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, Lee H-S *et al.* (2003). Understanding mechanisms of novel gene expression in polyploids. *Trends Genet* **19**: 141–147.

Otto SP (2007). The evolutionary consequences of polyploidy. *Cell* **131**: 452–462.

Parchman T, Geist K, Grahnen J, Benkman C, Buerkle CA (2010). Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* **11**: 180.

Parisod C, Salmon A, Zerjal T, Tenaillon M, Grandbastien M, Ainouche M (2009). Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in *Spartina*. *New Phytol* **184**: 1003–1015.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H *et al.* (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**: 551–556.

Poroyko V, Hejlek LG, Spollen WG, Springer GK, Nguyen HT, Sharp RE *et al*. (2005). The maize root transcriptome by serial analysis of gene expression. *Plant Physiol* **138**: 1700–1710.

Querné J, Ragueneau O, Poupart N (2011). *In situ* biogenic silica variations in the invasive salt marsh plant, *Spartina alterniflora*: A possible link with environmental stress. *Plant Soil* **352**: 157–171.

R Development Core Team (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria, ISBN 3-900051-07-0, Available at. http://www.R-project.org/.

Ramana Rao MV, Weindorf D, Breitenbeck G, Baisakh N (2011). Differential expression of the transcripts of *Spartina alterniflora* Loisel. (smooth cordgrass) induced in response to petroleum hydrocarbon. *Mol Biotechnol* **51**: 18–26.

Raybould AF, Gray AJ, Lawrence MJ, Marshall DF (1991). The evolution of *Spartina anglica* CE Hubbard (Gramineae): Origin and genetic-variability. *Biol J Linn Soc Lond* **43**: 111–126.

Salmon A, Ainouche ML, Wendel JF (2005). Genetic and epigenetic consequences of recent hybridization and polyploidy in *Spartina* (Poaceae). *Mol Ecol* **14**: 1163–1175.

Salmon A, Flagel L, Ying B, Udall JA, Wendel JF (2010). Homoeologous nonreciprocal recombination in polyploid cotton. *New Phytol* **186**: 123–134.

Song Y, Zhang C, Ge W, Zhang Y, Burlingame AL, Guo Y (2011). Identification of NaCl stress-responsive apoplastic proteins in rice shoot stems by 2D-DIGE. *J Proteomics* **74**: 1045–1067.

Srinivasachary S, Dida M, Gale M, Devos K (2007). Comparative analyses reveal high levels of conserved colinearity between the finger millet and rice genomes. *Theor Appl Genet* **115**: 489–499.

Subudhi PK, Baisakh N (2011). *Spartina alterniflora* Loisel, a halophyte grass model to dissect salt stress tolerance *in vitro*. *Cell Dev Biol Plant* **47**: 441–457.

Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J *et al.* (2010). *De novo* sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. *BMC Genomics* **11**: 262.

Swarbreck SM, Lindquist EA, Ackerly DD, Andersen GL (2011). Analysis of leaf and root transcriptomes of soil-grown *Avena barbata* plants. *Plant Cell Physiol* **52**: 317.

Takeda H, Kotake T, Nakagawa N, Sakurai N, Nevins DJ (2003). Expression and function of cell wall-bound cationic peroxidase in asparagus somatic embryogenesis. *Plant Physiol* **131**: 1765–1774.

Tate J, Joshi P, Soltis K, Soltis P, Soltis D (2009). On the road to diploidization? Homoeolog loss in independently formed populations of the allopolyploid *Tragopogon miscellus* (Asteraceae). *BMC Plant Biol* **9**: 80.

Tian L, Zhang L, Zhang J, Song Y, Guo Y (2009). Differential proteomic analysis of soluble extracellular proteins reveals the cysteine protease and cystatin involved in suspension-cultured cell proliferation in rice. *Biochim Biophys Acta* **1794**: 459–467.

Travis SE, Proffitt CE, Ritland K (2004). Population structure and inbreeding vary with successional stage in created *Spartina alterniflora* marshes. *Ecol Appl* **14**: 1189–1202.

Vega-Arreguin J, Ibarra-Laclette E, Jimenez-Moraila B, Martinez O, Vielle-Calzada J, Herrera-Estrella L *et al.* (2009). Deep sampling of the Palomero maize transcriptome by a high throughput strategy of pyrosequencing. *BMC Genomics* **10**: 299.

Wheat CW (2008). Rapidly developing functional genomics in ecological model systems *via* 454 transcriptome sequencing. *Genetica* **138**: 433–451.

Wicker T, Schlagenhauf E, Graner A, Close T, Keller B, Stein N (2006). 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* **7**: 275.

Yang H, Hu L, Hurek T, Reinhold-Hurek B (2010). Global characterization of the root transcriptome of a wild species of rice, *Oryza longistaminata*, by deep sequencing. *BMC Genomics* **11**: 705.

Yannic G, Baumel A, Ainouche M (2004). Uniformity of the nuclear and chloroplast genomes of *Spartina maritima* (Poaceae), a salt-marsh species in decline along the Western European Coast. *Heredity* **93**: 182–188.

Yoo M-J, Szadkowski E, Wendel JF (2012). Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* (doi:10.1038/hdy.2012.94).

Yu J, Wang J, Lin W, Li S, Li H, Zhou J *et al*. (2005). The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol* **3**: e38.

Zhu JK (2001). Plant salt tolerance. *Trends Plant Sci* **6**: 66–71.

Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB *et al*. (2004). Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res* **32**: e37.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)

Plant Mol Biol
DOI 10.1007/s11103-013-0111-7

Exploring the genome of the salt-marsh *Spartina maritima* (Poaceae, Chloridoideae) through BAC end sequence analysis

J. Ferreira de Carvalho · H. Chelaifa · J. Boutte · J. Poulain · A. Couloux · P. Wincker · A. Bellec · J. Fourment · H. Bergès · A. Salmon · M. Ainouche

Received: 28 February 2013 / Accepted: 13 July 2013

Abstract *Spartina* species play an important ecological role on salt marshes. *Spartina maritima* is an Old-World species distributed along the European and North-African Atlantic coasts. This hexaploid species (2n = 6x = 60, 2C = 3,700 Mb) hybridized with different *Spartina* species introduced from the American coasts, which resulted in the formation of new invasive hybrids and allopolyploids. Thus, *S. maritima* raises evolutionary and ecological interests. However, genomic information is dramatically lacking in this genus. In an effort to develop genomic resources, we analysed 40,641 high-quality bacterial artificial chromosome-end sequences (BESs), representing 26.7 Mb of the *S. maritima* genome. BESs were searched for sequence homology against known databases. A fraction of 16.91 % of the BESs represents known repeats including a majority of long terminal repeat (LTR) retrotransposons (13.67 %). Non-LTR retrotransposons represent 0.75 %, DNA transposons 0.99 %, whereas small RNA, simple repeats and low-complexity sequences account for 1.38 % of the analysed BESs. In addition, 4,285 simple sequence repeats were detected. Using the coding sequence database of *Sorghum bicolor*, 6,809 BESs found homology accounting for 17.1 % of all BESs. Comparative genomics with related genera reveals that the microsynteny is better conserved with *S. bicolor* compared to other sequenced Poaceae, where 37.6 % of the paired matching BESs are correctly orientated on the chromosomes. We did not observe large macrosyntenic rearrangements using the mapping strategy employed. However, some regions appeared to have experienced rearrangements when comparing *Spartina* to *Sorghum* and to *Oryza*. This work represents the first overview of *S. maritima* genome regarding the respective coding and repetitive components. The syntenic relationships with other grass genomes examined here help clarifying evolution in Poaceae, *S. maritima* being a part of the poorly-known Chloridoideae sub-family.

Keywords *Spartina maritima* · Chloridoideae · BAC-end sequences · Repetitive DNA · Genome evolution

Electronic supplementary material The online version of this article (doi:10.1007/s11103-013-0111-7) contains supplementary material, which is available to authorized users.

J. Ferreira de Carvalho · H. Chelaifa · J. Boutte · A. Salmon · M. Ainouche (✉)
UMR CNRS 6553 ECOBIO, OSUR, University of Rennes 1, Bât 14A Campus Scientifique de Beaulieu, 35042 Rennes Cedex, France
e-mail: malika.ainouche@univ-rennes1.fr

J. Poulain · A. Couloux · P. Wincker
Genoscope, 2 rue Gaston Crémieux, 91000 Evry, France

A. Bellec · J. Fourment · H. Bergès
Centre National de Ressources génomiques végétales-INRA, 24 Chemin de Borde Rouge, CS 52627, 31326 Castanet Tolosan Cedex, France

Introduction

The grass (Poaceae) genus *Spartina* is member of the Chloridoideae subfamily, an important group with more than 400 species in approximately 140 genera exhibiting a worldwide distribution (Peterson et al. 2010), but remarkably poorly-investigated with regard to genomic information. Chloridoideae belong to the PACMAD (Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae and Danthonioideae) clade (Grass Phylogeny Working Group GPWG II 2012). So far, genomic efforts have concentrated on three economically important grass subfamilies,

Published online: 23 July 2013

the Panicoideae (containing maize, sorghum, and sugarcane), the Ehrhartoideae (rice) and Pooideae (wheat, *Brachypodium*). Divergence times between Chloridoideae and Panicoideae were estimated about 34.6–38.5 million years ago and about 40–60 million years ago between Chloridoideae and the Erhartoideae–Pooideae respectively (Christin et al. 2008; Kim et al. 2009; Prasad et al. 2011). Phylogenetic relationships among Chloridoideae genera are not fully resolved and still under debate (Hilu and Alice 2001; Peterson et al. 2010). Most species exhibit C4-type metabolism, which confers higher productivity under warm, saline or arid conditions (Christin et al. 2009). Common base chromosome number is x = 10, sometimes 9 (Roodt and Spies 2003a), with widespread polyploidy and hybridization (Roodt and Spies 2003b). Genomic organization in Chloridoideae is particularly poorly known: Only a few studies have resulted in genetic maps for tropical crops such as finger millet *Eleusinecoracana* (Dida et al. 2006; Srinivasachary et al. 2007) or *Eragrostis tef* (Zhang et al. 2001; Yu et al. 2006). Recent but still limited transcriptome analyses have contributed to expressed sequence databases and gene annotation in the turfgrass *Cynodon dactylon* (Kim et al. 2008), or the saltmarsh species *Spartina alterniflora* (Baisakh et al. 2008; Ferreira de Carvalho et al. 2013), *Spartina maritima* (Ferreira de Carvalho et al. 2013) and the prairie cord grass *Spartina pectinata* (Gedye et al. 2010).

The *Spartina* genus is attracting a growing interest for various fundamental and economical perspectives. *Spartina* species play an important ecological role in the saltmarsh dynamics by protecting the coastline from erosion and modifying the physical structure of intertidal coastal zones where they are considered as "ecosystem engineers". Some species (Larher et al. 1977; Otte et al. 2004) are able to produce DMSP (dimethylsulfoniopropionate). This putative osmoprotectant molecule plays an important ecological role as it is a precursor of DMS (dimethylsulfide) released in the atmosphere where it contributes to cloud formation. Moreover, some *Spartina* species have gained attention as suitable crop with high cellulosic biomass for producing biofuel (Gonzalez-Hernandez et al. 2009). They also proved to be useful for phytoremediation purposes: they are able to tolerate heavy metal pollution and hydrocarbon (Lee 2003; Cambrollé et al. 2008; Ramanarao et al. 2011). Also, electricity production using *Spartina* microbial fuel cells seems promising as a new sustainable technology (Timmers et al. 2010).

From a fundamental perspective, the *Spartina* genus offers many opportunities in evolutionary ecology, in studies on polyploid speciation (Ainouche et al. 2004a) and to understand biological invasion processes following interspecific hybridization (Ayres et al. 2004; Ainouche et al. 2009). This genus is composed of 13–15 perennial species, (Mobberley 1956) with ploidy levels ranging from tetraploid (2n = 40) to dodecaploid (2n = 120–24) levels (reviewed in Ainouche et al. 2012). In recent molecular phylogenies, *Spartina* appears closely related to *Sporobolus* and *Calamovilfa* representatives (Peterson et al. 2010). The genus evolved through two main lineages respectively tetraploid and hexaploid (Baumel et al. 2002a; Fortuné et al. 2007) that diverged sometimes between 7–11 MYA, as estimated from chloroplast sequences (Bellot et al. in prep). Recurrent events of hybridization and polyploidy have arisen within and between these two lineages, and include one of the best documented example of recent allopolyploid speciation (reviewed in Ainouche et al. 2004b, 2009). The unintentional introduction of the native American species *Spartina alterniflora* (hexaploid, 2n = 62) to Western Europe and its subsequent hybridization (as maternal genome donor, Ferris et al. 1997; Baumel et al. 2001, 2003) with the native European *S. maritima* (hexaploid 2n = 60), resulting in two independently formed hybrids. In England, hybridization resulted in *Spartina* x *townsendii,* a perennial sterile hybrid first recorded around 1870 (Groves and Groves 1880), and still forming a vigorous population (Renny-Byfield et al. 2010) that gave rise (by chromosome doubling) around 1890 to a fertile and highly invasive allo-dodecapolyploid species *Spartina anglica,* which is now introduced on several continents. In South-west France, hybridization between *S. alterniflora* and *S. maritima* resulted in another sterile hybrid, *S. x neyrautii* which is still surviving in spite of severe habitat destruction (Baumel et al. 2003). This system is now used to explore early evolutionary changes following interspecific hybridization and whole genome duplication, and the genomic determinants of biological invasion (Ainouche et al. 2004a, b, 2009, 2012 and references therein).

In the perspectives of exploring the genome of these species, we have first chosen the Euro-African native hexaploid species *Spartina maritima,* which is involved in the paternal parentage of the hybrids and newly formed invasive allopolyploid *S. anglica*. *Spartina maritima* is usually confined to open habitat of short and long-established salt marshes, but also soft mud of low-marsh flooded at every high tide (Marchant 1967). Therefore, *S. maritima* is able to tolerate a wide range of substrates including lower marshes and long period of flooding (Marchant 1967; Castillo et al. 2000). Studies on the role of *S. maritima* in phytostabilization show a high potential to retain heavy metals such as cobalt, chromium and nickel in the rhizosphere (in Spanish estuaries: Luque et al. 1999; Cambrollé et al. 2008; and Portuguese salt marshes: Caetano et al. 2008). Moreover, *S. maritima* is able to accumulate cobalt in roots as well as copper, zinc and iron in leaves (Cambrollé et al. 2008). *Spartina* species function as excluders (Alberts et al. 1990) through external or internal exclusion mechanisms to delay translocation of heavy metals in the leaves (Hansel et al. 2001). In Southern

England and Brittany, native populations are currently regressing in its northern range limit. This is interpreted as a consequence of climate change and anthropogenic habitat disturbance (Raybould et al. 1991) but has also to be related with its biological and morphological traits. S*partina maritima* is a non-rhizomatous, genetically depauperate species (Yannic et al. 2004) with very low seed production (Marchant and Goodman 1969; Castellanos et al. 1994; Castillo et al. 2010).

Complementing ongoing studies at the transcriptome level (Ferreira de Carvalho et al. 2013), we take here advantage of a BAC (Bacterial Artificial Chromosome) library constructed for *S. maritima* by analyzing 40,641 BES to provide a first glimpse on the *Spartina* genome composition. This study represents the first large genomic investigation performed for *Spartina* species. The analyses focused on the detection of repeated elements, microsatellite and protein coding regions content (Fig. 1). Additionally, comparisons with related plant lineages of the grass family (rice, *Sorghum* and *Brachypodium)* provide new insights into the evolution of a Chloridoideae subfamily representative, then contributing filling a gap regarding this poorly investigated lineage.

Materials and methods

BAC library construction

Spartina maritima individuals were sampled on the Etel river marshes (Presqu'île du Verdon, Morbihan, France) and transferred into pots in the greenhouse. As *S. maritima* populations are genetically depauperate in Western Europe with low inter-individual genetic variation and predominant vegetative propagation (Yannic et al. 2004), the sampled plants are expected to represent the same genetic background. About 40 g of etiolated young leaves were collected, kept in liquid nitrogen and stored at −80 °C until DNA extraction for the construction of the BAC library at the Centre National des Ressources Génomiques Végétales (CNRGV, Toulouse, France). High Molecular Weight (HMW) DNA was prepared from leaves of *Spartina maritima*. Approximately 20 g of frozen leaf tissue was ground to powder in liquid nitrogen with a mortar and pestle used to prepare megabase-size DNA embedded in agarose plugs. HMW DNA was prepared as described by Peterson et al. (2000) and modified as described in Gonthier et al. (2010). Embedded HMW DNA was partially digested with HindIII (New England Biolabs, Ipswich, Massachusetts), subjected to two size selection steps by pulsed-field electrophoresis, using a BioRad CHEF Mapper system (Bio-Rad Laboratories, Hercules, California), and ligated to pIndigoBAC-5 HindIII-Cloning Ready vector (Epicentre Biotecnologies, Madison, Wisconsin). Pulsed-field migration programs, electrophoresis buffer, and ligation desalting conditions were performed according to Chalhoub et al. (2004). To evaluate the average insert size of each library, BAC DNA was isolated from about 384 randomly selected clones in each library, restriction enzyme digested with the rare cutter NotI, and analyzed by Pulsed-Field Gel Electrophoresis (PFGE). All fragments generated by NotI digestion contained the 7.5 kb vector band and various insert fragments. In total, 44,544 clones with a mean insert size of 110 kb were retained, representing 4,900 Mb or 1.5X the genome of *S. maritima* (3,700 Mb, estimated from Fortuné et al. 2008). As this genome is hexaploid, the BAC library would represent 8X the basic genome (estimated as 616 Mb if we assume equivalent genome size of the 3 duplicated homoeologous genomes). More than 20,000 paired BAC-ends were sequenced by the Genoscope (Evry, France) using the BigDye Termination kit on Applied Biosystems 3730xl DNA Analysers.

Organellar DNA content

To identify organellar DNA sequences, BESs were first compared to the *Oryza sativa indica* and *Sorghum bicolor* chloroplast and mitochondrial genomes (NC_008155.1, NC_007886.1, NC_008602.1 and NC_008360.1 downloaded from the NCBI website) using BLASTn with a stringent threshold of 10^{-6} and a minimum hit length of 70 bp. BESs were also compared to the assembled chloroplast genome of *S. maritima* from 454 Roche pyrosequencing data (Bellot et al. in prep).

Identification of repetitive sequences

A survey of the composition in repeat sequences of *Spartina maritima* was performed using RepeatMasker version 3.2.9 (http://www.repeatmasker.org/) with *Oryza sativa* as the query species in Repbase (Jurka et al. 2005). BESs were annotated based on their best match to the repeat database and categorized according to the reference database used.

All BESs containing retro-elements were extracted and aligned (BLASTx with an e-value of 10^{-6}) to the Repbase database (Jurka et al. 2005) including Reverse Transcriptase (RT) protein sequences from *Copia*-like and *Gypsy*-like elements. *Spartina maritima* RT sequences were then translated into proteins and aligned against Repbase RT sequences. The alignments were conducted using MUSCLE (Edgar 2004) and a maximum number of iterations of 8. *Copia*-like and *Gypsy*-like elements were analysed separately because of the high divergence between their RT

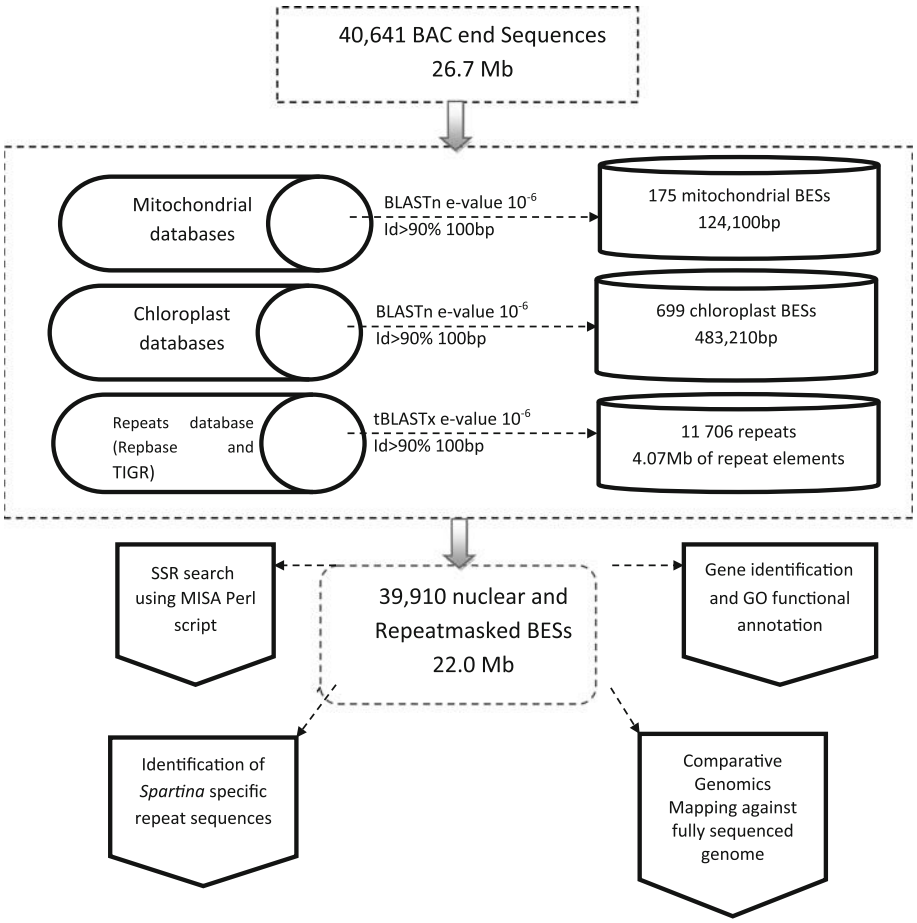


Fig. 1 Analyses conducted on the BAC-end sequences

domains. Phylogenetic analyses were performed using Geneious tree builder (Biomatters) with the Jukes-Cantor model and the Neighbour-joining method.

