Évolution du génome des spartines polyploïdes envahissant les marais salés : apport des nouvelles techniques de séquençage haut-débit

Julie Ferreira de Carvalho

N° d'ordre: ANNEE 2013

UNIVERSITÉ DE RENNES 1

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

Ferreira de Carvalho Julie

Préparée à l'unité de recherche UMR–CNRS 6553 ECOBIO
SCIENCES DE LA VIE ET DE L'ENVIRONNEMENT

Evolution du génome des Spartines polyploïdes envahissant les marais salés : apport des nouvelles techniques de séquençage haut-débit

Thèse à soutenir à Rennes le 19 février 2013

devant le jury composé de :

Angélique D'HONT
DR CIRAD Montpellier / *rapporteur*
Boulos CHALHOUB
DR URGV - INRA-CNRS Evry / *rapporteur*
Anne-Marie CHEVRE
DR INRA APBV Le Rheu / *examinatrice*
Jonathan WENDEL
PR Iowa State University / *examinateur*
Armel SALMON
MC-U. Rennes 1 / *co-encadrant de thèse*
Malika AINOUCHE
PRU –U. Rennes 1 / *directrice de thèse*

REMERCIEMENTS

Ce travail de thèse a été financé par la Région Bretagne (ARED) dans le cadre du projet 'EVOSPART : Evolution du génome des spartines polyploïdes envahissant les marais salés ». Les travaux de recherche ont bénéficié du soutien du Génoscope à travers le projet « GENOSPART : Genomics of *Spartina* ». Le travail s'est déroulé au sein de l'équipe Mécanismes à l'Origine de la Biodiversité (MOB, responsable Malika Ainouche), dans l'UMR CNRS 6553 Ecobio (« Ecologie, Evolution, Biodiversité ») dirigée par Jean-Sebastien Pierre puis Françoise Binet que je remercie pour leur accueil et leur soutien dans les projets de développement des ressources transcriptomiques et génomiques des spartines.

J'adresse toute ma reconnaissance à Malika Aïnouche pour m'avoir accueillie dans l'équipe et m'avoir fait confiance afin d'entreprendre le passage à l'ère de la génomique avec les Spartines. Merci de m'avoir accompagnée tout au long de cette thèse, des balbutiements des premières analyses bioinformatiques aux dernières corrections de ce manuscrit. J'ai beaucoup appris de vos qualités scientifiques mais aussi humaines. Merci à Armel Salmon d'avoir accepté dès son arrivée dans l'équipe en septembre 2010, le co-encadrement de ce travail, de m'avoir consacré du temps pour les analyses (notamment bioinformatiques) et aidée si efficacement à mener à bien cette étude. Vous avez fait de ces trois années et quelques mois, une expérience scientifique et personnelle enrichissante, et m'avez fourni les outils et compétences nécessaires pour continuer dans cette voie et, je l'espère, réussir dans la recherche académique.

Je tiens à remercier Angélique d'Hont et Boulos Chalhoub d'avoir accepté malgré leur emploi du temps chargé de rapporter cette thèse. Je remercie également Anne-Marie Chèvre et Jonathan Wendel (qui a fait un long trajet de l'Iowa) de faire partie de mon jury de thèse et d'examiner ce travail.

J'exprime aussi ma reconnaissance aux membres de mes comités de thèse qui m'ont apporté de précieux conseils et permis le bon déroulement de cette thèse : Hadi Quesneville, Jérôme Salse, Claude Rispe, Jonathan Wendel et Joseph Jahier qui durant trois ans a joué le rôle de tuteur de thèse pour l'école doctorale Vie Agro Santé que je remercie également.

Je remercie Kader Aïnouche pour ses conseils précieux en analyses phylogénétiques et pour les discussions enrichissantes sur les éléments transposables. Kader a aussi été d'une aide précieuse dans l'organisation de l'emploi du temps des enseignements de botanique auxquels j'ai contribué dans le cadre du monitorat.

Les principes de l'extraction d'ARN et de la réalisation des banques d'ADNc chez les Spartines n'avaient plus de secret pour Houda Chelaifa quand je suis arrivée au laboratoire. Je la remercie infiniment pour l'aide qu'elle m'a apportée au début de cette thèse, et pour tous les petits conseils que seule une 'ancienne' doctorante peut donner !

Je remercie les étudiants en Master et/ou doctorat de notre équipe avec qui j'ai plus particulièrement interagi lors de cette thèse : Julien Boutte, Antoine Rochetaux, Pierre Bourdau, Sidonie Bellot, Guillaume Martin et les autres membres de l'équipe MOB (plus particulièrement Paula Dias, Mathieu Rousseau et Marie-Thérèse Misset) pour les échanges enrichissants et relectures à l'impression de ce manuscrit.

Une partie importante de ce travail concerne l'échantillonnage sur le terrain, qui s'est déroulé sur la côte Atlantique et le sud de l'Angleterre et n'aurait été possible sans l'aide de Louis Parize tout au long de cette thèse, d'Ales Kovarik durant les étés 2011 et 2012 et d'Andrew Leitch au mois d'Août 2012. Ces moments partagés en dehors du laboratoire et/ou des congrès permettaient des échanges simples et amicaux à l'image de ces deux chercheurs.

Je remercie les membres de la plateforme de Génomique environnementale et fonctionnelle de l'OSUR, Philippe Vandenkoornhuyse, Oscar Lima, Delphine Naquin, Sophie Michon-Coudouel, Alexandra Dheilly et les membres du Génoscope, Patrick Wincker, Julie Poulain et Corinne Da Silva, Arnaud Couloux et Sophie Mangenot, et pour leur contribution au développement de ressources transcriptomiques et génomiques que j'ai analysées chez les spartines. L'espèce native européenne *S. maritima* dispose désormais d'une banque BAC réalisée au CNRGV (Centre National de Ressources Génomiques Végétales) à Toulouse. Je remercie particulièrement Hélène Bergès, Joëlle Fourment et Arnaud Bellec pour leur aide dans la réalisation et le screening de cette banque. Merci également à Olivier Garsmeur, Carine Charron et Angélique D'Hont pour leurs conseils avisés concernant le screening de régions génomiques d'intérêt de la banque BAC.

Merci à la Plateforme de bioinformatique de l'université de Rennes 1, Genouest et plus particulièrement à Olivier Collin et tous les modérateurs/développeurs du genocluster où mes blasts et autres assemblages ont tourné pendant des semaines.

Merci à Jiri Macas et Alex Kovarik de m'avoir accueillie dans leurs laboratoires respectifs en République Tchèque et de m'avoir permis mes premières explorations du compartiment répété chez les Spartines. Ces collaborations se sont déroulées dans le cadre d'un programme Hubert Curien 'BARRANDE' franco-tchèque.

Bien évidemment, derrière ce manuscrit se cachent les personnes de l'ombre qui permettent de surmonter les difficultés administratives et logistiques de l'administration et de la gestion. Je remercie chaleureusement Jocelyne Beven et Fabienne Defrance, ainsi que Sandra Rigaud, Valérie Haubertin, Tifenn Donguy. Merci également à Valérie Briand et Isabelle Picouays pour leur efficacité (et aussi pour les chocolats !). Merci à Fouad Nassur et Thierry Fontaine pour la culture des Spartines à la serre et leur maintien hors des attaques de pucerons.

Je remercie 'The Genetics Society and Nature Publishing Group' pour leur aimable autorisation à reproduire dans le cadre de cette thèse, notre article publié dans la revue *Heredity* (version originale disponible à l'adresse suivante, http://www.nature.com/hdy/journal/v110/n2/pdf/hdy201276a.pdf).

Merci à toute l'équipe « Tassili » de l'Université d'Alger (USTHB), je me souviendrai de tous les bons moments, partagés en France ou ailleurs pendant les réunions, congrès polyploïdie et sorties. Plus particulièrement, un grand merci à Malika Ourari pour son soutien et son aide pendant ma thèse.

Je souhaite aussi remercier tous les membres de l'équipe pédagogique de Botanique (en plus des personnes de notre équipe citées ci-dessus, Michèle Tarayre, Abdel Elamrani, Cécile Sulmon), j'ai vraiment beaucoup appris à vos côtés et apprécié chaque heure de TD/TP au cours de mes activités d'enseignement.

Je remercie également toutes les personnes qui ont rendu mon quotidien rempli de rire, de gâteaux, de thé, de chaussures, de restaurants et de bonnes bouteilles de vin : Marie-Thérèse, Agnès, Paula, Alex, Benjamin, Aurore, Anne-Sophie, Julien, Mathieu, Philippe, Sarah…

Et il y a le reste du monde, les personnes qui ne comprennent ~~rien~~ pas grand-chose à ce que je fais mais avec qui je partage ma passion de l'équitation (et avec qui je mange beaucoup de gâteaux aussi !), mes amis de lycée, de licence, d'Afrique du Sud et de master (merci à Anaïs qui a été présente dans les derniers instants du manuscrit et qui a su me donner l'énergie nécessaire à l'aide de gaufres maison !). Sans le soutien de mes proches et de ma famille je n'en serai pas là aujourd'hui, merci infiniment de m'avoir donné un environnement propice à l'accomplissement de soi. Merci à mes parents d'avoir toujours cru en moi, de m'avoir poussé à aller toujours plus loin et d'être aujourd'hui les rois du déménagement ! Enfin, Pierre, merci pour ton soutien sans faille, tes encouragements, la confiance que tu m'accordes et permet de poursuivre l'aventure en dehors de nos frontières. Vous avez tous vécu ma thèse intensément, une page se tourne, une nouvelle va s'écrire dans un décor différent mais je l'espère, avec les mêmes protagonistes.

Et… Binnenkort in Nederland !

SOMMAIRE

Introduction

We wish to discuss a structure for the salt of desoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biologic interest.

— Rosalind Franklin

Rosalind Franklin & R. G. Gosling, 'Molecular Structures of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid', *Nature*, 1953, **171**: 737.

Les progrès rapides effectués au cours de la dernière décennie sur la connaissance des génomes de plantes (Flagel & Blackman, 2012) ont montré leur complexité, leur nature souvent redondante, fruit de la superposition d'évènements récurrents de duplications partielles (segmentaires) ou totales (polyploïdie) et de l'accumulation ou de la perte de séquences répétées (Leitch & Leitch, 2012). Les publications des premiers génomes de plantes séquencés chez différentes lignées Eudicotylédones (*Arabidopsis thaliana* : The Arabidopsis Genome Initiative, 2000 ; *Populus trichocarpa* : Tuskan *et al.,* 2006 ; *Vitis vinifera* : Jaillon *et al.,* 2007 ; Papaya : Ming *et al*., 2008 ; Apple : Velasco *et al.*, 2010 ; Soybean : Schmutz *et al*., 2010 ; *Gossypium raymondii* : Wang *et al.,* 2012 et *Gossypium hirsutum* : Paterson *et al*., 2012) ou Monocotylédones (*Oryza sativa :* Yu *et al*., 2005 ; *Sorghum bicolor :* Paterson *et al*., 2009 ; *Zea mays* : Schnable *et al.*, 2009 ; *Brachypodium distachyon* : Vogel *et al.*, 2010 ; *Musa acuminata* : D'Hont *et al.*, 2012) ont permis de détecter les traces de plusieurs évènements anciens de duplication (paléopolyploïdie) dont certains s'avèrent remonter à l'ancêtre des Angiospermes (Cui *et al*., 2006) et même des plantes à graines, il y a plus de 300 millions d'années (Jiao *et al*., 2011). Les évènements de polyploïdisation plus récents, plus faciles à détecter par les dénombrements chromosomiques variables selon un multiple de base (séries euploïdes) sont connus depuis longtemps (Stebbins, 1950) et certaines familles (comme les Poacées) renferment plus de 80% d'espèces polyploïdes plus ou moins récentes.

La notion de génome diploïde est donc aujourd'hui devenue relative: on parle ainsi de génome « diploïdisé » cytologiquement (appariement des chromosomes par paires à la méiose) ou génétiquement (une paire de copies par gène), et la détermination des nombres de chromosomes de base dans un groupe d'espèces n'est pas toujours aisée (*e.g.* Cusimano *et al*., 2012). L'âge du polyploïde (« paléopolyploïde », « meso-polyploïde » ou « neopolyploïde ») est un paramètre important à prendre en compte dans ces définitions.

Les modes de formation des polyploïdes peuvent être variés (Doyle *et al*., 2008) ; il est admis que la polyploïdie résulte le plus souvent de la formation de gamètes non-réduits (Ramsey & Schemske, 1998) et de leur union au sein de la même espèce (autopolyploïdes) ou suite à une hybridation interspécifique (allopolyploïdes), ce qui conduit à la réunion de génomes plus ou moins divergents appelés homéologues qui se retrouvent dupliqués. Dans ce dernier cas, la duplication du génome contribue à la restauration de la fertilité et favorise l'établissement de la nouvelle lignée polyploïde. Du fait des échanges génétiques plus ou moins limités entre le polyploïde et ses parents diploïdes, la polyploïdie contribue à la formation de nouvelles espèces et représente un processus important au plan évolutif, particulièrement fréquent chez les plantes.

Les conséquences adaptatives de la spéciation par polyploïdie peuvent être perçues à court terme chez les polyploïdes récemment formés qui montrent souvent une amplitude écologique plus large que celle de leurs parents diploïdes ou qui montrent des capacités envahissantes (te Beest *et al*., 2012), ou à plus long terme par leur capacité à survivre aux grands bouleversements environnementaux (Van de Peer *et al.*, 2009). L'agriculture a largement tiré profit des potentialités adaptatives fournies par la polyploïdie, puisque la majeure partie des espèces domestiquées est composée d'espèces polyploïdes relativement récentes comme les blés tétraploïdes et hexaploïdes, le colza tétraploïde, le coton tétraploïde, la canne à sucre dodécaploïde, ou le café tétraploïde.

Ces dernières années, une attention particulière a été accordée à l'analyse des génomes polyploïdes et à leur évolution, ce qui a rapidement mis en évidence leur caractère particulièrement dynamique (Wendel, 2000 ; Comai, 2000 ; Chen, 2007). Une avancée importante dans la compréhension de cette dynamique a été effectuée grâce à l'utilisation de quelques systèmes expérimentaux pour lesquels il a été possible de re-synthétiser artificiellement des hybrides et allopolyploïdes : chez des espèces des genres *Brassica* (Song *et al.*, 1995 ; Pires *et al.*, 2004 ; Albertin *et al.*, 2006 ; Gaeta *et al.*, 2007; Szadkowski *et al.*, 2010 ; Marmagne *et al.*, 2010), *Triticum* (Liu *et al.*, 1998 ; Ozkan *et al.*, 2001 ; Chagué *et al.*, 2010 ; Mestiri *et al.*, 2010 ; Feldman *et al.*, 2012) et *Arabidopsis* (Madlung *et al.*, 2002 ; Chen *et al.*, 2004 ; Wang *et al.*, 2006 ; Lackey *et al.*, 2010), ce dernier modèle ayant pu, le premier, tirer parti des ressources génétiques disponibles de son génome séquencé. Ces systèmes ont permis de révéler, dans des contextes génétiques connus, la nature des changements qui

pouvaient intervenir dès les premières générations suivant la formation d'une espèce allopolyploïde.

Par ailleurs, les quelques cas connus de formation très récente (au cours des 19-20^{ième} siècles) d'hybrides et d'allopolyploïdes naturels fournissent dans ce contexte une opportunité particulière d'analyser à l'échelle des populations naturelles, les processus évolutifs résultant de la réunion de génomes divergents dans un même noyau et les conséquences de la duplication simultanée de l'ensemble des gènes, en comparant les hybrides et polyploïdes à leurs parents. Ces nouveaux hybrides ou polyploïdes se sont tous formés suite à l'introduction (par l'Homme) d'espèces en dehors de leur aire native. Chez les salsifis sauvages (genre *Tragopogon*, Astéracées), 3 espèces diploïdes (*T. dubius*, *T. pratensis* et *T. porrifolius*) ont été introduites au début des années 1900 dans le nord-ouest de l'Amérique et se sont hybridées à de multiples reprises pour former deux nouvelles espèces allotétraploïdes, *T. miscellus* et *T. mirus* qui se sont rapidement propagées en occupant préférentiellement les habitats rudéralisés (Soltis *et al.*, 2004). Dans le genre *Senecio* (Asteraceae), une nouvelle espèce allohexaploïde, *S. cambrensis* s'est formée en Angleterre (Pays de Galles) vers 1948, suite à plusieurs évènements indépendants d'hybridation entre une espèce diploïde introduite, *S. squalidus* et une espèce tétraploïde native, *S. vulgaris* (Abbott & Lowe, 2004). Dans le genre *Spartina* (Poaceae), l'introduction accidentelle de l'espèce américaine hexaploïde *S. alterniflora* dans l'ouest de l'Europe a permis une hybridation avec l'espèce hexaploïde native *S. maritima* pour former deux hybrides F1 stériles dans le sud de l'Angleterre (*S.* x *townsendii*, Groves & Groves, 1880) et dans le sud-ouest de la France (*S.* x *neyrautii*, Foucaud, 1897). La duplication du génome de l'hybride anglais a conduit à la formation d'une nouvelle espèce allohexaploïde vers 1890, *S. anglica* qui a rapidement colonisé les marais salés européens et se retrouve aujourd'hui introduite sur plusieurs continents (Ainouche *et al.*, 2004a).

Les travaux de recherches sur la polyploïdie ont été particulièrement fructueux durant la dernière décennie comme en témoignent la tenue de trois conférences internationales : à Londres (Leitch *et al.*, 2004), Saint-Malo (Ainouche & Jenczweski, 2010), et Prague (numéro spécial de la revue Heredity à paraître en 2013). Ces travaux ont ouvert de nouvelles perspectives dans la compréhension de l'évolution des polyploïdes, du point de

vue de leur mode de formation, de l'évolution structurale et fonctionnelle du génome et dans certains cas des conséquences phénotypiques ou écologiques associées.

Une composante majeure de la dynamique structurale des génomes polyploïdes s'avère résider dans les recombinaisons entre génomes homéologues, la conversion génique et l'évolution concertée qui limitent l'évolution indépendante des séquences dupliquées par polyploïdie (Nieto Felliner & Rossello, 2012), particulièrement bien mises en évidences dans les genres *Brassica* (Nicolas *et al.*, 2007 ; Gaeta & Pires, 2010 ; Szadkowsky *et al.*, 2010), *Nicotiana* (Lim *et al.*, 2000 ; Kovarik *et al.*, 2008) ou *Gossypium* (Wendel *et al.*, 1995 ; Salmon *et al.*, 2010, Flagel *et al.*, 2012). Les éléments transposables, qui constituent une composante importante des génomes des plantes (Kejnovsky *et al.*, 2012), jouent également un rôle dans cette dynamique, par leur implication dans les recombinaisons non homologues, leur activité transcriptionnelle (qui peut affecter l'expression de gènes voisins) ou transpositionnelle le plus souvent régulée par des mécanismes épigénétiques (Kashkush *et al.*, 2003 ; Madlung *et al.*, 2005 ; Chantret *et al.*, 2005 ; Parisod *et al.*, 2010 ; Petit *et al.*, 2010 ; Yaakov & Kashkush, 2011). Le « choc génomique » résultant de la réunion de deux génomes différents au sein d'un même noyau (McClintock, 1984) induit différents types d'altérations épigénétiques affectant notamment la méthylation de l'ADN (Madlung *et al.*, 2002 ; Salmon *et al.*, 2005 ; Lukens *et al.*, 2006 ; Parisod *et al.*, 2009 ; Hegarty *et al.*, 2011) et/ou faisant intervenir les petits ARNs (Chen *et al.*, 2008 ; Ha *et al.*, 2009a ; Kenan-Eichler *et al.*, 2011 ; Shivaprasad *et al.*, 2011) intervenant dans la régulation de l'expression des gènes.

L'évolution de l'expression des gènes (et des processus de régulation associés) est une des conséquences importantes de la spéciation allopolyploïde, conduisant à la mise en place de nouveaux phénotypes (Osborn *et al.*, 2003 ; Adams *et al.*, 2003 ; Adams & Wendel, 2005 ; Chen, 2007). En l'absence de changement, on attendrait une expression « additive » des profils d'expression des gènes parentaux chez un polyploïde, alors qu'une déviation de cette additivité a été souvent observée. Si on compte aujourd'hui plusieurs études réalisées chez différents modèles polyploïdes naturels ou re-synthétisés artificiellement, les genres *Arabidopsis* (Wang *et al.*, 2006 ; Ha *et al.*, 2009b ; Pignatta *et al.*, 2010 ; Chang *et al.*, 2010 ; Kim & Chen, 2011) et *Gossypium* (Adams *et al.*, 2003 ; Adams *et al.*, 2004 ; Udall *et al.*, 2006 ; Hovav *et al.*, 2008 ; Flagel *et al.*, 2008 ; Flagel & Wendel, 2010 ; Rapp *et al.*, 2009 ; Chaudhary *et al.*, 2009 ; Gong *et al.*, 2012) restent à ce jour les mieux explorés de ce point de vue. Ces

travaux ont montré différentes situations de non-additivité des profils d'expression des génomes parentaux chez les hybrides et allopolyploïdes, pouvant présenter une dominance d'expression maternelle ou paternelle (niveau d'expression global du gène mimant celui d'un des deux parents), transgressive (sur-expression ou sous-expression par rapport au niveau d'expression global du gène observé chez les parents), ou une expression préférentielle d'une des copies parentale (« biaisée ») par rapport à l'autre, pouvant aller jusqu'à la mise sous silence d'une des copies homéologues (Small & Wendel, 2002 ; Grover *et al.*, 2012a). Au niveau fonctionnel, ces changements ouvrent la voie à de nouvelles capacités adaptatives résultant de la sous-fonctionalisation (expression différentielle de chaque copie homéologue selon les tissus ou le stade du développement) ou la néofonctionalisation (une des copies dupliquées assure une nouvelle fonction), ce qui va déterminer la nature et l'intensité des pressions de sélection agissant sur les copies redondantes et leur rétention ou non à plus long terme dans le génome polyploïde (Lynch & Force, 2000). La plupart des études de l'expression des gènes a été réalisé au niveau du transcriptome, mais quelques études ont également bien documenté les conséquences de cette évolution au niveau du protéome, dans les genres *Brassica* (Albertin *et al.*, 2005 ; 2006), *Gossypium* (Hu *et al.*, 2011), *Arabidopsis* (Ng *et al.*, 2012) et *Tragopogon* (Koh *et al.*, 2012), démontrant ainsi les conséquences des changements associées à l'hybridation et la polyploïdie sur les produits finaux des gènes (les protéines) et donc sur le métabolisme cellulaire.

Aujourd'hui, les technologies de séquençage en masse (Next Generation Sequencing ou NGS) offrent de nouvelles possibilités d'explorer de façon particulièrement exhaustive la structure des génomes, le transcriptome, et les différentes marques épigénétiques (Ekblom & Galindo, 2011) des plantes à génomes complexes. En particulier, une ère nouvelle commence pour les systèmes biologiques non-modèles disposant jusqu'alors de ressources limitées et d'une faible connaissance de leur génome qui limitait leur investigation à l'usage de technologies de marquage multilocus anonymes (essentiellement dérivées de l'emploi des enzymes de restriction comme l'AFLP : *e.g*. Wang *et al.*, 2005) ou l'utilisation d'outils (exemple: microarrays) désignés sur un génome connu phylogénétiquement assez proche lorsque c'était possible (*e.g.* Chelaifa *et al*., 2010a).

Au cours de ce travail, nous avons commencé à initier les bases du séquençage (partiel) du génome et du transcriptome des Spartines polyploïdes qui représentent un excellent système pour analyser l'évolution polyploïde à différentes échelles de temps (Ainouche *et al.*, 2012) en explorant l'apport des nouvelles technologies de séquençage massif parallèle à la compréhension de l'évolution des génomes polyploïdes, en nous focalisant sur les espèces hexaploïdes *S. maritima* et *S. alterniflora*, leurs hybrides F1 *S.* x *neyrautii* et *S.* x *townsendii* et la nouvelle espèce allo-dodécaploïde *S. anglica*. Cette démarche représente un défi particulier, compte tenu du niveau de duplication et de la taille du génome des espèces concernées : 2C = 3,8 pg et 2C = 4,3 pg pour les parents hexaploïdes *S. maritima* et *S. alterniflora* respectivement (Fortuné *et al*., 2008) et 2C = 8,1 pg pour l'allo-dodécaploïde *S. anglica* (M. Ainouche, données non publiées). De plus, contrairement aux autres systèmes « modèles » utilisés dans les études de l'évolution des génomes polyploïdes, il n'existe pas chez les Spartines d'espèces diploïdes susceptibles d'aider à l'identification des génomes parentaux dupliqués à l'origine des espèces tétraploïdes et hexaploïdes actuelles. Les questions suivantes ont été plus particulièrement examinées au cours de cette thèse :

1- Quels sont les gènes écologiquement importants (*e.g*. impliqués dans la tolérance au stress, la croissance et la vigueur, et donc les capacités envahissantes) affectés par l'hybridation et la polyploïdie et comment l'expression de ces gènes évolue-t-elle en conditions naturelles ? Nous avons commencé, dans cette perspective, par réaliser un premier « transcriptome de référence » chez les Spartines, à partir de séquençage massif (Roche 454) de banques d'ADN complémentaire de différents organes (feuilles et racines) chez les 5 espèces de Spartines (parents, hybrides et allopolyploïde). Les assemblages et annotations de ce transcriptome ont permis de sélectionner un lot de gènes pour lesquelles des amorces spécifiques ont été désignées, permettant l'analyse de la variation de leur expression dans les populations naturelles.

2- Quelles sont les parts relatives des fractions codantes et non-codantes chez les Spartines (en particulier les séquences répétées dont les éléments transposables qui représentent un compartiment potentiellement dynamique dans le génome des hybrides et allopolyploïdes)? Pour répondre à cette question, nous avons analysé l'ADN génomique de l'espèce hexaploïde européenne *S. maritima* à travers (a) 40000

séquences d'extrémités de BAC (Bacterial Artificial Chromosomes) ou BES (BAC End Sequences) et (b) d'un run de pyroséquençage (Roche 454).

3- Comment la connaissance des gènes de Spartines nous éclaire-t-elle sur l'histoire plus ancienne des Spartines issues de la superposition d'évènements anciens et récents de polyploïdisation, et plus globalement, sur la lignée des Chloridoideae (sous famille à laquelle appartiennent les Spartines) qui est particulièrement peu connue dans la famille des Poacées, par rapport aux autres sous-familles plus étudiées, contenant les espèces modèles cultivées ? Dans les limites du temps imparti à la réalisation de cette thèse, nous avons commencé à aborder cette perspective en effectuant de premières comparaisons des données de séquençage d'ADN génomique et de transcriptome de Spartines aux génomes séquencés de la famille des Poacées, et proposé de futures perspectives de recherches et d'exploitation des ressources génomiques que nous avons contribué à générer dans le genre *Spartina*.

Le manuscrit est organisé de la façon suivante :

- Un premier chapitre présente un point bibliographique sur l'apport des nouvelles technologies de séquençage à haut débit (« NGS ») en s'intéressant plus particulièrement au contexte des études sur les polyploïdes.
- Un second chapitre présente le système biologique étudié (les Spartines dans la famille des Poacées).
- Un troisième chapitre décrit le matériel végétal et les méthodes employées.
- Le chapitre 4 concerne la construction de transcriptomes de références pour les Spartines hexaploïdes (article sous presse dans la revue *Heredity*), les hybrides F1 et l'allododécaploïde.
- Le chapitre 5 analyse la variation de l'expression des gènes dans les populations naturelles des parents hexaploïdes, leurs hybrides F1 et l'allopolyploïde *S. anglica.*
- Le chapitre 6 est consacré à l'analyse du génome de *Spartina maritima* d'une part à travers les données de pyroséquençage (Roche 454) et d'autre part à travers l'analyse des BES (présentée sous forme d'un article en préparation).
- Le chapitre 7 est consacré à une discussion générale des résultats et aux perspectives ouvertes par ce travail.

Chapitre 1

Apport des nouvelles technologies de séquençage haut-débit dans l'étude des espèces polyploïdes et des génomes complexes : applications et limites actuelles.

Les avancées récentes dans le domaine du séquençage d'ADN ont révolutionné le champ de la génomique rendant le séquençage massif de séquences plus rapide et de moins en moins coûteux. Ces nouvelles technologies de séquençage haut-débit (ou NGS : Next Generation Sequencing) permettent d'obtenir des données génomiques et transcriptomiques en masse pour des modèles biologiques variés et notamment chez des espèces possédant des génomes complexes avec des niveaux de ploïdie élevés. Aux débuts du séquençage, la méthode Sanger (Sanger *et al*., 1977) ne permettait de générer qu'un faible nombre de nucléotides à la fois (environ 400-500 pb). La première révolution fût le développement du séquençage par capillaire (Zagursky & McCormick, 1990) et la mise au point de la méthode de détection du marquage par fluorescence (Kim & Morris, 1996). Depuis 2005 et la mise sur le marché du premier séquenceur GS20 de chez Roche 454 Life Sciences (Margulies *et al.*, 2005) le rendement nucléotidique journalier a été multiplié d'un facteur de 100 à 1000 et a permis la réduction du coût moyen d'un million de nucléotides à seulement 0,1-4% du coût d'un séquençage Sanger (Kircher & Kelso, 2010). Depuis, d'autres technologies comme le séquençage par synthèse (illumina, ABI SOLID) ont émergé et permettent une réduction des coûts de séquençage encore plus drastique. Ainsi, il est aujourd'hui possible de séquencer et d'assembler *de novo* (sans référence) des séquences issues d'un ou de plusieurs séquenceurs « nouvelle génération » (454 ou Illumina par exemple) pour étudier la structure et la composition des génomes polyploïdes et le devenir des gènes dupliqués. De vrais défis restent toutefois à relever concernant la mise au point de méthodes d'analyses adaptées à ces génomes complexes.

Plusieurs revues sur les techniques de séquençage haut-débit sont parues (*e.g.* Kircher & Kelso, 2010 ; Metzker, 2010 ; Egan *et al*., 2012). L'apport des NGS a été étudié dans l'étude des plantes (pour revue : Deschamp & Campbell, 2009) et des espèces non-modèles (Ekblom &

Galindo, 2011). Nous rappellerons dans un premier temps, les avantages et les inconvénients de chaque plateforme technologique de séquençage afin de choisir les stratégies les mieux adaptées à nos questions de recherche. Dans un second temps, nous présenterons les applications des NGS dans l'analyse des génomes polyploïdes.

I. Présentation des principales techniques de séquençage à haut-débit

Avec l'essor des nouvelles technologies de séquençage haut-débit, les protocoles de préparation des acides nucléiques à séquencer ont aussi dû s'adapter et s'améliorer. Les extractions d'ADN et d'ARN sur colonnes sont fiables, rapides et efficaces. Les enzymes de rétro-transcription des ARN messagers (ARNm) fournissent de l'ADN complémentaire (ADNc) de qualité et en quantité suffisante. Cependant, suivant la question biologique, le protocole expérimental peut varier, et prendre en compte par exemple une phase de normalisation ou de labellisation des séquences (pour le multiplexage) ou encore une phase d'enrichissement des banques. Aujourd'hui, plusieurs technologies de séquençage sont disponibles. Les avantages et les limites de chacun seront discutés afin de choisir la technologie la mieux adaptée à chaque projet de séquençage.

Les technologies Roche 454 Life Sciences, Illumina et Ion Torrent utilisent la détection de signaux chimiques ou lumineux lors de l'incorporation des bases nucléotidiques pendant la synthèse du brin complémentaire afin de déterminer la composition en bases de la séquence. Dans ce cas, l'ADN est fragmenté à la taille appropriée, chaque fragment lié à des adaptateurs est cloné pour amplifier le signal fluorescent ou chimique (Figure 1a et b). Les trois méthodes diffèrent par les réactifs et les substrats utilisés et par la technologie de détection des signaux fluorescent ou chimique d'incorporation des bases.

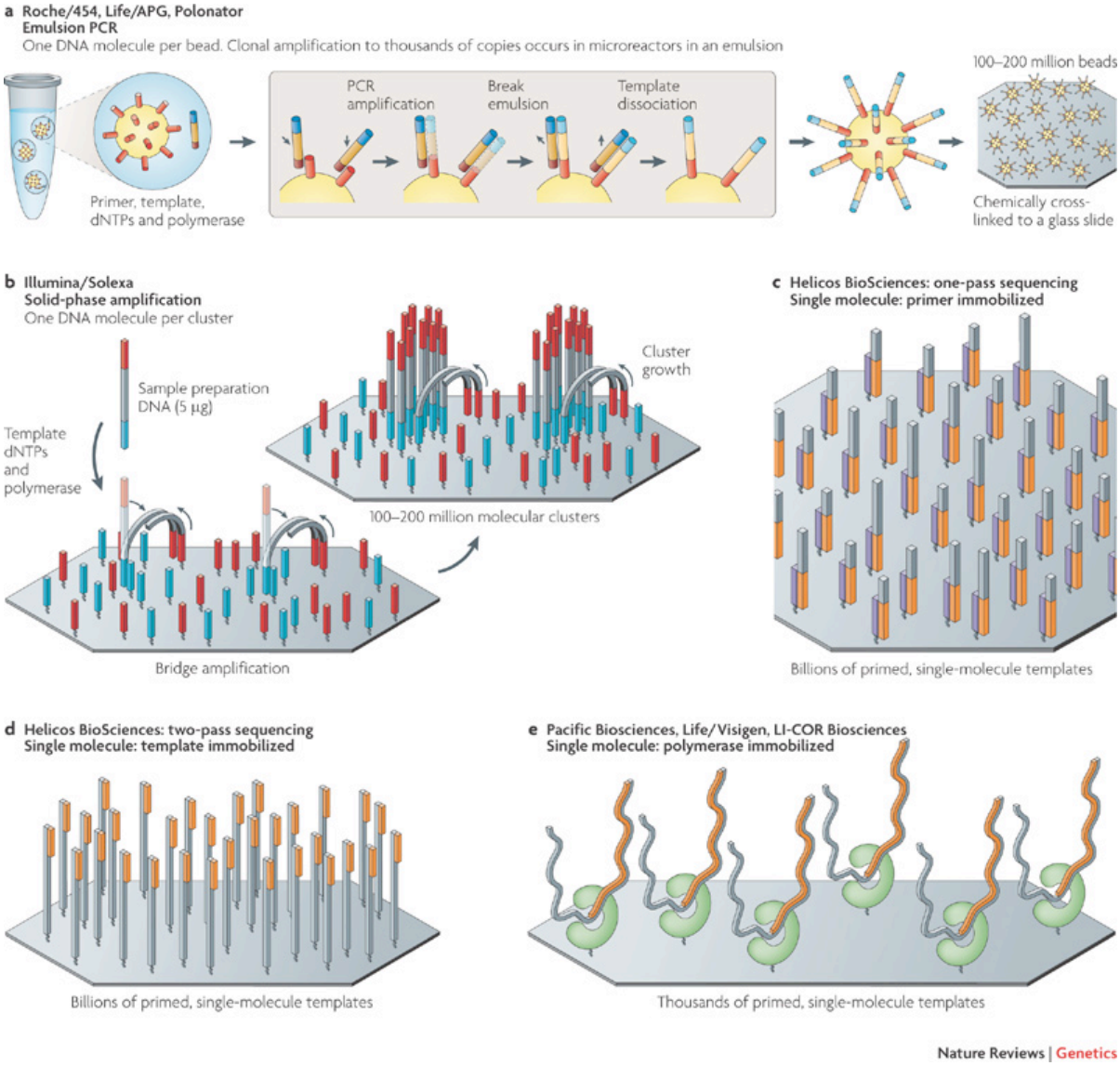

Figure 1: PCR en émulsion (a) une réaction d'émulsion oléo-aqueuse est créée afin d'encapsuler un complexe bille et fragment d'ADN dans une seule gouttelette d'eau. L'amplification est ensuite réalisée sur ces gouttelettes afin qu'elles contiennent plusieurs centaines de copies de la même séquence d'ADN. **(b)** La phase d'amplification peut aussi être réalisée sur une plaque dans ce cas, deux phases sont observées: une phase initiale d'hybridation et de synthèse du premier brin ADN, puis l'amplification grâce à des ponts et des amorces formant des clusters. Il existe plusieurs méthodes afin d'immobiliser les séquences d'ADN sur un support solide: **(c)** ancrage des amorces, **(d)** ancrage des séquences ou **(e)** ancrage de la polymérase. D'après Metzker, 2010.