The BESs were also compared with Gramineae v3.3, *O. sativa* v3.3 and *S. bicolor* v3.0 databases downloaded from TIGR Plant repeat Databases (plantrepeats.plantbiology.msu.edu: Ouyang and Bell 2004). BLASTn analyses were conducted using an e-value cut off of 10^{-6} and a minimum hit length of 100 bp.

Simple sequence repeat (SSR) detection and primer design

Microsatellites were detected using the MISA perl script (MIcroSatellite research tool, Thiel et al. 2003). Parameters were set to find all SSRs with a motif length from one to six nucleotides (i.e. mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats). SSR parameters were at least ten nucleotide long for mononucleotides, 12 for dinucleotides, 15 for trinucleotides, 20 for tetranucleotides, 25 for pentanucleotides and 30 for hexanucleotide motifs. The maximal number of bases interrupting two SSRs was set to 100 bp. The fasta file with BES containing SSR sequences was uploaded on the BatchPrimer3 web interface to design specific SSR primers (You et al. 2008). The criteria used for designing primer pairs included an optimum annealing temperature of 55°C, amplicon size range of 100–300 bp with an average of 150 bp, primer length optimum of 21 ± 2 bp and GC % 50 ± 5 as suggested for SSR primer design (Bohra et al. 2011).

De novo identification of *Spartina* repeats

The masked file output from RepeatMasker (containing 39,910 sequences excised from repetitive elements and representing 26.2 Mb) was self-blasted with a highly-stringent e-value (10^{-50}) to find potential novel uncharacterized repeat sequences from *S. maritima* genome. Sequences with at least six hits and a minimum of 90 % identity were then blasted against the NCBI GenBank non-redundant nucleic acid sequence database, the SwissProt database and a Poaceae EST database (including ESTs from *Zea mays*, *Brachypodium distachyon*, *Sorghum bicolor* and *Oryza sativa*) to find *Spartina* specific sequences. We also compared these sequences to different repeat databases namely TIGR Plant Repeat Databases

including Gramineae v3.3, *Zea mays* v3.0, *Oryza sativa* v3.3 and *Sorghum bicolor* v3.0 repeat sequences, RepBase (Jurka et al. 2005) and TREP database (wheat.pw.usda.gov/ITMI/Repeats/) using BLASTn and an e-value cut off of 10^{-6} to assess their unique nature. BESs with no blast hits were then assembled using the Roche software (GS De Novo Assembler v. 2.5.3, Roche) with the following parameters: 90 % identity and a minimum overlap of 40 nucleotides.

Gene content and functional annotation

BESs were masked for repeat sequences and low-complexity sequences with RepeatMasker v3.2.9 as described above. The masked BESs (39,910 sequences) were then compared to coding sequences of *Oryza sativa* and *Sorghum bicolor* (version 120 and 79 respectively, downloaded from www.phytozome.com). For all tBLASTx searches, an e-value cut off of 10^{-6} was used. The BESs showing homology with *Sorghum bicolor* transcripts were then analysed with the BLAST2GO software (Conesa et al. 2005; Götz et al. 2008) to assign GO terms. BLASTx alignments were conducted using the non-redundant database of NCBI and a 10^{-6} stringency. In parallel, the BESs were compared against the reference transcriptome of five *Spartina* species (Ferreira de Carvalho et al. 2013; Ferreira de Carvalho et al. *unpublished*). The reference transcriptome was built using 454 technology cDNA sequencing from 5 species of *Spartina*: *S. maritima*, *S. alterniflora*, *S.* x *townsendii*, *S.* x *neyrautii* and *S. anglica*. From the 420 Mb sequenced, 52,347 contigs were assembled using the Roche Software GS De Novo Assembler and annotated following the method described in Ferreira de Carvalho et al. (2013).

Comparative genome mapping

To explore areas of potential microsynteny between *Spartina maritima* and selected model plants, all 39,910 masked BESs were mapped to the sequenced genomes of *Arabidopsis thaliana*, *Brachypodium distachyon*, *Oryza sativa* and *Sorghum bicolor* (Athaliana_167.fa, Bdistachyon_192_hardmasked.fa, Sbicolor_79_RM.fa and Osativa_120_RM.fa downloaded from www.phytozome.com). The e-value cut off was set to 10^{-6} and best blast hits were retained if they had a minimum identity of 70 %. A given BAC was then considered collinear to the targeted genome if both ends were correctly orientated within 15–250 kb of each other on the same chromosome. Otherwise, the region was considered rearranged between the two species. The synteny between *Spartina maritima* and *Sorghum bicolor*, and between *Spartina maritima* and *Oryza sativa* was visualized using the CIRCOS program (V.0.55, Krzywinski et al. 2009). BESs showing a hit with the repeatmasked genome of *Sorghum bicolor* were mapped onto the 10 chromosomes using BLASTn (e-value of 10^{-6} and a minimum identity of 70 %). Similar comparisons were performed between *Spartina maritima* and the 12 chromosomes of *Oryza sativa.*

Results

After trimming BES for vector and low read quality sequences, 40,641 BAC ends were retained for further analyses. Among those, 37,354 sequences were paired-end (Table 1). The BESs ranged in size from 57 to 938 bp with an average of 656 bp corresponding to a total of 26,682,959 nucleotides that would represent about 4.3 % of the basic genome of *Spartina maritima* (x = 10,616 Mb assuming equivalent genome size of the 3 duplicated genomes in this hexaploid species). The GC content is of 45.6 %.

On the 40,641 BAC end sequences aligned against chloroplast databases, 699 found a match with the *Spartina maritima* chloroplast genome (representing 1.72 % of the BESs) (Table 1). Respectively, 683 (1.68 %) and 668 (1.64 %) BESs matched with the *S. bicolor* and the *O. sativa* chloroplast genomes. Regarding the mitochondrial genome, 175 (0.43 %) and 91 (0.22 %) BESs were found in comparison with the *S. bicolor* and *O. sativa* genomes, respectively (Table 1). When combining the two largest sets of blasted sequences (chloroplast sequences from *S. maritima* and mitochondrial sequences from *O. sativa*), 731 sequences are retrieved representing 1.80 % from the original BESs database. In total, 39,910 BESs were analysed in the following steps (Fig. 1).

Table 1 Summary of BAC end sequencing

Total number of BES	40,641
Number of paired BES	37,354
Number of non-paired BES	3,287
Total number of nucleotides (bp)	26,682,959
Mean length (bp)	656
Range size (bp)	57–938
GC content	45.62 %
Chloroplast matches (Nb of hits)	
Spartina maritima	699 (1.72 % of BES)
Sorghum bicolor	683 (1.68 % of BES)
Oryza sativa indica	668 (1.64 % of BES)
Mitochondrion matches (Nb of hits)	
Sorghum bicolor	175 (0.43 % of BES)
Oryza sativa indica	91 (0.22 % of BES)

Repetitive DNA content and composition

The 39,910 *Spartina maritima* BESs were compared to different databases of known repeat elements to identify repeat sequences from similarity searches. The first analysis was conducted with RepeatMasker. Class I (retrotransposons) elements are predominant among the *Spartina* repeat sequences and represent a significant portion (14.42 %) of the BESs analysed (Table 2). Class I elements can be subclassified into long terminal repeat (LTR elements) and non-LTR retrotransposons. LTR retrotransposons represent 13.67 % of the BESs analysed. Non-LTR retrotransposons represented by short interspersed elements (SINEs, 0.02 %) and long interspersed elements (LINEs, 0.73 %) are less abundant, accounting for 0.75 % of the BESs.

As LTR elements represent a large proportion of the repeat sequences present in the genome of *Spartina maritima,* we conducted a phylogenetic analysis of the different families of *Copia* and *Gypsy*-like elements. Respectively, 739 and 884 protein sequences were extracted from the *Copia*-like and *Gypsy*-like dataset of BESs. Sequences of at least 400 bp long were retained to build the trees. In the *Copia* analysis, 211 *Spartina maritima* sequences are aligned with 722 RT protein sequences from RepBase (Fig. 2). *Spartina maritima* RT sequences identified with red branches are present in the Ivana-Oryco, Maximus, and Hopscotch clades and at the base of the lineage including the Angela, Tar and Tork clades. The larger number of repeats is in the Hopscotch clade with the *Hopscotch* (previously found in *Oryza sativa*), *Shacop20* (*Medicago truncatula*), *Castor* (*Arabidopsis thaliana*) and *Retrofit* (*Oryza longistamina*) elements. The Maximus clade is also well-represented with a specific branch of *S. maritima* RT sequences. In the *Gypsy* tree, 123 sequences are aligned with 163 RT protein sequences from Repbase. The tree is partitioned into three clades including *Athila*, *Tat-Ogre* and Chromovirus elements (Fig. 3). *Spartina maritima* RTs are predominantly present in the *Tat* lineage, with *Grande1* and *ACinful* elements previously found in the genus *Zea*. The second most represented lineage is composed of *Tekay* chromoviruses including *Sukkala* (*Hordeum vulgare*) and RIRE3 (*O. sativa*) elements.

Among the Class II DNA transposons (0.99 %) the most abundant elements are from the sub-class *En-Spm* corresponding to 0.58 % of the BESs. The Superfamily *Tc1-IS630-Pogo* is represented by 198 sequences accounting for 0.13 % of the BESs. The *hobo-activator* superfamily is also represented, accounting for 0.10 % of the BESs, as well as the *MuDR-IS905* superfamily (0.09 %). A total of 65 Miniature Inverted Repeat transposable elements (MITEs) from the Superfamily *Tourist/Harbinger* are identified in the dataset representing 0.06 % of the genomic sequences analysed. With other repetitive elements present in the Repbase database, such as small RNA (0.78 %), simple repeats (0.21 %) and low complexity sequences (0.39 %), the total of known repeat elements in the genomic sequences of *Spartina maritima* corresponds to 16.91 %.

In parallel, the 39,910 BESs were also aligned against the TIGR databases using tblastx and a cut-off e-value of 10^{-6}. We found 12,481 hits against the Gramineae database, 10,479 against the *O. sativa* database and 6,440 against the *S. bicolor* database (data not shown). This is consistent with the number of repeat elements found using RepeatMasker.

To identify de novo repetitive sequences in the *Spartina maritima* genome a self-blast analysis was conducted on the sequences first filtered with RepeatMasker. Self-blastn analysis of repeatmasked BESs revealed 8,146 sequences (20.4 % of BESs) with at least six hits (Fig. 4). This dataset was then blasted against the non-redundant GenBank database and 1,915 BESs found a hit. Among those, 196 sequences found also a hit in the Uniprot protein database. Then, homologies were searched against known repeat elements databases. In total, 79 BESs correspond to known repeat sequences and 22 BESs show homology with the ESTs Poaceae database. At the end, 6,145 BESs (representing 14.97 % of nucleotides) remained with unknown annotation representing potential novel repeat sequences from the *Spartina maritima* genome. Among these, 4,324 (representing 2.7 Mb) BESs were assembled into 272 contigs (containing 1,826 BESs and representing 858,686 bp) and 2,498 BESs resulted as singletons.

A total of 4,285 simple sequence repeats (SSRs) were detected in the 26.18 Mb of *Spartina maritima* BESs,

Table 2 Classification and distribution of known plant repeats in the BAC end sequences

Class	Number of elements	% Of nucleotides	Length (bp)
Retroelements	10,582	14.42	3,774,403
SINEs	34	0.02	5,267
LINEs (L1/CIN4)	590	0.73	191,083
LTR elements	9,958	13.67	3,578,053
Ty1/Copia	4,122	5.45	1,425,752
Gypsy/DIRS1	5,630	8.16	2,137,673
DNA transposons	1,019	0.99	258,029
Unclassified	105	0.13	33,285
Total interspersed repeats		15.53	4,065,717
Small RNA	332	0.78	203,886
Simple repeats	1,043	0.21	54,101
Low complexity	2,114	0.39	102,517

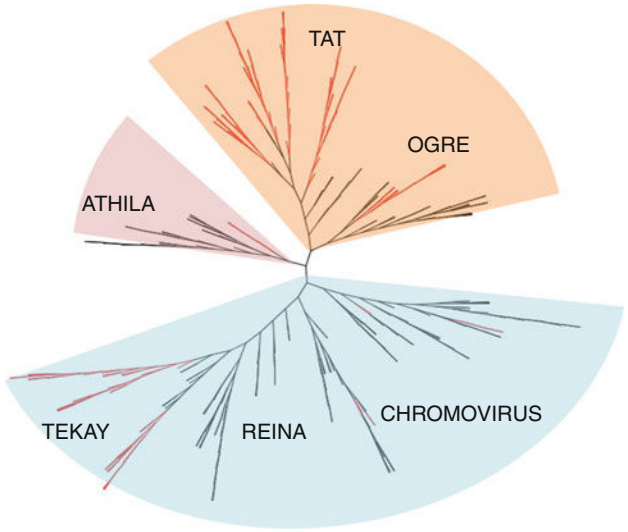


Fig. 2 Phylogenetic tree (Neighbour Joining analysis) of *Ty3-Gypsy* elements based on Reverse Transcriptase sequence alignments of *Spartina maritima* repeats (*red branches*) and the Repbase (*black branches*)

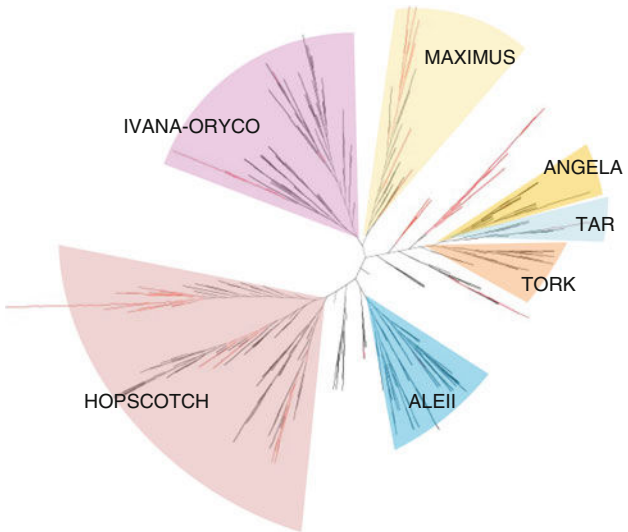


Fig. 3 Phylogenetic tree (Neighbour Joining analysis) of *Ty1-Copia* elements based on Reverse Transcriptase sequence alignments of *Spartina maritima* repeats (*red branches*) and the Repbase (*black branches*)

representing 64,643 bp or 0.25 % of the BESs sequenced (Supplementary Table 1) which is equivalent to one microsatellite every 6.1 kb (Table 3). Mononucleotides (60.9 %) are the most abundant motifs, followed by dinucleotides (21.6 %), trinucleotides (16.1 %), tetra, penta and hexanucleotides (1.42 %) (Table 3). A list of 200 SSR designed primer pairs is provided in Supplemental Table 2.

Gene content and functional annotation

The 39,910 masked for repeats BESs were first compared against the CDS databases of *O. sativa* and *S. bicolor* downloaded from the phytozome.net website using tBLASTx and a cut-off e-value of 10^{-6}. Among the BESs analyzed, 7,305 sequences were found matching at least one coding sequence of the *Oryza sativa* CDS database, representing 18.3 % of the analysed BESs. A total of 6,809 BESs were homologous to at least one coding sequence of the CDS database of *Sorghum bicolor*, representing 17.1 % of the total BESs. Using CDSs from *O. sativa* and *S. bicolor,* 4,070 and 4,098 different coding sequences were annotated. When comparing the BESs against the *Spartina* reference transcriptome (Ferreira de Carvalho et al. 2013), we found 8,968 best blast hits (e-value of 10^{-6} and a minimum identity of 90 %) representing 22.4 % of the BESs.

Among the 6,809 BESs of *S. maritima* showing significant homology with the coding sequence database of *S. bicolor*, 4,108 were associated with at least one GO term. Among the sequences assigned to a biological process category, most terms are associated with metabolic process (5,072 sequences: including primary, cellular macromolecules and nitrogen compound metabolic process), biosynthetic process (618 sequences) and regulation of biological process (337 sequences) (Fig. 5a). Among the BESs in the molecular function category, 2,362 sequences correspond to binding activities (including nucleic acid, nucleotide, ion and protein binding). Finally, 796 sequences are associated with transferase and 580 to hydrolase activities (Fig. 5b).

A summary of the *Spartina maritima* BES composition is presented in Fig. 6. Annotation of 52.31 % of the BESs is performed and provides a first overview of the composition of *Spartina maritima* genome. Cytoplasmic sequences account for 2.15 % of the sequences. Low complexity regions, small RNA and Simple sequence repeats occurred in 1.41 % of the BESs. Overall, interspersed repeats represent 15.48 % of the genome including LTR-*Copia* elements (5.45 %), LTR-*Gypsy* elements (8.16 %), LINEs and SINEs (0.75 %), unclassified repeats (0.13 %) and DNA transposons (0.99 %). Potential uncharacterized highly repeated sequences in the genome represent 14.97 %. Coding regions account for 22.40 % of the genome based on homology with ESTs data from close-related *Spartina* species. Nevertheless, unknown genomic regions still represent 43.59 % of the dataset.

Comparative genome mapping

The synteny between *Spartina* BES and other plants was characterized by searching for paired BES (1) on the same

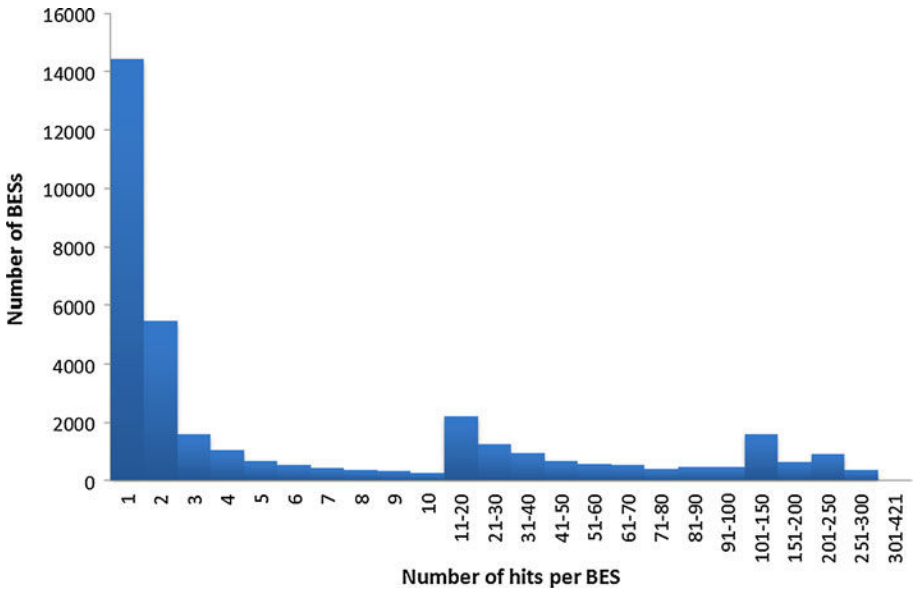


Fig. 4 Frequency of BESs showing similarity to other sequences in the same dataset for de novo identification of repeated regions

Table 3 Distribution and frequency of simple sequence repeats detected in *Musa acuminata*, *Oryza sativa* and *Zea mays* (from Hsu et al. 2011) compared to *Spartina maritima* using the MISA software

	Musa acuminata	*Oryza sativa*	*Zea mays*	*Spartina maritima*
Total Nb of BES analyzed	6,376	78,427	54,960	39,910
Total sequence length (bp)	4,517,901	69,423,321	37,410,959	26,182,878
Mononucleotides	0.8 % (6)	9.1 % (696)	7.2 % (167)	60.9 % (2,610)
Dinucleotides	47.7 % (350)	19.9 % (1,531)	15.4 % (358)	21.6 % (926)
Trinucleotides	20.6 % (151)	28.9 % (2,219)	35.7 % (831)	16.1 % (688)
Tetranucleotides	9.0 % (66)	10.2 % (783)	8.3 % (193)	0.72 % (31)
Pentanucleotides	13.1 % (96)	21.4 % (1,642)	21.6 % (504)	0.19 % (8)
Hexanucleotides	8.9 % (65)	10.5 % (804)	11.9 % (276)	0.51 % (22)
Total Nb of SSRs	734	7 675	2 329	4 285
SSR frequency (kb)	6.2	9.0	16.1	6.1
Most frequent SSR motif	AT/TA	CCG/CGG	AGC/GCT	A/T

chromosome, (2) within a 15–250 kb region and (3) orientated correctly with respect to each other and the homologous region. To assess the right distance between paired BESs, a histogram showing the distribution of the distance between paired BESs (using the *Sorghum bicolor* genome as a reference) was built (Supplementary Figure 1). Most BESs are comprised in a distance range from 15 to 250 kb; BESs out of this range are thought to be rearranged. Syntenic relationships between *S. maritima* repeat-masked BESs and other plant species were identified using BLASTn searches against the full-sequenced genomes of *Arabidopsis thaliana*, *Oryza sativa*, *Brachypodium distachyon* and *Sorghum bicolor*. As shown in Table 4, 3.2 % of the *Spartina maritima* BESs only matched the non-Poaceae genome (*A. thaliana*) with the retained parameters (70 % identity, e-value 10^{-6}), revealing the high divergence between the two taxa. The other Poaceae genomes matched *Spartina* BESs on levels ranging from 13.6 to 16.2 % (Table 4).

According to these parameters, *Arabidopsis thaliana* does not show syntenic relationships with *S. maritima* whereas about half of paired BESs are collinear with other Poaceae species (Table 4). The higher number of homologous BESs and synteny is found between *Spartina maritima* and *Sorghum bicolor* as expected from their phylogenetic relationships in the grass family. Among the 1,394 paired BESs, 826 are localized on the same *S. bicolor* chromosome, with 270 BESs situated outside the 15–250 kb distance "micro-synteny" range (i.e. rearranged) and 556 BESs within the distance window of 15–250 kb. Most of these (524) are collinear with *Sorghum,* whereas 32 exhibit a shift in the orientation of one of the BESs (Table 4). A substantial proportion of the paired BESs (568 representing 40.75 %) match to different *Sorghum* chromosomes (Table 4).

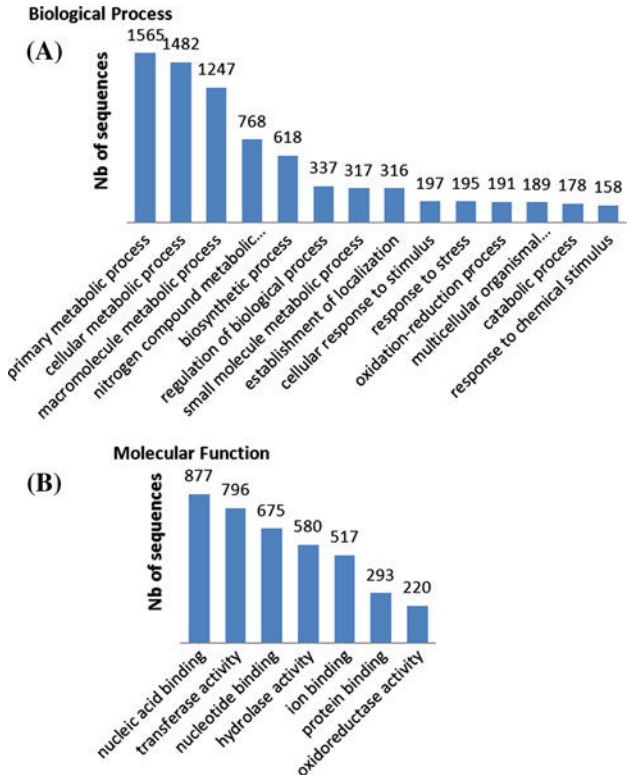


Fig. 5 Classification of GO annotations, **a** for biological process and **b** molecular function

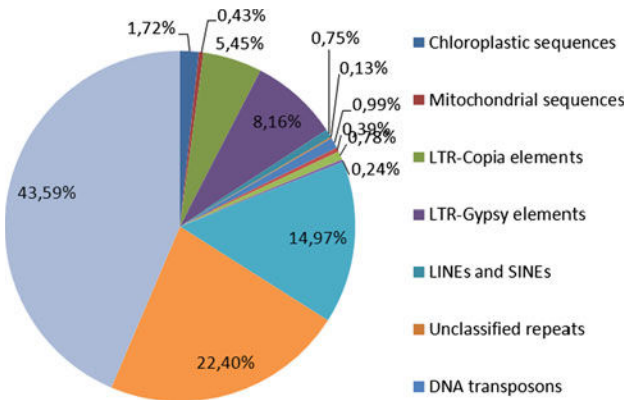


Fig. 6 Summary of *Spartina maritima* BES functional annotations by homology searches

The *Spartina* BESs mapped on the ten *Sorghum bicolor* chromosomes and on the twelve *Oryza sativa* chromosomes are represented in Fig. 7a, b, respectively. We chose to represent rearranged paired BESs for collinear regions including at least two pairs of rearranged BES (Fig. 7a, b). These putative orthologous regions involve both rearrangements on the same chromosomes or paired BESs matching different chromosomes. Collinear paired BESs show a high concentration on *Sorghum* chromosomes 1, 3, 4 and 6. Eight intrachromosomic and 6 interchromosomic rearrangements were detected (Fig. 7a). More rearrangements occurred between *Spartina* and *Oryza* than between *Spartina* and *Sorghum*, as 8 interchromosomic and 11 intrachromosomic could be detected (Fig. 7b).

Discussion

This study provides a first overview of the composition and structure of the *Spartina* genome. A set of 39,910 high quality genomic sequences of *Spartina maritima* (2n = 6x = 60, c.a. 3,700 Mb) was analysed to improve our knowledge on the repetitive and coding components of its genome.

Repetitive DNA in *Spartina*

The analyses of BAC-end sequences provided estimations of the repetitive sequence component, representing a proportion of 30.45 % of the sequences analysed, with 15.48 % showing homology to known repeat elements and 14.97 % potential highly repeated sequences specific to *Spartina maritima.* Repetitive DNA content in *Spartina* is intermediate between rice (35 %, 2n = 2x = 24, 1C = 420 Mb; IRGSP 2005) and *Brachypodium distachyon* (28.1 %, 2n = 2x = 10, 1C = 270 Mb; IBI 2010). However, regarding the *Spartina maritima* basic genome size (x = 10, c.a. 616 Mb), a larger number of repeat sequences would be expected: *Sorghum bicolor* has a genome size of 740 Mb (2n = 2x = 20) and a repeat element fraction of 62 % (Paterson et al. 2009). The proportion of repeats in *S. maritima* is most likely underestimated regarding the dataset analysed.

Transposable elements (TEs) are known to have important consequences on genome structure and functions (reviewed in Kejnovsky et al. 2012). Therefore, it is important to identify and evaluate the importance of the different families of repetitive elements in the genome. Identification of transposable elements in *Spartina maritima* is also essential to explore the effects of hybridization and genome duplication in *S. anglica* since *S. maritima* was the paternal genome donor to that species. Previous studies have shown no transposition burst in the allododecaploid *Spartina anglica* (Baumel et al. 2002b) most likely as a result of important methylation changes in regions flanking transposable elements (Parisod et al. 2009). In this study, analysis of TE distribution revealed that Class I TEs are significantly predominant in the genome of *Spartina maritima* compared to Class II TEs, with 14.42 % (10,582 elements) and 0.99 % (1,019 elements) of BESs, respectively. This contrasts from *Oryza sativa* for which Class II outnumbered Class I TEs with 61,900 and 163,800 TEs respectively. However, the nucleotide contribution of Class I elements in rice is larger

Table 4 Blastn hits and comparative genomics between *Spartina maritima* BESs (39,910 masked for repeats) and the *Arabidopsis thaliana*, *Brachypodium distachyon*, *Oryza sativa* and *Sorghum bicolor* genomes

	No. of hits (% of BESs)	Single BESs	Paired BESs			
			Localized on different chromosomes	Co-localized on the same chromosome		
				Distance inf. 15 kb or sup. 250 kb	Distance comprised between 15 and 250 kb	
					Orientation of BESs different	Correct orientation of both ends
A. thaliana	1,297 (3.2)	1,225	54	18	0	0
B. distachyon	5,421 (13.6)	4,389	398	242	68	324
O. sativa	6,115 (15.3)	4,863	600	196	52	404
S. bicolor	6,447 (16.2)	5,053	568	270	32	524

than Class II due to the largest size of LTR retrotransposons compared to DNA transposons (IRGSP 2005). Nonetheless, our results are consistent with the contents observed in *Brachypodium distachyon* (Pooideae) and *Sorghum bicolor* (Panicoideae) where Class I elements outnumber and cover a larger fraction of the genome than Class II TEs. Indeed, in *Brachypodium*, Class I and Class II elements occupy 23.33 and 4.77 % of the genome respectively (IBI 2010). In *Sorghum bicolor*, transposable elements account for 62 % of the genome including 54.52 % of Class I TEs (Paterson et al. 2009). The comparison of TE composition in a broad range of species suggests no phylogenetic explanations but radical changes associated with TE proportions (Kejnovsky et al. 2012).

In Class I elements, LTR retrotransposons are the most abundant with a larger percentage of Ty3-*Gypsy* elements compared to Ty1-*Copia* elements, 8.16 and 5.45 %, respectively. A similar pattern is observed in other Grass genomes such as *Sorghum bicolor* (Ty3-*Gypsy* 19.00 % and Ty1-*Copia* 5.18 %; Paterson et al. 2009), *Brachypodium distachyon* (Ty3-*Gypsy* 16.05 % and Ty1-*Copia* 4.86 %; IBI 2010) and *Oryza sativa* (Ty3-*Gypsy* 10.90 % and Ty1-*Copia* 3.85 %; IRGSP 2005). To identify and annotate the detected elements, we performed a phylogenetic analysis including annotated elements from various databases. In the *Gypsy*-like element tree, all clades are represented with a larger number of sequences corresponding to the TAT clade (*Spartina* sequences are related to *RIRE2* elements from *Oryza sativa*) and the Tekay clade (*Spartina* sequences are related to *RIRE3* elements from *O. sativa*). In the *Copia*-like element tree, a larger number of *Spartina* repeats are present in clade 8 (corresponding to *Hopscotch* and *Retrofit* elements in *Oryza*) and repeats are phylogenetically close to elements of clade 3 including *BARE1* and *RIRE1* elements previously found in *Hordeum* (Manninen and Schulman 1993) and *Oryza*. These abundant retrotransposons have most likely undergone amplification events in *Spartina maritima* and now represent the largest component of repetitive DNA. Indeed, large-scale amplification rounds can lead to TE high copy number in plant genomes over short evolutionary timescales (Bennetzen 2005). Particularly, LTR retrotransposons contribute in genome size expansion (Vitte and Bennetzen 2006). One example of LTR retrotransposon family proliferation in *Oryza australiensis* shows a two-fold increase in genome size compared to *O. sativa* in less than 3 million years (Piegu et al. 2006).

Few genomic resources are available for the *Spartina* genus and more generally in the Chloridoideae sub-family. As a consequence, the identification of repetitive DNA using closely related species databases is challenging. In this study, we used an approach to identify *Spartina maritima* lineage-specific highly repeated sequences, which proved to be useful and efficient in other studies (Huo et al. 2007; Cavagnaro et al. 2008; Ragupathy et al. 2011). Such lineage-specific repetitive DNA comprised 14.97 % of the DNA analysed. In other studies, the same analysis provided also a large proportion of novel repetitive elements. As a comparison, Ragupathy et al. (2011) found 7.4 % of unique *Linum usitatissimum* repeats; Cavagnaro et al. (2008) found 8.45 % of carrot-specific repeat sequences and Huo et al. (2007) discovered 7.4 % of unique *Brachypodium* repeat sequences. These estimations are due to the high nucleotide divergence between species specific TEs and annotated TEs in databases. Indeed, most LTR-retrotransposons older than 5 million years are severely fragmented or deleted in rice (Ma et al. 2004). Nevertheless, these proportions can be underestimated as we only analyzed a small sample of the genome of *Spartina maritima* and some repeats located in centromeres and telomeres are frequently under-represented in BAC libraries (Zhong et al. 2002; Osoegawa et al. 2007).

SSR markers are widely used for polymorphism analyses within species. In our study, a total of 4,285 SSR regions (representing 64,643 bp) have been identified from the 26.7 Mb of genomic DNA analyzed. Also, a list of SSR marker primer pairs was designed and can be used for further genetic diversity analyses in *Spartina*. The density

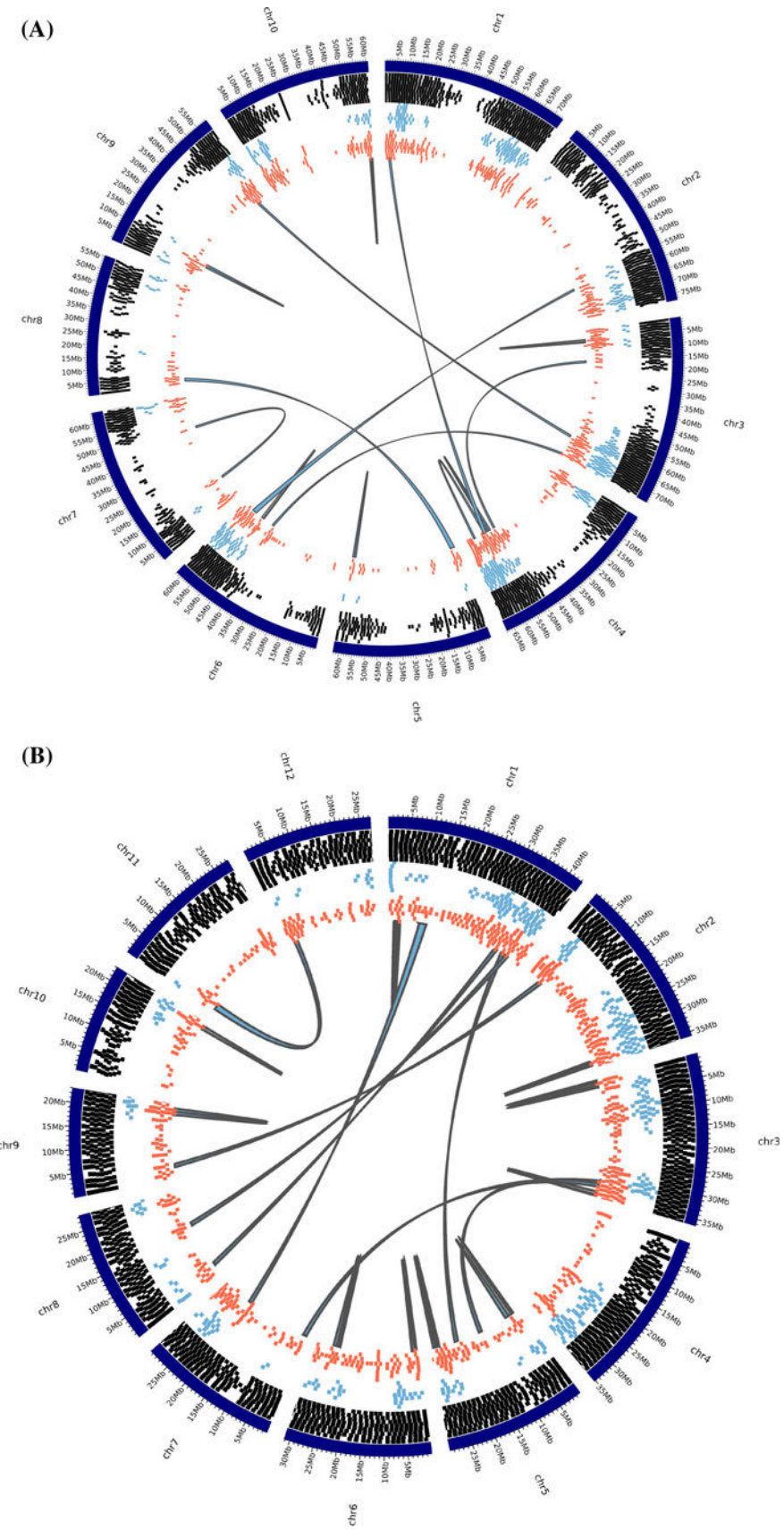


Fig. 7 BES sequences mapped to the **a** *Sorghum* and **b** *Oryza* genomes. The 10 (*Sorghum*) and 12 (*Oryza*) individual chromosomes are shown in the *outer circle*. From *outer* to *inner circles*, all homologous BESs are mapped: single BESs (*black tiles*), collinear paired BESs (*blue tiles*) and finally rearranged paired BESs (*orange tiles*). Paired BESs are linked to each other with *grey links*

found is of one SSR every 6.1 kb in *Spartina maritima*, mononucleotides being the most abundant with 60.9 % of all SSRs and A/T motif the most frequent. This pattern is also most frequent in *Arabidopsis thaliana* (Hsu et al. 2011). The SSR frequency is consistent with the observations in *Musa acuminata* (1 SSR every 6.2 kb; Cheung and Town 2007) and *A. thaliana* (1 SSR every 6.4 kb); but lowest than *O. sativa* (1 SSR every 9.0 kb) and *Z. mays* (1 SSR every 16.1 kb) (Hsu et al. 2011). These findings are in agreement with Morgante et al. (2002), who found relationships between SSRs and low-copy DNA fraction. Indeed, SSR frequency is inversely correlated to the proportion of repetitive DNA and especially LTR retrotransposons in plants.

A previous study was performed by Gedye et al. (2010) in *Spartina pectinata*, where they found 841 SSRs in ESTs longer than 500 bp representing 3.2 % of their dataset. GC-rich trinucleotide repeats were the most abundant in the dataset and accounted for 18.5 % of all SSRs. Although SSR discovery by genome sequencing is easier, the development of microsatellite resources through transcriptome has many advantages as it gives the possibility to find associations with functional genes and phenotypes (Li et al. 2002). Moreover, the mutation rate in coding sequencing being lower, the numbers of SSRs and polymorphisms are expected to be lower (Blanca et al. 2011) which increases transferability of SSR markers across species (Zalapa et al. 2012).

Spartina coding sequences

Comparison of BES sequences with the non-redundant protein database of *S. bicolor* suggested that 6,809 are transcribed sequences representing 17.1 % of the dataset. Proportion of coding sequences identified using the *Spartina* reference transcriptome based on 5 *Spartina* species (Ferreira de Carvalho et al. 2013; Ferreira de carvalho et al. unpublished) suggests that 22.4 % of the BES sequences are coding sequences. In order to find homology despite presence of introns, the stringency must be lowered, thus increasing the possibility to find false positives. Difference of 5.3 % of putative genes between the *Spartina* and the *Sorghum* databases suggests unique transcripts and probably *Spartina*-specific genes (or genes that are lost in *Sorghum bicolor*). In the flax genome, Ragupathy et al. (2011) observed a proportion of 5.6 % unique flax transcripts, with 21.1 % of BESs showing homology to NCBI-ESTs and 26.8 % showing similarity to flax transcripts. The proportion of BESs with potential coding regions (22.4 %) is comparatively higher than the assessment of coding regions in most BES-based studies: carrot, 10 % (Cavagnaro et al. 2008); apple, 8.6 % (Han and Korban 2008); *Musa*, 11 % (Cheung and Town 2007) and comparable or lower than the coding fractions reported in walnut (24.9 %; Wu et al. 2011), *Brachypodium* (25.3 %; Huo et al. 2007) and *Citrus clementina* (36.0 %; Terol et al. 2008).

Based on the number of BESs matching at least one coding sequence of *Sorghum bicolor* in the CDS database (6,809), the mean sequence size of BESs (656 bp) and the total size of BESs sequenced (26.7 Mb), we estimated a percentage of 16.7 % of BESs containing potentially coding genes. Considering the basic genome size of *S. maritima* (616 Mb) and the mean size of an *Oryza sativa* gene (2.7 kb; IRGSP 2005), we estimated the transcriptome size of *S. maritima* to be around 103.21 Mb, representing 38,229 genes. This estimation is consistent with the gene number found in fully sequenced Poaceae such as *Sorghum bicolor* (34,008 genes; Paterson et al. 2009) and *Oryza sativa* (41,046 genes; Yu et al. 2005). Gene density predicts that a gene occurs every 16.1 kb based on the fact that we might expect 38,229 genes in the basic genome of *Spartina maritima* (estimated as 616 Mb). By comparison, *S. bicolor* has a gene density of one gene every 24.0 kb (Paterson et al. 2009) and *O. sativa* of one gene every 9.9 kb (IRGSP 2005). *Musa acuminata* is predicted to have one gene every 14.3 kb (D'Hont et al. 2012) and *A. thaliana* has one gene every 4.5 kb (AGI 2000).

Comparative genomics

Genus *Spartina* is part of the Chloridoideae subfamily, a poorly studied taxon of the Poaceae. The syntenic relationships remain unclear between *Spartina* and related grass species. Therefore, the comparative analysis of homologous regions facilitates the investigation of genome evolution and dynamics. In comparison with *S. maritima*, *Arabidospis thaliana* shows no syntenic paired BESs as they diverged 140–150 MYA (Chaw et al. 2004). Moreover, *A. thaliana* has undergone a recent duplication followed by the loss of 70 % of the duplicated genes (Bowers et al. 2003). The majority of the microsyntenic regions in grasses that existed before the duplication event have disappeared due to the contraction and diploidization of the genomes. *Sorghum bicolor* is the most comparable fully sequenced genome with an equivalent basic chromosome number (x = 10, 730 Mb for *S. bicolor* and 616 Mb for *S. maritima*) and similar gene density.