Le pyroséquençage

Dans la technologie mise au point par Roche 454 Life Sciences, une seule séquence est liée à une microbille par un adaptateur et amplifiée grâce à une PCR en émulsion. Chaque microbille est ensuite placée dans un micropuit avec des billes supplémentaires couplées de sulphurylase et de luciférase (Figure 2c et d). Une amorce est hybridée à l'ADN puis chacune des bases est ajoutée séquentiellement et dans un ordre prédéfini à partir de l'extrémité 3' de l'amorce. Chaque base est marquée par un fluorophore différent et le signal est mesuré uniquement lorsque la base est intégrée à la séquence cible par la polymérase et libère un pyrophosphate. L'ATPsulfurylase vient alors transformer ce pyrophosphate en ATP qui est ensuite utilisé (d'où le terme « pyroséquençage ») et couplé à une luciférine par une luciférase. La réaction libère une oxyluciférine et un signal lumineux. La caméra CCD (Charge Coupled Device) permet de capturer les signaux émis et de déduire la séquence en fonction de l'ordre d'incorporation des bases et de l'intensité lumineuse sur le chromogramme. Cette technologie produit de longues séquences qui facilitent l'assemblage *de novo* de génome sans utiliser de séquence de référence. En quelques années, la longueur des séquences est passée d'environ 300pb à 500pb et aujourd'hui, la mise à jour de l'appareil permet d'atteindre des séquences d'environ 1 kb. Ainsi, une expérimentation sur le GS-FLX+ génère 1Gb de nucléotides en 23h (par exemple, le génome du riz estimé à 372 Mb pourrait être séquencé 2,7 fois pour une couverture de 1X). Ces capacités technologiques en font aujourd'hui un outil indispensable dans les études génomiques et transcriptomiques qui requièrent un assemblage *de novo*.

Le séquençage par synthèse

Initialement développée par Solexa, la plateforme Illumina Genome Analyzer (mise sur le marché en 2007) marque une autre révolution dans le séquençage haut-débit et l'utilisation de séquences courtes dans les applications liées à la génomique. Cette technologie utilise une phase solide d'amplification (Figure 1b) où les adaptateurs sont liés à chaque extrémité de la séquence-cible ; celle-ci est ensuite amplifiée par multiplications clonales des séquences (et la

formation de cluster grâce à des ponts d'amplification). Les quatre types de bases sont ajoutés successivement et celles non-hybridées sont éliminées. Ce type de séquençage (contrairement au pyroséquençage) permet d'incorporer un nucléotide à la fois ; ainsi, le cycle est répété jusqu'à la taille attendue du fragment. Le premier séquenceur permettait de générer 1 Gb de données d'une taille de 35pb. Aujourd'hui, le système HiSeq 2000 possède un rendement de 6 milliards de séquences pairées (les extrémités d'un insert d'une longueur connue sont séquencés) pour un total d'environ 600 Gb en 11 jours (avec une taille de séquences de 100pb) et permet un large spectre d'applications : du reséquençage, à l'assemblage *de novo* en passant par l'analyse des niveaux d'expression des gènes.

Le séquençage par nano-mesure de pH

Le système Ion torrent (Rothberg *et al.*, 2011) est unique puisqu'il utilise les variations de pH afin de détecter l'incorporation des nucléotides. Il détecte plus particulièrement la libération des protons à chaque incorporation de nucléotide au cours de l'élongation. En ajoutant séquentiellement les nucléotides, cette méthode permet de détecter ceux qui ont été incorporés au brin complémentaire. Le « Personal Genome Machine » ou PGM 318 peut ainsi générer 1 Gb de séquences (avec une taille moyenne de fragments séquencés d'environ 250pb) en moins de deux heures. L'avantage de cette technologie se situe essentiellement dans la préparation peu coûteuse des banques (en réactifs) et le prix d'achat du séquenceur. Ces séquenceurs sont en effet des « séquenceurs de paillasse » comme les dernières générations Roche 454 (454 GS Junior) et Illumina (MiSeq). Dans cette catégorie, le système Ion Torrent PGM 318 est très compétitif, la longueur moyenne des séquences issues du MiSeq étant plus courte (150bp) et le GS Junior produisant le plus faible rendement (35Mb).

Le séquençage par ligation

Cette technologie utilise une ligase et des amorces marquées afin de déterminer l'ordre des nucléotides d'un brin (Landegren *et al.*, 1988). Des amorces de différentes longueurs contenant un dinucléotide connu et identifiées avec des sondes fluorescentes sont utilisées conjointement avec de l'ADN ligase (Figure 2a et b). Les amorces qui ne se sont pas hybridées

sont éliminées et s'ensuit une détection par fluorescence des amorces liguées à la séquence-cible. Ce cycle peut se répéter en utilisant des amorces clivées permettant à la sonde fluorescente de se détacher et de régénérer un groupe phosphate en 5' de l'amorce (Figure 2a). Ainsi, les processus de ligation, détection et de cassure se répètent plusieurs fois jusqu'à une longueur prédéterminée. La méthode par ligation du séquenceur SOLiD (Life Technologies/Applied Biosystems) produisant des fragments courts (autour de 75pb) est surtout utilisée dans les projets de reséquençage de génome et de transcriptome. En particulier, la distance importante entre les « mate-paired » permet d'ancrer des contigs (par rapport aux séquences « paired-end » issues de 454 ou d'Illumina par exemple) afin d'améliorer un assemblage.

Les méthodes de « troisième génération » : le séquençage « single-molecule »

Ces méthodes utilisent un signal chimioluminescent à chaque incorporation de nucléotide afin de séquencer directement les fragments d'ADN un à un, rendant l'amplification de l'ADN inutile. Ces innovations diminueraient la durée et simplifieraient la préparation des banques d'ADN. De plus, des ADN en faible concentration ou même dégradés pourraient être utilisés (Orlando *et al.*, 2011) et la phase de PCR pouvant causer des biais importants (recombinaisons par PCR ou d'erreurs de polymérisation) serait éliminée (Kircher & Kelso, 2010). Par exemple, le système HeliScope de chez Helicos Genetic Analysis est aujourd'hui commercialisé. Il produit entre 600 millions et 1 milliards de séquences d'une longueur de 35pb en 30h. Cet appareil permet aussi le multiplexage de 96 échantillons par tunnel, c'est à dire 4800 échantillons par expérimentation. Ainsi, cette technologie apparaît parfaitement adaptée dans les études transcriptomiques quantitatives (Ozsolak *et al.*, 2009). Une autre technologie récente a été développée par les laboratoires Pacific Biosciences qui a mis en test depuis 2010 son appareil (PacBio RS SMS) à quelques laboratoires. Cette méthode permet l'incorporation continue de nucléotides marqués pendant la synthèse du brin d'ADN. L'ADN polymérase est attachée à la surface d'un détecteur ZMW (Zero Mode Waveguide) qui acquiert l'ordre des nucléotides phosphatés quand ceux-ci sont incorporés au brin généré (Figure 1d). Cette technologie permet d'atteindre des longueurs de séquences de plus de 10 000pb. Cependant, la

polymérase peut se dégrader avant d'atteindre cette taille à cause des lectures au laser (contrairement à l'HeliScope). La plateforme PacBio possède le taux d'erreur le plus élevé de toutes les plateformes NGS (environ 15%) mais il est contrebalancé par le séquençage multiple d'un même fragment d'ADN (Metzker, 2010). Ainsi, le plus récent System PacBio RS et le nouveau protocole SMRT cell produisent 35 à 45 Mb de données par cellule d'une longueur moyenne de 1500pb et offrent une fiabilité de 99,999% pour une couverture de 30X (Egan *et al.*, 2012).

Les technologies émergentes

Une quatrième vague de séquenceur va voir le jour, s'arguant des qualités de chacune des plateformes citées plus haut. Ainsi, le Starlight (Life technologies) pourrait surpasser la longueur des séquences de la plateforme de Pacific Biosciences en remplaçant les polymérases dégradées au cours de l'élongation des séquences ADN. La qualité du séquençage sera potentiellement meilleure que celle de la plateforme PacBio en améliorant le protocole de reséquençage du même brin d'ADN faisant ainsi diminuer le taux d'erreur. Mais l'innovation qui offre le plus de promesses est à l'heure actuelle le séquençage par nanopores, développé par la firme Oxford Nanopore Technologies sur la plateforme GridION (Clarke *et al*., 2009 ; Check Hayden, 2012). Brièvement, quand une membrane contenant des nanopores est immergée dans une solution conductrice et un voltage est appliqué, un courant électrique est observé à travers le nanopore dû au passage d'ions. Si des nucléotides d'ADN passent à travers ce pore, l'amplitude du courant électrique à travers le pore est affectée et la séquence ADN reconstituée (Figure 1e).

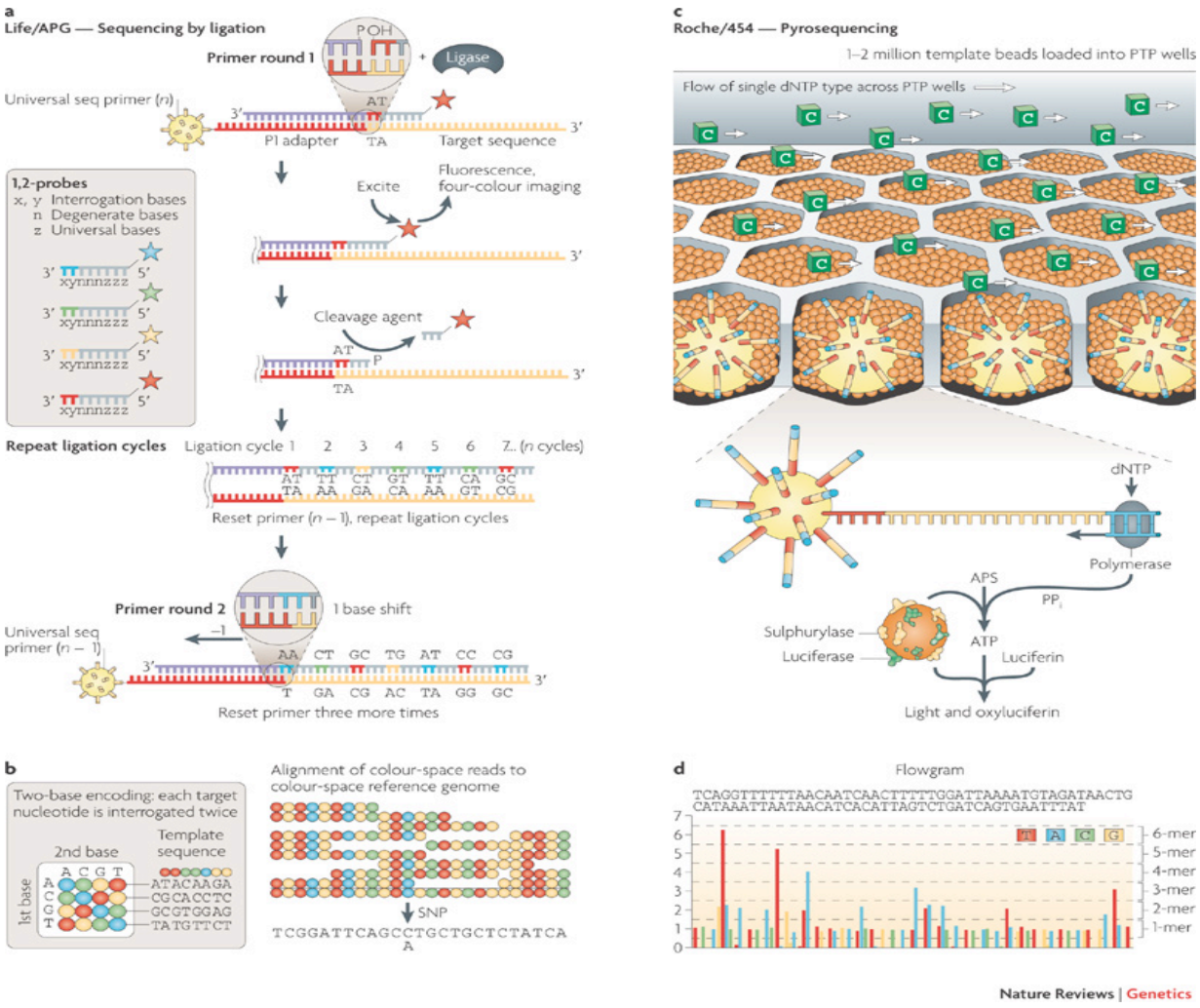

Figure 2: Séquençage utilisant la PCR en émulsion (a) La méthode par ligation est utilisée sur la plateforme SOLiD (Support Oligonucleotide Ligation Detection). Une amorce universelle et des amorces marquées sont ajoutées à la réaction. Les amorces hybridées en position complémentaire sont ensuite soumises à une détection de fluorescence. Après la détection des quatre sondes fluorescentes, les amorces sont chimiquement clivées à l'aide d'ions argent afin de générer un groupe 5'-PO4. Le cycle est alors répété neuf fois**. (b)** L'identification de la séquence est basée sur un encodage de tous les dinucléotides. Chaque séquence est interrogée deux fois et les résultats sont compilés et alignés sur une séquence de référence afin de décoder la séquence-cible. **(c) La méthode par pyroséquençage est utilisée sur la plateforme Roche 454 Life Sciences.** Les microbilles liées à l'ADN sont incorporées dans des puits de la plaque "PicoTiterPlate" et des billes couplées à de la sulphurylase et de la luciférase sont ajoutées. A l'intégration de la base complémentaire adéquate un signal lumineux est émis et capté par une caméra. **(d)** La lumière générée par les réactions enzymatiques est enregistrée en une série de pics appelée chromogramme. D'après Metzker, 2010.

Comparaisons et challenges

Dans le Tableau 1, nous présentons un résumé des caractéristiques techniques de chaque technologie de séquençage. Les différentes plateformes fournissent des longueurs moyennes de séquences variables, en particulier depuis le lancement du MinION (Oxford Nanopore Technology). Le nombre de séquences générées, le type et le taux d'erreur associé sont très différents d'une plateforme à l'autre et sont autant de critères pour aider au choix de la technique de séquençage (Tableau1). Suivant la problématique du projet et le modèle biologique, il est essentiel de choisir la technologie la mieux adaptée. Dans le cas d'un séquençage de génome *de novo*, d'une espèce non-modèle par exemple, il est essentiel de générer de longues séquences qui seront plus aisées à assembler. Dans ce cas, les technologies 454 Life Sciences, IonTorrent, PacBio et MinION pourraient être des choix intéressants. Néanmoins, il faut produire une profondeur de séquençage suffisante au vu de la taille supposée du génome à séquencer (pour un assemblage de bonne qualité la profondeur doit se situer entre 15 et 30X). Par exemple, dans le cas d'un génome de 300 Mb, le séquenceur doit fournir au moins 5Gb de données pour atteindre une profondeur de lectures de 15X ; ce qui équivaut à plusieurs run de 454 GS FLX Titanium XL+ ou de Ion Torrent PGM 318. Dans le cas du séquençage de transcriptome, le même raisonnement est appliqué et les tissus de l'espèce analysée doivent être multipliés afin de séquencer le plus grand nombre de gènes différents. Dans ce but, une phase de normalisation peut être utile (avant le séquençage) afin de limiter la sur-représentativité des gènes très exprimés dans les banques des tissus analysés. Il existe des méthodes permettant d'écarter les séquences ribosomiques ou d'autres gènes de ménage fortement exprimés (Shcheglov *et al.*, 2007). Dans le but de réduire l'effort de séquençage et d'augmenter la profondeur pour chaque gène-cible ou voie métabolique, il peut s'avérer utile d'enrichir les banques en gènes d'intérêt. Trois méthodes d'enrichissement de banque de gènes sont discutées dans la revue de Mamanova *et al.* (2010) : la capture de séquences par hybridation, l'amplification et le séquençage d'amplicons et l'amplification basée sur l'inversion de sonde moléculaire. La méthode par capture de séquences (Grover *et al.,* 2012b) est une des nouvelles approches prometteuses, ayant été utilisée récemment avec succès chez le coton polyploïde (Salmon *et al.,* 2012).

Dans une optique de quantification des niveaux de transcription, il est préférable d'avoir au préalable réalisé un premier transcriptome de référence pour l'espèce d'intérêt. Les transcrits des individus testés suivant différentes conditions peuvent ensuite être séquencés grâce à des technologies produisant des séquences de faible longueur mais permettant une profondeur très importante. Dans les études quantitatives, la profondeur est d'autant plus importante qu'elle sera le support des analyses statistiques d'expressions différentielles. Dans ce cas, les séquenceurs Illumina, SOLiD et HeliScope peuvent être de bons candidats. Le choix se faisant ensuite sur le rendement nécessaire pour chaque modèle biologique.

Dans l'étude des espèces polyploïdes, la taille importante des génomes exige en général un effort de séquençage important. La redondance d'informations (copies dupliquées, homéologues ou paralogues) au sein de ces génomes est une autre caractéristique à prendre en compte dans l'analyse de ces génomes. Différentes approches sont utilisées suivant le type de données de séquençage à traiter en terme de contrôle qualité, d'assemblage et d'analyses de résultats. Ainsi, je tenterai dans la suite de cette revue bibliographique, de décrire les outils utiles à l'étude des espèces polyploïdes et de remettre en perspective les avantages et les limites actuelles des nouvelles techniques de séquençage et d'analyse des génomes complexes.

Tableau 1 : Propriétés des séquenceurs de Nouvelle Génération disponibles. D'après Henson *et al.* (2012), Check Hayden (2012) et le site du fabricant www.helicosbio.com.

Séquenceur	Longueur moyenne des séquences	Rendement par run	Durée d'un run	Coût du séquenceur (US$)	Coût des réactifs par Gb (US$)	Type d'erreur	Taux d'erreurs
HiSeq 2000	100pb	600Gb	11 jours	690 000	40	Substitution	1 à 2% au-delà de 100pb
SOLiD 4	75pb	100Gb	12 jours	475 000	<110	biais A/T	0,06%
SOLiD 4hq	75pb	300Gb	14 jours	595 000	70	biais A/T	0,01%
SOLiD PI	75pb	77Gb	8 jours	349 000	80	biais A/T	0,01%
454 GS FLX Titanium XL+	700pb	700Mb	23h	500 000	7000	Indel	0,50%
IonTorrent PGM 316	200pb	100Mb	env. 2h	50 000	<7500	Indel	1,2% au-delà de 150pb
IonTorrent PGM 318	200pb	1Gb	env. 2h	50 000	<925	Indel	1,2% au-delà de 150pb
MiSeq	150pb	>1Gb	27h	125 000	740	Substitution	1 à 2% au-delà de 100pb
454 GS Junior	400pb	35Mb	12h	108 000	22000	Indel	1%
PacBio RS	2700pb	90Mb par cellule	<1 jour?	695 000	11 000-340 000	déletion C/G	13%
HeliScope	35pb	410Mb par jour	NA	1 000 000	500	Indel	1 à 3%
MinION	100 000pb	100kb/sec/cellule	NA	900	35	Substitution	4%

II. L'assemblage et l'analyse des génomes complexes

De nombreux logiciels existent aujourd'hui pour assembler les séquences issues des différentes technologies NGS. Plusieurs articles de revue ont déjà listé les logiciels disponibles pour l'assemblage de génome (Henson *et al.*, 2012) et de transcriptome (Kumar & Blaxter, 2010). Leur principale différence réside dans leur modèle mathématique utilisé pour traiter les données : la stratégie des graphes de De Bruijn (ou des k-mers) ou celle des graphes des séquences chevauchantes (méthode overlap ou layout ; Myers, 1995) (pour revue : Miller *et al.*, 2010). Le repérage des chevauchements peut se faire en alignant deux à deux chaque nouvelle séquence ajoutée à chaque contig déjà assemblé. Si le score d'alignement est supérieur à un seuil donné, les deux séquences sont considérées comme chevauchantes. La méthode des k-mers utilise des fragments de séquences d'une longueur k et la séquence consensus est créée en ajoutant les éléments se partageant un k-1 identique (Imelfort, 2009). Cette approche est donc plus appropriée pour le traitement d'un grand nombre de séquences de petites tailles (type données Illumina). Ainsi, pour choisir un assembleur adapté, il faut considérer la quantité des séquences à traiter (et disposer d'une puissance de calcul adaptée), leur nature (seules ou pairées), leur longueur et le type d'erreur de séquençage susceptible de se retrouver dans les données. Enfin, les fichiers de sortie doivent être clairs et faciles à parcourir pour disposer des informations utiles.

Deux approches distinctes peuvent être utilisées pour les espèces non-modèles : l'assemblage peut être réalisé à l'aide d'un génome phylogénétiquement proche séquencé par *mapping* ou à l'aide d'une approche *de novo* (sans génome de référence) ; l'avantage de cette dernière approche est aussi d'identifier des séquences spécifiques à l'organisme analysé. Dans cette catégorie, les assembleurs les plus utilisés dans l'assemblage de génome ou de transcriptome sont gsAssembler (communément désigné sous le nom de Newbler et développé par Roche-454 Life Sciences), CAP3 (Huang & Madan, 1999), MIRA (Chevreux *et al*., 1999 ; 2004) qui peut effectuer des assemblages hybrides quelle que soit l'origine des séquences (Sanger et NGS), Velvet (Zerbino & Birney, 2008) et Trinity (Grabherr *et al.*, 2011), ces trois derniers logiciels permettant de traiter des données pairées. De même, SOAPdenovo (Li *et al.*, 2009a), ABySS (Simpson *et al.*, 2009) et AllPATHS (Butler *et al*., 2008)

sont utilisés pour assembler des séquences pairées issues de génomes larges et redondants (pour revue voir Henson *et al.*, 2012 et Miller *et al.*, 2010).

Dans le but d'assembler un génome entier, il est indispensable de combiner plusieurs types de données NGS uniques ou pairées avec différentes tailles d'insert (Henson *et al*., 2012). Les séquences pairées (mate-paired ou paired-end se différencient par la distance entre les deux extrémités séquencées) permettent l'ancrage des séquences et améliorent sensiblement un assemblage *de novo*. Plusieurs articles tentent de comparer les différents assembleurs (Kumar & Blaxter, 2010 ; Brautigam *et al.*, 2011) sur des jeux de données réels mais aucun d'entre eux ne remplit parfaitement tous les critères. Chacun possède ses spécificités, ses atouts et ses limites (Tableau 2). Le projet de l'assemblathon (Earl *et al.*, 2011) fait intervenir plusieurs équipes internationales autour de la question de l'assemblage de fragments courts à partir d'un génome diploïde simulé (séquences Illumina pairées de 100pb) et testent différents assembleurs (ceux cités plus haut et d'autres développés dans les équipes). Les résultats sont encourageants et mettent en évidence la possibilité d'assembler un génome avec à la fois une couverture importante et une grande exactitude. Selon les auteurs, les critères les plus importants à prendre en compte concernent la taille et la couverture des contigs générés. Bien qu'aucun assembleur ne sorte particulièrement du lot (chacun possédant des limites algorithmiques propres), les logiciels ALLPATHS, SOAPdenovo et SGA (pour String Graph Assembler) montrent de bons résultats pour la plupart des critères choisis. Une autre étude, cette fois sur des données réelles de plusieurs espèces (séquences Illumina pairées de 101pb), compare différents assembleurs disponibles (Salzberg *et al*., 2011). Les auteurs mettent en avant l'importance de la qualité des séquences initiales (séquençage et nettoyage) plutôt que l'importance de l'assembleur en lui-même. Dans cet article, les critères principaux de contrôle seraient le nombre de contigs générés et leur taille ce qui rejoint les résultats de l'Assemblathon. Salzberg et ses collaborateurs (2011), sur leurs jeux de données et leurs critères d'assemblage, privilégient le programme ALLPATHS.

Malgré ces études récentes, aucun critère standard pour évaluer la fiabilité et l'efficacité d'un assemblage n'est communément admis (Strickler *et al.*, 2012). Martin & Wang (2011) ont proposé de comparer l'assemblage à un sous-ensemble de transcrits dont la taille et l'abondance serait connues afin d'évaluer le contigage, la présence de chimères et

de variants d'épissage. Dans la plupart des études, les contigs sont alignés contre des bases de données de gènes (ou des génomes complets) afin d'estimer la proportion de transcrits séquencés. Si un génome phylogénétiquement proche séquencé est disponible, l'alignement par « mapping » des séquences obtenues sur un génome (ou transcriptome) de référence permet de tester la précision de l'assemblage réalisé. Pour ces analyses, un logiciel capable de créer des « gaps » dans les séquences (si les transcrits sont alignés sur un génome de référence) et capable d'identifier des variants d'épissage dans le cas de données de séquences transcriptomiques (variables d'un tissu à un autre, d'une espèce à une autre) est nécessaire. Le Tableau 2 répertorie ces programmes. GsMapper (454 Life Sciences) pour les données 454, Bowtie (Langmead *et al.*, 2009) pour les données de séquences courtes et Mosaik (Smith *et al*., 2008) font partie des suites logicielles les plus utilisées. Enfin, l'inspection visuelle des contigs et des alignements est indispensable afin de vérifier l'intégrité des contigs créés avec par exemple, le logiciel Tablet (Milne *et al.*, 2010).

La plupart des articles s'appuient sur des données issues de génomes diploïdes et ne prennent pas en compte les caractéristiques des génomes polyploïdes ou paléopolyploïdes qui compliquent les assemblages de séquences. En plus de la distinction des paralogues (présents aussi chez les espèces diploïdes), une couche additionnelle de complexité est apportée par la présence de copies dupliquées (ou homéologues) issues de la polyploïdisation. Selon l'âge du polyploïde et son origine (allopolyploïde ou autopolyploïde), l'homologie entre les homéologues à chaque locus va diverger de façon plus ou moins importante. La stratégie d'assemblage consiste alors à réaliser un assemblage consensus à chaque locus qui ne distingue pas les copies homéologues mais dont les paramètres prennent soin de maximiser les chances d'intégrer toutes les copies (paralogues et homéologues). Ensuite, les copies homéologues peuvent être différenciées par deux approches : si les parents diploïdes sont identifiés, ils peuvent servir de référence pour distinguer les sous-génomes homéologues réunis au sein de l'allopolypoïde. Grâce à la présence de polymorphismes nucléotidiques spécifiques aux parents, les copies homéologues peuvent être affiliées à leur génome parental d'origine (comme décrit chez les cotons polyploïdes : Udall *et al*., 2006 ; Flagel & Wendel, 2010 ; Salmon *et al.*, 2010 ; Yoo *et al*., 2013, ou le soja tétraploïde : Ilut *et al*., 2012). Dans le cas où les parents diploïdes ne

sont pas disponibles, la détection des copies homéologues chez les polyploïdes est plus difficile et représente un défi méthodologique.

Tableau 2 : Programmes utilisés dans les assemblages de génome et de transcriptome (modifié d'après Henson *et al.*, 2012 et Strickler *et al.*, 2012).

Outil	Type d'assemblage	Type de données	Caractéristiques principales
CAP3 (Huang *et al.*, 1999)	Assemblage *de novo*	454	Utile pour assembler des contigs, méthode de clustering
gsAssembler (454 Life Sciences)	Assemblage *de novo* et mapping	454	Nouvelle version améliorée et rapide, précis et efficace
MIRA (Chevreux *et al.*, 2004)	Assemblage *de novo* et mapping	454 et Illumina	Performant et rapide dans les assemblages hybrides et pairés, bien documenté
TGICL (Pertea *et al.*, 2003)	Assemblage *de novo*	454	Méthode de clustering
Trinity (Grabherr *et al.*, 2011)	Assemblage *de novo*	Illumina	Traite les données pairées
Bowtie (Langmead *et al.*, 2009)	Mapping	Illumina	Traite à la fois les données pairées et non-pairées
BWA (Li & Durbin, 2009)	Mapping	Illumina	Alignement de courtes séquences rapide et fiable, traite les données pairées
BWA-SW (Li & Durbin, 2009)	Mapping	454	Alignement rapide des séquences, permet les gaps
GSNAP (Wu & nacu, 2010)	Mapping	Illumina	Détecte les SNPs et les variants d'épissage
MAQ (maq.sourceforge.net)	Mapping	Illumina	Détecte les SNPs
Mosaik (bioinformatics.bc.edu/marthlab)	Mapping	454, Illumina, SOLiD, Helicos et Sanger	Traite les données pairées et non-pairées, rapide, détecte les indels et permet les gaps
Velvet-Oases (Zerbino & Birney, 2008)	Assemblage *de novo*	454, Illumina et SOLiD	Traite les données pairées (paired-end et mate-pair), peut combiner des données de 454 et d'Illumina, bon support, grande variété de fichiers de sortie
SOAPdenovo (Li *et al.*, 2009a)	Assemblage *de novo*	Illumina	Traite les données pairées et non-pairées, facile d'utilisation, fiable et rapide
ABySS (Simpson *et al.*, 2009)	Assemblage *de novo*	Illumina, 454, SOLiD et Sanger	Traite les données pairées et non-pairées, rapide, besoin de peu de RAM
ALLPATHS (Butler *et al.*, 2008)	Assemblage *de novo*	Illumina	Traite les données pairées (paired-end et mate-paired), possibilité d'ajouter des séquences 454

III. NGS et évolution des génomes polyploïdes

Les technologies de séquençage en masse, en fournissant une grande profondeur de lecture, offrent donc des perspectives nouvelles dans l'analyse des génomes polyploïdes redondants, pour lesquels plusieurs copies de gènes sont attendues à chaque locus, et plus particulièrement pour les systèmes non modèles ne disposant pas de ressources génomiques importantes. Les premiers travaux (que nous présenterons ci-dessous) utilisant les NGS dans l'analyse des polyploïdes, qui commencent à être publiés depuis 2010, montrent leur utilité dans l'analyse de l'évolution des séquences répétées, la dynamique évolutive des gènes et l'évolution de l'expression des gènes.

1. Analyse des séquences répétées

Les séquences répétées constituent une part majeure des génomes eucaryotes et leur expansion ou contraction représente un paramètre clé dans les variations de la taille des génomes de base, conférant une plasticité particulière aux génomes de plantes, pour la plupart polyploïdes (Leitch et Leitch 2008). Parmi les séquences répétées jouant un rôle important dans la dynamique des génomes polyploïdes, on compte les éléments transposables et les familles multigéniques (comme les gènes ribosomiques). Il existe à ce jour plusieurs revues présentant les différentes catégories d'éléments transposables et leur classification (*e.g.* Deragon *et al.*, 2008 ; Wicker *et al.,* 2007 ; Grandbastien & Casacuberta, 2013 ; Kejnovsky *et al*., 2012). Chez les Poacées, les rétrotransposons à LTR (Long Terminal Repeat) de la classe I sont les plus représentés dans les espèces analysées à ce jour (Devos, 2010).

Si les séquences répétées compliquent les analyses d'assemblage de séquences orthologues (Treangen & Salzberg, 2011), les NGS offrent aujourd'hui une possibilité plus rapide de détecter les séquences répétées présentes dans un génome complexe et d'évaluer leurs proportions relatives par rapport au compartiment codant. C'est ainsi que des génomes de base de taille notoirement importante (*e.g.* le pois : Macas *et al.*, 2007 ; le blé et l'orge : Wicker *et al.*, 2009) ont bénéficié de la technologie 454 Roche, qui a permis de détecter plusieurs familles d'éléments répétés, d'évaluer leur proportion respective et d'analyser leur évolution. L'équipe de J. Macas (de l'Institut of Plant Molecular Biology, Ceske Budejovice, CZ) a mis au point une méthode (présentée dans Novak *et al.*, 2010) basée

sur le regroupement de séquences (« clustering ») selon leur similarité nucléotidique. Les séquences de chaque groupe de similitude (ou « cluster ») sont assemblées et annotées (en utilisant un pipeline faisant intervenir différentes bases de données d'éléments répétés). L'abondance des éléments peut aussi être estimée grâce aux nombres de séquences présentes dans le « cluster » correspondant. Plusieurs études ont utilisé cette méthode (chez le soja : Swaminathan *et al.*, 2007 ; l'orge : Wicker *et al*., 2008 ; le bananier : Hribova *et al.*, 2010 et dans le genre *Silene*, Macas *et al.*, 2011). Renny-Byfield *et al.* (2011) ont employé cette méthode chez le tabac (*Nicotiana)*, afin de comparer les génomes de l'allopolyploïde *N. tabacum* (formé il y a environ 200 000 ans) et de ses deux parents : *N. sylvestris* (génome S maternel) et *N. tomentosiformis* (génome T paternel), à partir d'un jeu de données de pyroséquençage (Roche 454). Ces auteurs montrent que les génomes de ces 3 espèces ont suivi des trajectoires évolutives différentes : l'espèce maternelle a subi une dynamique plus importante (expansion / homogénéisation) des séquences répétées que l'espèce paternelle ; dans le génome de l'allopolyploïde, les séquences du génome paternel semblent préférentiellement éliminées, conduisant à une réduction de la taille du génome (« genome downsizing ») par rapport à la taille (additive) attendue au vu de celles des espèces parentales. En combinant séquençage à haut débit (Roche 454 et Illumina) et analyses de cytogénétique moléculaire (Fluorescent *In Situ* Hybridization), Chester *et al.* (2012) ont identifié et localisé un groupe de séquences hautement répétées apparentées aux Chromovirus (Eléments transposables de classe I de type *Gypsy*) qui semblent être la cible particulière d'élimination préférentielle chez le tabac allotétraploïde ; ce phénomène est aussi observé chez le tabac synthétique, indiquant un processus rapide et dirigé d'élimination.

Chez les espèces allotétraploïdes récemment formées dans le genre *Tragopogon* (Soltis *et al.*, 2012), l'utilisation des NGS (séquençage Roche 454) a conduit à identifier des séquences répétées en tandem spécifiques de chaque génome parental diploïde, qui ont permis de montrer (par hybridation *in situ*) la dynamique particulière (aneuploïdie, monosomie ou trisomie) des jeux de chromosomes d'origine paternelle et maternelle dans les populations naturelles de l'allopolyploïde (Chester *et al.*, 2012).