Grass genomes largely benefited from the high-throughput technologies. The sequencing of the *Sorghum* genome provided new insights into the synteny of cereal lineages (Paterson et al. 2009). Despite their divergence time (around 50 MYA; Christin et al. 2008), sorghum and rice are largely collinear with 57.8 % of *Sorghum* gene models assigned to blocks collinear with rice (Paterson et al. 2009). Kim et al. (2009) have compared *Cynodon dactylon* (Chlorioideae) ESTs to other grass subfamily

representatives and have estimated that Chloridoideae and Panicoideae diverged about 34.6–38.5 million years ago. To our knowledge, the only physical comparative study involving a Chloridoideae member was performed by Srinivasachary et al. (2007) who compared a finger millet (*Eleusine coracana*, 2n = 4x = 36) genetic map with rice (2n = 2x = 24) and found that 30 % of millet BES end sequenced genomic clones and 73 % of millets ESTs identify putative rice orthologs. The recombination rate is increased in the distal chromosome regions (such as in wheat and rice, Akhunov et al. 2003; See et al. 2006) and can be caused by translocation and retention of duplicated gene copies in highly-recombinant regions. Moreover, six of the nine millet chromosomes correspond to six single rice chromosomes and the remaining three millet chromosomes are orthologous to rice chromosomes, each with one rice chromosome inserted in the centromeric region of a second rice chromosome to form a millet chromosomal conformation. Interestingly, homologous regions were identified between chromosome 2 of millet and chromosomes 2 and 10 of rice; chromosome 5 of millet and chromosomes 5 and 12 of rice; and chromosome 6 of millet and chromosomes 6 and 9 of rice. According to the known chromosome structures of rice and sorghum (Salse et al. 2008) chromosomes 1, 4, 8 and 9 of *Eleusine* are similar to chromosomes 3, 6, 7 and 5 of *Sorghum*, respectively and the synteny is potentially conserved as no major rearrangements are observed between *Eleusine* and rice regarding these four chromosomes. The other chromosomes seem to have undergone rearrangements since the divergence between Panicoideae and Chloridoideae 45–50 MYA. Therefore, those four conserved chromosomes should be less rearranged than the others in the Chloridoideae subfamily including *Spartina* species. We did not observe large macrosyntenic rearrangements using the mapping strategy employed in the present manuscript but some regions (respectively 8 among chromosomes and 6 between chromosomes) appeared to have experienced rearrangements between genera *Spartina* and *Sorghum*. We detected more rearrangements between *Spartina* and *Oryza* which is consistent with divergence times between Chloridoideae and these two respective lineages. Using BAC-End Sequence survey in sugarcane, Kim et al. (2013) also detected rearrangements between *Saccharum* and *Sorghum* that diverged about 7.8 MYA. These rearrangements were interpreted as a result of genome duplication rounds that occurred independently in the *Saccharum* lineage.

Among the Chloridoideae, *Eleusine* (x = 9) and *Spartina* (x = 10) have evolved separately into two sister clades: The Cynodonteae and the Zoysieae (Peterson et al. 2010). Furthermore, even though base chromosome number in the sub-family is x = 10, aneuploidy is frequent and lower base chromosome numbers (x = 7, 8, 9) are reported (Peterson et al. 2010). Duplication events are also frequent with ploidy levels ranging from diploid to 20-ploid (in *Pleuraphismutica* Buckley) with many of them allopolyploids as a consequence of extensive hybridization which complicates comparative analyses among genera (Roodt and Spies 2003a). Chloridoideae genome history needs definitely further investigation; The BAC library constructed and analysed in this study may provide more physical information on the putative rearrangements that occurred during Chloridoideae evolution.

In conclusion, this study represents the first overview of the *Spartina maritima* genome regarding the respective coding and repetitive components. This information will be particularly useful to explore genome evolution in hybrids and allopolyploid species deriving from *S. maritima*. The syntenic relationships with other grass genomes examined here help clarifying evolution in Poaceae, *Spartina maritima* being a part of the poorly-known Chloridoideae sub-family.

Acknowledgements This work was supported by CNRS INEE and University of Rennes 1, the Partner University Funds and by the Genoscope (GENOSPART Project). The analyses benefited from the Genouest (Bioinformatics) Plateform facilities. J. Ferreira de Carvalho benefited from a Ph.D. Grant (ARED EVOSPART) from the Regional Council of Brittany.

References

Ainouche ML, Baumel A, Salmon A (2004a) *Spartina anglica* C. E. Hubbard: a natural model system for analysing early evolutionary changes that affect allopolyploid genomes. Biol J Lin Soc Lond 82:475–484

Ainouche ML, Baumel A, Salmon A, Yannic G (2004b) Hybridization, polyploidy and speciation in *Spartina* (Poaceae). New Phytol 161:165–172

Ainouche ML, Fortuné PM, Salmon A, Parisod C, Grandbastien M-A, Fukunaga K et al (2009) Hybridization, polyploidy and invasion: lessons from *Spartina* (Poaceae). Biol Invasions 11:1159–1173

Ainouche M, Chelaifa H, Ferreira J, Bellot S, Ainouche A, Salmon A (2012) Polyploid evolution in *Spartina*: dealing with highly redundant hybrid genomes. In: Soltis DE, Soltis PS (eds) Polyploidy and genome evolution. Springer, Berlin, pp 225–243

Akhunov ED, Goodyear AW, Geng S, Qi L–L, Echalier B, Gill BS et al (2003) The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. Genome Res 13:753–763

Alberts J, Price M, Kania M (1990) Metal concentrations in tissues of *Spartina alterniflora* (Loisel) and sediments of georgia salt marshes. Estuar Coast Shelf Sci 30:47–58

Ayres DR, Smith DL, Zaremba K, Klohr S, Strong DR (2004) Spread of exotic cordgrasses and hybrids (*Spartina* sp.) in the tidal marshes of San Francisco Bay, California, USA. Biol Invasions 6:221–231

Baisakh N, Subudhi PK, Varadwaj P (2008) Primary responses to salt stress in a halophyte, smooth cordgrass (*Spartina alterniflora* Loisel.). Funct Integr Genomics 8:287–300

Baumel A, Ainouche ML, Levasseur JE (2001) Molecular investigations in populations of *Spartina anglica* C.E. Hubbard (Poaceae) invading coastal Brittany (France). Mol Ecol 10:1689–1701

Baumel A, Ainouche ML, Bayer RJ, Ainouche AK, Misset MT (2002a) Molecular phylogeny of hybridizing species from the genus *Spartina* Schreb. (Poaceae). Mol Phylogenet Evol 22:303–314

Baumel A, Ainouche M, Kalendar R, Schulman AH (2002b) Retrotransposons and genomic stability in populations of the young allopolyploid species *Spartina anglica* CE Hubbard (Poaceae). Mol Biol Evol 19:1218–1227

Baumel A, Ainouche ML, Misset MT, Gourret JP, Bayer RJ (2003) Genetic evidence for hybridization between the native *Spartina maritima* and the introduced *Spartina alterniflora* (Poaceae) in South-West France: *Spartina* × *neyrautii* re-examined. Plant Syst Evol 237:87–97

Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15:621–627

Blanca J, Cañizares J, Roig C, Ziarsolo P, Nuez F, Picó B (2011) Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). BMC Genomics 12:104

Bohra A, Dubey A, Saxena R, Penmetsa RV, Poornima KN, Kumar N et al (2011) Analysis of BAC-end sequences (BESs) and development of BES-SSR markers for genetic mapping and hybrid purity assessment in pigeonpea (*Cajanus spp*.). BMC Plant Biol 11:56

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438

Caetano M, Vale C, Cesário R, Fonseca N (2008) Evidence for preferential depths of metal retention in roots of salt marsh plants. Sci Total Environ 390:466–474

Cambrollé J, Redondo-Gómez S, Mateos-Naranjo E, Figueroa ME (2008) Comparison of the role of two *Spartina* species in terms of phytostabilization and bioaccumulation of metals in the estuarine sediment. Mar Pollut Bull 56:2037–2042

Castellanos E, Figueroa M, Davy A (1994) Nucleation and facilitation in salt-marsh succession—interactions between *Spartina maritima* and *Arthrocnemum perenne*. J Ecol 82:239–248

Castillo JM, Fernández-Baco L, Castellanos EM, Luque CJ, Figueroa ME, Davy AJ (2000) Lower limits of *Spartina densiflora* and *S. maritima* in a Mediterranean salt marsh determined by different ecophysiological tolerances. J Ecol 88:801–812

Castillo JM, Ayres DR, Leira-Doce P, Bailey J, Blum M, Strong DR et al (2010) The production of hybrids with high ecological amplitude between exotic *Spartina densiflora* and native *S. maritima* in the Iberian Peninsula. Divers Distrib 16:547–558

Cavagnaro PF, Chung S-M, Szklarczyk M, Grzebelus D, Senalik D, Atkins AE et al (2008) Characterization of a deep-coverage carrot (*Daucus carota* L.) BAC library and initial analysis of BAC-end sequences. Mol Genet Genomics 281:273–288

Chalhoub B, Belcram H, Caboche M (2004) Efficient cloning of plant genomes into bacterial artificial chromosome (BAC) libraries with larger and more uniform insert size. Plant Biotechnol J 2:181–188

Chaw S-M, Chang C–C, Chen H-L, Li W-H (2004) Dating the monocot? Dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol 58:424–441

Cheung F, Town C (2007) A BAC end view of the *Musa acuminata* genome. BMC Plant Biol 7:29

Christin PA, Besnard G, Samaritani E, Duvall MR, Hodkinson TR, Savolainen V et al (2008) Oligocene CO2 decline promoted C4 photosynthesis in grasses. Curr Biol 18:37–43

Christin PA, Petitpierre B, Salamin N, Büchi L, Besnard G (2009) Evolution of C4 phosphoenolpyruvate carboxykinase in grasses, from genotype to phenotype. Mol Biol Evol 26:357–365

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676

D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, Garsmeur O et al (2012) The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature 488:213–217

Dida MM, Srinivasachary Ramakrishnan S, Bennetzen JL, Gale MD, Devos KM (2006) The genetic map of finger millet, *Eleusine coracana*. Theor Appl Genet 114:321–332

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Ferreira de Carvalho J, Poulain J, Da Silva C, Wincker P, Michon-Coudouel S, Dheilly A et al (2013) Transcriptome *de novo* assembly from next-generation sequencing and comparative analyses in the hexaploid salt marsh species *Spartina maritima* and *Spartina alterniflora* (Poaceae). Heredity 110:181–193

Ferris C, King RA, Gray AJ (1997) Molecular evidence for the maternal parentage in the hybrid origin of *Spartina anglica* C.E. Hubbard. Mol Ecol 6:185–187

Fortuné PM, Schierenbeck KA, Ainouche AK, Jacquemin J, Wendel JF, Ainouche ML (2007) Evolutionary dynamics of Waxy and the origin of hexaploid *Spartina* species (Poaceae). Mol Phylogenet Evol 43:1040–1055

Fortuné PM, Schierenbeck K, Ayres D, Bortolus A, Catrice O, Brown S et al (2008) The enigmatic invasive *Spartina densiflora*: a history of hybridizations in a polyploidy context. Mol Ecol 17:4304–4316

Gedye K, Gonzalez-Hernandez J, Ban Y, Ge X, Thimmapuram J, Sun F, Wright C, Ali S, Boe A, Owens V (2010) Investigation of the transcriptome of prairie cord grass, a new cellulosic biomass crop. Int J Plant Genomics 3:69

Gonthier L, Bellec A, Blassiau C, Prat E, Helmstetter N et al (2010) Construction and characterization of two BAC libraries representing a deep-coverage of the genome of chicory (*Cichorium intybus* L., Asteraceae). BMC Res Notes 3:225

Gonzalez-Hernandez JL, Sarath G, Stein JM, Owens V, Gedye K, Boe A (2009) A multiple species approach to biomass production from native herbaceous perennial feedstocks. In Vitro Cell Dev Biol Plant 45:267–281

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ et al (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36:3420–3435

Grass Phylogeny Working Group II (2012) New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. New Phytol 193:304–312

Groves H, Groves J (1880) *Spartina* × *townsendii* Nobis. Report of the Botanical Society and exchange club of the British Isles 1:37

Han Y, Korban SS (2008) An overview of the apple genome through BAC end sequence analysis. Plant Mol Biol 67:581–588

Hansel CM, Fendorf S, Sutton S, Newville M (2001) Characterization of Fe plaque and associated metals on the roots of mine-waste impacted aquatic plants. Environ Sci Technol 35: 3863–3868

Hilu KW, Alice LA (2001) A phylogeny of Chloridoideae (Poaceae) based on matK sequences. Syst Bot 26:386–405

Hsu C–C, Chung Y-L, Chen T-C, Lee Y-L, Kuo Y-T, Tsai W-C et al (2011) An overview of the *Phalaenopsis* orchid genome through BAC end sequence analysis. BMC Plant Biol 11:3

Huo N, Lazo GR, Vogel JP, You FM, Ma Y, Hayden DM et al (2007) The nuclear genome of *Brachypodium distachyon*: analysis of BAC end sequences. Funct Integr Genomics 8:135–147

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Kejnovsky E, Hawkins JS, Feschotte C (2012) Plant transposable elements: biology and evolution. In: Wendel J, Greilhuber J, Dolezel J, Leitch I (eds) Plant genome diversity volume 1: plant genomes, their residents and their evolutionary dynamics. Vienna, pp 17–34

Kim C, Jang CS, Kamps TL, Robertson JS, Feltus FA, Paterson AH (2008) Transcriptome analysis of leaf tissue from Bermudagrass (*Cynodon dactylon*) using a normalised cDNA library. Funct Plant Biol 35:585–594

Kim C, Tang H, Paterson AH (2009) Duplication and divergence of grass genomes: integrating the Chloridoids. Trop Plant Biol 2:51–62

Kim C, Lee T-H, Compton RO, Robertson JS, Pierce GJ, Paterson AH (2013) A genome-wide BAC end-sequence survey of sugarcane elucidates genome composition, and identifies BACs covering much of the euchromatin. Plant Mol Biol 81:139–147

Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D et al (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645

Larher F, Hamelin J, Stewart GR (1977) L'acide dimethylsulfonium propanoïque de *Spartina anglica*. Phytochemistry 16:2019–2020

Lee RW (2003) Physiological adaptations of the invasive cordgrass *Spartina anglica* to reducing sediments: rhizome metabolic gas fluxes and enhanced O2 and H2S transport. Mar Biol 143:9–15

Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol 11:2453–2465

Luque CJ, Castellanos EM, Castillo JM, Gonzalez M, Gonzalez-Vilches MC, Figueroa ME (1999) Metals in halophytes of a contaminated estuary (Odiel Saltmarshes, SW Spain). Mar Pollut Bull 38:49–51

Ma J, Devos KM, Bennetzen JL (2004) Analyses of LTR-Retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res 14:860–869

Manninen I, Schulman AH (1993) *BARE-1*, a *Copia*-like retroelement in barley (*Hordeum vulgare* L.). Plant Mol Biol 22:829–846

Marchant CJ (1967) Evolution in *Spartina* (Gramineae). I. The history and morphology of the genus in Britain. Bot J Linn Soc 60:1–24

Marchant C, Goodman P (1969) *Spartina maritima* (Curtis) Fernald. J Ecol 57:287–302

Mobberley DG (1956) Taxonomy and distribution of the genus *Spartina*. Iowa State Coll J Sci 30:471–574

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200

Osoegawa K, Vessere GM, Shu CL, Hoskins RA, Abad JP, de Pablos B et al (2007) BAC clones generated from sheared DNA. Genomics 89:291–299

Otte ML, Wilson G, Morris JT, Moran BM (2004) Dimethylsulphopropionate (DMSP) and related compounds in higher plants. J Exp Bot 55(404):1919–1925

Ouyang S, Bell R (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res 32:360–363

Parisod C, Salmon A, Zerjal T, Tenaillon M, Grandbastien M, Ainouche M (2009) Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in *Spartina*. New Phytol 184:1003–1015

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

Peterson DG, Tomkins JP, Frisch DA, Wing RA, Paterson AH (2000) Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide. J Agric Genomics 5. www.ncgr.org/research/jag.

Peterson PM, Romaschenko K, Johnson G (2010) A classification of the Chloridoideae (Poaceae) based on multi-gene phylogenetic trees. Mol Phylogenet Evol 55:580–598

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H et al (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res 16:1262–1269

Prasad V, Stromberg CAE, Leache AD, Samant B, Patnaik R, Tang L et al (2011) Late Cretaceous origin of the rice tribe provides evidence for early diversification in Poaceae. Nat Commun 2:480

Ragupathy R, Rathinavelu R, Cloutier S (2011) Physical mapping and BAC-end sequence analysis provide initial insights into the flax (*Linum usitatissimum* L.) genome. BMC Genomics 12:217

Ramanarao MV, Weindorf D, Breitenbeck G, Baisakh N (2011) Differential expression of the transcripts of *Spartina alterniflora* Loisel (Smooth Cordgrass) induced in response to petroleum hydrocarbon. Mol Biotechnol 51:18–26

Raybould AF, Gray AJ, Lawrence MJ, Marshall DF (1991) The evolution of *Spartina anglica* CE Hubbard (Gramineae)—origin and genetic variability. Biol J Linn Soc Lond 43:111–126

Renny-Byfield S, Ainouche M, Leitch IJ, Lim KY, Le Comber SC, Leitch AR (2010) Flow cytometry and GISH reveal mixed ploidy populations and *Spartina* nonaploids with genomes of *S. alterniflora* and *S. maritima* origin. Ann Bot 105:527–533

Roodt R, Spies JJ (2003a) Chromosome studies in the grass subfamily Chloridoideae. I. Basic chromosome numbers. Taxon 52:557–566

Roodt R, Spies JJ (2003b) Chromosome studies in the grass subfamily Chloridoideae. II. An analysis of polyploidy. Taxon 52:736–746

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM et al (2008) Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. Plant Cell Online 20:11

See DR, Brooks S, Nelson JC, Brown-Guedira G, Friebe B, Gill BS (2006) Gene evolution at the ends of wheat chromosomes. Proc Natl Acad Sci USA 103:4162–4167

Srinivasachary, Dida MM, Gale MD, Devos KM (2007) Comparative analyses reveal high levels of conserved colinearity between the finger millet and rice genomes. Theor Appl Genet 115:489–499

Terol J, Naranjo MA, Ollitrault P, Talon M (2008) Development of genomic resources for *Citrus clementina*: characterization of three deep-coverage BAC libraries and analysis of 46,000 BAC end sequences. BMC Genomics 9:423

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

The International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768

Thiel T, Michalek W, Varshney R, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106:411–422

Timmers RA, Strik DPBTB, Hamelers HVM, Buisman CJN (2010) Long-term performance of a plant microbial fuel cell with *Spartina anglica*. Appl Microbiol Biotechnol 86:973–981

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA 103:17638–17643

Wu J, Gu YQ, Hu Y, You FM, Dandekar AM, Leslie CA et al (2011) Characterizing the walnut genome through analyses of BAC end sequences. Plant Mol Biol 78:95–107

Yannic G, Baumel A, Ainouche M (2004) Uniformity of the nuclear and chloroplast genomes of *Spartina maritima* (Poaceae), a

salt-marsh species in decline along the Western European Coast. Heredity 93:182–188

You F, Huo N, Gu Y, Luo M, Ma Y, Hane D et al (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. BMC Bioinformatics 9:253

Yu J, Wang J, Lin W, Li S, Li H, Zhou J et al (2005) The genomes of *Oryza sativa*: a history of duplications. PLoS Biol 3:e38

Yu J-K, Sun Q, Rota ML, Edwards H, Tefera H, Sorrells ME (2006) Expressed sequence tag analysis in tef *Eragrostis tef* (Zucc) Trotter. Genome 49:365–372

Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E et al (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. Am J Bot 99:193–208

Zhang D, Ayele M, Tefera H, Nguyen HT (2001) RFLP linkage map of the Ethiopian cereal tef: *Eragrostis tef* (Zucc) Trotter. Theor Appl Genet 102:957–964

Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A, Nagaki K et al (2002) Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. Plant Cell 14:2825–2836

Molecular Phylogenetics and Evolution 93 (2015) 5–16

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

The chloroplast genome of the hexaploid *Spartina maritima* (Poaceae, Chloridoideae): Comparative analyses and molecular dating ☆

M. Rousseau-Gueutin [a,1,6], S. Bellot [a,2,6], G.E. Martin [a,3], J. Boutte [a], H. Chelaifa [a,4], O. Lima [a], S. Michon-Coudouel [b], D. Naquin [c,5], A. Salmon [a], K. Ainouche [a], M. Ainouche [a,*]

[a] UMR CNRS 6553 Ecobio, OSUR (Observatoire des Sciences de l'Univers de Rennes), Université de Rennes 1/Université Européenne de Bretagne, 35042 Rennes, France
[b] Plate-forme Génomique Environnementale et Fonctionnelle, OSUR-CNRS, Université de Rennes 1, 35042 Rennes, France
[c] Plate-Forme de Bioinformatique, Genouest INRIA/IRISA, Université de Rennes-1, 35042 Rennes, France

ARTICLE INFO

Article history:
Received 25 January 2015
Revised 10 June 2015
Accepted 18 June 2015
Available online 13 July 2015

Keywords:
Spartina maritima
Chloridoideae
Chloroplast genome
Polyploidy
Molecular phylogeny
Molecular dating

ABSTRACT

The history of many plant lineages is complicated by reticulate evolution with cases of hybridization often followed by genome duplication (allopolyploidy). In such a context, the inference of phylogenetic relationships and biogeographic scenarios based on molecular data is easier using haploid markers like chloroplast genome sequences. Hybridization and polyploidization occurred recurrently in the genus *Spartina* (Poaceae, Chloridoideae), as illustrated by the recent formation of the invasive allododecaploid *S. anglica* during the 19th century in Europe. Until now, only a few plastid markers were available to explore the history of this genus and their low variability limited the resolution of species relationships. We sequenced the complete chloroplast genome (plastome) of *S. maritima*, the native European parent of *S. anglica*, and compared it to the plastomes of other Poaceae. Our analysis revealed the presence of fast-evolving regions of potential taxonomic, phylogeographic and phylogenetic utility at various levels within the Poaceae family. Using secondary calibrations, we show that the tetraploid and hexaploid lineages of *Spartina* diverged 6–10 my ago, and that the two parents of the invasive allopolyploid *S. anglica* separated 2–4 my ago via long distance dispersal of the ancestor of *S. maritima* over the Atlantic Ocean. Finally, we discuss the meaning of divergence times between chloroplast genomes in the context of reticulate evolution.

1. Introduction

The chloroplast (cp) genome, or plastome, of most land plants is composed of two inverted repeats (IR) that separate a large (LSC) and a small (SSC) single copy regions (Jansen and Ruhlman, 2012; Kolodner and Tewari, 1979). It has an average size of about 150 kb, ranging from 117 kb in *Erodium Carvifolium* (Blazier et al., 2011) to 218 kb in *Pelargonium* × *hortorum* (Chumley et al., 2006), with a gene content and order relatively well conserved in angiosperms (Jansen and Ruhlman, 2012), although extensive modifications may be encountered in some lineages such as the Campanulacae, Fabaceae or Geraniaceae families (Guisinger et al., 2010; Jansen et al., 2007; Martin et al., 2014). Chloroplast coding and non-coding regions have been employed to infer plant phylogenetic relationships at various taxonomical levels (Clegg and Zurawski, 1992; Jansen et al., 2007; Moore et al., 2010; Shaw et al., 2007, 2014). Their haploid state, uniparental inheritance and general absence of recombination (Jansen et al., 2007; Moore et al., 2010) make such sequences particularly useful for phylogenetic and phylogeographic studies in the contexts of reticulate evolution (*i.e.* hybridization) and polyploidy that characterize the history of most plant lineages (Fawcett and Van de Peer, 2010; McKinnon, 2004; Wendel and Doyle, 2005).

☆ This paper was edited by the Associate Editor Timothy Evans.

* Corresponding author at: UMR CNRS 6553 Ecobio, OSUR (Observatoire des Sciences de l'Univers de Rennes), Université de Rennes 1, Bât. 14A Campus de Beaulieu, 35042 Rennes Cedex, France.
E-mail address: malika.ainouche@univ-rennes1.fr (M. Ainouche).
[1] Current address: INRA, Institut de Génétique, Environnement et Protection des Plantes (IGEPP) UMR1349, BP35327, 35653 Le Rheu Cedex, France.
[2] Current address: Systematic Botany and Mycology, University of Munich, Menzinger Straße 67, D-80638 Munich, Germany.
[3] Current address: CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement), UMR AGAP, F-34398 Montpellier, France.
[4] Current address: INRA/Université d'Evry Val d'Essone, Unité de Recherche en Génomique Végétale, UMR1165, Organization and Evolution of Plant Genomes, 2 rue Gaston Crémieux, 91057 Evry, France.
[5] Current address: Plate-forme IMAGIF, FRC3115 CNRS, 91198 Gif sur Yvette Cedex, France.
[6] The two first authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ympev.2015.06.013
1055-7903/

This study focuses on genus *Spartina* in the Poaceae family. A worldwide phylogenetic analysis of the Poaceae based on molecular and morphological data obtained from more than 12,000 grasses was recently performed (Soreng et al., 2015) and clarified the subfamily classification. *Spartina* Schreb. represents a strongly supported monophyletic clade in the Chloridoideae subfamily (tribe Zoysieae), where it was found embedded in the paraphyletic genus *Sporobolus* (Peterson et al. 2010a, 2014b). This led to the proposal to conserve *Sporobolus* against *Spartina*, at the genus level (Peterson et al., 2014c). The *Spartina* clade (syn: *Sporobolus* sect. *Spartina* Peterson et al., 2014c) is comprised of 17 species colonizing mostly coastal saltmarshes (Mobberley, 1956; Clayton et al., 2006). Recurrent hybridization and polyploidization have particularly affected this lineage, with important evolutionary and ecological consequences (Ainouche et al., 2009; Strong and Ayres, 2013). The basic chromosome number in this group is $x = 10$ (Marchant, 1967) with ploidy levels ranging from tetraploid to dodecaploid species (Ainouche et al., 2012). *Spartina* evolved in two main lineages (Baumel et al., 2002): the first one containing American tetraploid species and the second one containing two American hexaploids (*S. alterniflora* Loisel. and its sister species *S. foliosa Trin.*) and the Old-World hexaploid species *S. maritima* Curtis. These two lineages are now recognized as *Sporobolus* subsections *Spartina* and *Alterniflori* respectively (Peterson et al., 2014b). Phylogenetic relationships in the hexaploid clade were well resolved using nuclear (ITS and Waxy) and chloroplast (*trnL-trnF*, *trnT-trnL*, *rpl32-trnL*, and *rps16-trnK* spacers, *trnL*, *ndhA* and *rps16*, introns) sequences (Baumel et al., 2002; Fortune et al., 2007; Peterson et al., 2014b), whereas relationships among the tetraploids remain to be resolved. Inter-specific crosses within and between these two main lineages produced various homoploid and allopolyploid (heptaploid, nonaploid and dodecaploid) taxa (Ainouche et al., 2012). In particular, hybridization between the hexaploid species *S. alterniflora* and *S. maritima* at the end of the 19th century resulted in the formation of a new vigorous and invasive allododecaploid species *S. anglica*, which has become a textbook example of recent allopolyploid speciation and a model for studying the evolution of polyploid genomes (Ainouche and Wendel, 2013). We focus here on its hexaploid ($2n = 6x = 60$) parent *Spartina maritima*, which is distributed along the Atlantic European and South-African coasts and was repeatedly involved in hybridization events. In Europe, *S. maritima* hybridized with the introduced American hexaploid *S. alterniflora*, which provided the maternal genome (and plastome) to the resulting homoploid hybrids *S. x neyrautii* in France and *S. x townsendii* in England. The latter hybrid has given rise to the above-mentioned allododecaploid *S. anglica* by genome doubling (Baumel et al., 2003; Ferris et al., 1997). In Spain, bidirectional hybridization occurred between *S. maritima* and the introduced heptaploid *S. densiflora* resulting in individuals that inherited alternatively the plastome of *S. maritima* or of *S. densiflora* (Castillo et al., 2010).

High ploidy levels and frequent hybridization events make phylogenetic analyses in *Spartina* particularly challenging using nuclear sequences, and reinforce the need to develop additional informative chloroplast markers to decipher the complex history of this genus. However no chloroplast genome of *Spartina* have been sequenced so far and despite that Chloridoideae contain ca. 1600 species in 31 genera and exhibit a worldwide distribution (Peterson et al., 2007; Soreng et al., 2015), only one Chloridoideae plastome has been sequenced: *Neyraudia reynaudiana* (Wysocki et al., 2014), which belongs to another tribe (Triraphideae, Soreng et al., 2015). This contrasts with other Poaceae subfamilies such as the Pooideae or Bambusoideae for which twelve and eight plastomes are available respectively (NCBI Organelle Genome Resources, www.ncbi.nlm.nih.gov/genomes/ accessed on 30 November 2014). The sequencing and analysis of the plastome of *Spartina* will thus expand the current understanding of chloroplast genome evolution in Poaceae. It will also provide new genetic markers to better resolve the phylogenetic relationships in Chloridoideae (Hilu and Alice, 2001; Peterson et al., 2014a; Soreng et al., 2015), and will allow estimating divergence times of Chloridoideae lineages, including *Spartina*.

In this study, we assembled the plastid genome of *Spartina maritima* and inferred its molecular evolution by comparing its structure and gene content to those of other published Poaceae plastid genomes. We identified variable coding and non-coding regions with potential phylogenetic and taxonomic utility in Chloridoideae, and used some of them to reassess phylogenetic relationships in *Spartina*. Finally, we dated the divergence between Chloridoideae and other grass lineages and between the different *Spartina* clades, providing the first estimate of the plastome divergence times in the polyploid *Spartina* species.

2. Material and methods

2.1. Plant material and DNA isolation

Samples from *Spartina maritima* were collected at the Etel river estuary (Morbihan, France), transplanted and maintained in controlled conditions in the greenhouse at the University of Rennes 1 (France). Total genomic DNA was isolated from fresh young leaves using the extraction kit Nucleospin Plant II (Macherey Nagel), following instructions provided by the manufacturer.

For comparative and molecular dating analyses (see below), genomic DNA was also extracted from other *Spartina* species including the hexaploid *S. alterniflora* (Landerneau, Finistère, France), the following tetraploid species: *S. bakeri* (Florida, USA), *S. arundinacea* (Amsterdam island), *S. patens* (New Jersey, USA), and representatives of related lineages from the Chlorioideae subfamily: *Sporobolus heterolepis* (Iowa, USA) and *Cynodon dactylon* (Ille et Vilaine, France).

2.2. High throughput sequencing, plastome assembly and annotation

Genomic DNA of *S. maritima* was subjected to one run of pyrosequencing using a GS-FLX 454 pyrosequencer (Life Sciences – Roche) at the Environmental and Functional Genomics platform (Biogenouest, Rennes). This run generated 993,229 reads (average length: 450 bp) after removal of low quality sequences. Reads corresponding to plastid DNA were extracted using a BLASTn (*E*-value: 10^{-6}) search against the plastome sequences of 23 Poales: *Anomochloa marantoidea* (Genbank accession: NC_014062), *Bambusa emeiensis* (NC_015830), *Bambusa oldhamii* (NC_012927), *Dendrocalamus latiflorus* (NC_013088), *Ferrocalamus rimosivaginus* (NC_015831), *Indocalamus longiauritus* (NC_015803), *Phyllostachys nigra* (NC_015826), *Phyllostachys edulis* (NC_015817), *Triticum aestivum* (NC_002762), *Festuca arundinacea* (NC_011713), *Lolium perenne* (NC_009950), *Coix lacryma-jobi* (NC_013273), *Saccharum officinarum* (NC_006084 and NC_005878), *Sorghum bicolor* (NC_008602), *Zea mays* (NC_001666), *Oryza nivara* (NC_005973), *Oryza sativa japonica* (NC_001320), *Oryza sativa indica* (NC_008155), *Brachypodium distachyon* (NC_011032), *Hordeum vulgare* (NC_008590), *Typha latifolia* (NC_013823) and *Phoenix dactylifera* (NC_013991). A total of 35,976 reads were recovered and assembled using Newbler (v. 2.6, Roche. Inc.). Thirty-two contigs ranging in size from 261 to 45,711 bp were obtained. Only the contigs covered by more than 50 reads were taken into account and organized using the plastome of *Saccharum officinarum* as a reference. All the genomic regions located at the junction between two contigs were verified by Sanger sequencing. The primers were designed using Primer 3 Plus (Rozen and Skaletsky, 2000) and are provided in supplementary Table S1. Additionally, a quarter of one

flow-cell lane containing genomic DNA of *S. maritima* was sequenced using an Illumina HiSeq 2000 platform (BGI, Hong-Kong), yielding 2 * 48.75 million of 100 bp paired-end reads from a library of ca. 500 bp DNA fragments. From this dataset, 1.02 million of reads were mapped on the draft *Spartina maritima* plastome using Bowtie (Langmead et al., 2009), and allowed to verify the plastome sequence obtained from the 454 reads. The *S. maritima* plastome sequence was deposited in Genbank (accession number: KP176438).

Plastome annotation was first performed using DOGMA (Dual Organellar GenoMe Annotator, http://dogma.ccbb.utexas.edu) (Wyman et al., 2004). The annotation was then verified and if necessary manually corrected by aligning *S. maritima* with other Poaceae plastomes and by comparing their annotation using Geneious R6 (http://www.geneious.com) (Drummond et al., 2010). A graphical representation of the plastome was realized using OGDRAW (http://ogdraw.mpimp-golm.mpg.de/) (Lohse et al., 2013).

2.3. PCR amplification

PCR amplifications were carried out in a total volume of 50 µl, containing 1X Green GoTaq Reaction Buffer (Promega), 0.2 mM of dNTP, 0.4 µM of each primer, 1 Unit of GoTaq DNA polymerase (Promega), mqH2O and 20 ng of template DNA. Cycling conditions were 94 °C for 2 min followed by 30 cycles at 94 °C for 30 s, xx–xx °C for 90 s, 72 °C for 90 s and a final extension at 72 °C for 7 min (the annealing temperature of each primer is indicated in Supplementary Table S1). PCR products were purified using the Spin Column PCR products purification kit (Bio basic Inc.) and sequenced directly on a capillary sequencer ABI 3730XL (Eurofins MWG Operon).

2.4. Repeat elements and indels detection

Repeated elements in the plastome of *S. maritima* (excluding one copy of the IR) were identified using REPuter (Kurtz et al., 2001). We looked for microsatellites (mono, di and trinucleotides), tandem, dispersed and palindromic repeats. As previously performed in other Poaceae species (Zhang et al., 2011), we focused on tandem repeats having a minimal size of 15 bp, as well as dispersed and palindromic repeats that had a size of at least 20 and 30 bp respectively (with a maximum distance of 3 kb between the repeats). Only repeated elements presenting at least 90% of sequence similarity between the two repeat copies were taken into account. Overlapping repeats were considered as one repeat motif and only the longest repeat was retained. We investigated if the repeated elements identified in the plastome of *S. maritima* were also present in the Chloridoid *Neyraudia reynaudiana* (NC_024262) and in the two related Panicoideae species (*Sorghum bicolor* and *Zea mays*) by aligning their plastomes using Geneious.

To identify large indels (>30 bp) that occurred in *Spartina* since its divergence from other grass lineages, we aligned the plastomes of *S. maritima* and of representatives of the following Poaceae subfamilies: Chloridoideae (*Neyraudia reynaudiana*), Panicoideae (*Coix lacryma jobi*, *Saccharum officinarum*, *Sorghum bicolor* and *Zea mays*) and Pooideae (*Brachypodium dystachion*). The alignment of these plastomes was performed using Geneious (Drummond et al., 2010). Genbank accessions of the chloroplast genomes used in this analysis are the same as presented earlier.

2.5. Sequence divergence between Spartina maritima and other Poaceae plastomes

Sequence divergence between the plastome of *Spartina maritima* (Chloridoideae) and those of other Poaceae subfamilies was evaluated independently for intergenic spacers, introns, exons, rRNAs and tRNAs by calculating pairwise distance between homologous regions. The plastomes were selected to represent each major group of sequenced Poaceae: one Anomochlooideae (*Anomochloa marantoidea*), two Bambusoideae (*Bambusa emeiensis*; *Phyllostachys edulis*), one Oryzoideae (*Oryza sativa japonica*), three Panicoideae (*Saccharum officinarum* hybrid cultivar NCo 310; *Sorghum bicolor*; *Zea mays*), and two Pooideae (*Hordeum vulgare* ssp. *vulgare*; *Triticum aestivum*). Genbank accessions of these plastomes are the same as in Section 2.2. Homologous sequences were aligned using MUSCLE (Edgar, 2004). Pairwise distances were calculated with the ***ape*** R-cran Package (Paradis et al., 2011. available at: http://cran.r-project.org/web/packages/ape/ape.pdf) using the K2p evolution model (Kimura, 1980) and distribution of sequence divergence along the chloroplast genome for each species relative to *Spartina maritima* was determined. All these steps were automated with custom scripts. The mean evolutionary rate of the different genetic categories (*e.g.* IGS, intron, exon, rRNA and tRNA) was compared using a Mann–Whitney test with Bonferroni correction. The same method was used to compare the mean evolutionary rates of the LSC, SSC and IR regions. Fast evolving regions were defined as regions displaying a mean sequence divergence higher than the mean sequence divergence plus one standard deviation of the concerned genetic category (IGS, intron, exon, rRNA and tRNA; LSC, SSC and IR). For each protein encoding gene, a Ka/Ks estimation between *S. maritima* and the other nine species was performed using the yn00 method (Yang and Nielsen, 2000) implemented in PAML (Yang, 2007).

2.6. Phylogeny and molecular dating analyses

We amplified four intergenic (*ndhC-trnV*, *trnL-trnF*, *trnT-trnL* and *trnY-trnD*) and three coding (*ndhF*, *matK*, *rbcL*) plastid regions in various Chloridoideae species in order to date divergence times within this subfamily as well as in other Poaceae (Table 1). For the three coding regions, the homologous sequences were also retrieved from 20 Poaceae plastomes including *Anomochloa marantoidea* (Anomochlooideae) used as outgroup. For the four non-coding regions, only the homologous sequences of a few representatives of each Poaceae subfamily were retrieved and used for subsequent analyses: two Bambusoideae (*Bambusa emeiensis* and *Phyllostachys edulis*), one Chloridoideae (*Neyraudia reynaudiana*), one Oryzoideae (*Oryza sativa*), three Panicoideae (*Saccharum officinarum*, *Sorghum bicolor* and *Zea mays*), two Pooideae (*Hordeum vulgare* and *Triticum aestivum*) and one Anomochlooideae (*Anomochloa marantoidea*, outgroup). Genbank accessions are the same as in Section 2.2. The sequences from each region were aligned using MAFFT (Katoh et al., 2002), and the coding and intergenic markers were respectively concatenated in two separate matrices. The best-fitted model of sequence evolution for each region (individual or concatenated) was determined using JModeltest (Posada, 2008) with the AICc as choice criterion. Maximum likelihood analyses were then performed for each matrix using PhyML (Guindon and Gascuel, 2003), with 1000 replicates of bootstrap. Molecular dating was achieved using BEAST v. 1.8.0 (Drummond et al., 2012). Bayesian analyses were run twice independently, for 60 million of generations with sampling every 1000 state and a burning phase corresponding to the first 6 millions of generations. The default settings were conserved, except for the topology building process (set to Yule process) and the molecular clock (set to relaxed lognormal clock). To get absolute divergence times, we calibrated the age of the most recent common ancestor of the Panicoideae and BOP clades (Bambusoideae–Oryzoideae–Pooideae, Soreng et al., 2015) and the divergence time between Chloridoideae and Panicoideae. Prasad et al. (2011) considered the effect of calibration with different fossil data on

Table 1
Chloridoideae species used for phylogeny and molecular dating analyses, and accession numbers of the plastid regions amplified in each species.

Species	Non coding regions				Coding regions			
	trnL-trnF	*trnT-trnL*	*trnY-trnD*	*ndhC-trnV*	*matK*	*rbcL*	*ndhF*
Spartina alterniflora	EU056306.1 (Fortune et al., 2008)	AF275667.1 (Baumel et al., 2002)	KJ882323	KJ882332	KJ882341	KJ88m2339	KJ882340
Spartina arundinacea	EU056303.1 (Fortune et al., 2008)	AF372631.1 (Baumel et al., 2002)	KJ882321	KJ882330	HE586084 (Aliscioni et al., 2012)	HE575832 (Aliscioni et al., 2012)	HE575783 (Aliscioni et al., 2012)
Spartina bakeri	EU056304.1 (Fortune et al., 2008)	AF372632.1 (Baumel et al., 2002)	KJ882326	KJ882334	–	–	–
Spartina maritima	EU056307.1 (Fortune et al., 2008)	AF372632.1 (Baumel et al., 2002)	KJ882322	KJ882331	KP176438	KP176438	KP176438
Spartina patens	EU056302.1 (Fortune et al., 2008)	AF372628.1 (Baumel et al., 2002)	KJ882327	KJ882335	–	–	–
Spartina pectinata	EF156731 (Columbus et al., 2007)	AF372625.1 (Baumel et al., 2002)	KJ882328	KJ882336	AF312353 (Hilu and Alice, 2001)	AJ784821 (Christin et al., 2008)	AF251465(Christin et al., 2008)
Cynodon dactylon	EU056300.1 (Fortune et al., 2008)	AF372633.1 (Baumel et al., 2002)	KJ882325	KJ882333	AF312331 (Hilu and Alice, 2001)	AM849393 (Christin et al., 2008)	AM849142 (Christin et al., 2008)
Sporobolus heterolepis	KJ882338	KJ882337	KJ882324	KJ882329	AF164429 (Hilu and Alice, 1999)	KJ740997 (Forrestel et al., 2014)	KJ740989 (Forrestel et al., 2014)

divergence time estimation in Poaceae, and found out that the addition of new Oryzeae fossils (phytoliths) yielded older ages, with a more or less pronounced effect depending on the lineage to which the phytoliths were attributed. The exact location of these fossils in the grass phylogeny being still under discussion (Bouchenak-Khelladi et al., 2010; Christin et al., 2014), we performed two dating analyses, both based on the calibration of the Panicoideae-BOP and of the Chloridoideae–Panicoideae clades using the ages estimated by Prasad et al. (2011), either (i) without or (ii) taking into account the phytoliths. In the latter case, we used the ages provided by the most likely placement of the phytoliths, which were also the oldest ages (Prasad et al., 2011) . The Panicoideae-BOP node was thus given an age of 72 or alternatively 86 my, and the Chloridoideae–Panicoideae divergence was calibrated as 50 or 57 my old (Prasad et al., 2011). To reflect the normal posterior probability distributions inferred by Prasad et al. (2011, Fig. S2) around these mean values, the prior probability distribution of these ages was set to a normal distribution with standard deviations of 8 and 6 my respectively. In order to evaluate the impact of using secondary calibrations, another dating analysis, based on the primary calibration of the Panicoideae-BOP clade using the fossil described in Crepet and Feldman (1991) with a lognormal prior distribution (mean 55 my, standard deviation 0.3 in log space, offset 25 my) was performed. This calibration was applied alone or together with the primary calibration of the stem age of *Oryza* by the phytoliths, using a lognormal prior distribution (mean 65 my, standard deviation 0.3 in log space; Prasad et al., 2005, 2011). Since the ages obtained with primary calibration were comprised in the results given by the secondary calibrations (not shown), only ages obtained from the more conservative secondary calibration analyses are presented.

3. Results

3.1. The plastid genome of Spartina maritima

The *Spartina maritima* plastome has a length of 135,592 bp and presents a quadripartite structure consisting of a pair of Inverted Repeats (21,011 bp) separated by a Large Single Copy (80,858 bp) and a Small Single Copy (12,712 bp) regions (Fig. 1, Table 2). Coding regions represent 60.6% whereas the intronic and intergenic spacer regions (IGS) correspond to 13.2% and 26.2% of the plastome (Table 2). The overall G + C content is of 37.4% and it is higher in the rRNAs (54.6%) and tRNAs (52.8%) than in the protein-encoding (38.4%) or intronic regions (36.2%).