Les séquences répétées en tandem, encore appelées microsatellites ou Short Sequence Repeat (SSR) sont abondantes et distribuées sur l'ensemble du génome

(Buschiazzo & Gemmell, 2006). Le nombre d'unités et donc la longueur des microsatellites varient fortement entre les génotypes des individus et sont de ce fait d'un usage courant en génétique des populations où ils sont utilisés comme marqueurs co-dominants. Certaines études utilisent les nombres d'allèles microsatellites pour déduire le niveau de ploïdie en comparant génotypes diploïdes et polyploïdes (*e.g*. chez *Panicum virgatum* : Okada *et al*., 2011). Les études antérieures utilisaient la méthode de Sanger et plusieurs étapes étaient nécessaires pour la réalisation d'une banque de microsatellites (enrichissement, clonage, séquençage). Aujourd'hui, les NGS offrent des méthodes rapides pour la détection d'un grand nombre de marqueurs microsatellites (Ekblom & Galindo, 2011). Les technologies 454 et Illumina sont toutes les deux utilisées dans ce but, mais le séquenceur Roche apparaît plus performant dans le nombre de séquences trouvées et permet surtout de générer de plus longues séquences. La probabilité de séquencer des microsatellites complets et de pouvoir dessiner des amorces dans les régions flanquantes est donc augmentée même pour une profondeur de séquençage moins élevée (95 études chez les plantes, passées en revues par Zalapa *et al.*, 2012). Il existe aujourd'hui plusieurs logiciels comme MISA (Microsatellite research tool, Thiel *et al*., 2003) mreps (Kolpakov *et al.*, 2003) ou WebSat (Martins *et al.*, 2009) permettant d'identifier des microsatellites présentant des motifs tri ou tétra-nucléotidiques sur au moins 12pb (Zalapa *et al.*, 2012). Si les microsatellites sont le plus souvent étudiés à partir d'ADN génomique, l'identification de microsatellites dans les régions codantes permet de retrouver des associations avec des fonctions géniques intéressantes (Li *et al.*, 2002), d'avoir un nombre de séquences moins élevées que dans les régions non-codantes (Blanca *et al*., 2011), d'augmenter la transférabilité des banques et surtout de s'assurer que le microsatellite isolé soit présent dans un seul locus (Zhu *et al*., 2011). Chez l'espèce tétraploïde *Spartina pectinata*, le séquençage de transcriptome à partir de la technologie Roche 454 a permis d'identifier 841 sites microsatellites (Gedye *et al.*, 2010).

2. Evolution des gènes dupliqués : Identification, rétention – pertes, polymorphismes

Les duplications partielles (géniques ou segmentaires) ou génomiques (polyploïdie) créent une redondance d'information génétique qui peut évoluer de différentes façons (Ohno, 1970 ; Lynch et Conery 2000 ; Adams & Wendel, 2004 ; Flagel & Wendel, 2009 ; Schnable *et al.*, 2012). La perte de gènes dans les génomes d'anciens polyploïdes résulte du phénomène appelé « fractionation » (Langham *et al.*, 2004) qui contribue à la « diploïdisation » génétique des paléopolyploïdes dans la mesure où leur ancêtre peut être qualifié de « diploïde » (Freeling, 2009). La perte de ces gènes n'est pas forcément aléatoire et depuis la publication des premiers génomes de plante séquencés, différents travaux ont exploré la question du lien entre gènes retenus, leur catégorie fonctionnelle, et les mécanismes de régulation de leur expression (e.g. *Arabidopsis thaliana* : Blanc & Wolfe, 2004 ; Thomas *et al.*, 2006). Cette dynamique de perte physique de gènes dupliqués est aussi observée dès les premières générations suivant la formation des polyploïdes (Doyle *et al.*, 2008).

Le séquençage massif parallèle permet à présent de documenter à plus grande échelle cette dynamique de la structure des génomes polyploïdes. Dans le genre *Tragopogon,* Buggs et ses collaborateurs (2010) ont utilisé des données de séquençage de transcriptome issues des plateformes 454 et Illumina afin d'identifier plusieurs milliers de SNPs entre les parents et pouvoir ainsi différencier les copies parentales homéologues chez l'allotétraploïde récemment formé *T. miscellus*. Cette première étude a permis d'identifier 7782 SNPs, dont 92 ont été validés sur ADN génomique par la technologie Sequenom MassARRAY iPLEX Genotyping (Gabriel *et al.*, 2009). Ces ressources ont permis de détecter différents dosages alléliques parentaux dans les populations et la perte récurrente de copies homéologues pour certaines catégories fonctionnelles de gènes dans des populations naturelles de *T. miscellus*, issus d'évènements indépendants d'hybridation entre les parents diploïdes. Ces auteurs notent que ce sont les mêmes catégories de gènes (sensibles aux effets de dosage) dont la copie dupliquée a été perdue suite à la polyploïdie plus ancienne dans la famille des Astéracées. Flagel *et al.* (2012) ont réalisé chez *Gossypium* un assemblage de transcriptome basé sur des données de séquençage en Sanger, 454 et Illumina leur permettant de détecter plus de 250 000 SNPs différenciant les parents diploïdes et leur

permettant de distinguer les sous-génomes homéologues chez le coton allotétraploïde. Ils montrent ainsi une augmentation des taux de substitutions au sein du génome des espèces allotétraploïdes et une proportion non négligeable de cas de recombinaisons homéologues non-réciproques.

Le séquençage à haut débit couplé aux approches d'enrichissement en séquences cibles par captures de gènes ou de régions génomiques (Grover *et al*., 2012b) commence à être utilisé en biologie évolutive dans les analyses de génomes polyploïdes (comparaisons de copies homéologues, études de polymorphismes). Chez l'allotétraploïde *Gossypium hirsutum* sauvage et cultivé (coton) contenant deux génomes AD dupliqués, Salmon *et al*. (2012) ont pu analyser la diversité et les niveaux d'hétérozygotie de plusieurs centaines de paires de copies de gènes homéologues et ont montré que la réduction d'hétérozygotie chez la forme cultivée affectait préférentiellement le sous-génome A. Winfield *et al.* (2012) ont également utilisé la capture de séquences pour identifier plus de 500 000 SNPs distribués entre les génomes homéologues A, B et D du blé hexaploïde (*Triticum aestivum*) et différentes accessions de blé. Bundock *et al.* (2012) ont utilisé les informations du génome du sorgho pour capturer les séquences du génome (phylogénétiquement proche) de la canne à sucre (*Saccharum officinarum*), dans la perspective de générer des marqueurs SNPs pouvant être utilisés dans ce complexe polyploïde.

Enfin, la profondeur de lecture des séquences fournies par les NGS offre de nouvelles perspectives dans l'analyse de produits d'amplifications (« séquençage d'amplicons ») d'un grand nombre d'échantillons. Parmi les polyploïdes, ces approches ont été utilisées chez *Saccharum officinarum* (Bundock *et al.*, 2012), *Brassica napus* (Gholami *et al.*, 2012) et *Triticum* (Bérard *et al.*, 2009 ; Lai *et al.*, 2012 ; Edwards *et al.*, 2012). Le séquençage d'amplicons s'avère une excellente alternative au clonage de séquences dans la perspective de reconstructions phylogénétiques chez les espèces polyploïdes. Griffin *et al.* (2011) ont pu ainsi analyser les relations phylogénétiques d'espèces polyploïdes du genre *Poa*.

3. **Contribution des NGS à l'analyse de l'expression des gènes chez les polyploïdes**

Les nouvelles technologies de séquençage à haut débit offrent des possibilités particulières à l'analyse de l'expression des gènes chez les polyploïdes dans la mesure où elles permettent, non seulement d'évaluer les niveaux d'expression à partir de séquençage de transcriptome (RNA-Seq), mais parce qu'elles offrent un accès direct à l'expression des différentes copies homéologues. Auparavant, les niveaux d'expression respectifs de copies homéologues n'avaient pu être analysés à l'échelle du génome (plusieurs milliers de gènes) que dans de rares situations, comme chez le coton allotétraploïde (Hovav *et al*., 2008 ; Flagel *et al*., 2008), pour lequel des puces (microarrays) spécifiques des sous-génomes parentaux diploïdes avaient été mises au point à partir de SNPs différenciant les ESTs des deux espèces diploïdes parentales (Udall *et al*., 2006). Chez *Tragopogon*, Buggs *et al*. (2010) ont annoté une banque d'ADNc du parent diploïde, séquencée par la technologie Roche 454, qui a servi de référence pour aligner les jeux de données de séquence obtenues par la technologie Illumina à partir des deux espèces parentales diploïdes et de l'allotétraploïde *T. miscellus*. Ces auteurs ont pu faire le lien entre mise sous silence de certaines copies et leur perte physique détectée à partir d'ADN génomique. Ilut *et al.* (2012) ont comparé le transcriptome de feuilles de l'allotétraploïde *Glycine dolicharpa* comparée à ceux de ses parents diploïdes à l'aide de séquençage massif (RNA-Seq) par la technologie Illumina, et en se servant du génome du soja (*Glycine max*) comme référence (montrant une augmentation de l'activité photosynthétique chez le polyploïde). La contribution relative des homéologues parentaux au transcriptome de l'allopolyploïde a pu être ainsi évaluée, en s'intéressant plus particulièrement aux gènes impliqués dans la photosynthèse. Chez le tabac allotétraploïde (*Nicotiana tabacum*), Bombarely *et al*. (2012) ont combiné la détection de copies homéologues à partir des assemblages de séquences transcriptomiques (obtenues à l'aide de la technologie Roche 454) des parents diploïdes et d'une approche phylogénomique intégrant les séquences des diploïdes et de l'allotétraploïde.

Chez le coton allotétraploïde naturel et resynthétisé expérimentalement, Yoo *et al.* (2013) ont analysé par RNA-Seq (Illumina) le lien entre biais d'expression de copies homéologues (issues du génome A ou du génome D) et le phénomène de « dominance parentale », une situation de « non-additivité » d'expression parentale où l'expression globale du gène de l'allotétraploïde est similaire à celle de l'un des deux parents diploïdes.

Ces auteurs montrent de façon inattendue que lorsqu'un allotétraploïde « mime » le niveau d'expression d'un de ses parents (« dominant »), ce niveau d'expression peut en fait être atteint par sur-expression de la copie homéologue issue de l'autre parent non dominant.

L'expression des gènes est modulée par les mécanismes épigénétiques, et les nouvelles technologies de séquençage ouvrent également de nouveaux horizons dans ce domaine (Hirst & Marra, 2010). Les analyses de méthylome faisant intervenir le séquençage en masse après traitements de l'ADN au bisulphite publiées à ce jour restent encore essentiellement l'apanage des espèces modèles dont le génome est séquencé (*e.g*. *Arabidopis thaliana*, Becker *et al*., 2011 ; *Oryza sativa* : Li *et al*., 2012). Chez les polyploïdes, on note l'étude de Ha *et al*. (2009) par séquençage massif parallèle après immunoprécipitation de la chromatine chez *Arabidopsis thaliana*, *A. arenosa* et leurs descendants allotétraploïdes.

Dans le genre *Triticum*, les travaux de Kantar *et al.* (2012) reportent le séquençage massif de micro-ARNs du sous-génome A (chromosome 4) du blé polyploïde. Un screening par séquençage massif des petits ARN chez les hybrides et allopolyploïde synthétiques de blés a permis à Kenan-Eichler *et al.* (2011) de montrer que les miRNA augmentaient avec le niveau de ploïdie, tandis que les siRNA diminuaient ; cette diminution des siRNA a été associée à une potentielle mobilisation des éléments transposables suite à l'hybridation et l'allopolyploïdie. Récemment, Thiebaut *et al.* (2012) ont utilisé les NGS pour identifier de nouveaux miRNA chez *Saccharum officinarum* et leur rôle dans les réponses au stress biotique et abiotique.

Les prochaines années devraient donc voir s'accumuler rapidement de nouvelles données permettant d'explorer à plus large échelle et dans de nombreux systèmes biologiques, les réponses variées des génomes à la polyploïdie.

Chapitre 2

Présentation du système biologique: Les Spartines dans la famille des Poacées

Dans ce chapitre nous restituerons la place du genre *Spartina* dans la famille des Poacées en présentant plus particulièrement ce qui est connu dans la littérature concernant leur histoire évolutive.

I. Evolution du génome des Poacées

La famille des Poacées est représentée par plus de 700 genres et 10 000 espèces, et couvre plus de 20% de la surface de la Terre. Les différents mécanismes de reproduction, leur variabilité anatomique et génétique leur confèrent de grandes capacités d'adaptation aux principaux types d'habitats terrestres. Les espèces sauvages constituent une part dominante des milieux ouverts, prairies des régions tempérées, steppes des régions semi-arides ou arides, savanes des régions tropicales. C'est une famille agronomiquement importante autant pour l'alimentation humaine et le fourrage que par le nombre d'espèces cultivées à travers le globe.

L'histoire évolutive des Poaceae et leurs relations phylogénétiques sont aujourd'hui bien connues. Les monocotylédones ont divergés des eudicotylédones à la fin du Jurassique, il y a entre 140 et 150 millions d'années (Chaw *et al*., 2004). Tang et ses collaborateurs (2010) ont mis en évidence chez le riz et le sorgho un événement de duplication génomique datant de cette période. La famille des Poaceae a divergé il y a 80 à 85 millions d'années (Prasad *et al.*, 2005), cet évènement s'est accompagné d'une duplication génomique il y a 70 millions d'années (Paterson *et al.,* 2004). Par la suite, les deux grands clades actuels formant les Poaceae ont divergé (Figure 3) : le clade BEP est composé des sous-familles des Bambusoideae, Pooideae et Ehrhartoideae ; le clade PACMAD est composé des sous-familles des Panicoideae, Arundinoideae, Chloridoideae, Micraroideae, Aristidoideae et Danthonioideae. Ces deux clades ont divergés il y a 50 à 70 millions d'années (Paterson *et al*., 2004). Au sein du clade PAC, l'ancêtre commun du maïs et du sorgho serait apparu il y a entre 29 et 12 millions d'années, date de la divergence entre ces deux espèces (Paterson *et*

al., 2004 ; Swigonova *et al*., 2004). La sous-famille des Chloridoideae est encore peu étudiée, néanmoins une étude récente placerait la date de divergence entre les Chloridoideae et les Panicoideae entre 34,6 et 38,5 millions d'années (Kim *et al.,* 2009). Les relations phylogénétiques au sein des Chloridoideae ont fait l'objet de plusieurs études qui n'arrivaient pas toujours à résoudre de façon satisfaisante les liens entre les différents genres qui ont souvent été nommés de façon artificielle (Hilu & Alice 2001). Les phylogénies moléculaires récentes (Peterson *et al.,* 2010) placent le genre *Spartina* proche des genres *Calamovilfa* et *Sporobolus* au sein de la sous-tribu des Sporobolinae, et dans la tribu des Zoysieae. Cette tribu a pour lignée sœur la tribu des Cynodonteae contenant le genre *Eleusine* (dont *E. coracana* ou « finger millet ») qui aurait divergé des Spartines il y environ 22 millions d'années (Liu *et al.,* 2011).

La phylogénie des Poaceae a été intensément étudiée notamment à travers l'apparition de la physiologie en C4 dans la famille (Kellogg, 2001 ; Christin *et al*., 2008 ; Grass Phylogeny Working Group II, 2012). La photosynthèse en C4 permet de concentrer le CO_2 sur le site de fixation de la Rubisco pendant le cycle de Calvin (Sage, 2004 ; Edwards *et al*., 2010). Le métabolisme en C4 permet de réduire la photorespiration et de saturer la photosynthèse en CO_2. Ces modifications anatomiques et biochimiques permettent ainsi aux plantes C4 de coloniser des habitats ouverts et secs dans les régions tropicales et sub-tropicales (Osborne & Freckleton, 2009 ; Edwards & Smith, 2010). Ce métabolisme est apparu à plusieurs reprises au cours de l'histoire des Poaceae (Kellogg, 2001). Les Chloridoideae (dont le genre *Spartina)* sont en majorité de type C4 et utilisent la phosphoenol pyruvate carboxykinase (PCK) comme enzyme de décarboxylation (Christin *et al.*, 2008).

Malgré une histoire complexe et la présence de lignées ayant divergées depuis plus de 50 millions d'années, le nombre de gènes et la synténie au sein des Poaceae restent bien conservés. Les premiers travaux de cartographie génétique utilisant des marqueurs RFLP chez les blés ont montré la conservation de la synténie entre les espèces proches de la même sous-famille (Chao *et al.,* 1989). Des études comparatives entre les différentes espèces du clade des Triticeae (blé, orge, seigle et autres espèces sauvages) montrent des niveaux de synténie élevés entre ces espèces (Devos *et al*., 1993a ; 1993b ; Dubcovsky *et al.,* 1996). Le contenu génique et l'ordre des gènes sont aussi maintenus au sein des Poaceae

entre les espèces modèles (*e.g.* le riz, la canne à sucre, le sorgho, le maïs, le blé et l'orge) pour revue voir Gale & Devos (1998) et Devos (2005). Néanmoins, les études comparatives récentes utilisant des inserts de grandes tailles (BAC ou YAC) montrent des réarrangements et des délétions à une échelle plus faible. L'étude de la microsynténie révèle ainsi une diminution de la conservation de l'ordre des gènes (Keller & Feuillet, 2000).

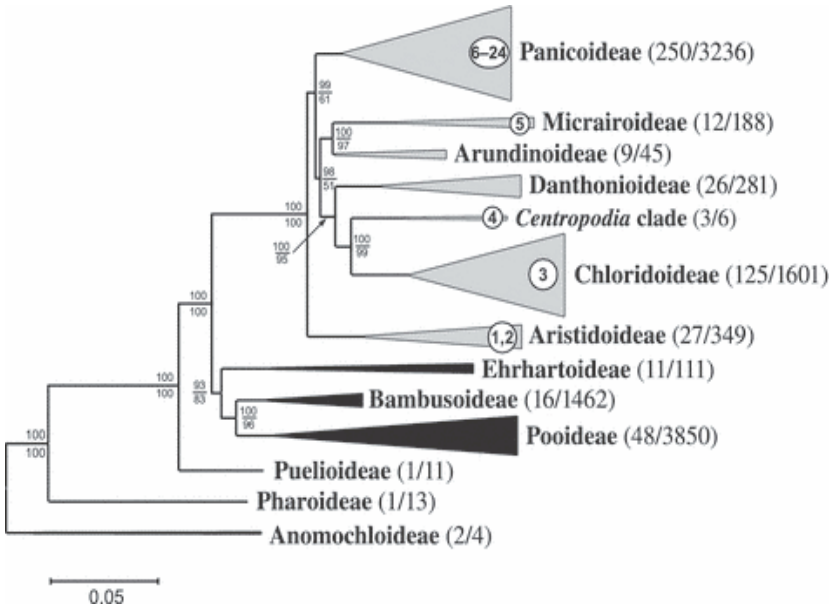

Figure 3 : Relations phylogénétiques entre les sous-familles de Poaceae d'après Grass Phylogeny Working Group II (2012). Les clades BEP (Bambusoideae, Ehrhartoideae et Pooideae) et PACMAD (Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae et Danthonioideae) sont représentés en noir et gris, respectivement. Les nombres en bout de ligne montrent le nombre d'espèces échantillonnées sur le nombre total d'espèces de la sous-famille. Les nombres entourés correspondent à des références d'espèces en C4 impliquées dans des études comparatives, telles qu'indiquées dans la publication GPWG II (2012)**.**

La région *Adh1* a été précédemment montrée comme colinéaire chez les Poacées (Avramova *et al.,* 1996), avec quelques remaniements détectés dans les études plus récentes, comme la mise en évidence de quatre gènes additionnels (gènes 3, 3.5, 8 et 9) chez le sorgho par rapport au riz. Les gènes 3 et 3.5 se sont transposés chez l'ancêtre commun au sorgho et au maïs (Tikhonov *et al*., 1999 ; Devos, 2005). Deux gènes additionnels (gènes 8 et 9) se sont transposés après la dernière divergence entre le sorgho et la canne à

sucre datant d'il y a 8-9 millions d'années (Jannoo *et al.,* 2007). Ainsi, l'organisation des gènes peut être bouleversée par des réarrangements chromosomiques, des pertes de gènes différentielles après des évènements de duplication génomique et l'effet mutagène des éléments transposables (Devos, 2010). Ces facteurs contribuent donc à l'érosion de la colinéarité et à la diversification des génomes.

L'augmentation du nombre de données génomiques générées pour ces espèces modèles a permis de mettre en évidence un grand nombre de duplications segmentaires et de translocations interchromosomiques. Salse et ses collaborateurs (2008) utilisent une méthode d'alignement de séquences orthologues afin de déduire les duplications entre chromosomes du riz, les chromosomes du blé et les relations de colinéarité entre ces deux génomes et la sous-famille des Panicoideae (avec le maïs et le sorgho). Le modèle d'évolution dans cette étude établit un ancêtre à 5 chromosomes ayant subi un événement de duplication génomique, il y a 50 à 70 millions d'années (Paterson *et al.,* 2004).

Suite aux duplications génomiques, deux translocations interchromosomiques et deux fusions des chromosomes (7 et 10, et des chromosomes 4 et 6) ont donné les chromosomes ancestraux 3 et 2, respectivement (Figure 4). L'ancêtre commun à toutes les Poaceae possèderait 12 chromosomes (Gaut, 2002 ; Salse *et al*., 2008). Un nombre variable de chromosomes de base est observé suivant les sous-familles, suggérant une histoire évolutive spécifique à chaque lignée. Dans le modèle proposé par Salse (2008), le riz aurait gardé le nombre de chromosomes de base de l'ancêtre après la duplication (n=12) alors que le nombre de chromosomes de base est réduit chez le blé et l'orge (n=7), *Brachypodium* (n=5) le maïs (n=10) et le sorgho (n=10). Chez les Triticeae (sous-famille des Pooideae, clade des Poideae), le modèle suggère la présence de cinq évènements de fusion. Les espèces actuelles possèdent un nombre chromosomique de base n=7 (Salse *et al.,* 2008 ; Abrouk *et al*., 2010). *Brachypodium distachyon* (sous-famille des Pooideae, clade des Brachypodieae) a une histoire évolutive différente des autres Pooideae analysées (blé et orge) avec sept évènements de fusions interchromosomiques et un nombre chromosomique de base n=5 (Abrouk *et al.,* 2010). Chez les Panicoideae, le maïs et le sorgho ont une histoire évolutive ancienne commune avec l'occurrence de deux fusions chromosomiques sur un génome ancestral à 12 chromosomes (Wei *et al*., 2007 ; Salse *et al.*, 2008). Ensuite, ces deux espèces ont évolué indépendamment. Alors que la structure génomique du sorgho est restée

similaire à celle du génome ancestral des Panicoideae fusionné (n=10) ; le maïs a subit une duplication génomique avec un nombre chromosomique intermédiaire à n=20 (Gaut & Doebley, 1997 ; Swigonova *et al*., 2004 ; Salse *et al.,* 2008). Après cette duplication, un nombre important de fusions entre les chromosomes (au moins 17 selon Salse *et al*., 2008) ont eu lieu, conférant un nombre de chromosomes de base de 10 dans le genre *Zea*.

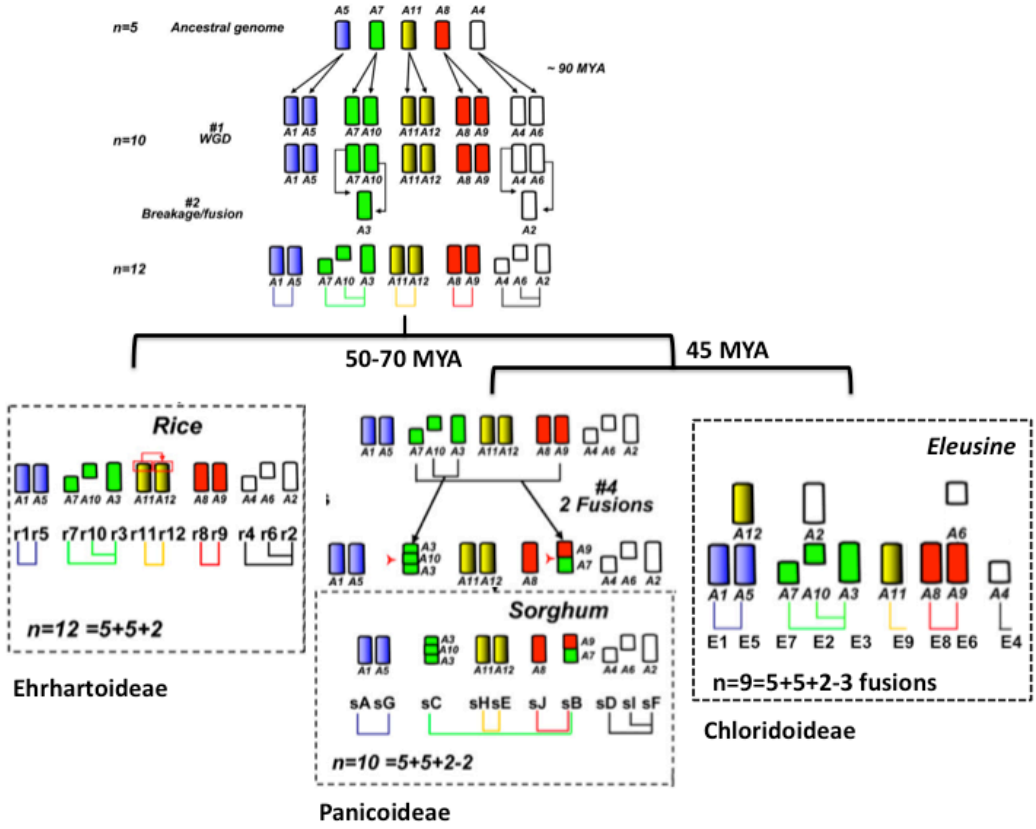

Figure 4 : Modèle d'évolution des génomes du sorgho et du riz modifié d'après Salse *et al.* (2008) à partir d'un ancêtre commun à n=5 chromosomes. La structure chromosomique d'*Eleusine coracana* (Chloridoideae) est rajoutée, par déduction des comparaisons de la carte génétique d'*Eleusine coracana* avec *Oryza sativa* (Srinivasachary *et al*., 2007).

Au sein du clade PACMAD, l'histoire évolutive des représentants des Panicoideae est bien connue (incluant le maïs, le sorgho, la canne à sucre et *Setaria*) alors que la sous-famille des Chloridoideae reste très peu explorée. Cette dernière présente des nombres chromosomiques de base très diversifiés (x=10, 9, 8, 7) et un grand nombre d'espèces polyploïdes avec des niveaux de ploïdie allant de diploïde à 20-ploïde (Peterson *et al*., 2010).

En 2009, Kim et ses collaborateurs ont séquencé et comparé des ESTs (Expressed Sequence Tags) de *Cynodon dactylon* avec les principaux génomes de Poaceae. La divergence entre les Chloridoideae et les Panicoideae a été datée entre 34,6 et 38,5 millions d'années. Cette datation est en accord avec la phylogénie réalisée par Christin *et al.* (2008) en utilisant le gène PCK. Au sein des Chloridoideae, certaines espèces présentant un intérêt agronomique comme le millet (*Eleusine coracana*) ou *Eragrostis tef* ont fait l'objet de cartes génétiques et disposent de quelques ressources génétiques (ESTs) (Zhang *et al*., 2001 ; Dida *et al*., 2006; Kim *et al*., 2009). En comparant *Eleusine coracana* (2n=4x=36) au génome du riz (2n=2x=24) Srinivasachary *et al.* (2007) ont montré des réarrangements et trois évènements de fusions chromosomiques chez cette espèce par rapport aux chromosomes de riz (Figure 4).

Il est aujourd'hui essentiel d'augmenter les ressources génomiques disponibles pour les Chloridoideae afin d'identifier et de dater plus précisément les restructurations interchromosomiques qui génèrent un nombre de base variable chez cette sous-famille et d'approfondir les relations phylogénétiques entre les Panicoideae, Ehrhartoideae et Chloridoideae. Le génome ancestral des Chloridoideae est-il semblable à celui des Panicoideae (n=10) dont il partagerait les particularités ou aurait-il gardé la structure ancienne, plus proche du génome ancestral des Ehrhartoideae (n=12) ?

II. Histoire évolutive des Spartines polyploïdes

Les Spartines (famille des Poacées, sous famille des Chloridoideae) sont des plantes vivaces tétraploïdes à dodécaploïdes (nombre de base x=10) colonisant les marais salés littoraux du Nouveau Monde (centre de diversité contenant 13 espèces d'après Mobberley, 1956) et de l'Ancien Monde (une seule espèce avant le 19ème siècle), où elles jouent un rôle écologique important : tolérantes au sel, elles occupent une position variable, selon les espèces, le long de l'estran où elles contribuent à la dynamique sédimentaire du marais (Figure 5). En fixant les sédiments, les espèces du bas shore augmentent les processus d'atterrissement, permettant l'installation de végétaux moins tolérants à l'immersion dans l'eau salée. Elles ont pour cette raison souvent été utilisées et délibérément introduites hors de leur aire d'origine pour stabiliser les berges ou créer des polders. Les introductions délibérées ou accidentelles hors de leur aire d'origine ont favorisé les hybridations

interspécifiques (avec les espèces locales), suivies ou non de polyploïdisation (allopolyploïdie) ayant conduit à la formation de nouveaux taxa souvent envahissants (Ainouche *et al*., 2009).

Les Chloridoideae ont divergé de la sous-famille des Panicoideae, il y a entre 34,6 et 38,5 millions d'années (Kim *et al*., 2009 ; Christin *et al*., 2008). Au sein des Chloridoideae, le genre *Spartina* (appartenant au clade des Zoysieae) a divergé des *Eleusine* (Eleusininae) il y a environ 22 millions d'années (Liu *et al*., 2011). Les phylogénies moléculaires des Spartines basées sur les séquences nucléaires *Waxy* et *ITS* et sur des séquences chloroplastiques, séparent le genre en deux clades (Baumel *et al*., 2002 ; Fortuné *et al.*, 2007). Le premier clade est formé des espèces hexaploïdes. Celui-ci aurait divergé du clade des espèces tétraploïdes (regroupant des espèces d'origine américaines) il y a 6 millions d'années (Bellot, 2010). Il contient aussi *Spartina pectinata*, une espèce tétraploïde ayant fait l'objet d'études transcriptomiques pour ses capacités à servir de biocarburant par la biomasse qu'elle produit (Gedye *et al*., 2010). Cette espèce montre des niveaux de ploïdie variable, de tétraploïde à octoploïde (Kim *et al*., 2012). De nombreux événements d'hybridations interspécifiques et de polyploïdisation ont jalonné l'histoire de ce genre (Figure 5, Ainouche *et al.*, 2012). L'hybridation entre une espèce tétraploïde (*S. arundinacea*) et une espèce hexaploïdes (*S. alterniflora*) a donné naissance à l'espèce 7x *S. densiflora* (Ainouche *et al*., 2009). Le clade hexaploïde est formé des espèces *S. maritima*, distribuée sur les côtes atlantiques européennes et africaines et de deux espèces américaines génétiquement peu divergentes, *S. alterniflora* (native des côtes Atlantiques) et *S. foliosa* (endémique des côtes Pacifiques de Californie et du Mexique). *Spartina maritima* et *Spartina alterniflora* ont divergé récemment il y a 3 millions d'années (Bellot, 2010).

Spartina alterniflora Loisel (2n=6x=62) occupe une large distribution, du Canada à l'Argentine. Elle a été introduite en Chine où elle a rapidement envahi de nombreux estuaires (An *et al.,* 2007). Elle a été également introduite sur les côtes Pacifiques américaines en Californie (baie de San Francisco) où elle s'est hybridée avec l'espèce native endémique *S. foliosa* (2n=6x=60). Les hybrides, en se rétro-croisant préférentiellement avec l'espèce introduite, forment de nouveaux génotypes à meilleure valeur adaptative (Ayres *et al.*, 2008) menaçant l'espèce endémique. Depuis plusieurs années un programme

d'éradication et de monitoring des Spartines envahissantes a été mis en place dans la baie de San-Francisco (www.spartina.org).

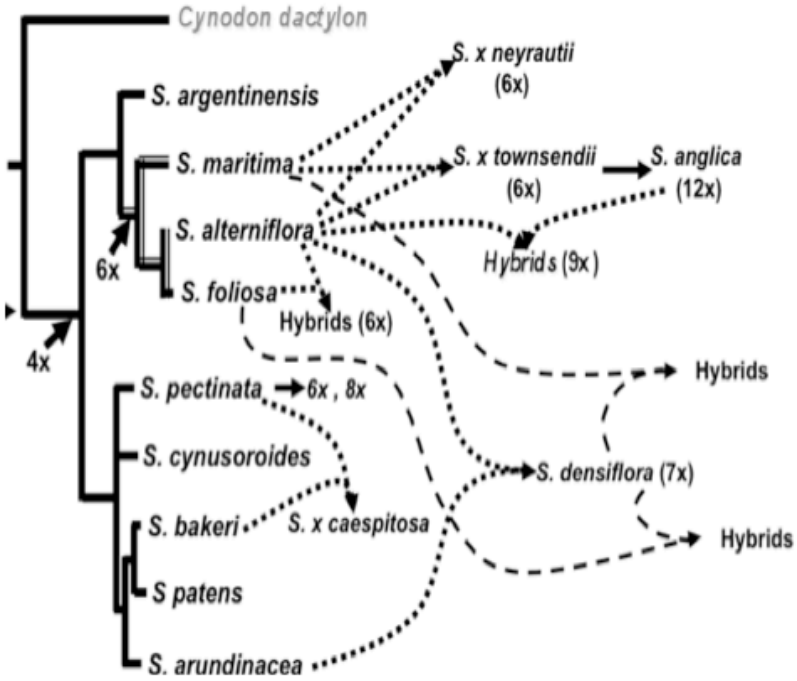

Figure 5 : Relations phylogénétiques au sein des Spartines, évènements récents d'hybridation et de polyploïdie (D'après Ainouche *et al.* 2012).

Spartina alterniflora a été introduite accidentellement en Europe à la fin du 19ième siècle, dans le sud de l'Angleterre et sur les côtes atlantiques françaises où elle s'est hybridée avec l'espèce native européenne *S. maritima* Fernald (2n=6x=60). La première hybridation a été découverte en 1870 dans l'estuaire de Southampton et a donné l'hybride F1 stérile *S.* x *townsendii* (Groves & Groves, 1880). Au pays basque, dans l'estuaire de la Bidassoa, un autre hybride stérile a été découvert en 1892 (Foucaud, 1987) et appelé *S.* x *neyrautii* (Jovet, 1941). La population de *S.* x *neyrautii* est aujourd'hui extrêmement réduite suite à l'urbanisation de l'aire d'hybridation (Baumel *et al*., 2003). A l'inverse, les populations anglaises de *S*. x *townsendii* sont vigoureuses sur les côtes anglaises (Renny-Byfield *et al*., 2010).

A la fin du 19ième siècle, une nouvelle espèce fertile et particulièrement vigoureuse a été décrite en Angleterre résultant du doublement chromosomique (2n=120, 122, 124) de *S.* x *townsendii* (Marchant, 1963). Cette nouvelle espèce allopolyploïde a été nommée *S. anglica* (Hubbard, 1968). Cette espèce a rapidement colonisé les côtes anglaises et françaises grâce à sa large amplitude écologique le long de l'estran, sa capacité à tolérer plusieurs heures d'immersion et ses différents modes de propagation (dispersion des graines et fragmentation des rhizomes) par les bateaux (ballasts), les courants marins ou les oiseaux. Suite à sa propagation naturelle et artificielle (introductions délibérées) *Spartina anglica* est aujourd'hui distribuée sur plusieurs continents (Amérique du Nord, Nouvelle Zélande, Australie, Chine). Elle fait l'objet de tentatives variées et infructueuses d'éradication (*e.g.* Cottet *et al.* 2007) et elle est listée parmi les 100 espèces les plus envahissantes au niveau mondial depuis 2000 (Union Internationale pour la Conservation de la Nature, IUCN). Inversement, certains traits physiologiques rendent cette espèce intéressante, comme les fortes capacités de dépollution et d'oxygénation du sol dans les perspectives de phytoremédiation des sédiments pollués (Lee, 2003).

III. Conséquences génétiques et génomiques de l'hybridation et de la duplication du génome chez les Spartines

D'un point de vue fondamental, les hybridations récentes entre les espèces hexaploïdes *S. maritima* et *S. alterniflora* en Europe offrent une opportunité assez rare d'examiner en conditions naturelles les mécanismes associés à la mise en place d'une nouvelle espèce allopolyploïde (Ainouche *et al*., 2004a) et de distinguer les effets de l'hybridation interspécifique (chez les deux hybrides F1 formés indépendamment *S.* x *townsendii* et *S.* x *neyrautii*) de la duplication du génome (chez *S. anglica*).

Au sein du clade hexaploïde, *S. maritima* et *S. alterniflora* apparaissent nettement divergentes des points de vue anatomique, écologique et génétique (ancêtre commun datant de 3 millions d'années ; Bellot, 2010), ce qui explique la stérilité des hybrides F1 formés entre ces deux espèces. De plus, ces deux espèces présentent un polymorphisme relativement important au niveau des régions non-codantes ou des introns (Baumel *et al*., 2002a ; Fortuné *et al*., 2007), et entre 94% et 99% d'identité nucléotidique dans les zones

codantes, selon les gènes examinés à ce jour (Chelaifa *et al*., 2010a). A chaque locus, jusqu'à six copies sont attendues. L'étude de l'évolution moléculaire du gène *Waxy* a permis de distinguer les copies paralogues, homéologues et orthologues : une à trois copies ont été détectées chez les espèces hexaploïdes suggérant ainsi l'origine allopolyploïde de ces espèces polyploïdes (Fortuné *et al*., 2007).

Les populations des espèces parentales dans l'Ouest de l'Europe (*S. alterniflora* et *S. maritima*) présentent une très faible variation génétique (Yannic *et al*., 2004). Ces éléments appuient l'idée d'un très fort goulot d'étranglement génétique à l'origine de *S.* x *townsendii* et de *S. anglica*, qui résulterait d'un événement unique d'hybridation, ou de multiples événements impliquant des génotypes parentaux très similaires (Ainouche *et al.,* 2004a). Les espèces parentales hexaploïdes (*S. maritima* et *S. alterniflora*) tout comme l'allo-dodécaploïde *S. anglica* montrent une méiose régulière avec comportement disomique des chromosomes (Marchant, 1968). Toutefois, sur le site initial de l'apparition de l'hybride anglais de nouveaux hybrides possédant différents niveaux de ploïdie ont été détectés et résulteraient du back-cross de l'allopolyploïde *S. anglica* avec son parent maternel *S. alterniflora* (Renny-Byfield *et al*., 2010). L'hybride *S.* x *neyrautii* possède le même génome chloroplastique que *S.* x *townsendii* et *S. alterniflora* (Ferris *et al*., 1997, Baumel *et al*., 2001 ; Baumel *et al*., 2003). Les deux hybrides F1 résultent donc du même croisement. Toutefois, deux génotypes différents de *S. maritima* semblent être impliqués dans les deux hybridations (Baumel *et al*., 2003).

Les populations de *S. anglica* montrent une diversité génétique inter-individuelle extrêmement réduite dans l'Ouest de l'Europe ainsi que dans les régions plus récemment colonisées comme l'Australie (Baumel *et al*., 2001). Toutes les populations de *S. anglica* montrent un génotype nucléaire multilocus quasi-identique à celui de l'hybride F1 *S.* x *townsendii*, et un génome chloroplastique identique à celui de *S.* x *townsendii* et du parent maternel *S. alterniflora* (Baumel *et al.,* 2001 ; Ainouche *et al.,* 2004b). Le génome de *S. anglica* correspond à l'addition des deux génomes parentaux (Baumel *et al*., 2001 ; 2002b ; Ainouche *et al*., 2004b). Peu de remaniements sont observés contrairement à ce qui a pu être observé chez des allopolyploïdes re-synthétisés artificiellement chez les blés (*e.g.* Ozkan *et al*., 2001), ou les Brassicacées (Song *et al*., 1995 ; Lukens *et al*., 2003). Toutefois, certains changements génétiques sont observés chez *S.* x *townsendii* et *S. anglica* : des fragments AFLP sont perdus affectant préférentiellement le génome originaire de *S. alterniflora*

(Salmon *et al.*, 2005; Ainouche *et al*., 2009). Les changements les plus importants notés chez *S. anglica* à ce jour concernent les altérations de la méthylation du génome, qui sont initiés suite à l'hybridation et qui sont transmis à l'allopolyploïde (Salmon *et al.,* 2005). Ces altérations de la méthylation s'avèrent particulièrement importantes dans les zones voisines d'éléments transposables (Parisod *et al*., 2009). Ainsi, suite à l'hybridation, la dynamique du génome est principalement affectée par des changements épigénétiques. Afin d'étudier les conséquences transcriptomiques, de récents travaux basés sur des puces à ADN ont montré que les espèces parentales (*S. alterniflora* et *S. maritima*) pourtant faiblement divergentes au niveau nucléotidique, diffèrent de façon importante au niveau de l'expression des gènes (Chelaifa *et al*., 2010a). De plus, les conséquences de l'hybridation et de la duplication du génome ont été examinées chez deux hybrides F1 naturels (tous deux formés à partir des mêmes parents dans deux sites distincts) et chez l'allopolyploïde *S. anglica* (Chelaifa *et al.,* 2010b). Cette étude a mis en évidence les effets importants mais différents de l'hybridation et de la duplication du génome sur le transcriptome, avec une dominance de l'expression maternelle (de *S. alterniflora*) chez les hybrides F1. Cette dominance s'atténue ensuite chez l'allopolyploïde qui se distingue par une surexpression de la majorité des gènes différentiellement exprimés. Il est donc aujourd'hui indispensable de développer les ressources génomiques et transcriptomiques afin d'éclairer l'histoire évolutive du génome des Spartines.

Chapitre 3

Matériel et méthodes

Les données génomiques utilisées au cours de ce travail sont obtenues dans le cadre d'un programme de recherche (projet Genomics of Spartina « GENOSPART ») faisant intervenir le Génoscope (Evry), le Centre National des Ressources Génomiques Végétales (CNRGV, Toulouse) et la plateforme de Génomique Environnementale de l'Observatoire des Sciences de Rennes (OSUR). L'espèce native *Spartina maritima* a été choisie comme espèce modèle pour le développement des ressources génomiques avec la réalisation d'une banque BAC, d'un séquençage d'extrémités de BAC et du pyroséquençage d'ADN génomique. Les analyses de transcriptome, qui renforceront les bases de données géniques et permettront la découverte de nouveaux gènes d'intérêts, sont effectuées chez les cinq espèces de Spartines constituant notre système d'analyse de la spéciation allopolyploïde, selon le schéma présenté en Figure 6.

La plateforme de séquençage à haut débit choisie est le séquenceur 454 de chez Life Sciences en raison de la longueur des séquences générées au moment du lancement de ces projets (2008-2009). C'est ce paramètre, important à considérer pour une espèce dont le génome n'est pas encore annoté, qui nous a conduit à privilégier cette technique par rapport à d'autres plateformes (par exemple Illumina ou SOLiD) qui offrent une plus grande profondeur de séquençage, mais des fragments plus courts donc plus difficiles à assembler et à annoter (Metzker, 2010). Les coûts et technologies évoluant rapidement, les analyses de transcriptome pourront dans une seconde étape, être approfondies par de telles approches. La plateforme 454 est en effet, de plus en plus employée dans l'analyse des génomes complexes de plantes (*e.g.* Velasco *et al.*, 2007), la caractérisation d'éléments répétés (Wicker *et al*., 2006 ; Swaminathan *et al*., 2007 ; Macas *et al*., 2007) l'annotation des gènes et la détection de SNPs (Deschamps & Campbell, 2009 ; Buggs *et al.*, 2010).

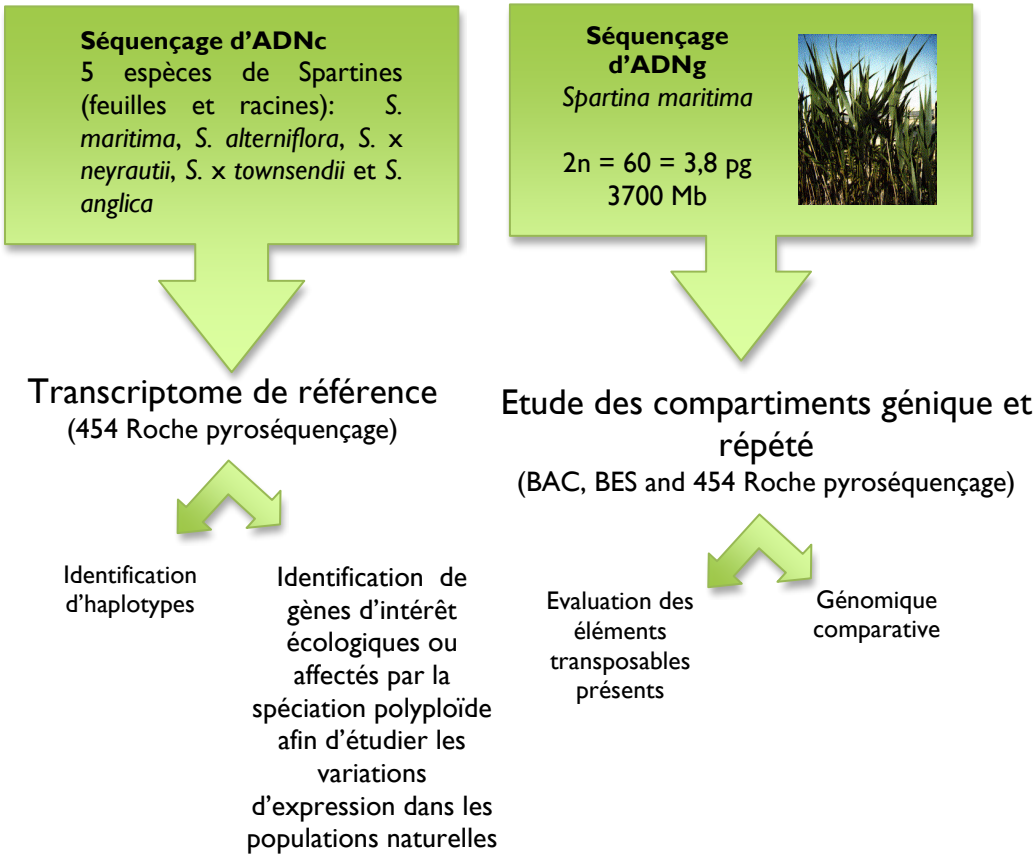

Figure 6 : Etapes méthodologiques du travail de thèse.

I. Matériel biologique

Les analyses sont centrées sur le clade des espèces hexaploïdes du genre *Spartina* formé de l'espèce native européenne *S. maritima* et de l'espèce américaine *S. alterniflora*, des deux hybrides F1 *S.* x *neyrautii* et *S.* x *townsendii* et de l'espèce néo-allopolyploïde *S. anglica* (Tableau 3).

L'espèce native européenne *Spartina maritima* a été échantillonnée sur le site de Quenouille (Morbihan) en 2009. Les individus ont été transplantés dans la serre expérimentale du Campus de Beaulieu de l'Université de Rennes 1. L'ADN des feuilles a ensuite été extrait pour le séquençage et la construction d'une banque BAC.

Concernant les analyses de transcriptomes et d'expression dans les populations naturelles, les individus ont été récoltés sur les mêmes sites suivant deux protocoles différents. Afin de réaliser le pyroséquençage d'ADNc (ADN complémentaire), des plantes entières ont été prélevées sur le terrain avec une motte de sol autour des racines, puis transplantées en serre expérimentale (Campus de Beaulieu, Université de Rennes 1) et maintenues dans les mêmes conditions. Les plantes sont transplantées dans des pots de 3 à 7 litres, contenant un mélange de 1/3 sable, 1/3 terre et de 1/3 terreau. L'arrosage (automatique) est quotidien et dure environ 6mn le matin. Le cycle d'éclairage est lié à la photopériode naturelle. Les feuilles ont ensuite été récoltées et stockées à -80°C jusqu'à l'extraction d'ARN. Dans l'optique des analyses de la variation d'expression dans les populations naturelles (Chapitre 5), différents transects ont été effectués sur les différents sites, selon un gradient d'immersion qui va du bas vers le haut de l'estran. Sur chacun de ces transects, les feuilles de 6 à 8 plantes de différents clones ont été récoltées et conservées dans de l'ARN later (permettant la conservation de l'intégrité des ARNs) jusqu'à l'extraction d'ARN.

Spartina alterniflora a été échantillonnée à Landerneau (Finistère) et *S. maritima* à Noirmoutier (Vendée) et sur le site de Quenouille (Morbihan) en Juillet 2010. *Spartina anglica* a été échantillonnée à Roscoff et dans l'Anse de Goulven (Finistère) en Août 2007. Les hybrides F1 ont été échantillonnés près des sites d'hybridation en France et en

Angleterre, *S.* x *neyrautii* et de *S.* x *townsendii* ont été respectivement récoltés à Hendaye (Pyrénées-Atlantiques) et à Hythe (Hampshire, Angleterre) en 2009.

Tableau 3 : Espèces, populations et analyses effectuées sur les Spartines étudiées.

Espèces	Populations	Niveaux de ploïdie	Analyses effectuées
S. maritima	Quenouille (Morbihan)	6x	- Pyroséquençage d'ADN génomique - Réalisation d'une banque BAC - Analyses transcriptomiques
	Noirmoutier (Vendée)	6x	- Analyses transcriptomiques (banque normalisée) - Expression dans les populations naturelles
S. alterniflora	Landerneau (Finistère)	6x	- Analyses transcriptomiques - Expression dans les populations naturelles
S. x *neyrautii*	Hendaye (Pyrénées-Atlantiques)	6x	- Analyses transcriptomiques - Expression dans les populations naturelles
S. x *townsendii*	Hythe (Angleterre)	6x	- Analyses transcriptomiques - Expression dans les populations naturelles
S. anglica	Roscoff (Côtes d'Armor)	12x	- Analyses transcriptomiques - Expression dans les populations naturelles
	Anse de Goulven (Finistère)	12x	- Expression dans les populations naturelles

II. Obtention de banques d'ADNc et séquençage haut-débit

Nous présenterons successivement les méthodes d'extraction d'ARN, de synthèse d'ADN complémentaire (ADNc) et de leur normalisation réalisées chez les parents, hybrides et allopolyploïde maintenus en conditions contrôlées, puis le séquençage haut-débit des banques d'ADNc non-normalisés et normalisés.

1. Extraction des ARN totaux

L'extraction des ARNs totaux a été faite selon un protocole préalablement mis au point au sein de l'équipe (Chelaifa *et al.*, 2010a et b). Environ 1g de matière fraîche ou conservée dans de l'ARN later est rincé plusieurs fois à l'eau du robinet puis à l'eau distillée avant d'être minutieusement broyée dans de l'azote liquide au moyen d'un mortier et d'un pilon. La première phase consiste en la séparation des ARN et des ADN à l'aide de TriReagent (Sigma Aldrich Inc.). La phase contenant les ARN est ensuite précipitée à l'aide d'isopropanol (Sigma Aldrich Inc.) et suivie par un lavage avec de l'éthanol à 75%. Un deuxième cycle de précipitation et de lavage à l'aide d'éthanol 100% et d'acétate de sodium est nécessaire pour maximiser le rendement et la qualité des ARN extraits. Les ARNs de racines étant plus difficiles à extraire, un troisième cycle de précipitation a été nécessaire. Les culots d'ARN obtenus sont resuspendus dans 50 µL d'eau ultrapure stérile puis dosés à l'aide d'un spectrophotomètre Nanodrop ND-1000 (Nanodrop Technologies, Inc.). Leur qualité est déterminée en utilisant le kit RNA 6000 Nano LabChip et le Bioanalyzer 2100 (Agilent Technologies). Les ARN extraits ont été aliquotés et stockés à -80°C.

2. Synthèse d'ADNc et normalisation

Afin de préparer les échantillons d'ADNc en vue du séquençage, les extraits d'ARN les plus concentrés et de qualité élevée (sans contamination d'ADN) sont rétrotranscrits en ADNc à l'aide du kit SMARTer cDNA synthesis (Clontech) suivant les instructions du fabricant. Brièvement, 1g d'ARN totaux est nécessaire à la procédure. Le premier brin d'ADNc est synthétisé en utilisant des amorces oligos(dT) modifiées du kit Clontech (3'SMART CDS Primer II A). Quand l'enzyme Reverse Transcriptase SMARTScribe atteint l'extrémité 5' de l'ARNm, celle-ci ajoute quelques nucléotides à l'extrémité 3' du brin d'ADNc. Après la synthèse du second brin d'ADNc, les fragments double-brin sont amplifiés (21 cycles en

utilisant les amorces 5' PCR Primer II A). Ce protocole permet de générer 2 à 6 µg d'ADNc qui sont ensuite purifiés en utilisant le kit QIAquick PCR purification (Qiagen). Un mélange équimolaire d'échantillons a été constitué pour chaque organe et chaque espèce afin d'atteindre une quantité de 10 µg totaux d'ADNc qui sont stockés à -20°C jusqu'au séquençage.

L'étape de normalisation a été effectuée sur les feuilles et racines de l'espèce *Spartina maritima* afin de séquencer les transcrits les moins abondants et ainsi identifier un nombre de gènes plus important. Un µg d'ADNc de chaque tissu (feuilles et racines) a été normalisé séparément grâce au kit TRIMMER de chez Evrogen en suivant les instructions du fabricant. Pour cela, 4 µl d'une solution 4X Hybridization buffer ont été ajoutés aux échantillons d'ADNc puis l'ensemble a été dénaturé à 95°C pendant 5 min puis hybridé à 68°C pendant 5 heures. Les réactifs suivants, préalablement réchauffés à 68°C ont été ajoutés à la préparation : 3,5 µl d'eau milliQ, 1 µl de 5X DNAse buffer, 1 µl d'enzyme double-strand nuclease (DSN) puis le tout est incubé à 68°C pendant 25 min. L'enzyme DSN a été inactivée en ajoutant 10 µl de solution DSN stop et en chauffant le tout à 68°C pendant 5 min. Les échantillons d'ADNc normalisés ont ensuite été dilués en ajoutant 40,5 µl d'eau milliQ et soumis à 2 réactions d'amplification PCR. La première PCR (50 µl) inclus 1 µl d'ADNc dilué, 5 µl 10X Advantage 2 PCR buffer (Clontech), 1 µl 50X dNTPs mix, 1,5 µl PCR primer M1 10 mM (Evrogen), 1 µl 50X Advantage 2 Polymerase mix (Clontech). L'amplification est réalisée dans les conditions suivantes : dénaturation initiale à 95°C pendant 1 min, suivie de 18 cycles (95°C pendant 15 sec, 66°C pendant 20 sec et 72 °C pendant 3 min). La deuxième réaction d'amplification a été réalisée en ajoutant aux 2 µl d'ADNc dilué et normalisé, 1 µl de 10X Advantage 2 PCR Buffer (Clontech), 2 µl de 50X dNTP mix, 4 µl de PCR Primer M2 10 mM (Evrogen), 2 µl de 50X Advantage 2 Polymerase mix (Clontech) (dans 100 µl) et a été amplifiée en suivant le protocole suivant : une étape de dénaturation initiale à 95°C pendant 1 min, puis 12 cycles (95°C pendant 15 sec, 64°C pendant 20 sec, 72 °C pendant 3 min, et une extension finale en deux étapes (64°C pendant 15 sec et 72°C pendant 3 min). La qualité des ADNc normalisés et non-normalisés a été vérifiée sur un Bioanalyzer 2100 (Agilent Technologies) à l'aide des DNA 7500 chip. Les ADNc sont alors quantifiés sur un Nanodrop ND-1000 (Nanodrop Technologies, Inc.) et stockés à -20°C jusqu'au séquençage.

3. Séquençage haut débit des banques d'ADNc

Les banques d'ADNc non-normalisés des deux organes des cinq espèces de Spartines ont été nébulisées et séquencées à la plateforme du Génoscope (Evry). Pour chaque banque, 500 ng d'ADNc totaux sont nécessaires afin de remplir une demi-plaque de séquenceur 454 GS XLR70 Titanium Genomic Sequencer (Roche Inc.). Pour chaque espèce, les banques d'ADNc de feuilles et de racines sont physiquement distinctes sur une plaque de séquençage.

Le séquençage des banques normalisées d'ADNc de feuilles et de racines pour *S. maritima* a été réalisé à la plateforme de Génomique Environnementale et Fonctionnelle de l'OSUR. Un total de 500 ng d'ADNc est nécessaire pour chaque banque, les ADNc de feuilles et de racines ont été séquencés séparément sur deux demi-plaques du séquenceur 454 GS XLR70 Titanium Genomic Sequencer (Roche Inc.).

III. Analyses bioinformatiques du transcriptome

Les analyses bioinformatiques ont porté sur les séquences d'ADNc de racines et de feuilles des cinq espèces du complexe. Les assemblages et annotations des transcriptome ont été réalisés sur les cinq espèces et sont explicités dans la suite de cette partie et sur la Figure 7. L'analyse porte plus précisément sur les analyses des deux espèces hexaploïdes : *S. alterniflora* et *S. maritima* (banques d'ADNc de feuilles et de racines, normalisées et non-normalisées) et a fait l'objet d'un article (Ferreira de Carvalho *et al.*, 2013) publié dans la revue *Heredity*.

1. Nettoyage des séquences et assemblages

Les séquences de faible qualité ainsi que les adaptateurs ont été supprimés pendant le traitement des données sur la plateforme 454. Le logiciel GS Assembler 2.3 (Roche, Inc.) a été utilisé pour assembler les lectures en contigs. Plusieurs assemblages ont été réalisés pour chaque banque prise séparément ou combinée par espèces, par tissu et par type de normalisation. Finalement, deux assemblages globaux ont été réalisés : un pour les espèces hexaploïdes ; l'autre pour toutes les espèces du complexe.

Les Spartines étudiées sont toutes polyploïdes (hexaploïdes à dodécaploïdes), ainsi de 6 à 12 allèles (3 à 6 paires d'homéologues) transcrits peuvent être attendus sur chaque locus. La stratégie d'assemblage suivie vise à assembler les séquences homologues (à la fois les orthologues et les homéologues) en contigs en utilisant une stringence faible. Dans cette perspective, différents pourcentages d'identité ont été testés (90, 95, 96, et 97%) afin de choisir le seuil le plus adapté. Le plus bas pourcentage (90% d'identité sur au moins 100pb) a été choisi car il maximise l'assemblage de toutes les séquences potentiellement homologues et homéologues, néanmoins nous ne pouvons pas écarter la possibilité d'assembler en même temps des paralogues peu divergents. Les autres paramètres utilisés sont ceux spécifiques au traitement de données transcriptomiques par défaut dans la version 2.3 du logiciel GS Assembler (Roche, Inc.).

Les informations utiles, telles que le nombre de séquences utilisées dans l'assemblage, le nombre de contigs et de singletons, ainsi que leur longueur moyenne et leur profondeur ont été extraites des fichiers de sortie .ace, .txt et .fasta, à l'aide de scripts Python développés au sein du laboratoire.

2. Recherche de similitudes et annotations fonctionnelles

BLASTn et tBLASTx contre des bases de données d'EST de Poaceae

Les alignements BLASTn et tBLASTx (Altschul *et al.*, 1990) ont été réalisés sur les contigs assemblés contre deux bases de données nucléotidiques : une base de données d'ESTs d'*Oryza sativa* (rapdb.dna.affrc.go.jp) et une base de données d'ESTs de Poaceae incluant des séquences issues des espèces de Poacées *Oryza sativa*, *Zea mays*, *Brachypodium distachyon* et *Sorghum bicolor* (www.gramene.org). Toutes les analyses par BLAST ont été réalisées avec une e-value de 10^{-5} et le Best BLAST Hit (BBH) pour chaque contig a été examiné pour l'annotation fonctionnelle et les ontologies.

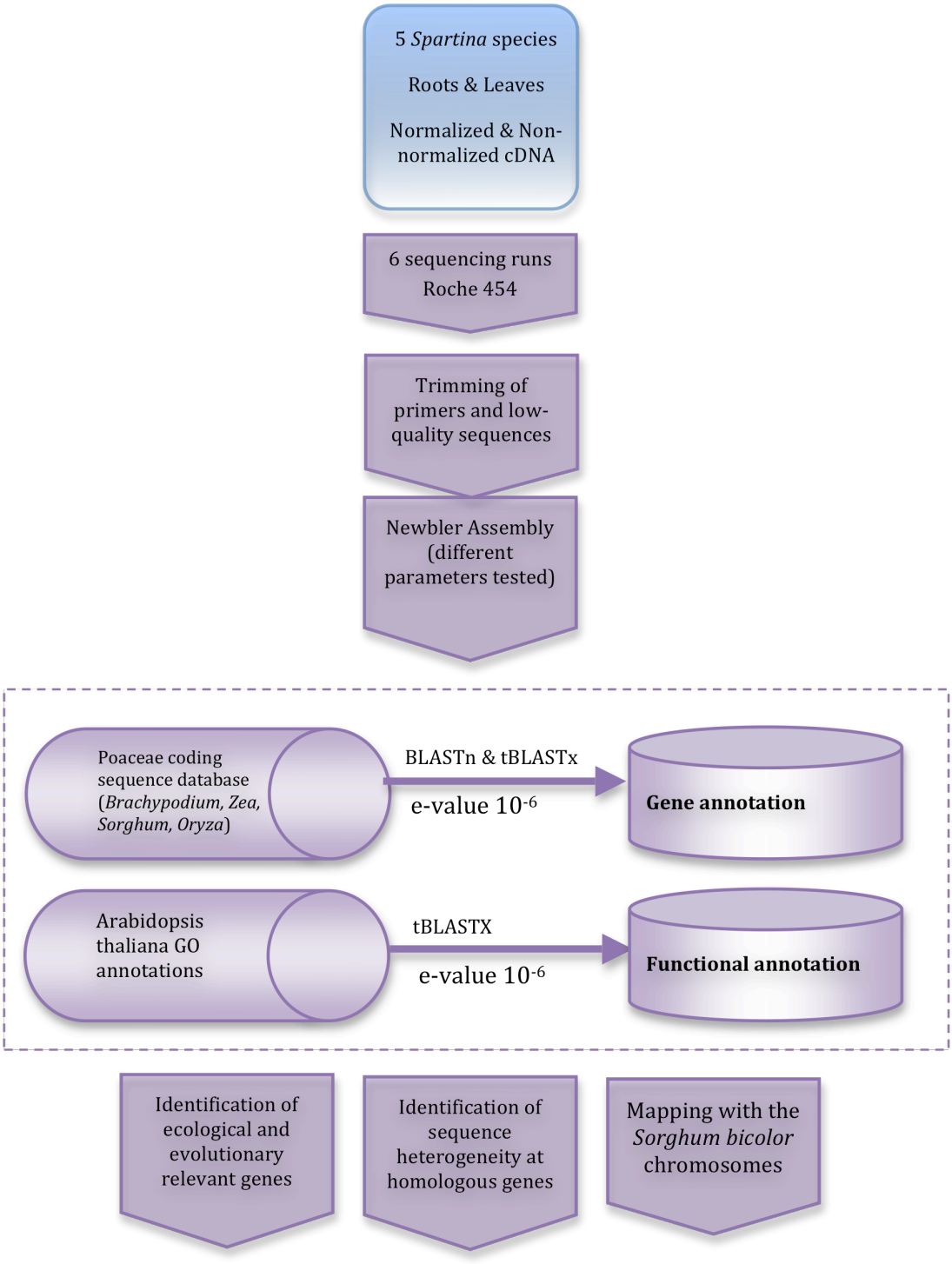

Figure 7 : Démarche méthodologique suivie lors des analyses transcriptomiques.

Annotations fonctionnelles

Les annotations issues de Gene Ontology (GO) ont été effectuées à l'aide du logiciel BLAST2GO (Conesa *et al.*, 2005; Götz *et al.*, 2008) en utilisant des analyses tBLASTx (e-value 10^{-6} et valeur de similitude maximum de 55) entre les contigs des Spartines hexaploïdes et la base de données protéiques d'*Arabidopsis thaliana* (TAIR, www.arabidopsis.org). Les annotations ont ensuite été examinées afin d'identifier des gènes potentiellement intéressants d'un point de vue de l'écologie des Spartines (*e.g.* gènes de réponse au stress salin, au stress oxydant, gènes de tolérance aux métaux lourds ou gènes de la croissance cellulaire). De plus, les gènes précédemment montrés comme différentiellement exprimés entre les espèces du complexe suite à l'hybridation et la polyploïdie à l'aide de puce hétérologues de riz ont été plus précisément analysés (Chelaifa *et al.*, 2010a, b). Les numéros d'accessions des amorces de riz utilisées sur la puce Agilent (44 K Agilent G2519F) ont été utilisés pour retrouver les séquences homologues dans le transcriptome de référence des Spartines hexaploïdes (BLASTn, e-value 10^{-5}).

Alignement sur le génome du sorgho

Les contigs issus des assemblages des espèces *S. maritima* et *S. alterniflora* ont été alignés sur le génome du sorgho, l'espèce séquencée et annotée la plus proche de notre modèle d'étude (Paterson *et al.*, 2009). Afin de comparer la distribution et la densité des gènes homologues entre les Spartines et le sorgho, les annotations de *S. bicolor* ont été téléchargées sur le site http://genome.jgi-psf.org/Sorbi1/Sbicolor_79_gene.gff3. Le nombre de gènes homologues de Spartines a été estimé par fenêtre de 100kb en les alignant (BLASTn, e-value 10^{-5}, BBH) sur le génome du sorgho. Les résultats sont visualisés à l'aide du logiciel CIRCOS V.0.55 (Krzywinski *et al.*, 2009). Enfin, dans le but d'évaluer la représentativité des contigs de Spartines sur le génome du sorgho, des corrélations de Pearson et des droites de régressions ont été calculées entre le nombre de gènes par fenêtre de 100kb chez le sorgho et les homologues putatifs assemblés et annotés chez les Spartines. Les analyses statistiques ont été effectuées pour le génome entier et pour chacun des 10 chromosomes du sorgho à l'aide du logiciel R (R Development Core Team, 2005).

3. Recherche de polymorphismes nucléotidiques

Une première évaluation de l'hétérogénéité de séquences et du nombre de copies dans chaque contig a été possible sur un nombre restreint de contigs présentant une couverture et une profondeur de séquences importantes et présents à la fois dans les jeux de données de *S. maritima* et de *S. alterniflora* afin d'étudier l'hétérogénéité de séquences. Pour cela, nous avons regardé plus spécifiquement les polymorphismes (SNPs : Single Nucleotide Polymorphisms) partagés entre les séquences utilisées pour assembler un contig donné pour chaque espèce, à l'aide du programme Ace.py du package Biopython (http://biopython.org/) et de programmes développés au sein du laboratoire (J. Boutte et A. Salmon). Les SNPs non-partagés entre les séquences et ceux localisés dans des régions homopolymériques sont écartés de l'étude. Les séquences sont ensuite assemblées entre elles avec 100% de similarité et en utilisant au moins un SNP partagé. La séquence consensus alors assemblée est considérée comme un haplotype pouvant représenter un allèle ou une copie potentiellement homéologue (Boutte *et al*., *en préparation*).

IV. Analyses d'expression dans les populations naturelles

1. Extraction ARN et Synthèse d'ADNc

Les ARNs ont été extraits selon un protocole mis au point au sein de l'équipe (Chelaifa *et al.*, 2010a et b) à partir de feuilles échantillonnées *in situ* et préservées dans de l'ARN later. Environ 1g de matière fraîche est broyé dans de l'azote liquide. La première phase consiste à séparer l'ARN de l'ADN à l'aide de TriReagent composé de 50% de Phénol et 30% de Guanidium thiocyanate (Sigma T9424) et de chloroforme 99% (C2432). La phase contenant les ARNs est ensuite précipitée à l'aide d'isopropanol 99% (Sigma I9516) et suivie par un lavage avec de l'éthanol à 75%. Un deuxième cycle de précipitation et de lavage à l'aide d'éthanol 100% et d'acétate de sodium est nécessaire pour maximiser le rendement et la qualité des ARN extraits. Les culots d'ARN obtenus sont resuspendus dans 50 µL d'eau ultrapure stérile puis dosés à l'aide d'un spectrophotomètre Nanodrop ND-1000 (Nanodrop Technologies, Inc.). La qualité des ARNs extraits est ensuite vérifiée sur gel d'agarose. Les ARN extraits ont été aliquotés et stockés à -80°C. La synthèse d'ADN complémentaire a été

réalisée à partir de 50 ng d'ARN totaux en suivant les instructions du fournisseur du kit InvitroGen (ThermoScript RT-PCR System). La synthèse du premier brin complémentaire est initiée en utilisant des amorces oligos(dT) et poursuivie en utilisant la Taq polymérase. La synthèse du second brin d'ADNc est amorcée par les amorces spécifiques des gènes d'intérêts (Tableau 1). La qualité des ADNc a été vérifiée par électrophorèse sur gel d'agarose. Les échantillons d'ADNc sont ensuite stockés à -20°C.

2. Identification des gènes d'intérêt et dessin des amorces

Grâce à la construction du premier transcriptome de référence pour les espèces hexaploïdes de Spartines (Ferreira de Carvalho *et al.*, 2013), des fonctions et des gènes d'intérêts ont pu être identifiés. Nous avons pu alors sélectionner des gènes correspondant à des fonctions jouant un rôle dans les réponses aux stress salin, oxydant, et de tolérance aux métaux lourds, ainsi que des gènes entrant dans le métabolisme de la lignine et de la cellulose. Enfin, à partir des listes de gènes différentiellement exprimés entre les parents hexaploïdes, les hybrides F1 et l'allododecaploïde mis en évidence par Chelaifa *et al.* (2010 a et b) et des séquences du transcriptome de référence, nous avons pu retrouver les séquences homologues aux sondes des puces de riz pour identifier et quantifier ces gènes dans les populations naturelles de Spartines. Un ensemble de 47 contigs présentant des fonctions intéressantes d'un point de vue écologique mais aussi évolutif a été utilisé pour le dessin des amorces à l'aide du logiciel Primer3 (Rozen & Skaletsky, 2000). Nous avons veillé à dessiner prioritairement les amorces sur des gènes en simple ou faible nombre de copies chez *Sorghum bicolor*, afin d'éviter le risque de co-amplification de familles multigéniques (BLASTn sur les chromosomes du sorgho). Les limites des exons ont été définies par alignement (MUSCLE : Edgar, 2004) sur le modèle séquencé le plus proche phylogénétiquement, *Sorghum bicolor*, afin de s'assurer que les amorces étaient sur un même exon et en région 3'-UTR du gène.

Les principales conditions liées à la technologie d'amplification par PCR quantitative ont été paramétrées lors du dessin des amorces : courte taille attendue des fragments amplifiés (150-180pb), faible teneur en GC (50%) et une température d'hybridation des paires d'amorces proche de 55°C (Tm à 60°C). Les amorces ne comportent pas plus de 4

nucléotides identiques consécutifs et pas plus de 2G ou 2C en extrémité 3'. Les amorces ont ensuite été validées par PCR classique puis par PCR quantitative afin de s'assurer qu'un seul fragment était amplifié.

Validant tous les critères, 13 gènes et 3 gènes de ménage (Sucrose synthase, alpha-tubulin et Glycéraldéhyde 3-Phosphate Dehydrogenase) ont été choisis (Tableau 4). Les annotations ont été vérifiées grâce à la base de données agBASE et leur outil de recherche d'homologie GoAnna en Août 2012 (McCarthy *et al.*, 2011) et sont présentées dans le Tableau 5.