The plastome of *S. maritima* encodes 110 different genes, including 76 protein-coding genes, 30 transfer RNAs (tRNA) and 4 ribosomal RNAs (rRNA; Table 2). Nineteen genes are duplicated in the IR: 7 protein-coding, 8 tRNA and 4 rRNA genes. Fifteen distinct genes contain one intron (five are duplicated in the IR) and one (*ycf3*) contains two introns.

3.2. Comparisons of the plastomes of S. maritima and other Poaceae: gene content, repeat elements and indels

The plastome of *S. maritima* (135,592 bp) presents a relatively similar size to other Poaceae plastomes, which range from 134,494 bp in *Oryza nivara* to 141,182 in *Saccharum*, the average plastome length in Poaceae being 138,187 bp. Plastid gene content and order are also similar in *Spartina* and in the other Poaceae.

Repeated elements within the plastome of *S. maritima* were identified using REPuter (Kurtz et al., 2001). In our search, we detected 34 repeated elements: 3 microsatellites, 15 tandem repeats of at least 15 bp, 12 palindromic sequences of at least 20 bp and four dispersed repeats of at least 30 bp (Table S2).

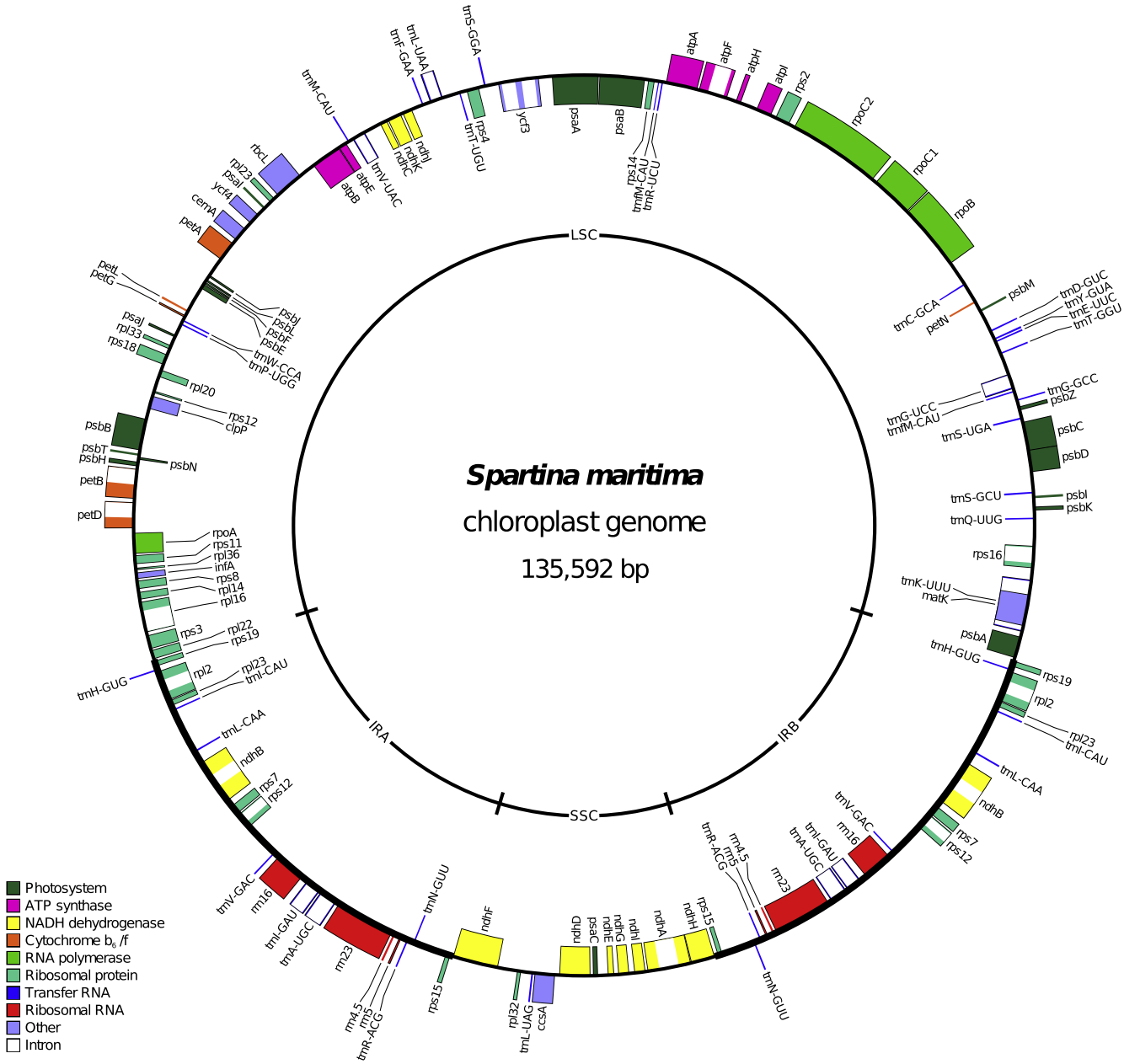


Fig. 1. The chloroplast genome of *S. maritima*.

They are mostly located in intergenic spacers and most of them are in the LSC region. Only eight repeats are present in the coding regions (e.g. *rpoC1*, *rpoC2*, *rps18* and *rpl16*) despite they totalize more than half of the plastome length. Within coding regions *rps18* displays the longest repeat (42 bp), which is present in the other Chloridoideae and in the Panicoideae analyzed. Respectively 53%, 40% and 33% of the palindromic sequences, tandem and dispersed repeats are also in the plastomes of the Chloridoideae *Neyraudia reynaudia* and of the Panicoideae *Zea mays* and *Sorghum bicolor*. Out of the 18 repeats conserved between *S. maritima* and *Neyraudia reynaudia*, seven (39%) are present in coding regions (Table S2). None of the three microsatellites identified in *S. maritima* are present in *Neyraudia reynaudia* or in the two Panicoideae species investigated.

The alignment of *S. maritima* and other Poaceae plastomes allowed the detection of 12 large indels (>30 bp) that occurred in *Spartina maritima* since its divergence from the Chloridoideae *Neyraudia reynaudiana* (Table 3). These indels (11 deletions and one insertion) range in size from 36 to 93 bp, are nearly all present in non-coding regions (except for one deletion in *rpoC2*) and 11 of them are located in the Large Single Copy, the remaining one being located in the IR. More than half of the deletions are flanked by direct repeats of 3–12 nucleotides (Table 3).

3.3. Sequence divergence

A comparison of pairwise distances (K2p) calculated between *S. maritima* and other Poaceae (Table S2) revealed that IGSs evolve significantly more rapidly than introns with a distance of 0.070 ± 0.002 substitutions/site (subst./site) against 0.050 ± 0.003, the latter still evolving significantly faster than CDSs (0.040 ± 0.001). Slowest evolving regions are tRNAs (0.012 ± 0.002) and rRNAs (0.002 ± 0.0006). Among introns (Table S3), the mean sequence divergence range from 0.008 (*rps12* intron) to 0.091 (*ndhA* intron). Nine of the 23 introns presenting a minimal size of 500 bp evolve particularly rapidly,

Table 2
Characteristics of the plastome of *Spartina maritima*.

Plastome characteristics	*Spartina maritima*
Size (bp)	135,592
LSC size in bp (%)	80,858 (59.6)
SSC size in bp (%)	12,712 (9.4)
IR length in bp (%)	21,011 (31)
Size in bp (%) coding regions	82,150 (60.6)
Size in bp (%) of protein-encoding regions	70,115 (51.7)
Size in bp (%) of introns	17,964 (13.2)
Size in bp (%) of rRNA	9190 (6.8)
Size in bp (%) of tRNA	2845 (2.1)
Size in bp (%) of IGS	35,478 (26.2)
Number of different genes	110
Number of different protein-encoding genes	76
Number of different tRNA genes	30
Number of different rRNA genes	4
Number of different genes duplicated by IR	19
Number of different genes with introns	16
Overall % GC content[a]	37.4
% GC content in protein-encoding regions[a]	38.4
% GC content in introns[a]	36.2
% GC content in rRNA[a]	54.6
% GC content in tRNA[a]	52.8

[a] The sequences of the two Inverted Repeats were taken into account for this analysis.

displaying a sequence divergence higher than the mean (+ standard deviation): the *ndhA* intron (K2p in subst./site = 0.091; 1035 bp), *rpl16* intron (0.081; 1085 bp), *trnK_UUU* intron (0.078; 676 bp), *trnL_UAA* intron (0.074; 554 bp), *rps16* intron (0.072; 826 bp), *atpF* intron (0.067; 837 bp), *petB* intron (0.067; 779 bp), *ycf3* intron 2 (0.067; 740 bp), *ycf3* intron 1 (0.053; 726 bp). Similarly, the 15 IGS regions presenting the highest sequence divergence (from 0 to 0.195) and a minimal size of 500 bp (Table S3) are: *rpl32-trnL_UAG* (0.151; 503 bp), *trnK_UUU-rps16* (0.144; 549 bp), *ndhC-trnV_UAC* (0.136; 835 bp), *trnT_UGU-trnL_UAA* (0.129; 811 bp), *petN-trnC_GCA* (0.127; 901 bp), *ndhF-rpl32* (0.127; 935 bp), *atpI-atpH* (0.127; 819 bp), *trnF_GAA-ndhJ* (0.126; 536 bp), *trnG_UCC-trnT_GGU* (0.124; 1012 bp), *petA-psbJ* (0.123; 909 bp), *ycf3-trnS_GGA* (0.119; 587 bp); *trnD_GUC-psbM* (0.117; 995 bp), *rps16-trnQ_UUG* (0.108; 1273 bp), *trnT_GGU-trnE_UUC* (0.106; 540 bp), *trnP_UGG-psaJ* (0.106; 542 bp). For the protein coding-regions, sequence divergence was evaluated by comparing the synonymous (Ks) substitution rates (Table S4). The mean sequence divergence ranges from 0.007 (*psbL*) to 0.273 (*rpl22*). Nine genes present a higher sequence divergence than the mean gene divergence rate and a minimum size of 500 bp (Table S4): *ndhH* (0.234; 1182 bp), *rps3* (0.217; 675 bp), *petA* (0.209; 963 bp), *ndhF* (0.208; 2220 bp), *ndhA* (0.186; 539 bp), *ccsA* (0.185; 969 bp), *petB* (0.184; 642 bp), *matK* (0.178; 1542 bp), *ndhI* (0.173; 543 bp). For all protein-coding genes except *rps19*, *psaI* and *rps12*, the Ka/Ks ratio is <0.5 (Table S6), indicative of high purifying selective constraint acting on the plastid genes.

These analyses of sequence divergence allowed identifying four fast evolving regions in the Poaceae plastome (Fig. 2): between the *trnK_UUU* and *trnQ_UGG* (5520 bp), the *trnG_UCC* and *trnC_GCA* (4754 bp), the *yfc3* and *trnV_UAC* (8805 bp) and the *ndhF-ccsA* regions (4976 bp). The remaining fast evolving sequences identified are dispersed in the chloroplast genome. No fast evolving regions have been identified in the IR, which evolves significantly more slowly than the LSC and SSC regions.

3.4. Phylogeny and molecular dating

Phylogenetic analyses were first performed separately for each marker; since no incongruence was encountered among the topologies, only the results from the two concatenated matrices of coding and of non-coding regions are presented here. Phylogenetic relationships obtained with the maximum likelihood approach (not shown) were identical to those obtained using the Bayesian analysis, which are presented in Fig. 3. The topologies obtained with matrices of non-coding (Fig. 3, left) and coding (Fig. 3, right) sequences agree with the previously known phylogenetic relationships among the grass subfamilies: Bambusoideae, Oryzoideae and Pooideae form a monophyletic lineage ("BOP" clade), with Oryzoideae (represented by *Oryza*) placed either as sister to the Pooideae–Bambusoideae sub-clade when using coding sequences (Fig. 3, right) or as sister to Bambusoideae when using the more variable non-coding regions (Fig. 3, left). The second lineage is composed of Panicoideae and Chloridoideae (including *Spartina*) as sister subclades (Fig. 3). In the Chloridoideae, *Neyraudia* is the first to diverge, followed by *Cynodon*, and *Sporobolus* is sister to *Spartina*. Within *Spartina*, tetraploid and hexaploid species form two monophyletic lineages, and among the tetraploids *S. pectinata* is sister to a clade formed by *S. patens* and *S. bakeri* (Fig. 3, left).

The divergence times are similar regardless of the fossils used for calibration (see Section 2.6), and they are also congruent between coding and non-coding markers, with largely-overlapping 95% highest posterior densities (HPD), as presented in Fig. 3. When considering the mean ages obtained with coding and non-coding sequences, the basal genus *Anomochloa* splitted from the other Poaceae 87–108 mya, whereas the BOP and PACMAD clades diverged 69–84 my ago. Within Pooideae, *Brachypodium* diverged from *Triticum* 42–50 mya, itself separated by 37–44 my from the *Festuca* lineage. The split between *Triticum* and *Hordeum* is estimated at 14–21 my, and the one between *Festuca* and *Lolium* is more recent (7–8 my). *Agrostis* seems to have diverged from the latter 27–33 mya. The Bambusoideae

Table 3
Identification of the indels that occurred in the plastome of *Spartina maritima* since its divergence with *Neyraudia reynaudiana*. The location, type of event (insertion/deletion) and length of the indels (in comparison with *Neyraudia reynaudiana*) is shown. The presence of direct repeats flanking the sequences deleted in *S. maritima* is indicated.

Genes surrounding the indel	Indel location in *S. maritima*	Type	Indel length in *S. maritima*	Repeated motif flanking the deletion
rps16-trnQ_UUG	5874–5875 (LSC)	Deletion	85 bp	AACAAATAAACTAT
trnS_GCU-psbD	8437–8438 (LSC)	Deletion	36 bp	TTCTTT
trnD_GUC-psbM	16,405–16,406 (LSC)	Deletion	54 bp	
psbM-petN	17,641–17,642 (LSC)	Deletion	55 bp	
rpoC2	27,369–27,370 (LSC)	Deletion	42 bp	ACCC
trnR_UCU-trnfM_CAU	36,737–36,738 (LSC)	Deletion	45 bp	AAAAAA
ndhC-trnV_UAC	50,920–50,921 (LSC)	Deletion	93 bp	CTA
rbcL-rpl23	56,565–56,566 (LSC)	Deletion	54 bp	ATAGA
psbE-petL	62,665–62,666 (LSC)	Deletion	82 bp	
trnP_UGG-psaJ	64,700–64,760 (LSC)	Insertion	61 bp	
rpl33-rps18	65,536–65,537 (LSC)	Deletion	63 bp	ATT
rps12-trnV_GAC	91,297–91,298	Deletion	46 bp	

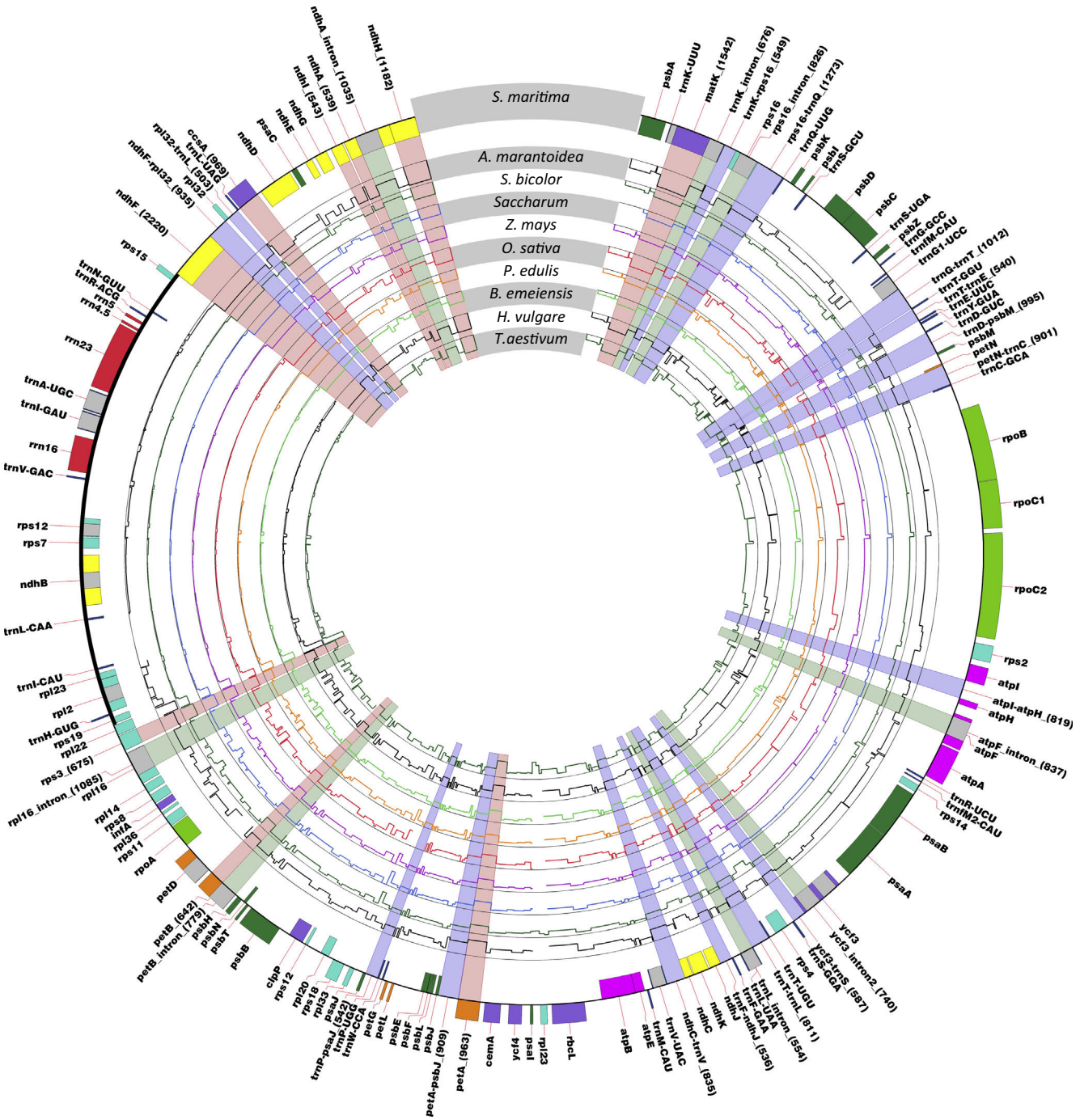


Fig. 2. Pairwise distance between *Spartina maritima* and other Poaceae orthologous plastomic regions. Pairwise distance for non-coding regions was calculated using the K2p evolution model (*e.g.* Kimura, 1980). For each protein encoding genes Ka/Ks estimation between *S. maritima* and the other nine species was performed using the yn00 method (Yang and Nielsen, 2000) implemented in PAML (Yang, 2007). The gene encoding, intronic or intergenic regions presenting the highest evolutionary rates in *S. maritima* in comparison with other Poaceae plastomes and a minimum length of 500 bp are highlighted in red, green and blue respectively. The size (in bp) of each highlighted region in *S. maritima* is indicated under brackets. Only one IR region is represented on each plastome map. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Dendrocalamus and *Bambusa* split 4 mya, and separated from the lineage of *Phyllostachys* 21–32 my ago. The latter diverged from *Ferrocalamus* 5–6 mya, and 1 mya from *Indocalamus* (Fig. 3). *Bambusa oldhamii* and *B. emeiensis* separated 2 mya. Within Chloridoideae, *Neyraudia reynaudiana* split from *Cynodon* 34–43 mya, *Cynodon* diverged from *Spartina* and *Sporobolus* 27–35 mya, and *Spartina* diverged from *Sporobolus* 12–20 mya (Fig. 3). The divergence between the chloroplast sequences of the hexaploid and the tetraploid *Spartina* lineages is estimated to have occurred 6–10 mya. In the hexaploid lineage, the divergence time between the plastid sequences of *S. maritima* and *S. alterniflora* is estimated at 2–4 mya. In the tetraploids, coding sequences suggest that *S. arundinacea* diverged from *S. pectinata* 3–4 mya (Fig. 3, right) whereas non-coding sequences suggest an age of 6–7 my (Fig. 3, left). *S. pectinata* diverged 3–4 mya from the *S. bakeri*-*S. patens* sub-clade, and the latter species separated 2 mya (Fig. 3, left).