3. PCR quantitative

Tous les échantillons d'ADNc ont été dilués à 0,5ng/µL avec de l'eau ultrapure et aliquotés dans une plaque 96 puits. A chaque échantillon dilué d'ADNc, 10 µL de mix SYBR Green 2X (BioRad) ont été ajoutés, 0,6 µL d'amorces spécifiques Forward et Reverse (5µM) pour chaque gène ainsi que de l'eau ultrapure stérile pour un volume réactionnel de 20 µL. Chaque plaque est préparée avec les 37 échantillons d'ADNc (trois réplicats techniques pour chaque échantillon) et une gamme de dilution de 5 points réalisée à partir d'un pool des ADNc des espèces analysées par dilutions en cascade (de 0,5 à 5.10^{-5} ng/µL).

En plus des 13 gènes étudiés, les 3 gènes de ménage sont testés pour choisir celui dont l'efficacité se rapproche de 100% et montre une expression constante entre échantillons grâce au logiciel GeNorm (BioRad).

Les fragments des gènes candidats amplifiés sont ensuite quantifiés sur un thermocycleur Chromo4 System (BioRad) selon le programme suivant : une incubation de 3 min à 95°C, puis 45 cycles avec une incubation de 15 sec à 95°C suivi de 60 sec à 60°C et d'une lecture optique. A la fin du programme, une courbe de fusion (avec lecture optique tous les 0,5°C de 65°C à 95°C) est réalisée afin de vérifier l'amplification d'un seul fragment spécifique.

Tableau 4: Séquences des amorces utilisées pour la PCR quantitative, pour le clonage et le séquençage des différentes copies dupliquées du gène Metal tolerance protein.

Gène	Amorce Forward	Amorce Reverse
Xanthine dehydrogenase 1	CGTAGGTGAGCCACCATTTT	TACTAAGCTCTGGCCGGAAA
Transcriptional adapter ADA2a	CGTACACTGGCTGAAGCAAA	CCAGAGTCTAAGCCAGTGCC
Zinc finger protein STOP1 homolog	CATGAGTCTTCAAGTCGGCA	TTCAGATCATGCACCGGTTA
WRKY24	GCCCCTTCTGTTGTGACATT	GTGGTCAAAGGAAACCCTCA
Metal tolerance protein A2	CTGTGAGATGGATGGTGTGG	ATTAGCCTACTGGCGCTCAA
Metal tolerance protein A2 (clonage)	TTGGCTTGCCATGTTACAAT	GGTTCCAAGCAAGAAGGTCA
Cinnamoyl-CoA reductase-like protein	AGGGTTGACCTGGACCTCTT	TGAATCCTTTGGACTGGGAG
Putative GDP-mannose pyrophosphorylase	AAACAAGCCCATGATTCTGC	TGGGAGCATGTGATTGTGAT
Cytochrome c oxidase subunit 6b-1	TCCGCTTCCCTACAACAAAC	CATTCTCCCTCTGCTCGTTC
Similar to Transcription elongation factor SPT6	AGCAAAATGATGAGAGGCGT	CTTCCTGCCTTCTTCAAACG
Hexokinase 1	TAGCCCAATGAGATTGGAGG	TTCCATCGTGTCCTCCTCTT
Transfactor, putative, expressed	GCAAGATGGCTTGCCTACTC	TGACACCTCCAAATCACCAA
Metallocarboxypeptidase inhibitor	TACCCCTTGTTGGTACGGAG	AGCCTCGTGGAGAAGCAATA
V-type proton ATPase catalytic subunit A	ATGATGGCTGACTCCACCTC	CCCAACAATTGTGACACTGC
Glyceraldehyde 3-Phosphate Dehydrogenase	TTTGAATGGCAAGCTCACTG	TAGCCCATGATCCCTTTGAG
Alpha-tubulin	AGCCTAGGGAGGCTTCAGTC	ATCAGGCGCTTGAAGAACAT
Sucrose synthase	CGTTCACGTCTTTAGGCACA	AGTTCCAGACCAACCTGGTG

4. Analyses statistiques

A la fin de chaque programme d'amplification, les niveaux d'expression sont calculés pour chaque échantillon en utilisant la courbe d'efficacité de la gamme étalon pour chaque gène analysé. Pour choisir le gène de référence le mieux adapté parmi les 3 gènes de ménage testés, le logiciel GeNorm (Biorad) a été utilisé. Le gène choisi, codant la GA3PDH, permet ensuite de normaliser les niveaux d'expression de tous les échantillons à l'aide logiciel Genex (BioRad). Pour certains échantillons, un des trois réplicats techniques n'a pas fonctionné. Dans la suite des analyses, deux réplicats techniques ont été pris en compte par échantillon (afin de ne pas biaiser le jeu de données et de pouvoir construire des matrices de données à 2 paramètres) et les 13 gènes ont été analysés indépendamment.

Afin d'estimer les variations d'expression des individus entre populations d'une même espèce et entre espèces, des ANOVA (Analyses de variance) à deux paramètres ont été réalisées (test paramétrique sur données appariées), aussi appelées tests de Wald. Pour cela, la construction de modèle linéaire mixte a été nécessaire. Pour effectuer les comparaisons des espèces deux à deux une matrice des contrastes a été réalisée. La vérification de la distribution des données a été réalisée avec un graphe quantile-quantile et le test de normalité de Shapiro pour chaque jeu de données. L'homogénéité des variances a été vérifiée avec un test de Bartlett par permutation pour les comparaisons interspécifiques et un test de Fisher pour les comparaisons interpopulations.

Pour étudier les variations d'expression entre individus de la même espèce le long d'un transect au sein de la même population, un test de Kruskal et Wallis a été utilisé. Toutes les analyses statistiques ont été réalisées à l'aide du logiciel R version 15.1 (R Development Core Team, 2011) et du package RVAideMemoire (Hervé, 2012).

Afin de comparer les niveaux d'expression des hybrides et de l'allopolyploïde à une valeur additive théorique des niveaux d'expression des parents, la Mid-Parent Value (MPV) est calculée en faisant la moyenne des niveaux d'expression des parents. L'intervalle de confiance (calculé avec le test de Wilcoxon) permet de prendre en compte les variations d'expression intraspécifique des parents et de situer le niveau théorique d'additivité à plus ou moins 5% de la MPV.

Tableau 5 : Récapitulatif des gènes utilisés pour étudier les variations d'expression, leur annotation fonctionnelle et le nombre de locus chez *Sorghum bicolor*.

Fonction biologique	Annotation	Identification	Nb de régions homologues chez *Sorghum bicolor*	N° du contig(longueur)
Réponse au stress salin				
	Xanthine dehydrogenase 1	At4g34890, Sb01g031520	1 région	contig01057 (652bp)
	Transcriptional adapter ADA2a	At3g07740, Sb01g007950	1 région	contig03573 (1012bp)
	Zinc finger protein STOP1 homolog	LOC_Os01g65080, Sb03g041170.1	2 régions	contig06019 (1241bp)
	WRKY24	LOC_Os01g61080, Sb03g038510	3 régions	contig04256 (1147bp)
Réponse aux métaux lourds				
	Metal tolerance protein A2	At3g58810, Sb09g002460,LOC_Os08g32650	1 région	contig01973 (780bp)
Métabolismes de la lignine et de la cellulose				
	Cinnamoyl-CoA reductase-like protein	At5g58490, Sb03g038630	1 région	contig10883 (333bp)
	Putative GDP-mannose pyrophosphorylase	LOC_Os08g13930, LOC_Os08g13930	3 régions	contig06944 (446bp)
Expression affectée par la spéciation réticulée et allopolyploïde				
	Cytochrome c oxidase subunit 6b-1	LOC_Os03g27290, Sb01g033560	2 régions	contig10492(320bp)
	Similar to Transcription elongation factor SPT6	At2g22080, LOC_Os04g54430	5 régions	contig32516(493bp)
	Hexokinase 1	At4g29130, Sb09g026080	2 régions	contig20419(1088bp)
	Transfactor, putative, expressed	LOC_Os03g20900, Sb01g036680	2 régions	contig13824(114bp)
	Metallocarboxypeptidase inhibitor	AES99445 (Medicago truncatula)	1 région	contig15367(883bp)
	V-type proton ATPase catalytic subunit A	At1g78900, Sb04g005040	3 régions	contig08620(1174bp)
Gènes de ménage				
	Glyceraldehyde-3-Phosphate Dehydrogenase	Sb07g002220	1 région	GA3PDH A
	Alpha-tubulin	Sb01g009560	1 région	Atub A
	Sucrose synthase	Sb10g006330	1 région	Sucrose B

5. Clonage et séquençage du gène codant la metal tolerance protein

Les espèces analysées étant polyploïdes, plusieurs copies dupliquées (homéologues) sont attendues par locus. Les analyses effectuées en PCRq mesurent donc l'expression globale (ensemble des copies homéologues) à un locus donné. Nous nous sommes posés la question du nombre de copies détectables par locus en prenant pour exemple un gène codant une metal tolerance protein. D'après les comparaisons avec les génomes proches séquencés du sorgho et du riz, ce gène est supposé en simple copie par génome de base (Tableau 5). Ce gène possède un seul exon (Figure 8). Ce gène, analysé par PCRq a ensuite été cloné et séquencé en vue de détecter les différentes copies potentiellement présentes à ce locus.

Les amorces ont été définies dans la région 3' du gène (en fin du dernier exon et début de la partie 3'-UTR) car elle est plus polymorphe. Ces informations sont issues d'un séquençage par la technologie 454 (Ferreira de Carvalho *et al.*, 2013) dont l'analyse a permis d'observer un polymorphisme à l'intérieur des zones codantes de ce gène d'intérêt. Les amorces dessinées ont servi à amplifier l'ADN génomique mais aussi l'ADN complémentaire des parents hexaploïdes *S. maritima* et *S. alterniflora* (Tableau 4). Le protocole d'amplification est le suivant : une incubation à 95°C pendant 2 min suivie de 30 cycles d'amplification (95°C pendant 30 sec, 58°C pendant 30 sec et enfin 72°C pendant 1 min) et enfin une phase d'élongation à 72°C pendant 7 min.

Les produits d'amplification ont été clonés pour séquençage (méthode de Sanger). Après purification par le Kit NucleoSpin Gel and PCR Clean-up (Macherey-Nagel), les produits PCR sont intégrés à des plasmides bactériens suite à une ligation (kit gGEM-T Easy, Promega). Les plasmides sont clonés à l'aide de bactéries thermocompétentes (*E. coli* DH5a, Invitrogen) transformées par choc thermique puis mises en culture sur milieu LB Agar. Après sélection des colonies ayant intégré le fragment, les plasmides sont extraits puis purifiés à l'aide du Kit Pure Yield Plasmid Miniprep System (Promega). Une PCR de contrôle est effectuée afin de vérifier l'insertion d'un bon fragment au sein du plasmide. Le mix réactionnel de la PCR de contrôle est le suivant: 7,4 µL d'eau, 5 µL de tampon, 2,5 µL de dNTPS (2mM), 2 µL de SP6 (5 µM), 2 µL de T7

(5 µM), 0,1 µL de Taq (5U/µL) et 1 µL de plasmide soit un total de 20 µL. Les échantillons sont ensuite envoyés pour séquençage au laboratoire Macrogen Europe (Amsterdam, Netherlands).

Les séquences obtenues (pour un même individu) sont comparées dans le but de détecter les différents haplotypes (séquences caractérisées par plusieurs sites polymorphes) grâce à l'alignement et l'inspection visuelle de la matrice à l'aide du logiciel BioEdit (Ibis Biosiences). De plus, une analyse phylogénétique (méthode de parcimonie) des séquences a été effectuée à l'aide du logiciel PAUP (Swofford, 2003). Cette partie du travail a fait intervenir la participation de Pierre Bourdau que nous avons co-encadré dans le cadre de la réalisation d'un stage de Master 1 au laboratoire (Bourdau, 2012).

Les polymorphismes nucléotidiques ont aussi été identifiés pour ce même locus, par une approche *in silico* des données générées grâce à la plateforme 454 (Roche Life Sciences). Pour chaque espèce parentale, le gène codant la metal tolerance protein a été identifié parmi les contigs assemblés. Puis, les séquences permettant de construire chaque contig ont été réassemblées avec une stringence plus élevée (selon une méthode mise au point dans le laboratoire ; Boutte *et al.*, *en préparation*) afin de retrouver les différents haplotypes de ce gène et les homéologues potentiels.

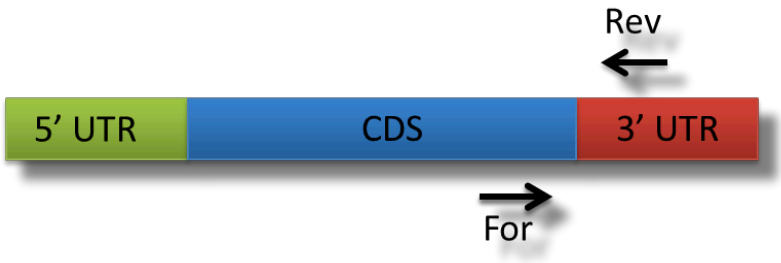

Figure 5 : Séquence génomique du gène codant la Metal Tolerance Protein et amorces utilisées pour l'amplification de l'ADN génomique et complémentaire chez *S. maritima* et *S. alterniflora*.

V. Analyses du génome de *Spartina maritima*

Le séquençage d'ADN génomique permet d'améliorer les connaissances actuelles sur l'organisation du génome nucléaire et cytoplasmique chez les Spartines et d'analyser le contenu des compartiments répétés (notamment les éléments transposables) et leur distribution par rapport aux zones codantes du génome. L'ADN génomique des Spartines a été extrait d'une part, en vue de la réalisation d'une banque BAC et d'autre part, en vue de séquençage massif (pyroséquençage).

1. Réalisation de la banque BAC

Pour la préparation de la banque BAC, des individus de *Spartina maritima* (Site de Quenouille, Rivière d'Etel, Morbihan) ont été étiolés. Les plantes ont été maintenues 4 jours à l'obscurité, puis 40g de matière fraîche de feuilles ont été récoltés et immédiatement conservés dans l'azote liquide. Les extractions d'ADN et la construction de la banque BAC ont été réalisées au Centre National des Ressources Génomiques Végétales (CNRGV) de Toulouse selon un protocole standard mis au point pour les plantes (Luo & Wing, 2003). L'ADN extrait est digéré à l'aide de deux enzymes de restriction (*HindIII* donnant la banque A et *BamHI* donnant la banque B) et les fragments digérés dont la taille est comprise entre 100 et 300 kb sont sélectionnés. Ces fragments sont ensuite clonés à l'aide du vecteur pIndigoBAC5 dans une souche hôte (*Escherichia coli* DH 10BT1 résistante) pour constituer la banque. Au total, 44 544 clones d'une taille moyenne de 110 kb ont été obtenus, soit environ 4900 Mb, ce qui représente une couverture d'environ 1.5 X du génome de *S. maritima* (3 700 Mb). On gardera toutefois à l'esprit que ce génome hexaploïde est redondant, et que cette couverture représente en fait environ 8 fois le génome de base (616 Mb) de *S. maritima*.

Cette banque est utilisée d'une part pour le séquençage d'extrémités de BACs (Bac End Sequences ou BES) et d'autre part, pour le séquençage de BACs entiers contenant des régions d'intérêt. Le séquençage des extrémités de BAC a été effectué par le Génoscope (selon la méthode de Sanger). La banque actuellement disponible est constituée de 40 641 séquences. J'ai contribué à la sélection des BACs contenant une région d'intérêt en vue du séquençage de

ces BACs entiers et de l'analyse des régions homologues chez cette espèce hexaploïde. Deux régions ont été choisies : le gène ADH-1, fréquemment utilisé en génomique comparative chez les Poaceae (Jannoo *et al.,* 2007) et le gène CAD-2 (Cinnamyl Alcohol Dehydrogenase 2) rentrant dans la voie de synthèse des parois végétales. Ces analyses encore en cours ne sont pas incluses dans ce manuscrit.

2. Pyroséquençage d'ADN génomique

Par ailleurs, en vue d'une analyse par pyroséquençage (454-Roche Applied Technology), l'ADN de *Spartina maritima* (Morbihan-Rivière d'Etel) a été extrait à partir de feuilles de plantes maintenues en serre sur le campus de Beaulieu (Université de Rennes 1). L'extraction de l'ADN a été réalisée avec le kit Nucleospin Plant II (Macherey Nagel) en suivant les recommandations du fabricant. Un total de cinq microgrammes d'ADN génomique a été nécessaire pour effectuer le séquençage à la plateforme de génomique environnementale de l'OSUR à l'aide d'un pyroséquenceur GS FLX (454 Life Science Roche). Un ensemble de 993 229 lectures d'une longueur moyenne de 275 pb ont été obtenues. Les séquences de faible qualité ainsi que leurs adaptateurs sont ensuite nettoyés par filtrage (développé sur la plateforme).

VI. Analyses bioinformatiques du génome de *Spartina maritima*

Le génome de *Spartina maritima* a été étudié à travers plusieurs jeux de données. Tout d'abord, le séquençage des extrémités de BAC vise à une première évaluation des compartiments codant et non-codant de ce génome en suivant la démarche présentée en Figure 9. Par ailleurs, le pyroséquençage d'ADN génomique à l'aide d'une plateforme 454 GS FLX Titanium (Roche, Inc.) vise à explorer, analyser un plus grand jeu de données et créer une base de données d'éléments répétés chez *S. maritima*.

1. Analyses des extrémités de BAC (BAC-End Sequences ou BES)

Identification des séquences des génomes cytoplasmiques

Afin d'identifier les séquences cytoplasmiques, les BES ont tout d'abord été comparés au génome chloroplastique assemblé à partir des données de séquençage par 454 chez *S. maritima* (Bellot, 2010 ; Bellot *et al.*, en préparation). Les BES ont aussi été comparés aux génomes chloroplastiques et mitochondriaux d'*Oryza sativa ssp indica* et de *Sorghum bicolor* (NCBI, décembre 2011) en utilisant le programme BLASTn et une valeur seuil de 10^{-6} sur une longueur d'alignement d'au moins 70pb.

Identification et évaluation du compartiment répété

L'identification des éléments transposables et autres séquences répétées dans les BES de *S. maritima* a été effectuée dans un premier temps à l'aide du logiciel RepeatMasker version 3.2.9 (http://www.repeatmasker.org/) en utilisant *Oryza sativa* comme espèce de référence dans Repbase (Jurka *et al.*, 2005). Les proportions relatives des différentes catégories dans les BES sont calculées. Le fichier masqué (séquences répétées excisées des BES) généré est ensuite utilisé pour l'annotation du compartiment génique. Par ailleurs, les BES sont aussi alignés contre des bases de données d'éléments répétés connus (Gramineae, *O. sativa* et *S. bicolor*) téléchargées sur le site TIGR Plant Repeat Databases (http://tigr.org/) en décembre 2011 (Ouyang & Bell, 2004). Dans cette analyse, le programme BLASTn a été utilisé avec une e-value de 10^{-6} et une longueur minimum d'alignement de 100bp. Les annotations se sont révélées plus nombreuses en utilisant la base de données d'*O. sativa*, les proportions des éléments répétés ont donc été calculées sur les résultats obtenus avec cette espèce.Spartines

Tous les BESs contenant des rétroéléments ont été extraits par alignement (tBLASTx avec une e-value de 10^{-6}) contre la base de données protéiques des reverse transcriptase (RT) de Repbase. Les séquences de *S. maritima* ont alors été traduites en protéines et alignées contre les mêmes séquences RT de Repbase classifiées en sous-familles pour chacun des éléments *Copia* et *Gypsy* à l'aide du programme MUSCLE (Edgar, 2004) implémenté au logiciel Geneious

(Biomatters). Ces séquences alignées ont fait l'objet d'une analyse phylogénétique selon la méthode du Neighbor-Joining en utilisant le modèle de Jukes-Cantor (1969).

Identification de novo des séquences répétées

Le fichier masqué contenant 39 910 séquences (soit 26,2Mb) a été aligné contre lui-même avec une stringence élevée (e-value 10^{-50}) afin de retrouver les séquences hautement répétées dans le génome de *S. maritima* et qui auraient trop divergé des éléments présents dans les bases de données actuelles pour être détectées. Les séquences présentant au moins 6 hits avec un pourcentage d'identité minimum de 90% ont été récupérées puis alignées (BLASTn) contre la base de données nucléotidiques non-redondante de GenBank (NCBI), la base de données SwissProt et la base de données d'ESTs de Poaceae (utilisée pour l'annotation des contigs et incluant les ESTs de *Zea mais*, *Brachypodium distachyon*, *Sorghum bicolor* et *Oryza sativa*). Ces séquences ont également été comparées à différentes bases de données de séquences répétées : TIGR Plant Repeat databases (Gramineae, *Zea mays, Oryza sativa* et *Sorghum bicolor*), RepBase et TREPdb en utilisant le programme BLASTn (e-value de 10^{-6}) afin de vérifier la spécificité de ces séquences chez *S. maritima*. Les BES correspondant à ces critères sont alors assemblés à l'aide du logiciel GS Assembler pour données génomiques selon les paramètres de 90% d'identité sur au moins 40 nucléotides chevauchants.

Microsatellites ou Simple sequence repeats (SSR)

Les microsatellites sont de courts motifs de séquences répétées en tandem et variable par leur nombre de répétitions. Le polymorphisme des microsatellites peut être utilisé comme marqueur génétique afin d'identifier les variations intraspécifiques ou construire des cartes génétiques. Les séquences microsatellites ont été détectées grâce au script Perl MISA (MIcroSatellite research tool, Thiel *et al.*, 2003). Les paramètres ont été établis afin de trouver tous les SSRs dont les motifs vont de 1 à 6 nucléotides (*i.e.* mono-, di-, tri-, tetra-, penta- et hexanucleotides) et leur longueur minimale varie selon le motif : 10 nucléotides pour les mononucléotides, 12pb pour les dinucléotides, 15pb pour les trinucléotides, 20pb pour les tétranucléotides, 25pb pour les pentanucléotides et 30pb pour les motifs hexanucléotidiques.

Identification et évaluation du compartiment codant et annotations fonctionnelles

Le fichier masqué des BES a ensuite été comparé aux séquences codantes (CDS) des espèces *O. sativa et S. bicolor* (version 120 et 79 téléchargées à partir du site www.phytozome.com en décembre 2011). Les analyses tBLASTx réalisées (e-value de 10^{-6}) ont montré une plus grande homologie de séquences avec *S. bicolor* ; les séquences homologues ont alors été annotées par Gene Ontology grâce au logiciel BLAST2GO (Conesa *et al.*, 2005; Götz *et al.*, 2008). Pour cela, des analyses tBLASTx (e-value 10^{-6}) entre les BESs et la base de données protéiques d'*Arabidopsis thaliana* (TAIR, www.arabidopsis.org) ont été effectuées. En parallèle, les BESs masqués ont aussi été comparés au transcriptome de référence assemblé et annoté pour les cinq espèces de Spartines *S. maritima*, *S. alterniflora*, *S.* x *townsendii*, *S.* x *neyrautii* and *S. anglica* (Ferreira de Carvalho *et al.*, 2013) dont les détails sont présentés dans le chapitre suivant.

Génomique comparative

Afin d'identifier les régions microsynténiques entre *S. maritima* et des espèces plus ou moins proches phylogénétiquement, les BES masqués ont été alignés sur les génomes séquencés d'*Arabidopsis thaliana*, *Brachypodium distachyon*, *Oryza sativa* et *Sorghum bicolor* (Athaliana_167.fa, Bdistachyon_192_hardmasked.fa, Sbicolor_79_RM.fa et Osativa_120_RM.fa téléchargés sur le site www.phytozome.org). La valeur seuil utilisée est de 10^{-6} et les Best BLAST Hits (BBH) ont été retenus si l'identité de l'alignement était supérieure à 70%. Un BES est alors considéré comme colinéaire sur le génome de référence si les deux extrémités sont correctement orientées sur le même chromosome et ancrées dans une région comprise entre 15kb et 250kb. Dans le cas contraire, la région ciblée est considérée comme réarrangée entre *S. maritima* et l'autre génome.

La synténie et les réarrangements entre *S. maritima*, *Sorghum bicolor* et *Oryza sativa* ont pu être visualisés grâce au programme CIRCOS (V.0.55, Krzywinski *et al.*, 2009). Les BESs homologues ont été alignés sur les 10 chromosomes du sorgho et les 12 chromosomes du riz en utilisant le programme BLASTn (e-value de 10^{-6} et une identité minimale de 70%).

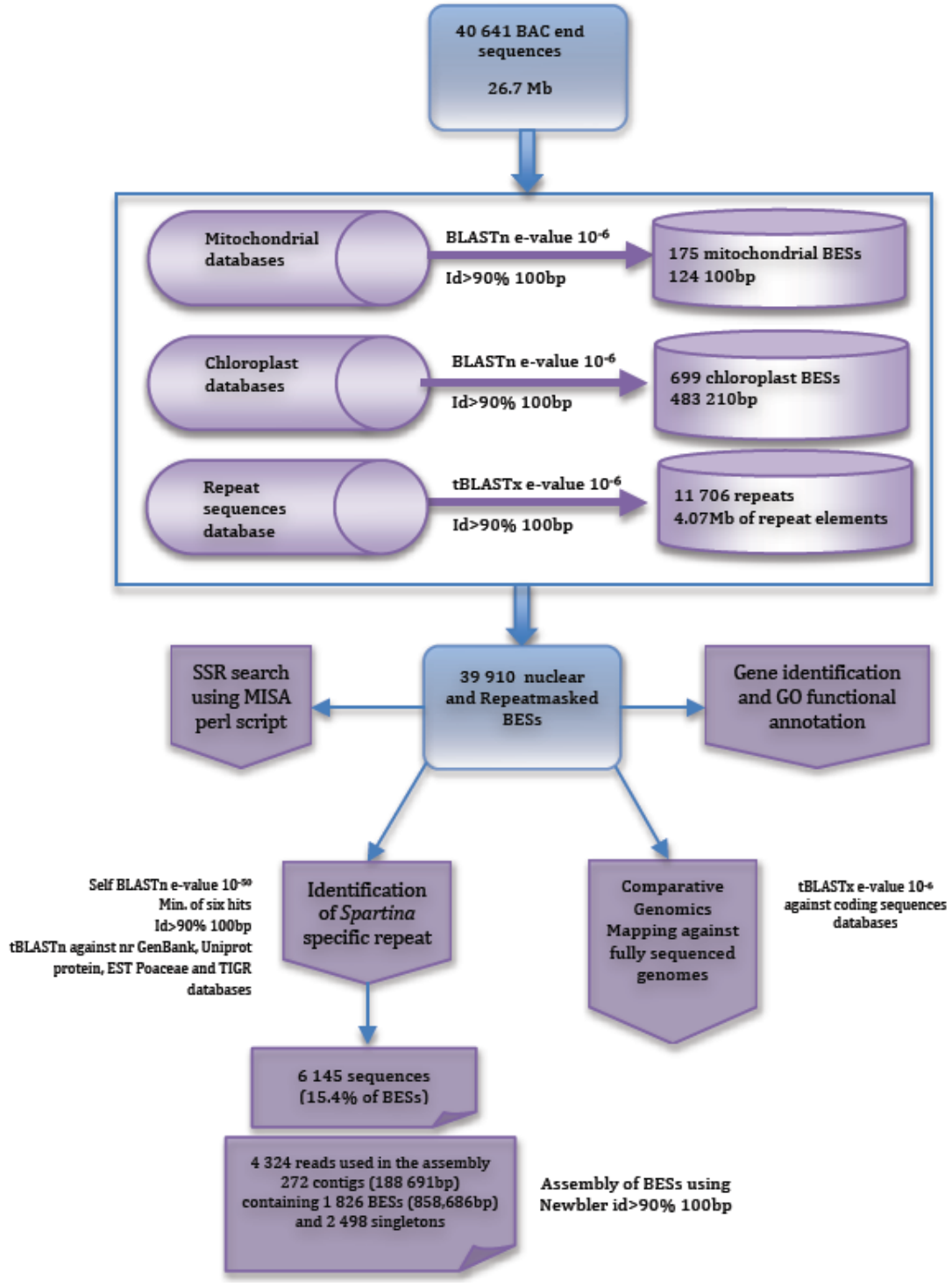

Figure 9 : Démarche suivie pour l'analyse des séquences d'extrémités de BAC (BES).

2. Analyses des données génomiques issues de la plateforme 454

Filtrage des séquences cytoplasmiques

La recherche des séquences chloroplastiques de *S. maritima* s'est effectuée à l'aide d'alignements par BLAST (Altschul *et al.*, 1990) contre les génomes chloroplastiques complets du riz et du maïs (BLASTn avec e-value inférieure 10^{-6} et un pourcentage d'identité supérieure à 90% sur plus de 100pb). La même démarche a été effectuée pour les séquences restantes par BLAST contre les génomes mitochondriaux complets du riz et du maïs.

Identification et évaluation du compartiment répété

L'identification des séquences répétées dans le génome de *Spartina maritima* a été effectuée suivant une méthode de regroupement de séquences similaires en « cluster ». Cette démarche utilise la méthode des graphes développée par Novak *et al.* (2010). Elle permet de regrouper les séquences interconnectées dans des sous-ensembles ou « clusters ». Les séquences présentes dans chaque sous-ensemble sont ensuite assemblées en contigs en utilisant l'assembleur CAP3 (Huang & Madan, 1999), avec comme critères un chevauchement minimum de 70 nucléotides et une identité de 85% entre deux séquences. L'annotation des différents clusters a été effectuée via plusieurs méthodes complémentaires. Tout d'abord, les séquences ont été soumises à une recherche d'éléments répétés en utilisant le programme RepeatMasker (Smit *et al.*, 1996-2010), en réalisant des alignements (BLASTn) contre les bases de données NCBI (http://www.ncbi.nlm.nih.gov/) d'éléments transposables et des alignements avec le logiciel FASTA (Pearson & Lipman, 1988) contre les parties codantes de domaines connus d'éléments transposables. Par ailleurs, une étude graphique de l'organisation des régions codantes des éléments transposables grâce au logiciel SeqGrapheR (Novak *et al.*, 2010) a été réalisée. Enfin, les séquences satellites ont été détectées à l'aide de dot-plots (Sonnhammer & Durbin, 1995). Les types d'éléments présents dans les séquences annotées en tant que microsatellites ont été identifiées grâce au script MISA (Microsatellite Identification Tools, http://pgrc.ipk-gatersleben.de/misa). La proportion en éléments répétés dans le génome nucléaire a ensuite été estimée en fonction du nombre de séquences correspondant aux

séquences répétées dans chaque cluster. Dans cette étude, une attention particulière a été portée aux clusters contenant des séquences codant des ARN ribosomiques. La séquence complète codant l'ARN nucléaire ribosomique a pu être entièrement reconstituée en un seul contig et a servi pour d'autres études au sein du laboratoire. De plus, les rétroéléments de classe I étant les plus abondants au sein des génomes des Poaceae, ceux-ci ont été plus précisément identifiés aux sous-familles de retrotransposons les plus représentés. Pour cela, le domaine codant la RT (transcriptase inverse ou reverse transcriptase) de chaque élément a été localisé et extrait des contigs. Ces séquences, converties en protéines, ont ensuite été alignées avec MUSCLE (Edgar, 2004) contre les séquences de Repbase (Jurka *et al.*, 2005) en utilisant les critères par défaut. Une analyse phylogénétique a ensuite été réalisée avec SeaView (Gouy *et al.*, 2010) en utilisant la méthode du Neighbor-Joining sur les distances observées (bootstrap de 1000 répétitions).

Afin de détecter de potentiels phénomènes d'amplification récente, les séquences considérées comme répétées sont blastées contre elles-mêmes (SelfBLASTn, e-value 10^{-6}) pour identifier les groupes de séquences proches et de calculer l'identité de ces séquences entre elles. Une distribution des fréquences a été réalisée pour les séquences de plus de 80pb.

Chapitre 4

Vers la compréhension du transcriptome des Spartines hexaploïdes et dodécaploïdes : Premiers transcriptomes de référence

Introduction et démarche générale

La compréhension des conséquences de l'hybridation interspécifique et de la spéciation allopolyploïde sur l'évolution de l'expression des gènes, la physiologie, l'adaptation et l'écologie des espèces requiert la connaissance du compartiment codant rendue aujourd'hui plus facilement accessible grâce aux nouvelles technologies de séquençage. Au début de ce travail, les données d'EST (Expressed Sequence Tags) des Spartines disponibles dans les bases de données représentaient environ 1300 séquences dont 15 de *S. maritima* (Chelaifa *et al*., 2010a) et 1266 ESTs de *S. alterniflora* (Baisakh *et al*., 2006 ; 2008 ; 2012 ; Chelaifa *et al*., 2010a). Une étude récente de séquençage de transcriptome chez *S. pectinata* a été publiée par Gedye *et al.* (2010). Cette espèce, tétraploïde, pour laquelle des cytotypes hexaploïdes et octoploïdes ont été également rapportés (Kim *et al.*, 2010 ; 2012), est actuellement considérée comme source potentielle de biofuel (Gonzales-Hernandez *et al*., 2009).

La construction *de novo* d'un transcriptome de référence chez les espèces hexaploïdes (*S. maritima* et *S. alterniflora)* est la première étape vers la compréhension du compartiment codant des Spartines impliquées dans les évènements récents d'hybridation et de spéciation allopolyploïde en Europe. Ce travail a fait l'objet d'une publication dans la revue *Heredity*, présentée dans la partie A de ce chapitre.

La partie B de ce chapitre présente des résultats complémentaires (non inclus dans la publication), de données obtenues à partir d'assemblages de données transcriptomiques chez les 5 espèces de Spartines (les deux parents hexaploïdes *S. maritima* et *S. alterniflora*, les deux hybrides F1 *S. x townsendii* et *S. x neyrautii*, et l'allopolyploïde *S. anglica*).

Afin d'identifier un maximum de gènes exprimés différents, plusieurs prélèvements dans des conditions différentes ont été effectués. Au total, cinq runs de pyroséquençages d'ADNc (Roche-454) ont été réalisés sur les cinq espèces du complexe polyploïde, sur deux organes différents : racines et feuilles récoltées en conditions contrôlées. De plus, un sixième séquençage a été réalisé sur une banque d'ADNc normalisée de feuilles et de racines, chez

l'espèce native européenne *S. maritima* échantilonnée en conditions contrôlées et naturelles (à travers un transect le long du shore).

Les séquences obtenues ont été nettoyées et assemblées suivant l'espèce, l'organe et le type de normalisation. Au total, 27 assemblages ont été réalisés.

Partie A. Le transcriptome de référence des espèces hexaploïdes *S. maritima* et *S. alterniflora*

A travers cet article, nous analysons les transcriptomes de feuilles et racines des deux espèces parentales hexaploïdes *S. maritima* (native des côtes atlantiques euro-africaines) et *S. alterniflora* (originaire des côtes atlantiques est-américaines). Ces deux espèces présentent des phénotypes et des écologies contrastées sur les marais salés. *Spartina alterniflora* présente une distribution plus large sur son aire native (du Canada à l'Argentine) et des capacités envahissantes dans la plupart des régions où elle a été introduite (Californie, Chine, Europe de l'Ouest). A l'inverse, les populations de *S. maritima* sont actuellement en train de régresser dans le nord-ouest de l'Europe. Ce phénomène pourrait résulter des changements climatiques et des perturbations anthropogéniques, ou être relié aux différences biologiques et morphologiques entre les deux espèces hexaploïdes. *Spartina alterniflora* possède de longs rhizomes facilitant son expansion et la sédimentation du sol, l'espèce remonte le long des estuaires et tolère des eaux plus saumâtres que *Spartina maritima*. Cette dernière occupe des zones pionnières du bas de l'estran et tolère de plus longues heures d'immersion. Elle présente peu de rhizomes et une production de graines très faible.

Les assemblages réalisés ont tout d'abord été comparés au génome complet le plus proche phylogénétiquement : *Sorghum bicolor*. Les contigs ont ensuite été annotés grâce à des banques d'EST. Finalement, les stratégies de séquençage, d'assemblage et d'annotation sont discutées.

Les assemblages *de novo* des séquences issues du pyroséquençage ont généré un total de 38 478 contigs dont une grande part (~99%) ont pu être annotés fonctionnellement à partir de base de données de séquences annotées chez les Poacées. Cette étape d'annotation fonctionnelle a permis d'identifier 16 753 gènes différents. Les contigs (ou « gènes ») identifiés

ont ensuite été alignés sur le génome de *Sorghum bicolor,* où ils se distribuent sur les bras chromosomiques sous-télomériques avec une forte corrélation de la densité de gènes. L'étape de normalisation de banques d'ADNc a, comme attendu, augmenté le nombre de transcrits séquencés et assemblés. La construction de ces nouveaux transcriptomes de Spartines a permis d'identifier des gènes aux fonctions écologiques potentiellement importantes dans la réponse adaptative aux stress salin et aux métaux lourds.

Les Spartines étudiées étant hexaploïdes, nous avons cherché à identifier le nombre de copies de gènes exprimés au sein de chaque espèce pour un sous-ensemble des gènes assemblés. Nous avons ainsi observé jusqu'à 4 haplotypes par gène pour chaque espèce, qui pourraient correspondre à la présence de deux copies homéologues exprimées (et d'un ou deux variants allèliques).

Ces nouveaux transcriptomes de référence de *S. maritima* et *S. alterniflora* réprésentent une ressource particulièrement importante pour les études ultérieures d'analyse d'expression de gènes spécifiques dans des conditions particulières, ou encore l'estimation de l'expression des différentes copies de gènes à l'échelle du transcriptome (par RNA-seq ou par puce spécifique).

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).

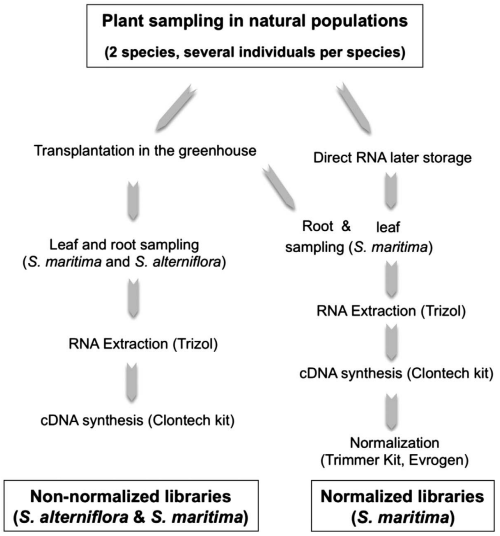

Figure 1 Sampling strategy and construction of the normalized (*S. maritima*) and non-normalized (*S. maritima* and *S. alterniflora*) cDNA libraries.

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).

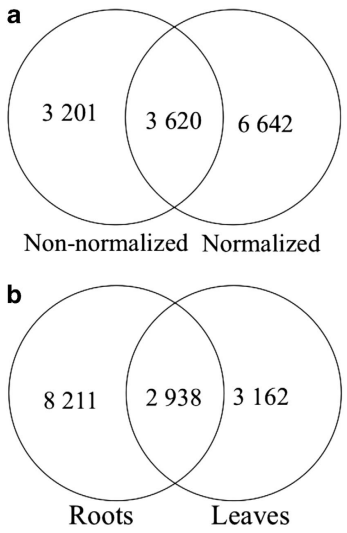

Figure 2 Common annotated contigs (using the Poaceae database) of *S. maritima* (**a**) between non-normalized and normalized cDNA libraries (leaves + roots) (**b**) between roots and leaves.

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

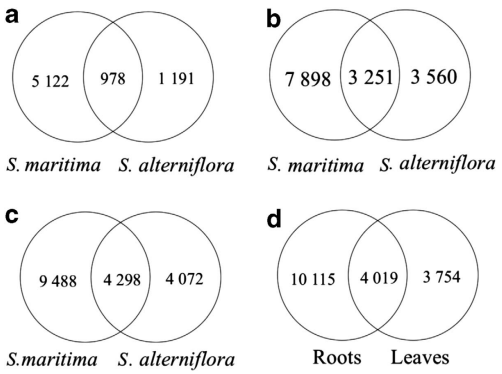

Figure 3 Common annotated contigs (using the Poaceae database) between *S. maritima* and *S. alterniflora*. (**a**) Comparison between *S. maritima* and *S. alterniflora* leaves. (**b**) Comparison between *S. maritima* and *S. alterniflora* roots. (**c**) Comparison between *S. maritima* and *S. alterniflora* (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*

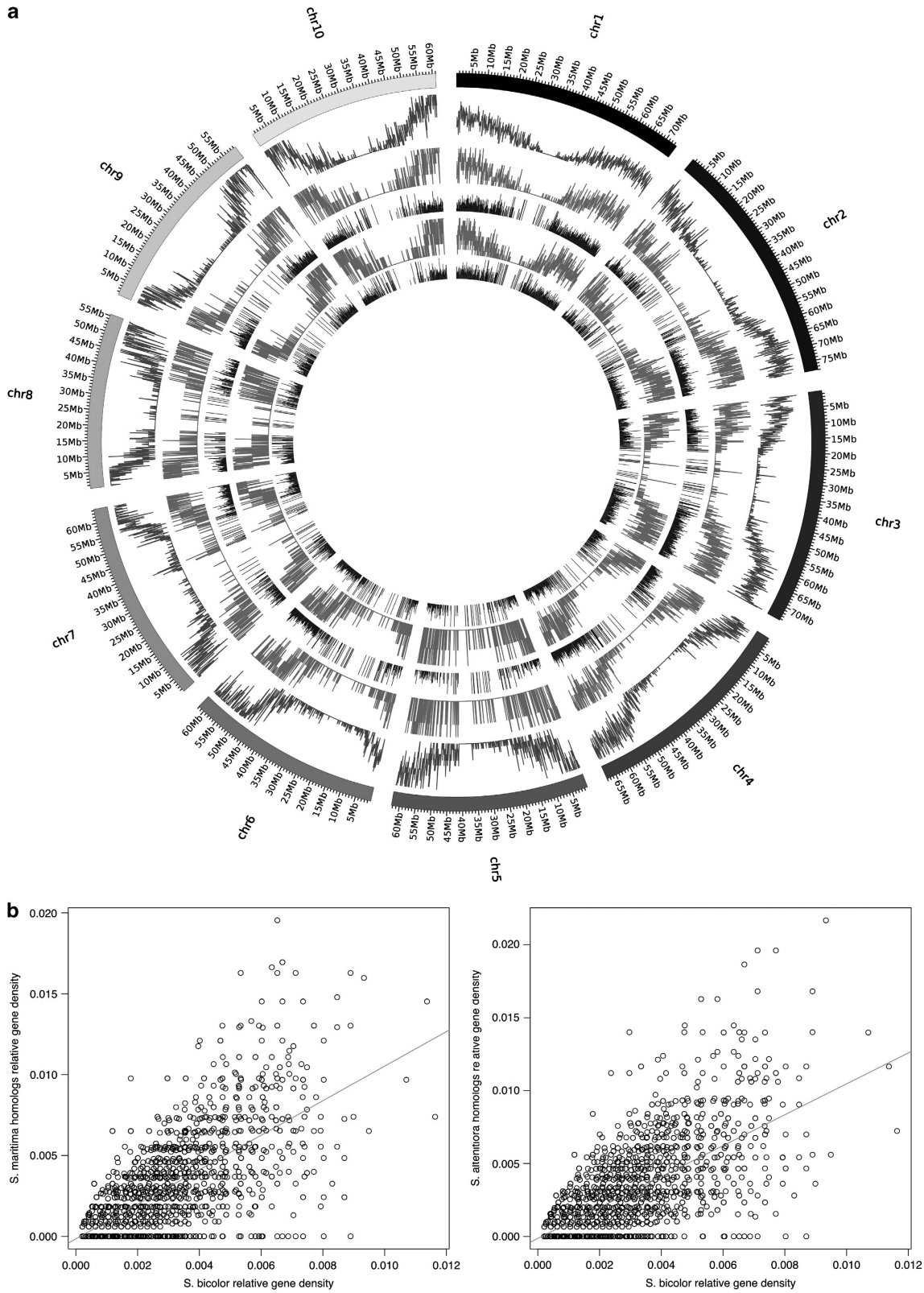

Figure 4 (**a**) *Spartina* contigs mapped to the *Sorghum* genome. The 10 individual chromosomes are shown in the outer circle. Relative gene densities on each chromosome are displayed successively inward as following: (i) gene density in *Sorghum bicolor*, (ii) gene density in *Spartina maritima*, (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 $< 2.2\,e^{-16}$).

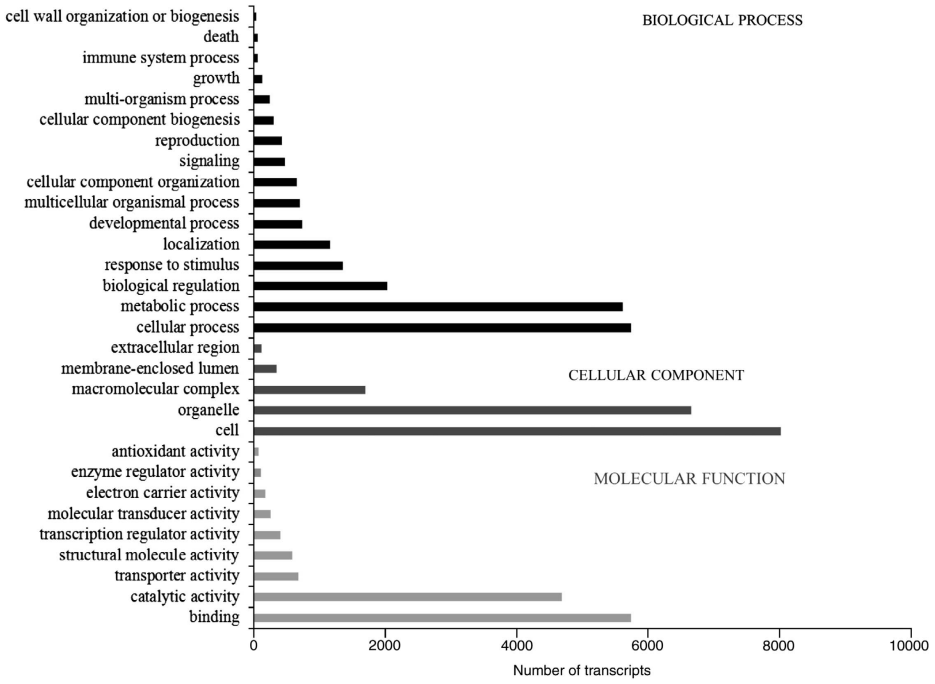

Figure 5 Functional classification of the leaf transcriptome of *S. maritima* and *S. alterniflora*. GO annotations were used for classification for GO cellular component, GO molecular function and GO biological process.

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)

Partie B. Résultats des assemblages des hybrides et de l'allopolyploïde, et des cinq espèces de Spartines

Dans cette partie, nous présentons les résultats d'assemblage et d'annotation pour les deux hybrides F1 (*S.* x *neyrautii* et *S.* x *townsendii*), l'allododécaploïde *S. anglica* et pour l'ensemble des cinq espèces du complexe *Spartina* dans le Tableau 1.

Le séquençage du transcriptome de l'hybride *S.* x *townsendii* pour les feuilles et les racines a généré respectivement 200 342 et 122 431 séquences (soit 57,9 et 26,4 Mb ; Tableau 6). La longueur moyenne des séquences pour chaque banque est de 289,3 pb pour les feuilles et 215,6 pb pour les racines. L'assemblage des deux banques individuellement, a permis la construction de 6419 et 5288 contigs pour les feuilles et les racines de *S.* x *townsendii* avec une longueur moyenne des contigs de 575,4 et 352,8 pb respectivement. Les séquences des deux organes ont ensuite été réunies afin d'obtenir un transcriptome de référence pour l'hybride *S.* x *townsendii.* Cet assemblage a permis de générer 12 696 contigs d'une longueur moyenne de 490.7pb. L'annotation fonctionnelle des séquences a été réalisée par alignement (BLASTn et tBLASTx) avec des banques d'EST de Poacées : tout d'abord avec une base de données d'*Oryza sativa*, puis une deuxième base de données englobant les ESTs de quatre Poaceae séquencées : *Oryza sativa*, *Brachypodium distachyon*, *Sorghum bicolor* et *Zea mays*. Cette dernière permet l'annotation d'un plus grand nombre de contigs à l'aide d'un alignement tBLASTx par rapport à l'utilisation de la base de données d'*Oryza sativa* seule. Ainsi, 8 712 contigs sont annotés (soit près de 70% des séquences consensus représentant 5265 unigènes), alors que 8071 contigs correspondent à un hit dans la base de données d'EST d'*Oryza sativa* (soit 4865 unigènes).

Le séquençage des transcriptomes de feuilles et de racines pour l'hybride F1 *S.* x *neyrautii* a généré, respectivement, 191 656 et 175 921 séquences soit 39,6 et 49,1 Mb de données (Tableau 6). La longueur moyenne des séquences pour chaque banque est de 206,7pb et 279,3pb. Chaque banque a été assemblée séparément, les ADNc de feuilles ont permis de construire 1820 séquences consensus alors que les ADNc de racines ont permis d'assembler 7301 contigs. Le transcriptome de référence de l'hybride *S.* x *neyrautii* incluant les séquences de feuilles et de racines comprend 11033 séquences consensus d'une longueur moyenne de 488,6

pb. Ces séquences ont ensuite été annotées à l'aide de bases de données de Poaceae connues : l'alignement (tBLASTx) sur les ESTs de riz a permis d'annoter 7737 contigs (représentant 70% des contigs assemblés) soit 4371 unigènes et l'alignement contre les EST de quatre Poaceae a permis d'annoter 7783 contigs (70% des séquences consensus) soit 4741 unigènes.

La duplication du génome chez *S.* x *townsendii* a donné naissance à l'allododécaploïde *S. anglica.* Le séquençage haut-débit de son transcriptome a permis de générer 191 499 séquences pour la banque d'ADNc de feuilles (soit 57,8 Mb de données) et 123146 séquences pour les racines (soit 24,5 Mb) (Tableau 6). La longueur moyenne des séquences pour ces deux banques est respectivement de 302,0 et 199,3 pb. L'assemblage de chaque banque a permis de construire 2383 et 3309 séquences consensus d'une longueur moyenne de 461,9 et 316,6 pb, respectivement. Ensuite les deux banques ont été assemblées conjointement afin de construire le transcriptome de référence pour *S. anglica* comprenant 6752 contigs d'une longueur moyenne de 383,2 pb. L'annotation fonctionnelle des séquences a été réalisée par alignement (BLASTn et tBLASTx) avec des banques d'EST d'*O. sativa* et de Poaceae : respectivement, 3966 (soit 59% de l'ensemble des contigs) et 3849 (soit 57%) séquences consensus ont pu être annotées, ce qui représente 2490 et 2617 unigènes.

Afin d'assembler et d'annoter le plus grand nombre de gènes chez les Spartines, les transcriptomes des cinq espèces du complexe ont été réunis. Concernant le transcriptome de référence issu des ADNc de feuilles, nous disposons d'un total de 1 336 894 séquences (soit 344,0 Mb) d'une longueur moyenne de 257,3 pb (Tableau 6). L'assemblage de ces séquences a permis la construction de 31123 contigs d'une longueur moyenne de 529,5pb. Parmi ces séquences consensus, 20 155 (soit 65%) et 21 086 (soit 68%) ont été annotées respectivement avec les ESTs de riz et de Poaceae, ce qui représente 10 006 et 10 832 unigènes. Pour le transcriptome de référence des racines de Spartines, 1 147 856 séquences (soit 300,6Mb) ont été générées, d'une longueur moyenne de 261,8 pb. De ces séquences, 49 945 contigs ont été assemblés d'une longueur moyenne de 716,2 pb. Les bases de données de riz et de Poaceae ont permis d'annoter respectivement 37 764 et 38 309 contigs soit 76% et 77% des séquences consensus. Les annotations fonctionnelles ont permis ainsi de mettre en évidence 14 881 et

16 315 unigènes. Enfin, les séquences générées pour les cinq espèces de Spartines sur les deux organes feuilles et racines ont été assemblées conjointement ce qui a généré 52347 contigs d'une longueur moyenne de 389,6pb. Parmi ces séquences consensus, 30 512 (soit 58%) ont été alignées sur des EST de riz et 30 205 (soit 60%) sur des EST de Poaceae. Les gènes différents trouvés dans ces bases de données sont au nombre de 19 806 et 20 583, respectivement.

Tableau 6 : Résumé des résultats de séquençage, d'assemblage et d'annotation chez les hybrides F1, l'allopolyploïde et l'ensemble des cinq espèces du complexe polyploïde étudié.

ANALYSIS	SEQUENCING		ASSEMBLIES				ANNOTATIONS			
	Number of reads sequenced (total number of nucleotides)	Mean size (bp)	Number of reads used in the assembly	Number of contigs	Mean size (bp)	Number of singletons	BlastN Oryza sativa	BlastN Poaceae	tBlastX Oryza sativa	tBlastX Poaceae
S. x *townsendii*										
Leaves	200342 (57,9Mb)	289,3	194924	6419 (3,7Mb)	575,4	50475	4430 (2243)	4816 (2786)	4573 (2409)	4810 (3027)
Roots	122431 (26,4Mb)	215,6	118742	5288 (1,9Mb)	352,8	44971	2834 (1421)	3356 (1987)	2713 (1449)	3164 (2123)
Leaves & roots	322773 (84,4Mb)	261,4	313666	12696 (6,2Mb)	490,7	77353	7902 (3657)	8071 (4865)	8850 (3879)	8712 (5265)
S. x *neyrautii*										
Leaves	191656 (39,6Mb)	206,7	180045	1820 (0,8Mb)	438,7	50935	905 (563)	1077 (733)	923 (591)	1024 (742)
Roots	175921 (49,1Mb)	279,3	163871	7301 (4,1Mb)	559,4	58641	5233 (2629)	5846 (3342)	5471 (2856)	5932 (3664)
Leaves & roots	367577 (88,8Mb)	241,5	343916	11033 (5,4Mb)	488,6	98004	6822 (3357)	7737 (4371)	7165 (3626)	7783 (4741)
S. anglica										
Leaves	191499 (57,8Mb)	302	183445	2383 (1,1Mb)	461,9	46665	1342 (793)	1482 (989)	1359 (846)	1445 (1037)
Roots	123146 (24,5Mb)	199,3	114585	3309 (1,0Mb)	316,6	49534	1515 (884)	1907 (1229)	1431 (889)	1750 (1256)
Leaves & roots	314645 (82,4Mb)	261,8	298030	6752 (2,6Mb)	383,2	80274	3432 (1890)	3966 (2490)	3463 (1970)	3849 (2617)
All *Spartina* species										
Leaves	1336894 (344,0Mb)	257,3	1 070 258	31123 (16,4Mb)	529,5	202 896	18439 (7 351)	20155 (10 006)	19090 (7 978)	21086 (10 832)
Roots	1147856 (300,6Mb)	261,8	1 000 179	49945 (35,8Mb)	716,2	176 022	34728 (10 196)	37764 (14 881)	36434 (11 300)	38309 (16 315)
Leaves & roots	2484750 (644,6Mb)	259,1	2 017 060	52347 (20,4Mb)	389,6	224 600	25995 (11 670)	30512 (19 806)	27414 (12 926)	31205 (20 583)

Discussion Générale sur l'assemblage de transcriptome et l'annotation

La technologie 454 (Roche-Life Sciences) offre une longueur moyenne de séquences plus grande que la plupart des plateformes concurrentes. Il est alors possible d'assembler et d'annoter *de novo* des ESTs d'espèces non modèles qui, jusqu'alors, présentaient des ressources génomiques limitées (Morozova *et al.*, 2009). Ce travail a ainsi permis l'assemblage du premier transcriptome de référence pour les espèces hexaploïdes et leurs hybrides et allopolyploïde récents.

L'assemblage et l'annotation de la totalité des séquences pour les cinq espèces polyploïdes permettent de générer 52 347 contigs incluant près de 90% des séquences utilisées lors de l'assemblage (le reste étant des singletons). Chez *S. pectinata*, 65% des séquences générées avaient été assemblées en contigs (Gedye *et al*., 2010) d'une longueur moyenne de 394 pb. Chez les Spartines hexaploïdes et leurs hybrides, la longueur moyenne des séquences consensus est de 390pb ce qui correspond aux résultats observés dans des études similaires (299 pb chez *Oryza longistaminata*, Yang *et al*., 2010). Parmi ces séquences consensus, près de 60% d'entre elles ont été annotées ce qui représente 20 583 unigènes (gènes codants différents). Cette proportion est en accord avec ce qui peut être observé chez des espèces non-modèles (69,8% chez *Panax quinquefolius*, Sun *et al*., 2010 ; 72,6% chez *S. pectinata*, Gedye *et al*., 2010) mais comparativement plus faible que ce qui est observé pour le transcriptome de référence des espèces hexaploïdes (69%, Ferreira de Carvalho *et al*., 2013 ; Partie A). Néanmoins, le transcriptome global permet d'annoter plus de gènes différents chez les Spartines grâce, notamment à l'alignement des séquences protéiques (tBLASTx) qui permet d'identifier un plus grand nombre de gènes homologues chez les Poaceae. De plus, l'utilisation de plusieurs bases de données combinées d'espèces phylogénétiquement proches améliore considérablement le nombre de gènes identifier comme cela a déjà été démontré chez d'autres espèces non-modèles (Barakat *et al*., 2009 ; Gedye *et al*., 2010 ; Franssen *et al*., 2011 ; Garg *et al*., 2011).

L'efficacité du séquençage peut être évaluée en comparant le nombre d'unigènes trouvés chez l'espèce non-modèle avec d'autres espèces séquencées phylogénétiquement proches (Parchman *et al*., 2010). Chez les Spartines nous disposons actuellement de 20 583 unigènes identifiés, ce qui représente entre 71% et 86% par rapport aux 28 236 gènes

codants trouvés chez *O. sativa* (RAP2, Rice Annotation Project, 2008) et ceux trouvés chez *Setaria italica* (entre 24 000 et 29 000 ; Bennetzen *et al*., 2012). *Sorghum bicolor* possède un nombre de gènes intermédiaire, estimé à 27 640 gènes codants (v1.4 Paterson *et al*., 2009). Ainsi, en utilisant une base de données de plusieurs Poacées séquencées, plus de la moitié des gènes potentiellement présents chez les Spartines hexaploïdes sont identifiés. Dans le but d'annoter un plus grand nombre de gènes, il serait nécessaire d'élargir l'échantillonnage en prenant des individus à des stades de développement (feuilles très jeunes) et des tissus (inflorescence, rhizomes) différents.

Cette nouvelle base de données d'EST de Spartines va permettre, outre le fait de pouvoir disposer des séquences pour des gènes ou des voies métaboliques d'intérêt, des études comparatives entre les différentes lignées de Spartines (hexaploïdes et tétraploïdes) et entre des représentants de la sous-famille peu connue des Chloridoideae et les autres sous-familles de Poaceae. Ce premier transcriptome va aussi pouvoir servir de référence afin d'aligner des données de RNA-Seq. Ces séquences de faible taille mais produites en masse vont ainsi permettre d'une part l'identification des copies homéologues chez les polyploïdes, et d'autre part leur expression relative chez les parents hexaploïdes, leurs hybrides et allopolyploïde récent en conditions contrôlées et naturelles. Les contigs assemblés pourront également être utilisés pour la conception de puces à ADN spécifiques des espèces étudiées.

Chapitre 5

Etude de la variation de l'expression globale de 13 gènes d'intérêt dans les populations naturelles de cinq espèces de Spartines

Introduction et démarche méthodologique

Cette étude vise à explorer l'amplitude de la variation de l'expression des gènes entre individus d'une même population, et entre populations d'une même espèce ou d'espèces différentes de Spartines en conditions naturelles. Compte tenu de la complexité des effets internes (variation et déterminismes génétiques, régulation de l'expression, stade de développement, état physiologique) ou externes (variabilité spatio-temporelle des conditions environnementales) déterminant les niveaux d'expression des gènes, la plupart des études transcriptomiques sont effectuées en conditions contrôlées.

Chez les Spartines, l'évolution de l'expression des gènes étudiés à ce jour a été réalisée à partir d'individus maintenus en même conditions de culture en serre expérimentale. En utilisant des puces à ADN constituées de sondes désignées sur le génome du riz, il a été ainsi possible de comparer les différences d'expression de plusieurs centaines de gènes (21 509 gènes sur une puce 44K Rice Agilent) dans les feuilles de *S. maritima* et *S. alterniflora* (Chelaifa *et al.,* 2010a) et d'analyser les niveaux d'expression de ces gènes chez leurs hybrides F1 *S.* x *neyrautii*, *S. x townsendii* et son descendant allopolyploïde *S. anglica*. Ces investigations ont permis une première évaluation des effets respectifs de l'hybridation interspécifique et de la duplication du génome sur le transcriptome (Chelaifa *et al.,* 2010b). Ces résultats ont montré des effets différents de l'hybridation et de la duplication du génome : chez les hybrides F1, une dominance de l'expression maternelle est observée, celle-ci s'atténue ensuite chez l'allododécaploïde. Ce dernier se distingue par un nombre élevé de gènes transgressivement surexprimés.

Des analyses transcriptomiques ont également été effectuées sur des individus de *S. alterniflora* soumis à un stress expérimental ce qui a permis de détecter par PCR quantitative l'expression de certains gènes impliqués dans la tolérance à la salinité (Baisakh *et al.,* 2008) ou aux hydrocarbures (Ramanarao *et al.*, 2012). En effet, un nombre important d'ESTs codants pour des facteurs de transcriptions, des transporteurs d'ions, des osmoprotecteurs, des antioxydants et des enzymes de détoxification ont été identifiés chez *Spartina alterniflora* et montrent des variations d'expression spatio-temporelles et tissu-dépendantes en fonction des conditions de salinité (Baisakh *et al.*, 2008 ; Subudhi & Baisakh, 2011). De plus, soumis à des hydrocarbures, *Spartina alterniflora* montre des gènes différentiellement

exprimés dans les feuilles et les racines (Ramanarao *et al.*, 2012). Ceux-ci présentent des niveaux réduits de transcrits dans les racines en conditions de stress mais une expression transgressive (réprimée ou activée) dans les feuilles suivant les concentrations en hydrocarbures.

Toutefois, les variations d'expression des gènes en conditions naturelles représentent un paramètre important à prendre en compte dans l'adaptation des espèces. Dans cette partie du travail de thèse, nous posons la question de savoir (1) si chaque espèce présente un profil d'expression caractéristique en dépit des fluctuations d'expression résultant de la variabilité des conditions stationnelles dans lesquelles les individus ont été récoltés, et (2) quelle peut être l'amplitude de la variation d'expression des gènes au sein des espèces d'origine hybride (*S.* x *townsendii* et *S.* x *neyrautii*) et allopolyploïde (*S. anglica*) par rapport à leurs parents (*S. alterniflora* et *S. maritima*).

Nous avons sélectionné pour cette étude 13 gènes d'intérêt présentant des fonctions métaboliques intéressantes (réponse au stress salin et oxydant, tolérance aux métaux lourds et métabolisme de la croissance cellulaire) et également des gènes montrant des niveaux d'expression variant au cours de la spéciation allopolyploïde qui avaient été étudiés en mêmes conditions de culture par microarrays de riz (Chelaifa *et al.,* 2010b). Ces niveaux d'expression ont été comparés chez plusieurs individus échantillonnés le long d'un transect amont-aval dans des populations naturelles des 5 espèces de Spartines étudiées.

De plus, les Spartines étant toutes polyploïdes, nous nous attendons à amplifier plusieurs copies homéologues du même gène. Dans cette étude, les polymorphismes nucléotidiques ont été identifés pour un gène (codant la metal tolerance protein) par une méthode classique de PCR et clonage (d(ADNc et ADN génomique) ainsi que par une analyse *in silico* des données transcriptomiques 454. Cette partie fera l'objet d'un article qui intégrera des données complémentaires sur la détection *in silico* des copies dupliquées sur tous les gènes analysés, selon une procédure mise au point au laboratoire (J. Boutte et A. Salmon).

Matériel et Méthodes

Grâce à la construction du transcriptome de référence pour les espèces hexaploïdes de Spartines (Ferreira de Carvalho *et al*., 2013, Chapitre 4), des fonctions et des gènes d'intérêt ont pu être identifiés. Nous avons alors sélectionné des gènes correspondant à des fonctions jouant un rôle dans les réponses aux stress salin, oxydant, et de tolérance aux métaux lourds et des gènes entrant dans le métabolisme de la lignine et de la cellulose. Enfin, à partir des listes de gènes différentiellement exprimés entre les parents hexaploïdes, les hybrides F1 et l'allododecaploïde mis en évidence par Chelaifa *et al.* (2010 a et b) et des séquences du transcriptome de référence, nous avons pu retrouver les homologues aux sondes correspondantes des puces de riz pour identifier et quantifier ces gènes dans les populations naturelles de Spartines. Validant tous les critères de la PCR quantitative, 13 gènes ont été choisis ainsi que 3 gènes de ménage (Sucrose synthase, alpha-tubulin et Glycéraldéhyde 3-Phosphate Dehydrogenase).

Les échantillons d'ADNc quantifiés proviennent de cinq espèces de Spartines récoltées sur différents transects selon un gradient d'immersion qui va du bas vers le haut de l'estran. Sur chacun de ces transects, les feuilles de 6 à 8 plantes de différents clones ont été échantillonnées. *Spartina anglica* a été échantillonnée à Roscoff et dans l'Anse de Goulven (Finistère). Des feuilles de *S.* x *neyrautii* et de *S.* x *townsendii* ont aussi été récoltées, respectivement à Hendaye (Pyrénées Atlantiques) et à Hythe (Hampshire, Angleterre). *Spartina alterniflora* a été échantillonnée à Landerneau (Finistère) et *S. maritima* à Noirmoutier (Vendée) et sur le site de Quenouille (Morbihan) en Juillet 2010.

Les espèces analysées étant polyploïdes, plusieurs copies dupliquées (homéologues) sont attendues par locus. Les analyses effectuées en PCRq mesurent donc l'expression globale (ensemble des copies homéologues) à un locus donné. Nous nous sommes posé la question du nombre de copies détectables par locus en prenant pour exemple un gène codant une metal tolerance protein. Ce gène a été analysé in silico à partir des séquences formant le contig et d'un programme de détection d'haplotypes mis au point au sein du laboratoire (J Boutte et A. Salmon ; Boutte *et al*., *en préparation*). Les séquences obtenues pour *S. maritima* et *S. alterniflora* sont comparées dans le but de détecter les différents haplotypes (séquences caractérisées par plusieurs sites polymorphes) grâce à l'alignement et

l'inspection visuelle de la matrice à l'aide du logiciel BioEdit (Ibis Biosiences). De plus, une analyse phylogénétique (méthode de parcimonie) des séquences a été effectuée à l'aide du logiciel PAUP (Swofford, 2003).

Résultats

Les variations d'expression des populations naturelles de Spartines ont été analysées pour les 13 gènes d'intérêt par PCR quantitative. Les niveaux moyens (deux réplicats techniques par individu) d'expression ainsi que l'erreur standard pour chaque espèce sont résumés dans le Tableau 7. Pour tous les gènes, aucune variation significative d'expression intrapopulation n'a été observée (Test de Kruskal et Wallis avec pvalue >0.05). Les individus au sein d'une même population montrent des niveaux d'expression homogènes (Figure 10). Dans la suite des résultats nous comparerons ces niveaux d'expression moyens entre populations d'une même espèce et entre espèces de Spartines.

Variations d'expression entre les populations d'une même espèce

Chez *S. maritima*, des différences significatives entre les deux populations étudiées sont observées pour 5 gènes : **Cytochrome c oxidase subunit 6b, V-type proton ATPase catalytic subunit, Transcription Elongation factor spt6, C2H2 Zinc Finger Protein A2** et le **WRKY24** (Figure 11, Tableau 7). Les trois premiers gènes avaient montré une variation d'expression au cours de la spéciation allopolyploïde ches les Spartines dans les analyses par microarrays, les deux derniers participent à la réponse au stress salin. Il est intéressant de noter que dans chacun des cas, la population de *S. maritima* provenant du site de Quenouille (rivière d'Etel, Morbihan) montre des niveaux d'expression significativement plus élevés que la population provenant du site de Noirmoutier (Vendée).

Chez *S. anglica*, des différences d'expression sont observées entre les populations pour 5 gènes: 3 gènes dont l'expression varie au cours de la spéciation allopolyploïde (**Transcription Elongation factor spt6, Hexokinase 1 et un MYB family transcription factor**) et 2 gènes réagissant au stress salin et aux métaux lourds (**WRKY24 et Metal tolerance protein A2**) (Figure 12, Tableau 7). Le site de Goulven (Finistère) présente des niveaux d'expression plus élevés que la population du site de Roscoff (Finistère) excepté pour le gène **WRKY24**.

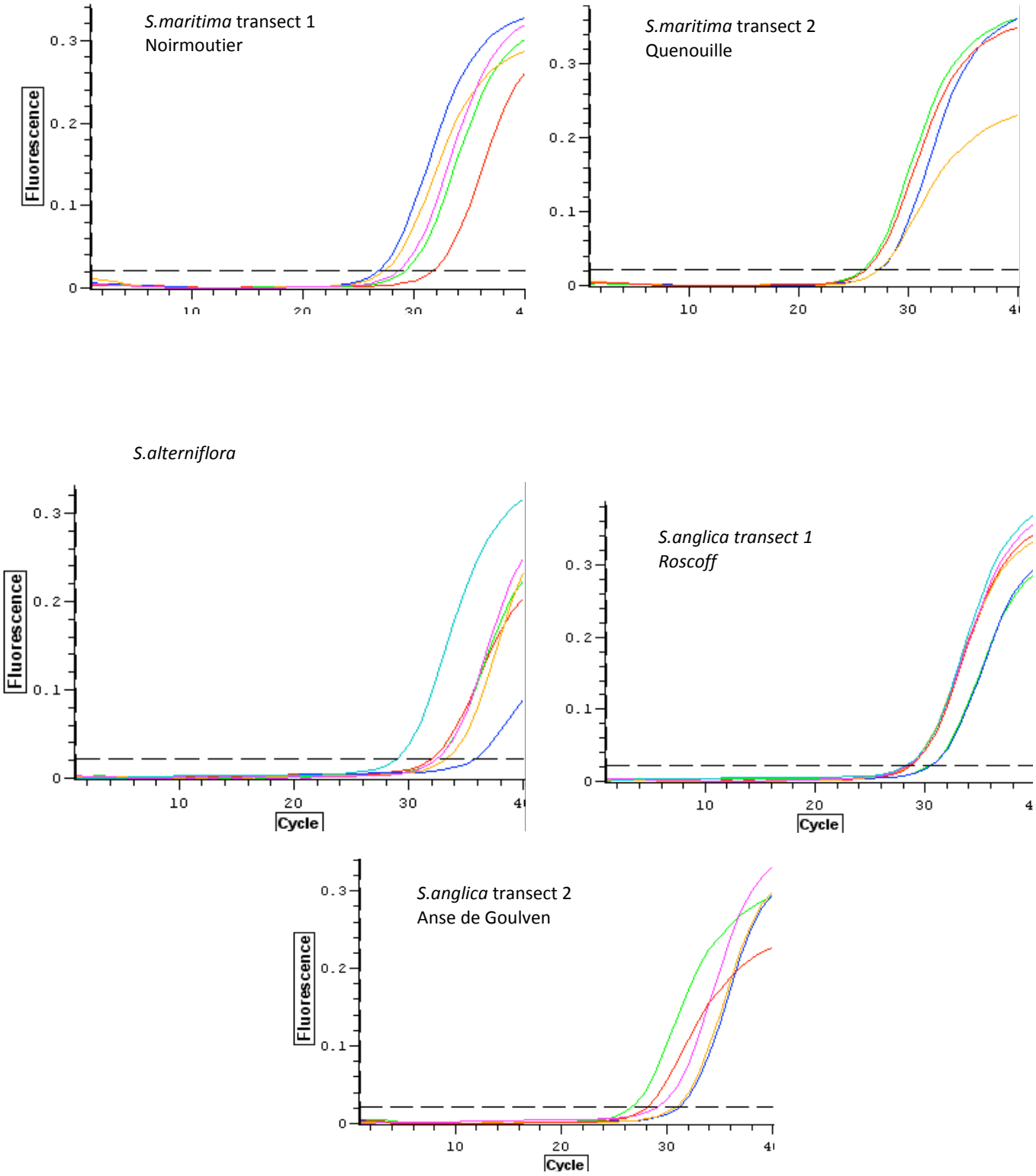

Figure 10 : Courbes d'amplification du gène codant la xanthine déhydrogénase 1 pour chaque transect analysé (chaque courbe représente un individu du transect) chez *Spartina maritima*, *S. alterniflora* et *S. anglica*.