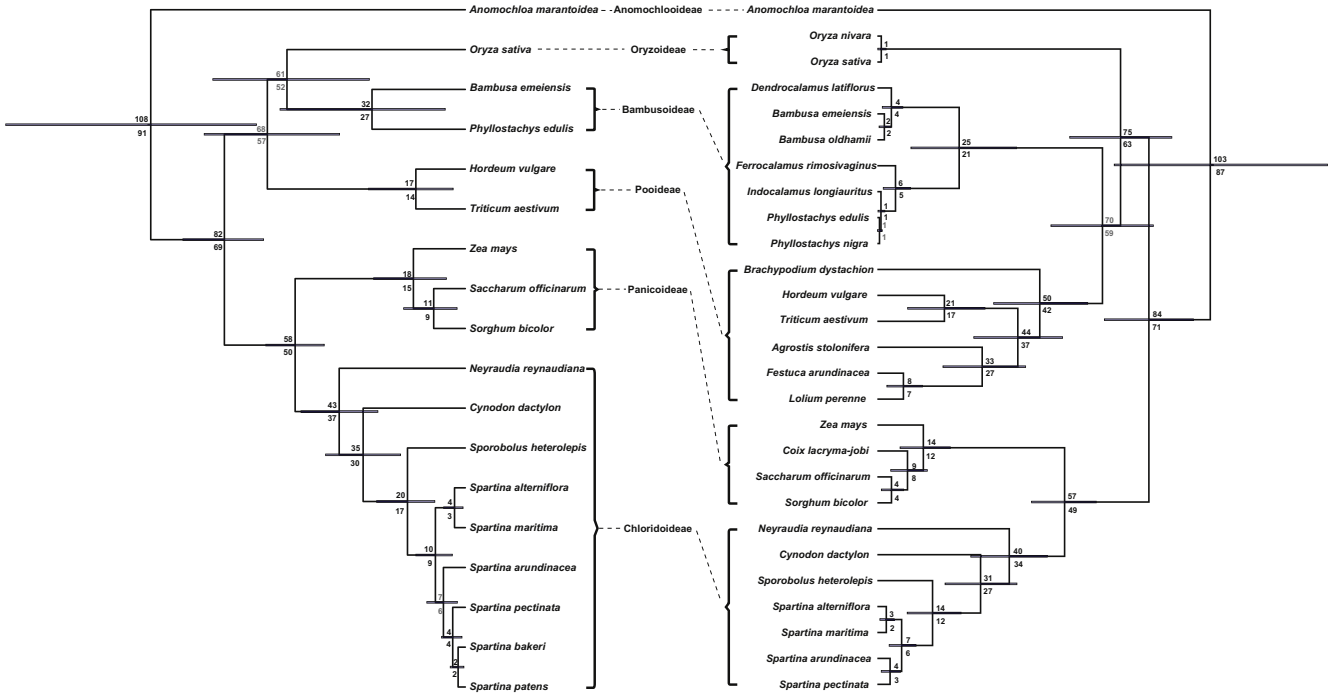


Fig. 3. Phylogenetic relationships and divergence times of Chloridoideae and other Poaceae based on the Bayesian analysis of concatenated non-coding (tree on the left) and coding (tree on the right) chloroplast sequences. Numbers above and below the branches are the ages in million years obtained respectively with the calibrations based on the new fossils and based only on the traditionally used fossils (see Section 2.6 of the *Material and methods*, and Prasad et al., 2011). Blue bars show the 95% highest posterior density intervals of the ages obtained with the calibration based on the new fossils. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

In this study, we sequenced the chloroplast genome of *Spartina maritima*, shedding light on the plastome structure within Chloridoideae, which comprise 1600 species in 31 genera (Soreng et al., 2015). The comparison of the chloroplast genomes of *Spartina maritima* and the recently sequenced Chloridoideae *Neyraudia reynaudiana* (Wysocki et al., 2014) contributed to fill an important gap in the knowledge of plastid genome evolution in this group and revealed variable regions of possible taxonomic and phylogenetic utility in this large subfamily. We used some of these markers to bring new insights on the evolutionary history and divergence times of the *Spartina* lineages, of particular interest in the context of the prominent reticulate history of this group.

4.1. The plastome of Spartina compared to other Poaceae plastomes

The size, gene content and organization of the plastome of *Spartina* are similar to that of other Poaceae. *Spartina* displays the typical gene set found in most angiosperms (reviewed in Wicke et al., 2011), but like other grasses, it lacks the *accD*, *ycf1* and *ycf2* genes. Since these genes are essential for the survival of photosynthetic plants (Drescher et al., 2000; Kode et al., 2005), they were most likely functionally transferred to the nucleus or functionally replaced by an eukaryotic gene, as observed for the *accD* plastid gene in other plant families (Babiychuk et al., 2011; Rousseau-Gueutin et al., 2013).

Large indels may be reliable markers to discriminate Chloridoideae tribes or genera. We identified 12 indels (>30 bp) that occurred since the divergence of *S. maritima* with *Neyraudia reynaudiana*. There was 11 deletions and 1 insertion, almost all located in intergenic regions of the LSC, as it is observed in other Poaceae (Yamane et al., 2006). Most of these indels are flanked by direct repeats, suggesting that replication slippage is responsible for the deletions (Bzymek and Lovett, 2001; Rousseau-Gueutin et al., 2011).

Considering the importance of repeat elements in plastome rearrangement (Weng et al., 2014), we investigated the presence of large tandem, dispersed and palindromic repeat sequences. We identified a similar number of repeats (31) than in other Poaceae species (Zhang et al., 2011), with about a half shared with the closely related Chloridoideae *Neynaudia reynaudiana*. This very low number of repeats, compared to other plant families such as Geraniaceae (Guisinger et al., 2011), is in accordance with the higher conservation of the Poaceae plastome organization (Zhang et al., 2011).

4.2. Chloroplast sequence evolution

By estimating sequence divergence between the plastomes of *S. maritima* and other Poaceae, we identified the most variable regions in different genetic categories (protein-coding genes, introns, intergenic spacers). These regions will be of high interest for phylogenetic studies as well as for DNA barcoding, especially in the context of intercontinental biological invasion involving several *Spartina* hybrid taxa (Ainouche et al., 2009; Strong and Ayres, 2013). Such variable markers can also be used to improve the still-debated phylogeny of Chloridoideae, of which the relationships between about one third of the genera still need to be explored (Peterson et al., 2012). As previously observed in other species (Martin et al., 2014), intergenic spacers evolved faster than introns, followed by protein coding genes. Some of the fast evolving regions identified in this study have been already used in previous Chloridoideae phylogenetic studies (at the subfamily or genus level): the *cssA*, *matK*, *ndhF* and *rps3* genes (Hilu and Alice, 2001; Peterson et al., 2010a, 2010b; Peterson et al., 2012); the

ndhA, *rps16* and *trnL* introns (Peterson et al., 2014b); and the *rpl32-trnL*, *trnK-rps16* and *trnT-trnL* intergenic spacers (Baumel et al., 2002; Columbus et al., 2007; Peterson et al., 2012, 2014b; Siqueiros-Delgado et al., 2013). Here, we provide an additional set of variable genic (*ndhA, ndhH, ndhI*, *petA*, *petB*), intronic (*atpF*, *petB*, *rpl16*, *trnK_UUU, petB* and *ycf3*) and intergenic (*ndhC-trnV, petN-trnC*, *ndhF-rpl32*, *atpI-atpH*, *trnF-ndhJ, trnG-trnT*, *petA-psbJ*, *ycf3-trnS*, *trnD-psbM*, *rps16-trnQ, trnT-trnE*, *trnP-psaJ*) markers that will facilitate tracking the evolutionary history of Chloridoideae. In recent Chloridoideae phylogenies, *Spartina* appears embedded in the polyphyletic genus *Sporobolus*, in the subtribe Sporobolineae, and tribe Zoysieae (Peterson et al., 2010a), which resulted in proposition of new taxonomic treatment of the genus (Peterson et al., 2014c). Additional variable regions detected in the chloroplast genome, combined to extensive species sampling should help the ongoing efforts to improve our understanding of the phylogenetic relationships and evolutionary history of this group.

4.3. Phylogeny and molecular dating

Our molecular phylogenies based on coding or non-coding plastid sequences agreed with the previously published phylogenetic trees (Baumel et al., 2002; Christin et al., 2008; Spriggs et al., 2014). Only the position of *Oryza* relative to other members of the BOP clade could not be resolved: it was either placed as sister to Bambusoideae (non-coding sequences) or, as commonly accepted (Wu and Ge, 2012), as sister to Pooideae and Bambusoideae (coding sequences). This incongruence (low support in both cases) can be explained by the rapid radiation of the BOP clade that occurred in a 4 my interval (Wu and Ge, 2012). The position of *Sporobolus heterolepis* as the sister taxon of the monophyletic *Spartina* clade is in agreement with other studies (Bouchenak-Khelladi et al., 2008; Peterson et al., 2014b, 2014c). Within *Spartina*, the hexaploid *S. maritima* and *S. alterniflora* on one hand, and the tetraploid *S. bakeri*, *S. patens* and *S. pectinata* on the other hand, form two clades that were recently named as *Sporobolus* sect. Spartina subsect. Alterniflori and Spartina respectively (Peterson et al., 2014b). The position of the tetraploid *S. arundinacea* as sister to the other tetraploids lacks support but does not contradict previous studies (Fortune et al., 2007, 2008).

We used our two Poaceae phylogenies to estimate the ages of various grass lineages, with focus on *Spartina*. Secondary calibrations were applied on the ages of the most recent common ancestors of Panicoideae and Chloridoideae as well as of Panicoideae and the BOP clade. It has been shown (Prasad et al., 2011) that when newly discovered Oryzeae fossils are used to calibrate the phylogeny of Poaceae, older ages than the widely accepted ones (*e.g.* Christin et al., 2008, 2014; Vicentini et al., 2008) are obtained. Prasad et al. (2011) also pointed out that these older estimates fit better with the ages recently found for other land plants as well as with the current biogeographical scenario of the origin and expansion of grasses. In order to facilitate comparisons with previous studies, we used two secondary calibration strategies based on alternatively the older (considering the new fossils) or the younger ages estimated by Prasad et al. (2011).

The ages of the major Poaceae splits obtained are in agreement with, or slightly older than, those previously published. According to our analyses, *Anomochloa* diverged from the other Poaceae subfamilies 87–108 mya, compared to 69–88 mya in Christin et al. (2014) and 107–127 mya in Prasad et al. (2011); *Oryza* separated from Bambusoideae and Pooideae 63–75 mya, compared to 53–73 mya in Christin et al. (2014); and the two latter clades diverged 59–70 mya compared to 50–68 in Christin et al. (2014). In Bambusoideae, we found that *Bambusa* diverged from *Dendrocalamus* 4 mya, as previously observed (Wu and Ge, 2012; 3 mya in Christin et al., 2014). Burke et al. (2014) estimated the same split at 6 mya, whereas they found the same age than us for the divergence between *Bambusa oldhamii* and *B. emeiensis* (2 mya) and similar estimates for the divergence of the other Bambusoideae *Ferrocalamus* and *Phyllostachys*, and for the divergence of this latter and *Indocalamus*, estimated here at maximum ca. 6 my and 1 my respectively (Fig. 3). The age of 5–6 my found here and in Burke et al. (2014) for the split between *Ferrocalamus* and *Phyllostachys* is puzzling because this divergence is supposed to be older than the split between *Chimonobambusa* and *Phyllostachys* (Triplett et al., 2010), which was estimated as 10 mya (Christin et al., 2008; Vicentini et al., 2008). In Pooideae, *Brachypodium* diverged from *Triticum* 42–50 mya, compared to 30–33 mya in Prasad et al. (2011) and 34–44 mya in Christin et al. (2014); *Triticum* diverged from *Agrostis* 37–44 mya and from *Hordeum* 14–21 mya compared to 31–39 and 14–16 mya in Christin et al. (2014), and *Agrostis* diverged from *Festuca* 27–33 mya compared to 27–32 mya in Christin et al. (2014). In the PACMAD clade, *Zea* diverged from *Sorghum* 12–18 mya, which diverged from *Saccharum* 4–11 mya ago (compared to 12–14 and 3–4 mya in Christin et al., 2014); *Neyraudia* (Triraphideae tribe: Peterson et al., 2011; Soreng et al., 2015) separated 34–43 mya from other Chloridoideae, compared to 32–41 in Christin et al. (2014), and *Cynodon* diverged from *Spartina* 27–35 mya compared to 26–33 in Christin et al. (2014). The *Spartina* clade (also named section *Spartina*, Peterson et al., 2014b) diverged from *Sporobolus heterolepis* (sect. *Calamovilfa*, Peterson et al., 2014b) 12–20 mya, whereas Prasad et al. (2011) estimated them to have diverged from the more distantly related (Peterson et al., 2014b) *Sporobolus festivus* 13–15 mya. We found the divergence between the tetraploid and hexaploid *Spartina* lineages (sections *Alterniflori* and *Spartina* respectively, Peterson et al., 2014b) dating back to 6–10 mya. This divergence is the same as the divergence between the tetraploid *S. pectinata* and the allododecaploid *S. anglica* (which inherited the chloroplast genome of the hexaploid *S. alterniflora*), and was previously estimated as 4 my old (Christin et al., 2008). This difference may result from the different calibration methods and taxon sampling used by Christin et al. (2008). Our taxon sampling focused on *Spartina*, whereas Christin et al. (2014) used more Poaceae representatives (and less *Spartina* species), and a different set of calibrations outside Poaceae, resulting in more constrained stem ages for the early-divergent lineages of the family. This could explain why the differences observed here between our estimations and those of Christin et al. (2014) decrease with the phylogenetic distance. Bouchenak-Khelladi et al. (2010) found ages in the order of those found by Christin et al. (2014), or younger, which again reflects their dense taxon sampling coupled with 4 calibrations, among which the phytoliths could only impact the age of the early-diverging lineages. The ages estimated by Spriggs et al. (2014) using similar calibrations than ours on a much broader taxon sampling are, as expected, older when including the phytoliths, but also older than the ages we estimated using the same calibration, whereas without including the phytoliths, they found ages younger than ours and even younger than those reported by Christin et al. (2014).

Spartina diversified on the American continent (Mobberley, 1956), and until the 19th century, *S. maritima* was the only known Old World species. Our molecular dating suggests that the ancestor of *S. maritima* separated from *S. alterniflora* and colonized the Old-World 2–4 million years ago. *S. maritima* has a disjunct distribution in Western Europe and South-Africa; although populations from its northern limit (southern England and Northern Brittany in France) are regressing (Marchant, 1967; Raybould et al., 1991), healthy populations are growing along the coasts of the Iberian Peninsula. *Spartina maritima* populations are also recorded in the Mediterranean region (e.g. Venice Lagoon: Silvestri et al., 2005), and colonized several estuaries of the eastern south-African coast

(Pierce, 1982). It is not clear yet whether *S. maritima* first arrived in western Europe and subsequently spread, was introduced southwards (Pierce, 1982), or has an African origin (Marchant, 1967). To date, phylogeographic studies were hampered by the lack of genetic variation among *S. maritima* populations (Yannic et al., 2004). The most variable regions that we found in its plastome will provide new opportunities to explore the history of this Old-World lineage.

Dating the divergence time between the hexaploid *S. alterniflora* and *S. maritima* is also of particular interest in the context of allopolyploid speciation, since hybridization between these species in Europe during the 19th century resulted in the formation of a perennial sterile hybrid (*S. x townsendii*), which gave rise to a new vigorous and invasive allododecaploid species *S. anglica* (Ainouche et al., 2009). This latter species thus contains two duplicated hexaploid homoeologous genomes, which appear to have diverged only 2–4 mya. It is then not surprising that the first generation hybrid (*S. x townsendii*) is sterile and exhibits meiotic abnormalities (Marchant, 1968), increasing the likelihood of producing unreduced gametes and the formation of an allopolyploid species (Buggs et al., 2008).

4.4. Molecular dating in polyploids using plastid markers

It is important to consider the meaning of the ages estimated for the divergence between chloroplast genomes of polyploid species, especially in the context of hybrid origins and maternal inheritance of plastomes. We assume that the hexaploid *Spartina* species have a reticulate (*i.e.* allopolyploid) origin according to the propensity of interspecific hybridization in *Spartina* and to previously published nuclear gene phylogenies in this genus (Ainouche et al., 2012; Fortune et al., 2007). The times we estimated from their plastid sequences have different possible meanings regarding the species history. Fig. 4 illustrates two possible scenarios that could happen in the case of an allotetraploid species deriving from two ancestral diploid species (AA and BB genomes) with the AA species as maternal parent, and of an allohexaploid deriving from this tetraploid (AABB) and from a third diploid species (CC genome). The first scenario assumes that the hexaploid AABBCC inherited its chloroplast genome from the tetraploid AABB and thus that both hexaploid and tetraploid exhibit an "AA-type" chloroplast genome (the diploid AA being the original plastome donor). Sequence divergence between the chloroplast genomes of the tetraploid and the hexaploid species will thus reflect the time passed since the divergence between these species (Fig. 4). The second possible scenario assumes that the hexaploid inherited its plastome from the diploid CC. In this case, the sequence divergence between the tetraploid and hexaploid chloroplast genomes will reflect the time passed since the divergence between the ancestral diploid genomes AA and CC, instead of the divergence time between the tetraploid and hexaploid species (Fig. 4). In natural populations, an additional complication may arise from bidirectional hybridization events resulting in different individuals exhibiting alternatively the plastome from one or the other parental species (Soltis et al., 2004). Altogether, these examples suggest that the divergence time estimated here from chloroplast sequences represent the upper age of the divergence between the hexaploid and tetraploid species. Dating the origin of polyploidy events is a non-trivial task when considering the various ways polyploids may form, and this is even more complicated when using nuclear sequence data (Doyle and Egan, 2010). The recurrence of polyploidy and the prominence of reticulate evolution in plants are well-documented (Jiao et al., 2011) but such issues are rarely considered in molecular dating studies, although it is recognized that molecular dating generally overestimate the actual species divergence times because ancient haplotype lineages (predating the speciation event) may have been sampled in the analyzed species (Middleton et al., 2014).

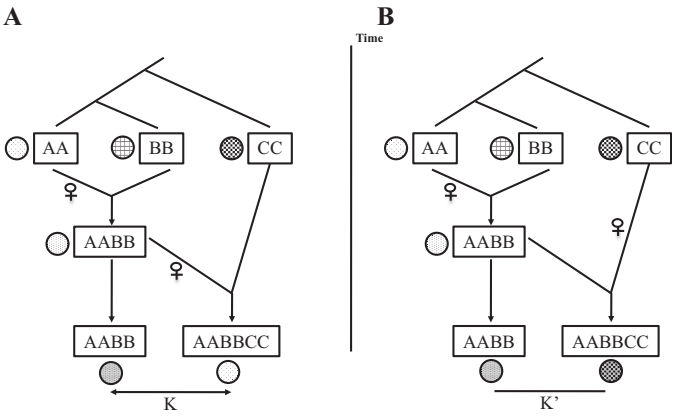


Fig. 4. Dating polyploid species divergence using maternally inherited chloroplast sequences. In this example, three different diploid species (with AA, BB, or CC nuclear genomes) have contributed to the formation of an allotetraploid (AABB nuclear genome, A-type maternal genome) and an allohexaploid (AABBCC) species. Two possible acquisitions of the allohexaploid chloroplast genome are considered. (A) The allotetraploid provided the maternal genome to the allohexaploid, which is then also A-type. Sequence divergence (K) between the plastomes of the allotetraploid and the allohexaploid species will then reflect the divergence time between these two species. (B) The allohexaploid inherited the chloroplast genome from the CC diploid genome. Sequence divergence between the plastomes of the allotetraploid and the allohexaploid species (K') will then reflect the time elapsed since the divergence of the diploid ancestors (AA and CC diploids).

In conclusion, using the reconstructed chloroplast genome of *S. maritima* we provided the first comparative analysis of Poaceae plastomes involving Chloridoideae and showed that *Spartina* displays autapomorphic indels even though the plastome structure is well conserved in the whole family. These indels as well as the highly variable coding and non-coding regions that we identified represent a valuable resource for future phylogenetic or phylogeographic studies in Chloridoideae, and especially in *Spartina* where reticulate evolution and polyploidy are prominent processes. Molecular dating based on plastid markers revealed that *S. maritima* colonized the Old-World sometimes 2–4 my ago, which is also the time of divergence between the two genomes reunited in the invasive allopolyploid *S. anglica*. Dating divergence in polyploids is facilitated by the use of haploid markers but even the ages derived from plastid sequences can be misleading when the precise modalities of hybridization and polyploidization in the clade of interest are not well understood.

Acknowledgments

This work was supported by UMR-CNRS Ecobio (Rennes, France), and benefited from facilities and support from the "Environmental and Functional Genomics" and "Genouest Bioinformatic" Platforms (University of Rennes 1). We acknowledge financial support from the "Région Bretagne", the Partner University Funds and the European Union Seventh Framework Programme [FP7-CIG-2013 – 2017; grant no. 333709 to M.R.-G.]. Dr. P.M. Peterson and an anonymous reviewer are thanked for helpful comments on an earlier version of this paper.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2015.06.013.

References

Ainouche, M.L., Fortune, P.M., Salmon, A., Parisod, C., Grandbastien, M.A., Fukunaga, K., Ricou, M., Misset, M.T., 2009. Hybridization, polyploidy and invasion: lessons from Spartina (Poaceae). Biol. Invasions 11, 1159–1173.

Ainouche, M.L., Chelaifa, H., Ferreira de Carvalho, J., Bellot, S., Ainouche, A.K., Salmon, A., 2012. Polyploid evolution in *Spartina*: dealing with highly redundant hybrid genomes. In: Soltis, P.S., Soltis, D.E. (Eds.), Polyploidy and Genome Evolution. Springer Verlag, Berlin, pp. 225–243.

Ainouche, M.L., Wendel, J.F., 2013. Polyploid speciation and genome evolution: lessons from recent allopolyploids. In: Pontarotti, P. (Ed.), 17th Evolutiuonary Biology Meeting. Springer Verlag, Marseilles, pp. 87–113.

Aliscioni, S., Bell, H.L., Besnard, G., Christin, P.A., Columbus, J.T., Duvall, M.R., Edwards, E.J., Giussani, L., Hasenstab-Lehman, K., Hilu, K.W., Hodkinson, T.R., Ingram, A.L., Kellog, E.A., Mashayekhi, S., Morrone, O., Osborne, C.P., Salamin, N., Schaefer, H., Spriggs, E., Smith, S.A., Zuloaga, F., 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. New Phytol. 193, 304–312.