Tableau 7 : Tableau récapitulatif des niveaux moyens d'expression (SE=Erreur Standard) des Spartines pour chacun des gènes analysés par PCR quantitative. Les résultats des tests statistiques ANOVA sont résumés pour chaque gène entre les populations de *S. anglica*, les populations de *S. maritima* et les 5 espèces de Spartines analysés (pvalue).

Gene annotation	*Spartina alterniflora*	*Spartina anglica*				*Spartina maritima*				*Spartina* x *neyrautii*	*Spartina* x *townsendii*	ANOVA 2p (pvalue)
		Species	Goulven	Roscoff	Anova (pvalue)	Species	Noirmoutier	Quenouille	Anova (pvalue)			
Salt Stress response												
Xanthine dehydrogenase1	96,65 (SE=30,89)	42,73 (SE=8,35)	55,21 (SE=17,64)	32,33 (SE=3,23)	0,11	148,72 (SE=27,64)	113,93 (SE=24,37)	192,21 (SE=52,29)	0,15	96,94 (SE=16,01)	42,53 (SE=9,59)	$4,48.10^{-05}$***
Transcriptional adpater ADA2	$0,29.10^{4}$ ($SE=0,08.10^{4}$)	$0,43.10^{4}$ ($SE=0,06.10^{4}$)	$0,33.10^{4}$ ($SE=0,04.10^{4}$)	$0,51.10^{4}$ ($SE=0,12.10^{4}$)	0,09	$0,42.10^{4}$ ($SE=0,10.10^{4}$)	$0,16.10^{4}$ ($SE=0,10.10^{4}$)	$0,41.10^{4}$ ($SE=0,10.10^{4}$)	0,77	$0,39.10^{4}$ ($SE=0,08.10^{4}$)	$0,13.10^{4}$ ($SE=0,07.10^{4}$)	0,13
C2-h2 zinc finger protein STOP1	$0,36.10^{4}$ ($SE=0,09.10^{4}$)	$0,47.10^{4}$ ($SE=0,06.10^{4}$)	$0,47.10^{4}$ ($SE=0,08.10^{4}$)	$0,46.10^{4}$ ($SE=0,09.10^{4}$)	0,92	$0,48.10^{4}$ ($SE=0,05.10^{4}$)	$0,32.10^{4}$ ($SE=0,03.10^{4}$)	$0,68.10^{4}$ ($SE=0,02.10^{4}$)	$2,2.10^{-16}$***	$0,05.10^{4}$ ($SE=0,01.10^{4}$)	$0,04.10^{4}$ ($SE=0,01.10^{4}$)	$5,30.10^{-10}$***
WRKY24	145,35 (SE=35,05)	135,57 (SE=14,92)	102,91 (SE=9,41)	162,78 (SE=23,88)	0,01**	93,42 (SE=6,57)	79,65 (SE=7,77)	110,62 (SE=7,90)	0,0002***	90,38 (SE=24,01)	25,36 (SE=4,78)	$1,26.10^{-05}$***
Heavy metal Stress response												
Metal tolerance protein A2	156,39 (SE=35,76)	19,59 (SE=2,50)	25,74 (SE=4,37)	14,46 (SE=1,93)	0,005**	45,66 (SE=9,01)	46,87 (SE=15,07)	44,14 (SE=8,86)	0,88	17,66 (SE=5,93)	13,28 (SE=2,26)	$3,4.10^{-14}$***
Lignin & Cellulose metabolism												
Cinnamoyl- CoA reductase	532,28 (SE=78,98)	587,99 (SE=57,24)	655,03 (SE=52,31)	532,12 (SE=94,75)	0,283	294,82 (SE=27,35)	308,67 (SE=42,22)	277,52 (SE=33,83)	0,84	60,13 (SE=7,23)	116,76 (SE=20,95)	$2,2.10^{-16}$***
GDP-mannose pyrophosphorylase	$4,73.10^{4}$ ($SE=8,97.10^{4}$)	$9,06.10^{4}$ ($SE=2,19.10^{4}$)	$6,93.10^{4}$ ($SE=3,31.10^{4}$)	$10,84.10^{4}$ ($SE=2,94.10^{4}$)	0,41	$4,89.10^{4}$ ($SE=1,07.10^{4}$)	$5,13.10^{4}$ ($SE=1,73.10^{4}$)	$4,58.10^{4}$ ($SE=1,21.10^{4}$)	0,1	$4,78.10^{4}$ ($SE=1,08.10^{4}$)	$1,06.10^{4}$ ($SE=0,24.10^{4}$)	0,020**
Expression altered during allopolyploid speciation												
Cytochrome c oxidase subunit 6b	$151,23.10^{4}$ ($SE=18,87.10^{4}$)	$103,16.10^{4}$ ($SE=8,97.10^{4}$)	$95,80.10^{4}$ ($SE=5,80.10^{4}$)	$109,29.10^{4}$ ($SE=15,84.10^{4}$)	0,24	$50,44.10^{4}$ ($SE=5,43.10^{4}$)	$48,66.10^{4}$ ($SE=6,83.10^{4}$)	$52,67.10^{4}$ ($SE=9,19.10^{4}$)	$2,2.10^{-16}$***	$15,78.10^{4}$ ($SE=1,43.10^{4}$)	$27,25.10^{4}$ ($SE=5,52.10^{4}$)	$2,2.10^{-16}$***
Transcription elongation factor spt6	$8,60.10^{4}$ ($SE=4,34.10^{4}$)	$0,46.10^{4}$ ($SE=0,17.10^{4}$)	$0,83.10^{4}$ ($SE=0,34.10^{4}$)	$0,15.10^{4}$ ($SE=0,07.10^{4}$)	0,005**	$0,97.10^{4}$ ($SE=0,52.10^{4}$)	$0,12.10^{4}$ ($SE=0,07.10^{4}$)	$2,03.10^{4}$ ($SE=1,08.10^{4}$)	0,02*	NA	NA	0,003***
Hexokinase 1	416,62 (SE=200,35)	16,47 (SE=1,87)	21,51 (SE=2,86)	12,27 (SE=1,78)	0,002**	15,41 (SE=3,25)	16,08 (SE=5,17)	14,57 (SE=3,83)	0,82	40,85 (SE=14,59)	22,51 (SE=2,99)	0,0001***
Myb family transcription - transfactor	$237,49.10^{4}$ ($SE=63,65.10^{4}$)	$21,31.10^{4}$ ($SE=3,24.10^{4}$)	$29,32.10^{4}$ ($SE=4,16.10^{4}$)	$14,63.10^{4}$ ($SE=4,01.10^{4}$)	0,005**	$62,31.10^{4}$ ($SE=20,95.10^{4}$)	$76,90.10^{4}$ ($SE=36,30.10^{4}$)	$44,06.10^{4}$ ($SE=13,92.10^{4}$)	0,33	$0,99.10^{4}$ ($SE=0,59.10^{4}$)	$8,91.10^{4}$ ($SE=3,50.10^{4}$)	$31,7.10^{-10}$***
Metallocarboxypeptidase inhibitor	59,16 (SE=9,51)	27,42 (SE=7,09)	22,58 (SE=3,84)	31,46 (SE=12,76)	0,37	31,23 (SE=6,69)	26,67 (SE=10,41)	36,94 (SE=7,88)	0,45	96,76 (SE=19,42)	115,63 (SE=43,88)	0,0008***
V-type proton ATPase catalytic subunit A	408,43 (SE=240,65)	1742,86 (SE=529,38)	2099,42 (SE=929,55)	1445 (SE=611,61)	0,2	1143,66 (SE=423,76)	89,09 (SE=57,57)	2461,87 (SE=727,38)	$3,5.10^{-05}$***	272,57 (SE=18,55)	208,02 (SE=53,77)	0,1642

Parmi ces gènes, deux d'entre eux **(Transcription Elongation factor spt6** et **WRKY24**) montrent des différences significatives d'expression entre les populations analysées des deux espèces *S. maritima* et *S. anglica*. Ces gènes montrent une variabilité d'expression très large chez les Spartines et sont probablement essentiels et induits dans une grande variété de signaux environnementaux.

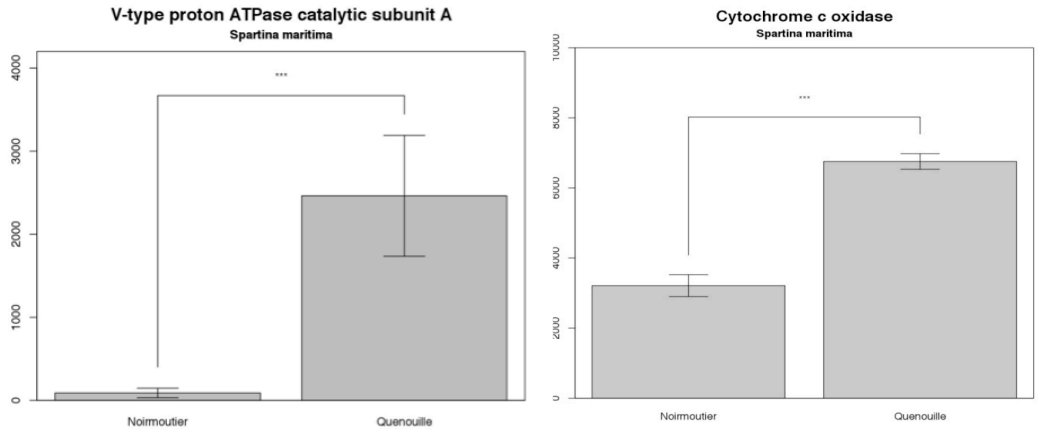

Figure 11 : Gènes montrant une expression différentielle entre les populations analysées de *Spartina maritima* (V-type proton ATPase catalytic subunit A pvalue<0,0001 ; Cytochrome c oxidase subunit 6b pvalue<0.001).

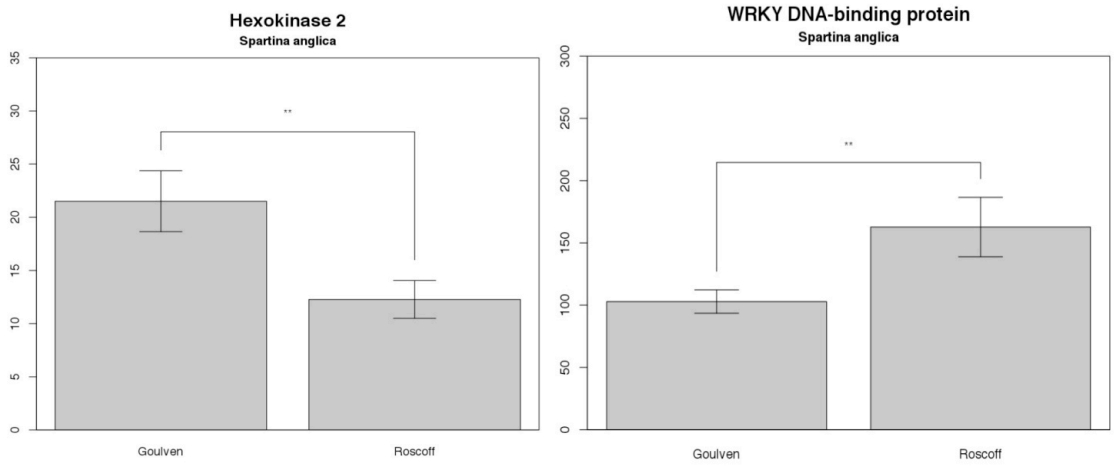

Figure 12 : Gènes montrant une expression différentielle entre les populations analysées de *Spartina anglica* (Hexokinase 1 pvalue<0,001 ; WRKY24 DNA-Binding Protein pvalue<0.001).

Variations d'expression entre les parents

Les niveaux d'expression sont significativement plus élevés chez *S. alterniflora* (par rapport à l'autre parent hexaploïde *S. maritima*) pour 7 des gènes analysés (**Metal Tolerance protein A2, MYB family transcription factor, Hexokinase 1, Transcription Elongation factor spt6, WRKY24, Cytochrome c oxidase subunit 6b, Cinnamoyl-CoA reductase**). Pour les 6 gènes restants, les tests statistiques ne montrent pas de différence significative entre les espèces mais un niveau d'expression équivalent entre les parents.

Variations d'expression entre les hybrides F1 et les parents hexaploïdes

Les différences d'expression entre *S.* x *neyrautii* et *S.* x *townsendii* ne sont pas significatives pour 12 gènes des 13 analysés. Seul le gène **WRKY24** (gène entrant dans la réponse au stress salin) montre une différence significative d'expression entre les hybrides F1, ce gène étant sous-exprimé chez *S.* x *townsendii* par rapport à l'hybride basque.

En observant les variations d'expression entre les hybrides F1 et les deux parents hexaploïdes nous pouvons distinguer 2 possibilités de non-additivité parmi les gènes étudiés : une expression identique à l'un des deux parents chez les hybrides ; dans ce cas, elle peut être identique au parent paternel (*Spartina maritima*), c'est le cas pour 5 des gènes (**Hexokinase 1, Cytochrome c oxidase subunit 6b, MYB family transcription factor, Metal tolerance protein A2** et **WRKY24** pour *S.* x *neyrautii*) ou maternel (*Spartina alterniflora*) pour 1 des gènes (**Xanthine dehydrogenase 1** pour *S.* x *townsendii*) (Figure 13). Dans un deuxième cas, nous observons aussi chez les hybrides, une expression extrême (plus faible ou plus élevée que celle des parents) dans 4 des gènes analysés: plus faible chez 3 gènes : **WRKY24** (chez l'hybride anglais)**, C2H2 Zinc Finger Protein STOP1 et Cinnamoyl-CoA reductase** ou plus elevée chez le gène **Metallocarboxypeptidase inhibitor** (Figure 14). Les différences entre les niveaux d'expression des hybrides et des parents sont non significatives dans le cas des gènes **Transcriptional adapter ADA2, GDP-mannose pyrophosphorylase et V-type proton ATPase** et du gène **Xanthine dehydrogenase 1** chez l'hybride F1 *S.* x *neyrautii***.**

En comparant les hybrides F1 et la Mid Parent Value ou MPV (correspondant à la moyenne d'expression des parents), les niveaux d'expression observés chez *S.* x *townsendii* et *S.* x *neyrautii* sont inférieurs ou égaux à cette MPV (±IC à 95%) sauf dans le cas du gène

codant une **Metallocarboxypeptidase inhibitor** où les niveaux d'expression de *S.* x t*ownsendii* et *S.* x *neyrautii* sont significativement supérieurs à la MPV (Tableau 8).

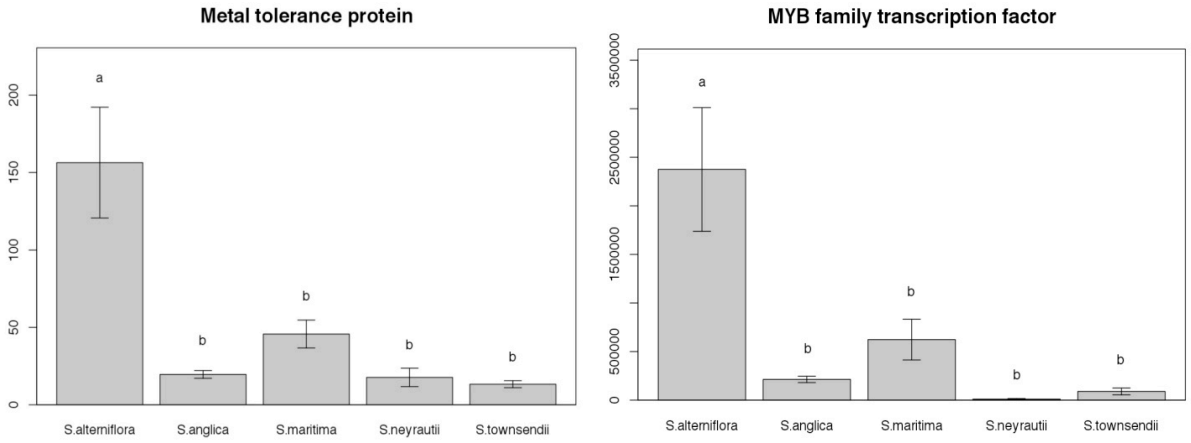

Figure 13 : Gènes montrant une expression différentielle entre les parents (*Spartina alterniflora*, *S. maritima)* et les hybrides F1 (*S.* x *townsendii* et *S.* x *neyrautii)* (Metal tolerance Protein A2 pvalue<0.001 ; MYB family transcription factor pvalue<0.0001).

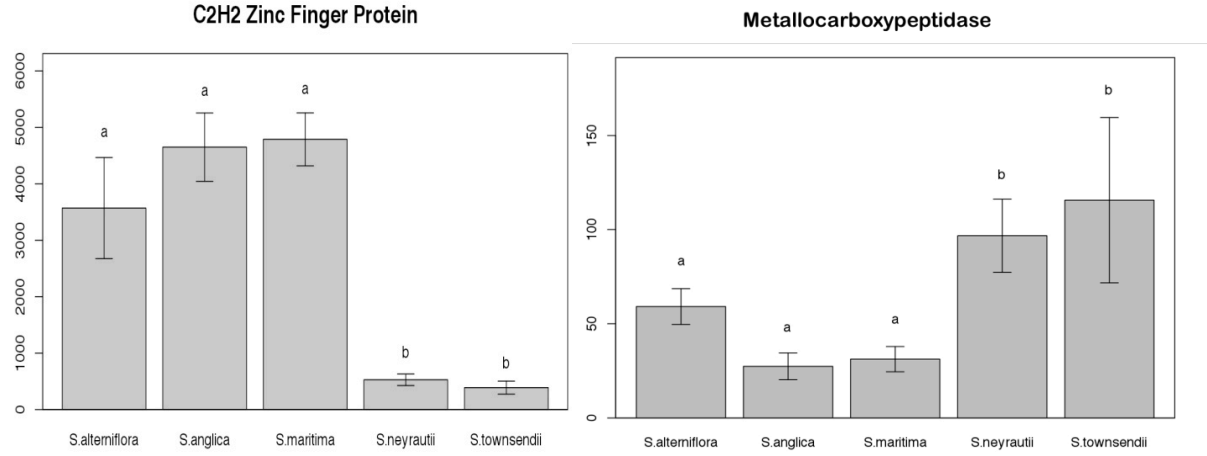

Figure 14 : Gènes montrant une expression différentielle entre les parents (*Spartina alterniflora*, *S. maritima)* et les hybrides F1 (*S.* x *townsendii* et *S.* x *neyrautii)* (C2H2 Zinc Finger Protein STOP1 pvalue<0.01 ; Metallocarboxypeptidase inhibitor pvalue<0.01).

Tableau 8 : Récapitulatif des données d'expression des hybrides F1, de la « Mid-Parent Value » et de l'intervalle de confiance à 95%. Les valeurs d'expression des hybrides sont surlignées en bleu quand elles se situent en dessous de la MPV (±IC à 95%) et surlignés en rouge quand les valeurs se situent au-dessus de la MPV (±IC à 95%).

Annotation	MPV	intervalle de confiance à 95%	*Spartina* x *neyrautii*	*Spartina* x *townsendii*
Xanthine dehydrogenase1	127,89	73,76 - 162,87	96,94 (SE=16,01)	42,53 (SE=9,59)
Transcriptional adapter ADA2	3767,37	1903,94-4579,00	$0,39.10^4$ (SE=$0,08.10^4$)	$0,13.10^4$ (SE=$0,07.10^4$)
C2-h2 zinc finger protein STOP1	4 325,84	3 264,06 - 5 209,03	$0,05.10^4$ (SE=$0,01.10^4$)	$0,04.10^4$ (SE=$0,01.10^4$)
WRKY24	114,19	77,46-137,74	90,38 (SE=24,01)	25,36 (SE=4,78)
Metal tolerance protein A2	89,95	39,10-114,50	17,66 (SE=5,93)	13,28 (SE=2,26)
Cinnamoyl- CoA reductase	389,81	301,80-424,92	60,13 (SE=7,23)	116,76 (SE=20,95)
GDP-mannose pyrophosphorylase	48 241,29	29 948,18-58 557,66	$4,78.10^4$ (SE=$1,08.10^4$)	$1,06.10^4$ (SE=$0,24.10^4$)
Cytochrome c oxidase subunit 6b	907 575,30	609 292,4-1 055 604,9	$15,78.10^4$ (SE=$1,43.10^4$)	$27,25.10^4$ (SE=$5,52.10^4$)
Transcriptional elongation factor Spt6	40 214,41	1 694,32- 32 059,69	NA	NA
Hexokinase 1	175,89	20,74-100,12	40,85 (SE=14,59)	22,51 (SE=2,99)
Myb family transcription factor	1 323 788	469 640,0 - 1 561 968,6	$0,99.10^4$ (SE=$0,59.10^4$)	$8,91.10^4$ (SE=$3,50.10^4$)
Metallocarboxypeptidase inhibitor	42,40	29,44-53,64	96,76 (SE=19,42)	115,63 (SE=43,88)
V-type proton ATPase catalytic subunit A	849,57	74,76 - 1 340,12	272,57 (SE=18,55)	208,02 (SE=53,77)

Variations d'expression entre *S. anglica* et *S.* x *townsendii*

L'expression chez *S. anglica* est surexprimée pour cinq des gènes analysés (**Cinnamoyl-CoA reductase, GDP-mannose pyrophosphorylase, Cytochrome c oxidase subunit 6b, WRKY24, C2H2 zinc finger protein STOP1**) par rapport aux niveaux d'expression chez l'hybride interspécifique *S.* x *townsendii* et six gènes montrent des niveaux d'expression similaires entre *S. anglica* et *S.* x *townsendii*. Seul le gène codant un **Metallocarboxypeptidase inhibitor** montre une expression significativement supérieure chez les hybrides par rapport à l'allopolyploïde.

Variations d'expression entre *S. anglica* et les deux parents

En comparant les niveaux d'expression entre l'allododécaploïde et les deux parents hexaploïdes, nous observons 4 cas :

1- les niveaux d'expression entre *S. anglica* et les parents hexaploïdes ne montrent pas de différence significative pour 5 gènes analysés (**C2H2 Zinc Finger Protein STOP1, Transcriptional adapter ADA2, Metallocarboxypeptidase inhibitor, GDP-mannose pyrophosphorylase et V-type proton ATPase**).

2- le niveau d'expression de l'allopolyploïde se situe entre les niveaux d'expression des deux parents hexaploïdes (dans deux cas **Cytochrome c oxidase subunit 6b, WRKY24** sur 13 gènes analysés).

3- Les niveaux d'expression chez *Spartina anglica* peuvent aussi correspondre au niveau d'expression d'un des deux parents : une expression similaire au parent maternel est observée pour un des gènes analysés (**Cinnamoyl-CoA reductase**) et une expression similaire au parent paternel pour 4 gènes (**Metal tolerance protein A2, Transcriptional adapter ADA2, MYB family transcription factor et Hexokinase 1**) (Figure 15).

4- Enfin, l'expression chez *S. anglica* est extrême dans 1 cas sur 13: elle est sous-exprimée par rapport aux deux parents pour le gène codant la **Xanthine dehydrogenase 1**.

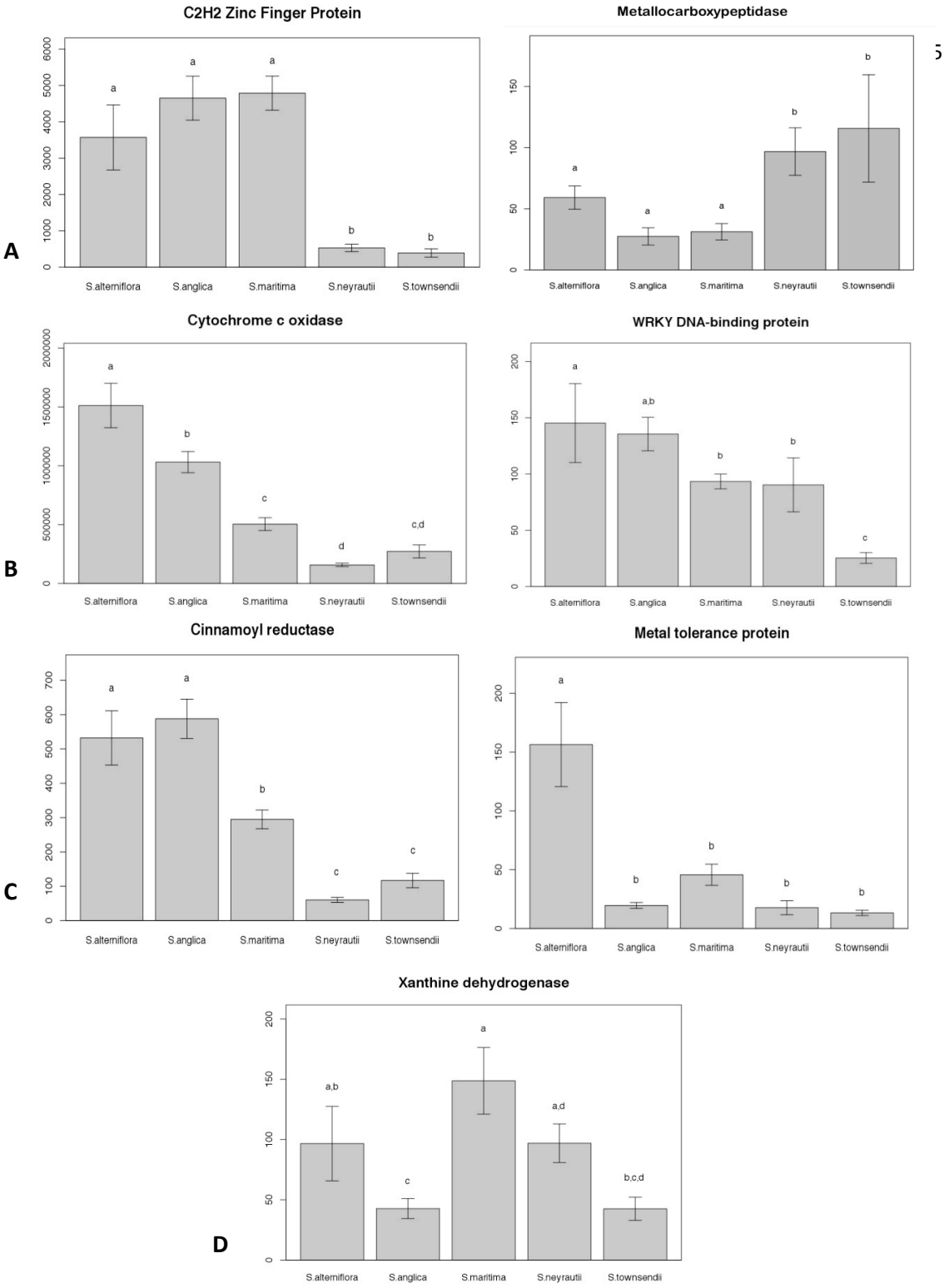

Figure 15 : Comparaisons des niveaux d'expression des gènes chez les cinq espèces de Spartines : (A) gènes ne montrant pas de différences significatives entre S. anglica et ses parents, (B) expression intermédiaire entre l'allopolyploïde et ses parents, (C) gènes montrant une expression similaire de S. anglica à l'un des parents, (D) gène montrant une expression extrême de *S. anglica* par rapport à ses parents (C2H2 Zinc Finger Protein STOP1 pvalue<0.01 ; Metallocarboxypeptidase inhibitor pvalue<0.01 ; Cytochrome c oxidase subunit 6b pvalue<0.001 ; WRKY24 DNA-Binding Protein pvalue<0.001 ; Cinnamoyl-CoA reductase pvalue<0.01 ; Metal tolerance Protein A2 pvalue<0.001 ; Xanthine dehydrogenase 2 pvalue<0.01).

Analyses des copies de gènes dupliqués chez les parents hexaploïdes pour le gène codant la Metal tolerance protein A2

Sur un total de 42 séquences initialement clonées, 30 séquences ont été analysées dont 11 séquences d'ADN génomique de *Spartina maritima*, 12 séquences d'ADNc de *Spartina maritima* et 7 séquences d'ADNc de *Spartina alterniflora*. Toutes les séquences d'ADN de *Spartina maritima* (ADN génomique et ADNc) obtenues ont une longueur de 454 pb après nettoyage et vérification visuelle des alignements. Six d'entre elles sont entièrement identiques, les autres présentent 1,2 ou 3 substitutions. Au sein des séquences d'ADNc de *Spartina alterniflora,* quatre d'entre elles sont de 467 pb, deux de 479 pb et une de 484 pb. Les deux séquences de 467 pb sont totalement identiques. L'analyse phylogénétique des séquences par la méthode de parcimonie (Figure 16) montre que les séquences isolées chez *S. maritima* forment deux clades (A et B) constitués respectivement de 14 et 8 clones. Au sein de chaque groupe, les séquences ne différent au maximum que d'un seul nucléotide. On note la présence d'une séquence (Smc3a, Figure 16) en position intermédiaire entre ces deux groupes. Après inspection de la matrice de données, cette séquence contient une portion typique du clade A et une autre typique du clade B, suggérant la présence d'une recombinaison (probablement *in vitro* lors de l'amplification) entre les types « A » et « B ». Les séquences amplifiées chez *S. alterniflora* montrent une plus grande hétérogénéité. On note un groupe de séquences très similaires (groupe C, Figure 16) qui montre une divergence nucléotidique importante des autres séquences, parmi lesquelles deux (sfc15d et sfc3d, groupe D, Figure 16) sont très similaires.

Dans le jeu de données de *S. maritima*, un contig (729 pb) a été identifié comme codant la Metal tolerance protein A2 formé de 15 séquences. Au total, 4 haplotypes ont pu être construits ; par fenêtre, un maximum de 3 haplotypes peut être observé (Tableau 9). Dans ce cas, deux des haplotypes sont très divergents (copies 2 et 3 dans le Tableau 9, 6 SNPs partagés) avec un troisième plus similaire à l'une des copies (copies 1 et 2 dans le Tableau 5, 1 SNP partagé). Dans le jeu de données de *S. alterniflora*, deux contigs (1437 et 1450 pb) ont été identifiés formés tous les deux de 29 séquences. Au total, 6 haplotypes ont pu être construits ; par fenêtre, un maximum de 4 haplotypes peut être observé (Tableau 9). Dans ce cas, deux des haplotypes sont très divergents (copies 2 et 3 dans le Tableau 9, 10 SNPs

partagés) et les deux autres présentent des copies ayant moins divergées pour chaque copie plus divergente (entre les copies 1 et 2 : 2 SNPs partagés et entre les copies 3 et 4 : 1 SNP partagé).

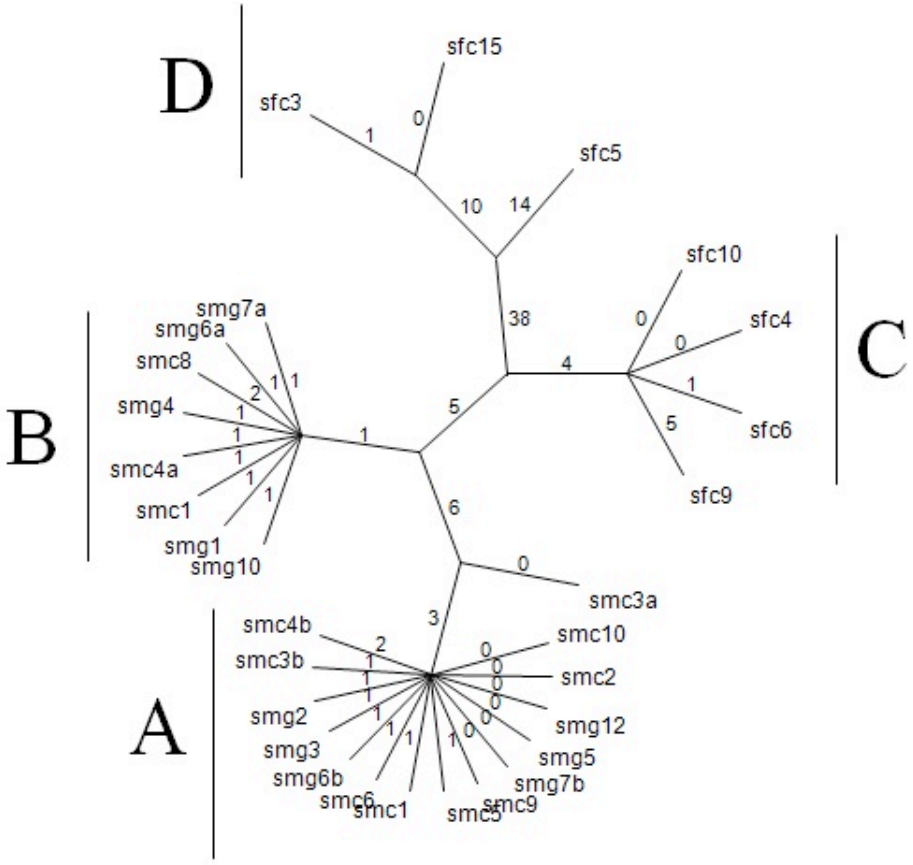

Figure 16 : Arbre phylogénétique des séquences clonées du gène Metal tolerance protein A2 (Méthode parcimonie). Les clades A et B incluent les séquences provenant de l'ADN génomique (g) et de l'ADN complémentaire de *Spartina maritima* (sm). Les clades C et D incluent les séquences provenant des ADN complémentaire de *Spartina alterniflora* (sf). Les chiffres au-dessus des branches représentent le nombre de changements nucléotidiques observés entre les séquences.

Tableau 9 : Polymorphismes nucléotidiques présents sur une partie du gène Metal Tolerance Protein alignée entre *Spartina maritima* et *S. alterniflora*. Les positions partagées entre les deux espèces sont alignées verticalement dans le tableau.