Babiychuk, E., Vandepoele, K., Wissing, J., Garcia-Diaz, M., De Rycke, R., Akbari, H., Joubes, J., Beeckman, T., Jansch, L., Frentzen, M., Van Montagu, M.C., Kushnir, S., 2011. Plastid gene expression and plant development require a plastidic protein of the mitochondrial transcription termination factor family. Proc. Natl. Acad. Sci. USA 108, 6674–6679.

Baumel, A., Ainouche, M.L., Bayer, R.J., Ainouche, A.K., Misset, M.T., 2002. Molecular phylogeny of hybridizing species from the genus Spartina Schreb. (Poaceae). Mol. Phylogenet. Evol. 22, 303–314.

Baumel, A., Ainouche, M., Misset, M., Gourret, J., Bayer, R., 2003. Genetic evidence for hybridization between the native *Spartina maritima* and the introduced Spartina alterniflora (Poaceae) in South-West France. Spartina × neyrautii re-examined. Plant Syst. Evol. 237, 87–97.

Blazier, C.J., Guisinguer, M.M., Jansen, R.K., 2011. Recent loss of plastid-encoded ndh genes within Erodium (Geraniaceae). Plant Mol. Biol. 76, 263–272.

Bouchenak-Khelladi, Y., Salamin, N., Savolainen, V., Forest, F., van der Bank, M., Chase, M.W., Hodkinson, T.R., 2008. Large multi-gene phylogenetic trees of the grasses (Poaceae): progress towards complete tribal and generic level sampling. Mol. Phylogenet. Evol. 47, 488–505.

Bouchenak-Khelladi, Y., Verboom, G.A., Savolainen, V., Hodkinson, T.R., 2010. Biogeography of the grasses (Poaceae): a phylogenetic approach to reveal evolutionary history in geographical space and geological time. Bot. J. Linn. Soc. 162, 543–557.

Buggs, R.J.A., Soltis, P.S., Evgeny, V., Mavrodiev, V., Symonds, V., Soltis, D.E., 2008. Does phylogenetic distance between parental genomes governs the success of polyploids? Castanea 73, 74–93.

Burke, S., Clark, L.G., Triplet, J.K., Grennan, C.P., Duval, M.R., 2014. Biogeography and phylogenomics of new world bambusoideae (Poaceae), revisited. Am. J. Bot. 101, 886–891.

Bzymek, M., Lovett, S.T., 2001. Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. Proc. Natl. Acad. Sci. USA 98, 8319–8325.

Castillo, J.M., Ayres, D.R., Leira-Doce, P., Bailey, J., Blum, M., Strong, D.R., Luque, T., Figueroa, E., 2010. The production of hybrids with high ecological amplitude between exotic *Spartina densiflora* and native *S. maritima* in the Iberian Peninsula. Divers. Distrib. 16, 547–558.

Christin, P.A., Besnard, G., Samaritani, E., Duvall, M.R., Hodkinson, T.R., Savolainen, V., Salamin, N., 2008. Oligocene CO_2 decline promoted C_4 photosynthesis in grasses. Curr. Biol. 18, 37–43.

Christin, P.A., Spriggs, E., Osborne, C.P., Strômberg, C.A.E., Salamin, N., Edwards, E.J., 2014. Molecular dating, evolutionary rates, and the age of the grasses. Syst. Biol. 63, 153–165.

Chumley, T.W., Palmer, J.D., Mower, J.P., Fourcade, H.M., Calie, P.J., Boore, J.L., Jansen, R.K., 2006. The complete chloroplast genome sequence of Pelargonium × hortorum: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol. Biol. Evol. 23, 2175–2190.

Clayton, W.D., Vorontsova, M.S., Harman, K.T., Williamson, H., 2006 onwards. GrassBase – The Online World Grass Flora. London The Board of Trustees, Royal Botanic Gardens, Kew <http://www .kew.org/data/grasses-db.html> (accessed 14.02.14).

Clegg, M.T., Zurawski, G., 1992. Chloroplast DNA and the study of Plant Phylogeny: present status and future prospects. In: Soltis, P.S., Soltis, D.E., Doyle, J.F. (Eds.), Molecular Systematics of Plants. Chapman and Hall, New York, pp. 1–13.

Columbus, J.T., Cerros-Tlatilpa, R., Kinney, M.S., Siqueiros-Delgado, M.E., Bell, H.L., Griffith, M.P., Refullo-Rodriguez, N.F., 2007. Phylogenetics of chloridoideae (Gramineae): a preliminary study based on nuclear ribosomal internal transcribed spacer and chloroplast trnL–F sequences. J. Syst. Evolut. Bot. 23, 565–579.

Crepet, W.L., Feldman, G.D., 1991. Earliest fossil evidence of grasses in the fossil record. Amer. J. Bot. 78, 1010–1014.

Doyle, J.F., Egan, A.N., 2010. Dating the origins of polyploidy events. New Phytol. 186, 73–85.

Drescher, A., Ruf, S., Calsa Jr., T., Carrer, H., Bock, R., 2000. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J. 22, 97–104.

Drummond, A.J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Heled, J., Kearse, M., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T., Wilson, A., 2010. Geneious v5.1 <http://www.geneious.com>.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29, 1969–1973.

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113.

Fawcett, J., Van de Peer, Y., 2010. Angiosperm polyploids and their road to evolutionary success. Trends Evol. Biol. 2, 16–21.

Ferris, C., King, R.A., Gray, A.J., 1997. Molecular evidence for the maternal parentage in the hybrid origin of *Spartina anglica* C.E. Hubbard. Mol Ecol. 6, 185–187.

Forrestel, E.J., Donoghue, M.J., Smith, M.D., 2014. Convergent phylogenetic and functional responses to altered fire regimes in mesic savanna grasslands of North America and South Africa. New Phytol. 203, 1000–1011.

Fortune, P.M., Schierenbeck, K.A., Ainouche, A.K., Jacquemin, J., Wendel, J.F., Ainouche, M.L., 2007. Evolutionary dynamics of Waxy and the origin of hexaploid *Spartina* species (Poaceae). Mol. Phylogenet. Evol. 43, 1040–1055.

Fortune, P.M., Schierenbeck, K., Ayres, D., Bortolus, A., Catrice, O., Brown, S., Ainouche, M.L., 2008. The enigmatic invasive *Spartina densiflora*: a history of hybridizations in a polyploidy context. Mol. Ecol. 17, 4304–4316.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.

Guisinger, M.M., Chumley, T.W., Kuehl, J.V., Boore, J.L., Jansen, R.K., 2010. Implications of the plastid genome sequence of Typha (Typhaceae, Poales) for understanding genome evolution in Poaceae. J. Mol. Evol. 70, 149–166.

Guisinger, M.M., Kuehl, J.V., Boore, J.L., Jansen, R.K., 2011. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol. Biol. Evol. 28, 583–600.

Hilu, K.W., Alice, L.A., 1999. Evolutionary implications of matK indels in Poaceae. Am. J Bot. 86 (12), 1735–1741.

Hilu, K.W., Alice, L.A., 2001. A phylogeny of Chloridoideae (Poaceae) based on matK sequences. Syst. Bot. 26, 386–405.

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., Depamphilis, C.W., Leebens-Mack, J., Muller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., Chumley, T.W., Lee, S.B., Peery, R., McNeal, J.R., Kuehl, J.V., Boore, J.L., 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl. Acad. Sci. USA 104, 19369–19374.

Jansen, R.K., Ruhlman, T.A., 2012. Plastid genomes of seed plants. In: Bock, R., Knoop, V. (Eds.), Genomics of Chloroplast and Mitochondria. Advances in Photosynthesis and Respiration Including Bioenergy and Related Processes, vol. 35. Springer, pp. 103–126.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., Soltis, D.E., Clifton, S.W., Schlarbaum, S.E., Schuster, S.C., Ma, H., Leebens-Mack, J., dePamphilis, C.W., 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473, 97–100.

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120.

Kode, V., Mudd, E.A., Iamtham, S., Day, A., 2005. The tobacco plastid accD gene is essential and is required for leaf development. Plant J. 44, 237–244.

Kolodner, R., Tewari, K.K., 1979. Inverted repeats in chloroplast DNA from higher plants. Proc. Natl. Acad. Sci. USA 76, 41–45.

Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J., Giegerich, R., 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 29, 4633–4642.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25.

Lohse, M., Drechsel, O., Kahlau, S., Bock, R., 2013. OrganellarGenomeDRAW – a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. 41, W575–W581.

Marchant, C., 1967. Evolution in Spartina (Gramineae): I. The history and morphology of the genus in Britain. Bot. J. Linn. Soc. 60, 1–24.

Marchant, C.J., 1968. Evolution in Spartina (Gramineae). II. Chromosomes, basic relationships, and the problem of S. × townsendii agg. Bot. J. Linn. Soc. 60, 381–409.

Martin, G.E., Rousseau-Gueutin, M., Cordonnier, S., Lima, O., Michon-Coudouel, S., Naquin, D., de Carvalho, J.F., Ainouche, M., Salmon, A., Ainouche, A., 2014. The first complete chloroplast genome of the Genistoid legume Lupinus luteus: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. Ann. Bot. 113, 1197–1210.

McKinnon, G., 2004. Reticulate evolution in higher plants. In: Henry, R. (Ed.), Plant Diversity and Evolution. CABI publishing, Wallingford, pp. 81–96.

Middleton, C.P., Senerchia, N., Stein, N., Akhunov, E.D., Keller, B., Wicker, T., Kilian, B., 2014. Sequencing of chloroplast genomes from wheat, barley, rye and their relatives provides a detailed insight into the evolution of the Triticeae tribe. PLoS ONE 9, e85761.

Mobberley, D., 1956. Taxonomy and distribution of the genus Spartina. Iowa State Coll. J. Sci. 30, 471–574.

Moore, M.J., Soltis, P.S., Bell, C.D., Burleigh, J.G., Soltis, D.E., 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc. Natl. Acad. Sci. USA 107, 4623–4628.

Paradis, E., Bolker, B., Claude, J., et al., 2011. Package 'ape'. <http://cran.r-project.org/web/packages/ape/ape.pdf>.

Peterson, P.M., Columbus, J.T., Pennington, S.J., 2007. Classification and biogeography of new world grasses: Chloridoideae. Aliso: J. Syst. Evol. Bot. 23, 43.
Peterson, P.M., Romaschenko, K., Johnson, G., 2010a. A classification of the Chloridoideae (Poaceae) based on multi-gene phylogenetic trees. Mol. Phylogenet. Evol. 55, 580–598.
Peterson, P.M., Romaschenko, K., Johnson, G., 2010b. A phylogeny and classification of the Muhlenbergiinae (Poaceae: Chloridoideae: Cynodonteae) based on plastid and nuclear DNA sequences. Am. J. Bot. 97, 1532–1554.
Peterson, P.M., Romaschenko, K., Barker, N.P., Linder, H.P., 2011. Centropodieae and Ellisochloa, a new tribe and genus in the Chloridoideae (Poaceae). Taxon 60, 1113–1122.
Peterson, P.M., Romaschenko, K., Snow, N., Johnson, G., 2012. A molecular phylogeny and classification of Leptochloa (Poaceae: Chloridoideae: Chlorideae) sensu lato and related genera. Ann. Bot. 109, 1317–1330.
Peterson, P.M., Romaschenko, K., Herrera, A.Y., 2014a. A molecular phylogeny and classification of the Cteniinae, Farragininae, Gouiniinae, Gymnopogoninae, Perotidinae, and Trichoneurinae (Poaceae: Chloridoideae: Cynodonteae). Taxon 63, 275–286.
Peterson, P.M., Romaschenko, K., Herrera, A.Y., Saarela, J.M., 2014b. A molecular phylogeny and subgeneric classification of *Sporobolus* (Poaceae: Chloridoideae: Sprobolineae). Taxon 63, 1212–1243.
Peterson, P.M., Romaschenko, K., Herrera, A.Y., Saarela, J.M., 2014c. Proposal to conserve Sporobolus against Spartina, Crypsis, Ponceletia, and Heleochloa (Poaceae: Chloridoideae: Sporobolinae). Taxon 63, 1373–1374.
Pierce, S.M., 1982. What is *Spartina maritima* doing in our estuaries? S. Afr. J. Sci. 78, 229–230.
Posada, D., 2008. JModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25, 1253–1256.
Prasad, V., Strömberg, C.A.E., Alimohammadian, H., Sahni, A., 2005. Dinosaur coprolites and the early evolution of grasses and grazers. Science 310, 1177–1180.
Prasad, V., Stromberg, C.A., Leache, A.D., Samant, B., Patnaik, R., Tang, L., Mohabey, D.M., Ge, S., Sahni, A., 2011. Late Cretaceous origin of the rice tribe provides evidence for early diversification in Poaceae. Nat. Commun. 2, 480.
Raybould, A.F., Gray, A.J., Lawrence, M.J., Marshall, D.F., 1991. The evolution of *Spartina anglica* CE Hubbard: variation and status of the parental species in Brittain. Biol. J. Linn. Soc. 44, 369–380.
Rousseau-Gueutin, M., Ayliffe, M.A., Timmis, J.N., 2011. Conservation of plastid sequences in the plant nuclear genome for millions of years facilitates endosymbiotic evolution. Plant Physiol. 157, 2181–2193.
Rousseau-Gueutin, M., Huang, X., Higginson, E., Ayliffe, M., Day, A., Timmis, J.N., 2013. Potential functional replacement of the plastidic acetyl-CoA carboxylase subunit (accD) gene by recent transfers to the nucleus in some angiosperm lineages. Plant Physiol. 161, 1918–1929.
Rozen, S., Skaletsky, H., 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. 132, 365–386.
Shaw, J., Lickey, E.B., Schilling, E.E., Small, R.L., 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Am. J. Bot. 94, 275–288.
Shaw, J., Shafer, H.L., Leonard, O.R., Kovach, M.J., Schorr, M., Morris, A.B., 2014. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: the tortoise and the hare IV. Am. J. Bot. 101, 1987–2004.
Silvestri, S., Defina, A., Marani, M., 2005. Tidal regime, salinity and salt mash plant zonation. Estuar. Coast. Shelf Sci. 62, 119–130.
Siqueiros-Delgado, M.E., Ainouche, M., Columbus, J.T., Ainouche, A., 2013. Phylogeny of the *Bouteloua curtipendula* complex (Poaceae: Chloridoideae) based on nuclear ribosomal and plastid DNA sequences from diploid taxa. Syst. Bot. 38, 379–389.
Soltis, D.E., Soltis, P.S., Pires, J.C., Kovarik, A., Tate, J.A., Mavrodiev, E., 2004. Recent and recurrent polyploidy in *Tragopogon* (Asteraceae): cytogenetic, genomic and genetic comparisons. Bot. J. Linn. Soc. 82, 485–501.
Soreng, R.J., Peterson, P.M., Romaschenko, K., Davidse, F.O., Zuloaga, G., Judziewicz, E.J., Filgueiras, T.S., Davis, J.I., Morrone, O., 2015. A worldwide phylogenetic classification of the Poaceae (Gramineae). J. Syst. Evol. 53, 117–137.
Spriggs, E.L., Christin, P.A., Edwards, E.J., 2014. C_4 photosynthesis promoted species diversification during the miocene grassland expansion. PLoS One 9, e97722.
Strong, D.R., Ayres, D.R., 2013. Ecological and evolutionary misadventures of Spartina. Annu. Rev. Ecol. Evol. Syst. 44, 389–410.
Triplett, J.K., Oltrogge, K.A., Clark, L.G., 2010. Phylogenetic relationships and natural hybridization among the North American woody bamboos (Poaceae: Bambusoideae: Arundinaria). Am. J. Bot. 97, 471–492.
Vicentini, A., Barber, J.C., Aliscioni, A.A., Giussani, L.M., Kellogg, E.A., 2008. The age of the grasses and clusters of origins of C4 photosynthesis. Glob. Change Biol. 14, 2693–2977.
Wendel, J.F., Doyle, J.J., 2005. Polyploidy and evolution in plants. In: Henry, R.J. (Ed.), Plant Diversity and Evolution. CABI Publishing, Wallington, U.K., pp. 97–117.
Weng, M.L., Blazier, J.C., Govindu, M., Jansen, R.K., 2014. Reconstruction of the ancestral plastid genome in geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Mol. Biol. Evol. 31, 645–659.
Wicke, S., Schneeweiss, G.M., de Pamphilis, C.W., Müller, K.F., Quandt, D., 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol. Biol. 76, 273–297.
Wu, Z.Q., Ge, S., 2012. The phylogeny of the BEP clade in grasses revisited: evidence from the whole-genome sequences of chloroplasts. Mol. Phylogenet. Evol. 62, 573–578.
Wyman, S.K., Jansen, R.K., Boore, J.L., 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20, 3252–3255.
Wysocki, W.P., Clark, L.G., Kelchner, S.A., Burke, S.V., Pires, J.C., Edger, P.P., Mayfield, D.R., Triplett, J.K., Columbus, J.T., Ingram, A.L., Duvall, M.R., 2014. A multi-step comparison of short-read full plastome sequence assembly methods in grasses. Taxon 63, 899–910.
Yamane, K., Yano, K., Kawahara, T., 2006. Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize and rice. DNA Res. 13, 197–204.
Yang, Z., Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17, 32–43.
Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.
Yannic, G., Baumel, A., Ainouche, M., 2004. Uniformity of the nuclear and chloroplast genomes of *Spartina maritima* (Poaceae), a salt-marsh species in decline along the Western European Coast. Heredity (Edinb) 93, 182–188.
Zhang, Y.J., Ma, P.F., Li, D.Z., 2011. High-throughput sequencing of six bamboo chloroplast genomes: phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). PLoS ONE 6, e20596.

Identification et évolution des séquences orthologues par séquençage massif chez les polyploïdes

Les nouvelles technologies de séquençage (NTS) offrent de nouvelles opportunités d'explorer les génomes et transcriptomes d'espèces polyploïdes. L'assemblage de transcriptomes et l'identification des copies de gènes dupliqués par allopolyploïdisation (homéologues) constituent cependant un véritable défi. Ce qui est plus particulièrement le cas dans un contexte de superposition de plusieurs évènements de polyploïdie et en l'absence de génome de référence diploïde. Les Spartines (Poaceae, Chloridoideae) représentent un excellent système pour étudier les conséquences à court terme des évènements d'hybridation et de polyploïdisation. En effet, *S. maritima* (hexaploïde) s'est hybridée à deux reprises avec *S. alterniflora* (hexaploïde) suite à son introduction récente en Europe, formant deux hybrides homoploïdes (*S. x townsendii* et S*. x neyrautii*). La duplication du génome de *S. x townsendii* a formé une nouvelle espèce allododécaploïde *S. anglica* (à la fin du XIX^{ème} siècle) qui a depuis envahi les marais salés de plusieurs continents. L'identification des gènes dupliqués au sein de *S. anglica* et de ses parents est importante pour la compréhension de son succès évolutif. Cependant, leurs niveaux de ploïdie, et l'absence d'espèce diploïde de référence chez les spartines nécessitent le développement d'outils adaptés. Dans ce contexte, nous avons développé et validé différents outils bioinformatiques permettant de détecter des polymorphismes afin d'identifier les différents haplotypes au sein de jeux de données NTS. Ces approches nous ont permis d'étudier l'hétérogénéité des domaines de l'ADN ribosomique 45S de *S. maritima*. Nous avons mis en évidence la perte de copies homéologues en conséquence de la diploïdisation en cours. Afin de développer les ressources transcriptomiques de ces espèces, cinq nouveaux transcriptomes de référence (110 423 contigs annotés pour les 5 espèces dont 37 867 contigs non-redondants) ont été assemblés et annotés. Les co-alignements des haplotypes parentaux et hybrides/allopolyploïdes nous ont permis d'identifier les homéo-SNPs discriminant les séquences homéologues. De plus, nous avons évalué la divergence entre les copies de gènes, identifié et confirmé les évènements de duplications récents au sein des Spartines. Au cours de cette thèse, nous avons également initié des approches de phylogénomique des spartines, qui permettront de préciser l'origine évolutive des copies dupliquées.

Identification and evolution of orthologous sequences in polyploid species by Next-Gen Sequencing

Next generation sequencing (NGS) technologies offer new opportunities to explore polyploid genomes and their corresponding transcriptomes. However, transcriptome assemblies and identification of homoeologous gene copies (duplicated by polyploidy) remain challenging, particularly in the context of recurrent polyploidy and the absence of diploid reference parents. *Spartina* species (Poaceae, Chloridoideae) represent an excellent system to study the short term consequences of hybridization and polyploidization in natural populations. The European *S. maritima* (hexaploid) hybridized twice with the American *S. alterniflora* (hexaploid) following its recent introduction to Europe, which resulted in the formation of two homoploid hybrids (*S. x townsendii* and *S. x neyrautii*). Whole genome duplication of *S. x townsendii* resulted in the fertile new allododecaploid *S. anglica* species (during the 19th century) that has now invaded saltmarshes on several continents. Identification of duplicated genes in *S. anglica* and its parental species is critical to understand its evolutionary success but their high ploidy levels require the development of adapted tools. In this context, we developed and validated different bioinformatics tools to detect polymorphisms and identify the different haplotypes from NGS datasets. These approaches enabled the study of the heterogeneity of the highly repeated 45S rDNA in *S. maritima*. In order to develop transcriptomic resources for these species, 5 new reference transcriptomes (110 423 annotated contigs for the 5 species with 37 867 non-redundant contigs) were assembled and annotated. Co-alignments of parental and hybrid/allopolyploid haplotypes allowed the identification of homoeoSNPs discriminating homoelogs. The divergence between duplicated genes was used to identify and confirm the recent duplication events in *Spartina*. Phylogenomic approaches on *Spartina* were also initiated in this thesis in the perspective of exploring the evolutionary history of the duplicated copies.