Metal Tolerance Protein														
Nucleotide position	745	746	749	750	753	772	777	778	779	781	782	787	788	818
***S. alterniflora* Isotig 01837 (length=1437, reads=29)**														
copy 1	A	T	C	A	T	C	T	G	T	C	T	A	A	C
copy 2	A	T	C	A	T	T	T	G	T	C	T	G	A	C
copy 3	T	A	A	G	C	T	T	T	T	T	C	A	T	C
copy 4	T	A	A	G	C	T	G	T	T	T	C	A	T	C
***S. maritima* Isotig 03435 (length=729, reads=15)**														
copy 1	C	A	A	G	C	T	G	T	A	T	G	A	A	C
copy 2	C	A	A	G	C	T	G	T	A	T	G	A	A	A
copy 3						T	T	G	T	A	T	A	A	C

Discussion

Les variations d'expression des populations naturelles de Spartines ont été analysées par PCR quantitative pour 13 gènes d'intérêt dans le but d'étudier l'amplitude des variations d'expression des espèces hexaploïdes *S. alterniflora*, *S. maritima*, des hybrides F1 *S.* x *townsendii et S.* x *neyrautii* ainsi que de l'allododécaploïde *S. anglica* en conditions *in situ* (Figure 17). A ce jour très peu d'études ont analysé les variations de l'expression des gènes en conditions naturelles. Signalons toutefois une étude récente sur deux accessions d'*Arabidopsis thaliana* en conditions naturelles de Richards et collaborateurs (2012) menée sur des populations *in situ*. Dans cette étude, l'expression de l'ensemble des gènes au cours du développement des plantes a été analysée sur microarray. Malgré les conditions environnementales hétérogènes, les replicats des deux accessions (Sha et Bay0 deux écotypes naturels provenant du Tadjikistan et d'Allemagne) présentent des résultats similaires démontrant l'homogénéité de la réponse des individus des génotypes d'*A. thaliana* aux conditions environnementales.

Une réponse homogène des populations de Spartines en milieu naturel

Au niveau intrapopulation, les tests statistiques ne montrent pas de variation significative d'expression intra-population pour aucun des gènes analysés. Les réponses intraspécifiques sont globalement assez homogènes concernant les 13 gènes étudiés. Plusieurs causes peuvent néanmoins expliquer ces faibles différences d'expression. Tout d'abord il peut exister un biais statistique, certaines populations présentent des variances élevées, notamment la population de *S. alterniflora* et les populations de *S. anglica*. La variance paraît corrélée avec le niveau d'expression. En effet, plus l'expression des individus est élevée plus la variance intrapopulation est élevée. Cette faible variation pourrait aussi être mise en rapport avec la faible diversité génétique intrapopulation connue en Europe chez ces espèces clonales (Ainouche *et al*., 2004a).

Au niveau interpopulation des variations significatives d'expression entre les populations d'une même espèce sont observées. En effet, deux populations ont été analysées pour chacune des espèces *S. maritima* et *S. anglica*. Chez *S. maritima*, tous les gènes sont surexprimés chez les individus du site de Quenouille par rapport aux individus du

site de Noirmoutier. Le premier site est plus haut sur l'estran, les individus doivent supporter des conditions de salinité plus élevées. En effet, pour le gène V-ATPase, il a été démontré une augmentation du nombre de transcrits après un traitement salin chez plusieurs espèces: *Lycopersicon esculentum* (Binzel, 1995) *Beta vulgaris* (Lehr *et al.*, 1999) et *Daucus carota* (Rausch *et al.*, 1996). Chez *S. anglica*, quatre des cinq gènes montrant une différence d'expression entre les populations, sont surexprimés chez les individus de l'anse de Goulven. Ces variations d'expression peuvent être expliquées par la disposition des individus sur l'estran ainsi que le type de substrat et la localisation des sites. En effet, le site de Goulven se situe en haut de l'estran d'une baie protégée et les individus sont dispersés suivant un gradient étendu alors que le site de Roscoff échantillonné est une plage avec des clones réduits et concentrés sur une surface réduite.

Divergence d'expression entre S. maritima et S. alterniflora

La méthode de PCR quantitative sur 13 gènes d'intérêts a mis en évidence des niveaux d'expression de gènes significativement plus élevés chez *S. alterniflora* par rapport à *S. maritima*, dans 7 cas sur 13 (Figure 17). Parmi ceux-ci, Le gène **WRKY24** contrôle plusieurs types de réponses aux stress. Chez *Arabidopsis*, ce gène est induit par différents facteurs : l'acide abscissique, H_2O_2, l'herbivorie et les infections biotrophiques et nécrotrophiques (Chen *et al.*, 2010). Ces effets conduisent à la production de reactive oxygen species (ROS) qui agissent comme des molécules 'signal' et causent du stress oxydant chez les plantes (Lamb & Dixon, 1997). L'expression du gène WRKY24 contribue à la défense de l'organisme dans un grand nombre de stress et régule des cascades de gènes liés aux métabolismes de l'acide salicique et de l'acide jasmonique. Un autre gène de stress montrant une surexpression chez *S. alterniflora* est le gène **Metal tolerance protein A2** (**MTP**). Il fait parti de la famille des transporteurs de zinc et contribue à la tolérance cellulaire : il contrôle l'accumulation de zinc dans les racines et le symplasme. La surexpression de ce gène chez *A. thaliana* entraîne une augmentation de l'accumulation de zinc et de cobalt dans les racines et les feuilles et ainsi, améliore la tolérance au zinc (Arrivault *et al.*, 2006). Les Spartines sont aussi étudiées pour leur forte biomasse et font d'excellents candidats pour la production de biofuel (Gonzalez-Hernandez *et al.*, 2009). Dans ce contexte, nous nous sommes intéressés à des gènes entrant dans les voies métaboliques de la lignine et la synthèse des membranes cellulaires végétales. La **Cinnamoyl-CoA reductase** ou **CCR** catalyse la synthèse des

monolignols servant à la formation de la lignine. La surexpression de ce gène chez *S. alterniflora* pourrait s'expliquer par la biomasse plus importante de cette dernière par rapport à *S. maritima*.

Les quatre autres gènes analysés montrant une surexpression chez *S. alterniflora* par rapport à *S. maritima* sont identifiés comme affectés par la spéciation dichotomique et/ou allopolyploïde (Chelaifa *et al.,* 2010a et b). Ces gènes permettent également de comparer les profils d'expression en conditions naturelles à ceux précédement analysés en même conditions de culture par microarrays. Ces gènes ont aussi des fonctions intéressantes. Il s'agit par exemple des hexokinases qui ont un rôle important dans le métabolisme des sucres (Harrington & Bush, 2003). Il existe 10 transcrits différents d'hexokinase chez *Oryza sativa*. Dans notre étude, le gène analysé correspond au transcrit du l'**Hexokinase 1** (HXK1). Chez la tomate, la surexpression de ce gène cause une réduction de la photosynthèse et de la croissance cellulaire, ainsi qu'une accélération de la sénescence (Dai *et al.*, 1999). Ainsi, les plantes utilisent l'HXK1 comme un détecteur de glucose qui permet les interactions entre les nutriments, la lumière et les hormones pour contrôler la croissance et le développement des plantes face à une grande variété de stress environnementaux (Moore *et al.*, 2003). On note également un gène codant un facteur de transcription de la famille Myb présentant une homologie plus forte avec le gène **Myb transcription factor (LOC_Os03g20900)** mis en évidence chez *A. thaliana* (Zhao *et al.*, 2011). Ce gène réprime la floraison quand l'intensité lumineuse diminue et régule négativement l'hyponastie foliaire et l'élongation cellulaire (Vandenbussche *et al.*, 2003 ; Mullen *et al*., 2006). En conditions contrôlées, ces gènes ne montrent pas de différence significative des niveaux d'expression entre les espèces parentales (Chelaifa *et al*., 2010b).

A l'inverse, nos résultats sur les gènes **Transcription Elongation Factor Spt6** et **Cytochrome c oxidase subunit 6b** obtenus *in situ* sont en accord avec les résultats de Chelaifa *et al.* (2010b). *Spt6* est une protéine très conservée au sein des Eucaryotes et initialement découverte chez *Saccharomyces cerevisiae* (Winston *et al.,* 1984). Ce gène est essentiel à la transcription *via* la modulation de la structure de la chromatine. Son expression est liée à la transcription active de certaines protéines et notamment des gènes codant des heat shock protein (*hsp*) (Winston, 2001). Un autre gène de comportement similaire code la **cytochrome c oxidase** de la sous-unité 6b qui catalyse la dernière étape du

transport d'électrons de la chaîne respiratoire (Barrientos *et al.*, 2002). Elle est composée de plusieurs sous-unités dont les gènes peuvent être mitochondriaux ou nucléaires. Les gènes mitochondriaux sont synthétisés en excès, ce sont donc les transcrits nucléaires qui vont réguler l'expression et l'activité de la cytochrome c oxidase (Giegé *et al.*, 2005). Dans ce sens, différentes études ont montré que plusieurs gènes codant les sous-unités présentaient une augmentation d'expression lorsque les plantes étaient soumises à des carbohydrates (Welchen *et al*., 2002 ; Curi *et al*., 2003 ; Giegé *et al*., 2005). Une étude protéomique réalisée par Yan et ses collaborateurs (2005) a permis de mettre en évidence la réponse de la sous-unité 6b du cytochrome c dans les racines de riz en présence d'un traitement salin. Cette protéine est impliquée dans la régulation des carbohydrates, de l'azote, du métabolisme énergétique et de la récupération des ROS.

Ainsi les différences observées entre *S. maritima* et *S. alterniflora* en même conditions de culture en serre par Chelaifa *et al.* (2010b) sont confirmées en conditions naturelles (pour 7 des 13 gènes analysés, Figure 10). La surexpression de certains gènes chez *S. alterniflora* pourrait s'expliquer grâce à la morphologie et écologie des espèces en question. En effet, des différences importantes existent entre les espèces concernant leur développement et leur morphologie respectives. *Spartina alterniflora* a une biomasse importante, de longues feuilles charnues et de long rhizomes produisant un grand nombre d'inflorescences (Liao *et al*., 2008) alors que les clones de *Spartina maritima* sont de faible taille. Les populations naturelles de *S. maritima* sont actuellement en train de régresser (Raybould *et al*., 1991) alors que *S. alterniflora* devient une espèce envahissante sur plusieurs continents (en Californie : Ayres *et al.,* 2004 ; Civille *et al.*, 2005 ; en Chine : Li *et al.*, 2009b et en Europe : Campos *et al.*, 2004 ; Querné *et al.*, 2011).

Variations de l'expression des gènes suite à l'hybridation interspécifique

Les niveaux d'expression entre les deux hybrides F1 *S.* x *neyrautii* et *S.* x *townsendii* sont identiques pour 12 gènes des 13 analysés. Seul le gène **WRKY24** est surexprimé chez *S.* x *neyrautii* par rapport à *S.* x *townsendii*.

En comparant les niveaux d'expression entre les hybrides et les parents, nous avons noté deux gènes (**V-type proton ATPase** et **Hexokinase 1**) qui ont une expression additive alors qu'ils montrent des différences d'expression en conditions expérimentales (Chelaifa *et*

al., 2010a). Parmi les gènes à expression non-additive, les gènes codant **le transcriptional adpater ADA2, le WRKY24, la Xanthine dehydrogenase 1** et la **GDP-mannose pyrophosphorylase (GMPase)** sont sous-exprimés chez l'hybride F1 *S.* x *townsendii* alors qu'ils présentent des valeurs additives chez l'espèce *S.* x *neyrautii*. Le gène **WRKY24** a été retrouvé parmi les gènes testés sur les puces de riz, en conditions contrôlées, ce gène est aussi sous-exprimé chez *S.* x *townsendii*. Les gènes codant **le transcriptional adpater ADA2** et le **WRKY24** ont été précédemment commentés, ce sont des gènes entrant dans la réponse au stress salin chez les Spartines tout comme le gène codant la **Xanthine dehydrogenase 1**. Chez *Arabidopsis*, la mise sous silence de ce gène entraîne de sévères conséquences sur le métabolisme de la tolérance aux stress, avec une réduction significative de la biomasse des graines et de la viabilité cellulaire suite à un épisode de sécheresse (Watanabe *et al.*, 2010).

Dans les autres cas (**Metal tolerance protein A2, C2H2 zinc finger protein STOP1, Cinnamoyl-CoA reductase, Cytochrome c oxidase subunit 6b** et **MYB transcription factor**) les gènes montrent une expression transgressive réprimée chez les deux hybrides par rapport aux parents hexaploïdes (Figure 17). L'expression des deux premiers gènes est plus directement liée au stress oxydant et à la présence de métaux lourds. Le deuxième gène code une protéine **C2h2 Zinc Finger** (STOP1). Chez un mutant de *A. thaliana* pour ce gène, les ions cadmium, cuivre, lanthanum, manganèse et sodium n'ont pas d'effets phénotypiques, alors que les plantes montrent une hypersensibilité aux ions aluminium (Iuchi *et al.*, 2007). Le **MYB transcription factor (LOC_Os03g20900)** a précédemment été identifié comme affecté par l'hybridation interspécifique en conditions contrôlées : dans ce cas, le niveau d'expression des hybrides correspond à la MPV. A l'inverse, les variations d'expression du gène codant la **Cytochrome c oxidase subunit 6b** ne montraient pas de différence significative sur les microarrays (Chelaifa *et al*., 2010a).

Seul le gène codant une **metallocarboxypeptidase inhibitor** est transgressivement activé chez les deux hybrides F1 par rapport aux parents hexaploïdes. Les gènes codant les metallocarboxypeptidase inhibitors servent à la séquestration des métaux lourds en formant notamment des complexes avec des chélateurs de métaux lourds comme les metallothionéines et les phytochélatines permettant d'exuder les métaux lourds du cytosol vers la surface foliaire (Sarret *et al.*, 2006). Il est intéressant de noter que les

metallothionéines sont très exprimées dans les feuilles des parents hexaploïdes (Ferreira de Carvalho *et al.*, 2013, Chapitre 4) ainsi que chez *Nicotiana tabacum* (Harada *et al.*, 2010). Dans cette étude, chez les plantes traitées au Cadmium (un métal lourd toxique pour les plantes et les animaux, Sanita di Toppi & Gabbrielli, 1999) de nombreuses protéines induis par ce stress ont été identifiées comme des metallocarboxypeptidase inhibitors. Il semblerait donc que l'expression de ces protéines soit directement liée à une contamination aux métaux lourds de l'environnement. A l'inverse, en conditions contrôlées, ce gène est sous-exprimé chez l'hybride anglais par rapport à ses parents. En transplantant les plantes en serre et les acclimatant, l'hybride F1 anglais montre une baisse du nombre de trasncrits de gènes liés à la tolérance aux métaux lourds. A noter aussi que les sites d'hybridation sont des zones portuaires potentiellement polluées (par rapport aux autres sites échantillonnés plus préservés) ce qui pourrait expliquer la forte expression du gène metallocarboxypeptidase inhibitor.

En conditions naturelles, un faible nombre de gènes montre une additivité des profils parentaux chez les hybrides. Dans les cas de non-additivité, la plupart des gènes sont sous-exprimés. Des phénotypes divergents comme ceux reportés chez les Spartines hybrides sont aussi fréquemment observés chez d'autres hybrides interspécifiques (Gross & Riesberg, 2005). L'expression transgressive est le plus souvent surexprimée chez les hybrides *Helianthus deserticola* (Lai *et al*., 2006) et *Senecio squalidus* (Hegarty *et al.*, 2009). L'hybridation entraîne des remaniements génomiques et transcriptomiques profonds, néanmoins ces effets ne sont pas clairement élucidés et leur mesure pourrait potentiellement varier en fonction des conditions expérimentales d'analyse et du type de tissu. Récemment, Yoo et ses collaborateurs (2013) ont montré chez un hybride F1 synthétique du genre *Gossypium*, une dominance d'expression vers le génome A (dans les feuilles) alors que l'inverse (dominance d'expression vers le génome D) est observé dans les travaux de Flagel & Wendel (2010) dans les pétales.

Effets de la duplication du génome hybride chez S. anglica

En comparant l'allododécaploïde S. anglica à l'hybride F1 *S.* x *townsendii*, six gènes ne montrent pas de différences significatives d'expression entre les deux espèces. Cinq gènes montrent chez *S. anglica* des niveaux d'expression supérieurs à ceux de *S.* x *townsendii* dont

deux gènes de réponse au stress salin et un gène précédemment montré comme affecté par la spéciation allopolyploïde (Chelaifa *et al*., 2010b). Les deux gènes liés au métabolisme de lignine et de la cellulose sont tous les deux surexprimés chez l'allododécaploïde par rapport à l'hybride anglais ce qui reflèterait les effets de la duplication sur le métabolisme des parois cellulaires. En conditions contrôlées, la comparaison entre les niveaux d'expression de *S. anglica* et *S.* x *townsendii* met en évidence 497 gènes (soit 4,6% des gènes) dont la majorité d'entre eux (310) sont surexprimés chez l'allopolyploïde. Chez l'hybride F1, la proportion de gènes montrant une déviation de l'expression par rapport aux parents hexaploïdes est plus importante que chez l'allododécaploïde. De même dans le complexe *Senecio*, la comparaison entre un hybride synthétique *Senecio* x *baxteri* et l'allohexaploïde *Senecio cambrensis* a montré un effet « tampon » de la duplication du génome comparativement à l'hybridation sur les changements liés à la spéciation réticulée (Hegarty *et al.*, 2006). De plus, Flagel *et al.* (2008) ont observé une expression biaisée des génomes parentaux homéologues chez les cotons allotetraploïdes synthétiques et naturels résultant principalement de l'hybridation comme chez *Arabidopsis suecica* (Madlung *et al*., 2002).

Un seul gène (**metallocarboxypeptidase inhibitor**) présente une expression significativement inférieure chez l'allododécaploïde par rapport à l'hybride anglais. Comme expliqué précédemment ce gène permet à la plante de tolérer les pollutions en métaux lourds de l'environnement. Il est très exprimé chez les hybrides F1 par rapport à la fois aux parents hexaploïdes mais aussi à l'allododécaploïde.

Analyse des copies de gènes dupliqués chez les parents hexaploïdes : Cas du gène Metal tolerance protein A2 (MTP)

Les analyses d'expression réalisées par PCR quantitative ont permis de mesurer l'expression globale de 13 gènes chez les parents hexaploïdes, les hybrides F1 et l'allododécaploïde. Cette expression est le reflet de l'addition de plusieurs copies dupliquées par polyploïdie. Les précédents travaux menés au sein du laboratoire ont suggéré une origine allopolyploïde des parents hexaploïdes *S. maritima* et de *S. alterniflora* (Fortuné *et al.*, 2007). Jusqu'à trois paires de copies homéologues par locus chez ces espèces seraient donc attendues, et jusqu'à six paires de copies chez l'allodécaploïde *S. anglica*. Chez les Spartines, Fortuné *et al.* (2007) ont montré une rétention différentielle de copies du gène

Waxy chez les hexaploïdes, avec la rétention d'une copie chez *S. maritima* et de 3 copies chez *S. alterniflora* par locus homologue.

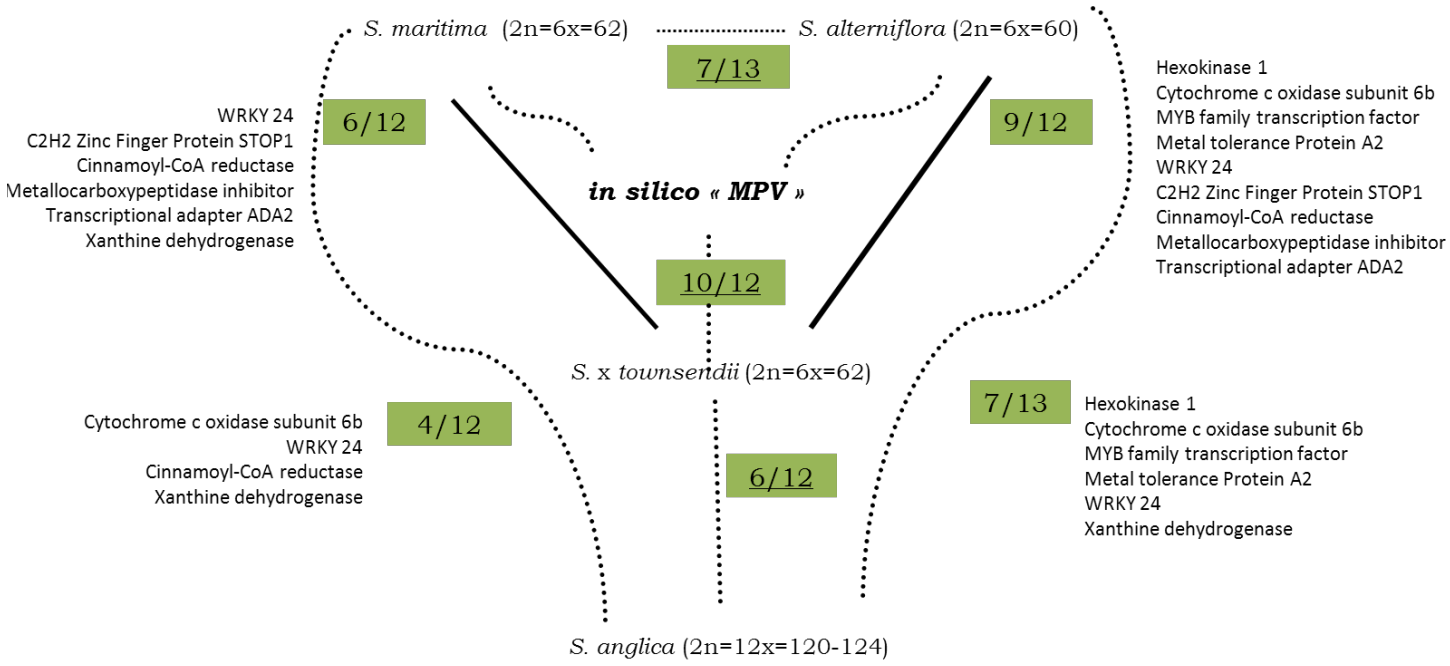

Figure 17 : Différences transcriptomiques entre les cinq espèces du complexe *Spartina* représentées par le nombre de gènes différentiellement exprimés entre les parents, l'hybride F1 anglais et l'allododécaploïde, sur le nombre de gènes analysés en PCR quantitative (adapté d'après Chelaifa *et al*., 2010a).

Dans cette étude de l'expression globale de gènes candidats, nous nous sommes intéressés à un gène en particulier codant une Metal Tolerance Protein (MTP) et avons tenté de détecter les différentes copies de ce gène potentiellement présentes d'une part dans le génome, et d'autre part, transcrites dans les feuilles des parents hexaploïdes. Le gène MTP est surexprimé chez *S. alterniflora* par rapport aux autres espèces analysées, il augmente la tolérance cellulaire aux métaux lourds et permet leur accumulation dans les tissus foliaire et racinaire (Arrivault *et al*., 2006).

Chez *Spartina maritima*, deux clades de séquences se détachent nettement reflétant la présence de deux copies, toutes les deux transcrites, pouvant représenter deux copies homéologues. Chez *S. alterniflora*, plusieurs types de séquences ont été identifiées dans le

transcriptome : un groupe de séquences identiques caractérise une copie (clade C, Figure 16), tandis que d'autres séquences (faiblement représentées dans nos résultats) pourraient représenter des copies homéologues mais aussi de potentiels paralogues au vu du nombre plus important de différences nucléotidiques observées. Des analyses complémentaires (augmentation de la profondeur de séquençage de ce gène, analyses phylogénétiques prenant en compte plusieurs espèces des clades tetraploïdes) seraient nécessaires pour préciser la nature et l'origine évolutive de ces séquences.

La détection de copies *in silico* du gène MTP à partir de données de pyroséquençage Roche-454 a permis de mettre en évidence entre 3 et 4 haplotypes par fenêtre pour chacune des espèces parentales comme cela a déjà été mis en évidence dans l'étude sur le transcriptome des espèces hexaploïdes (Ferreira de Carvalho *et al*., 2013, Chapitre 4). Dans les deux études, deux haplotypes présentent une divergence nucléotidique plus importante avec un ou deux variants additionnels pour chaque espèce. Les haplotypes plus divergents pourraient représenter des copies homéologues et les variants additionnels les allèles de chaque copie.

CONCLUSION

Ce travail représente la première analyse de l'expression des gènes en conditions naturelles des Spartines hexaploïdes, de leurs hybrides et de l'allododécaploïde. Les résultats ont permis de mettre en évidence les différences transcriptomiques entre les parents hexaploïdes liées à la spéciation dichotomique à mettre en relation avec l'écologie et la physiologie contrastées de ces deux espèces ayant divergé, il y a moins de 3MA. Deuxièmement, malgré leurs origines hybrides indépendantes, les hybrides F1 montrent des niveaux d'expression similaires en conditions naturelles. Néanmoins, les niveaux d'expression chez l'hybride F1 *S.* x *townsendii* sont en général significativement inférieurs à *S. anglica.* Seul le gène codant une metallocarboxypeptidase inhibitor augmentant la tolérance aux métaux lourds est surexprimé chez les deux F1 mettant clairement en évidence les effets des environnements pollués en métaux lourds sur leurs sites d'hybridation respectifs. Les effets de la spéciation allopolyploïde sont mis en évidence, notamment chez les deux gènes étudiés entrant dans le métabolisme de la lignine et de la cellulose. Plusieurs gènes impliqués dans les mécanismes de tolérance aux stress salin et

oxydant sont aussi fortement exprimés chez l'allododécaploïde. Ces effets transcriptomiques sont à mettre en relation avec la nature polyploïde de l'espèce et lui permettent de fabriquer une biomasse importante, des feuilles et des rhizomes vigoureux et d'être très tolérante à l'immersion.

Chapitre 6

Explorations du génome de *Spartina maritima*

Introduction et démarche générale

Les efforts de séquençage du génome de représentants de la lignée hexaploïde des Spartines par le laboratoire d'accueil se sont portés prioritairement sur l'espèce native européenne *S. maritima* (2n=6x=60, 2C=3,8 pg). Cette espèce (Figure 18), qui a joué le rôle de parent mâle dans les croisements avec *S. alterniflora* à l'origine des hybrides *S.* x *townsendii* (Ferris *et al.*, 1997 ; Baumel *et al.,* 2001) et *S.* x *neyrautii* (Baumel *et al.*, 2003), est distribuée sur les côtes atlantiques européennes et africaines où elle occupe une situation pionnière dans le bas shore, adaptée aux situations d'immersion dans l'eau salée. *Spartina maritima* est également tolérante à la pollution aux métaux lourds (comme sur les marais de l'Odiel en Espagne, un des estuaires les plus pollués au monde, Cambrollé *et al.*, 2008), ce qui en fait une bonne candidate en phytoremédiation des marais salés (Reboreda & Caçador, 2008). Si de vigoureuses populations sont connues dans le sud de la Péninsule Ibérique, elle semble en régression en limite nord de son aire de distribution (nord de la Bretagne et sud de l'Angleterre). En Angleterre, elle n'est plus présente sur le site d'origine de *S.* x *townsendii* (à Hythe), mais elle est encore signalée plus à l'ouest à Hayling Island (Raybould *et al.*, 2000). L'essentiel des données biologiques sur cette espèce provient des populations européennes où elle est connue pour une floraison produisant très peu de graines, et une propagation essentiellement végétative, ce qui expliquerait la très faible diversité génétique notée au niveau intra et interpopulation (Yannic *et al*., 2004).

Figure 18 : *Spartina maritima* dans les marais salés de Hayling Island (Angleterre) en août 2012.

Dans cette partie, nous analyserons tout d'abord le génome de *S. maritima* à travers un jeu de données de séquences d'extrémités de BAC (Bac End Sequences ou BESs), nous permettant d'avoir un premier aperçu de la proportion des régions codantes et non-codantes du génome. Les résultats sont présentés sous forme d'un projet d'article. Puis, nous analyserons les séquences répétées chez *S. maritima*, à travers un jeu de données de séquences génomiques obtenues par pyroséquençage (Roche 454).

Partie A. Analyse des régions codantes et non codantes à travers les séquences d'extrémités de BAC (BESs)

Exploring the genome of the salt-marsh *Spartina maritima* (Poaceae, Chloridoideae) through BAC end sequence analysis

Ferreira de Carvalho J.[1], Chelaifa H.[1], Boutte J.[1], Mangenot S.[2], Couloux A.[2], Wincker P.[2], Bellec A.[3], Fourment J.[3], Bergès H.[3], Salmon A.[1], Ainouche M.[1]*.

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

[2] Genoscope, 2 rue Gaston Crémieux, 91000 Evry (France)

[3]Centre National de Ressources génomiques végétales-INRA, 24 Chemin de Borde Rouge, CS 52627, 31326 Castanet Tolosan Cedex (France)

*Corresponding author: Malika L. Ainouche, UMR CNRS 6553 Ecobio, University of Rennes 1, Bât 14A Campus Scientifique de Beaulieu, 35 042 Rennes Cedex (France)

Telephone: +33 2 23 23 51 11

e-mail address : malika.ainouche@univ-rennes1.fr

Running Title: *Spartina maritima* BAC-end sequence analysis

Abstract

Spartina species play a critical role on coastal salt marshes. They are considered as 'ecosystem engineers' as they play an important ecological role in the marsh sedimentary dynamics. Most species are native to the New World, whereas *Spartina maritima* is and Old-World species distributed along the European and North-African Atlantic coasts. This hexaploid species (2n=6x=60) hybridized with different *Spartina* species introduced from the American coasts, which resulted in the formation of new invasive hybrids and allopolyploids. Thus, *Spartina maritima* raises evolutionary and ecological interests. However, genomic information is dramatically lacking in this genus.

In an effort to develop genomic resources, we generated and analysed 40,641 high-quality BAC-End Sequences (BESs), representing 26.7Mb or 4.3% of the basic genome of *Spartina maritima* (x=10, 616Mb). Level of chloroplast and mitochondrial contaminations were below 2% and 0.5%, respectively. The BESs ranged from 57bp to 938bp, with an average length of 656bp. BESs were searched for sequence homology against known databases. A fraction of 16.91% of the BESs represents known repeat elements including a majority of 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 SSRs were detected, dinucleotides A/T being the most represented in the dataset. Using the coding sequence database of *Sorghum bicolor*, 6,809 BESs (annotating 4,098 different coding sequences) found homology accounting for 17.1% of all BESs. Comparative genomics with related genera reveals that the microsynteny is better conserved with *Sorghum 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 (respectively 9 among chromosomes and 7 between chromosomes) appeared to have experienced rearrangements between genera *Spartina* and *Sorghum*.

This work represents the first overview of *Spartina maritima* genome regarding the respective coding and repetitive components. The syntenic relationships with other grass genomes examined here help clarifying the events and the dynamics that drove Poaceae evolution, *Spartina maritima* being a part of the poorly-known Chloridoideae sub-family.

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, 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; Petersen *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 salt-marsh 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 (deriving from the hexaploid lineage: Dauvergne and Ainouche unpublished) 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 (Mulholland &Otte, 2000) 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*., 2012). 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 to 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 the *Sporobolus* and *Calamovilfa* genera (Petersen *et al*., 2010). The genus evolved through two main lineages respectively tetraploid and hexaploid (Baumel *et al*., 2002a; Fortuné *et al*., 2007) that diverged less than 6 MYA, as estimated from chloroplast sequences (Bellot, 2010; 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; Ainouche *et al*., 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; Baumel *et al*., 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 & 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; 2004b; 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 (Figure 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.

MATERIAL 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 40g 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). DNA extraction for megabase-size DNA was performed using a standard protocol for plants (Luo & Wing, 2003). High-molecular weight DNA was partially digested with the enzymes *Hind*III and *Bam*HI to build two distinct libraries and was subjected to fragment size-selection. Two rounds of size-selection (digested DNA longer than 100kb but shorter than 300kb) were necessary to go through the isolation process. Size-selected DNA was ligated into the vector pIndigoBAC5 and DH10BT1 *E. coli* competent cells were electroporated with the ligation products. In total, 44,544 clones with a mean insert size of 110kb were retained, representing 4,900 Mb or 1.5X the genome of *S. maritima* (3700 Mb, estimated from Fortune *et al*., 2008). As this genome is hexaploid, the BAC library represents 8X the basic genome (616 Mb). More than 20 000 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 70bp. 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^{-06}) 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 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 & Bell, 2004). BLASTn analyses were conducted using an e-value cut off of 10^{-6} and a minimum hit length of 100bp.

Simple Sequence Repeat (SSR) detection

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 100bp.

***De novo* identification of *Spartina* repeats**

The masked file output from RepeatMasker (containing 39,910 sequences excised from repetitive elements and representing 26.2Mb) 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 420Mb 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 15kb to 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* 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%).

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 938bp with an average of 656bp corresponding to a total of 26,682,959 nucleotides equivalent to 4.3% of the basic genome of *Spartina maritima* (x=10, 616Mb). The GC content estimation 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) (Table1). 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 (Table1). 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 (Figure 1).

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 400bp long were retained to build the trees. In the *Copia* analysis, 211 *Spartina maritima* sequences are aligned with 722 RT protein sequences from RepBase (Figure 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 (Figure 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. This is consistent with the number of repeat elements found using RepeatMasker (data not shown).

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 (Figure 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 2.7Mb) 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 26.18Mb of *Spartina maritima* BESs, representing 64,643bp or 0.25% of the BESs sequenced (Table 3) which is equivalent to one microsatellite every 6.1kb (Table 4). Mononucleotides (60.9%) are the most abundant motifs, followed by dinucleotides (21.6%), trinucleotides (16.1%), tetra, penta and hexanucleotides (1.42%) (Table 4).

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) (Figure 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 (Figure 5B).

A summary of the *Spartina maritima* BES composition is presented in Figure 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 chromosome, (2) within a 15 to 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 (Figure 7). Most BESs are comprised in a distance range from 15 to 250kb; BESs out of this range are thought to be rearranged. Syntenic relationships between *S.*

maritima 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 5, 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 5).

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 5). 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 to 250 kb distance "micro-synteny" range (*i.e*. rearranged) and 556 BESs situated in the distance window of 15 to 250kb. Most of these (524) are collinear with *Sorghum,* whereas 32 exhibit a shift in the orientation of one of the BESs (Table 5). A substantial proportion of the paired BESs (568 representing 40.75%) match to different *Sorghum* chromosomes (Table 5).

The 5,053 *Spartina* BESs mapped on the ten *Sorghum bicolor* chromosomes are represented in Figure 8. Collinear paired BESs show a high concentration on *Sorghum* chromosomes 1, 3, 4 and 6. We chose to represent rearranged paired BESs for collinear regions including at least two pairs of rearranged BES (Figure 8). These putative orthologous regions involve both rearrangements on the same chromosomes or paired BESs matching different chromosomes.

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 representing 4.3% of the basic nuclear genome of *Spartina maritima* 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=420Mb; IRGSP, 2005) and *Brachypodium distachyon* (28.1%, 2n=2x=10, 1C=270Mb; IBI, 2010). However, regarding the *Spartina maritima* basic genome size (x=10=616Mb), a larger number of repeat sequences would be expected: *Sorghum bicolor* has a genome size of 740Mb (2n=2x=20) and a repeat element fraction of 62% (Paterson *et al.*, 2009). It seems that in Grass genomes, the proportion of repeats is conversely correlated with chromosome size.

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 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 & 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 (Ragupathy *et al.*, 2011; Cavagnaro *et al.*, 2008; Huo *et al*., 2007). 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 64,643bp) have been identified from the 26.7Mb of genomic DNA

analyzed. The density found is of one SSR every 6.1kb 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 6.2kb; Cheung and Town, 2007) and *A. thaliana* (1 SSR every 6.4kb); but lowest than *O. sativa* (1 SSR every 9.0kb) and *Z. mays* (1 SSR every 16.1kb) (Hsu *et al*., 2011). These findings are in agreement with Morgante and collaborators (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 and collaborators (2010) in *Spartina pectinata*, where they found 841 SSRs in ESTs longer than 500bp 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% (Cavagnero *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 (656bp) and the total size of BESs sequenced (26.7Mb), we estimated a percentage of 16.7% of BESs containing potentially coding genes.

Considering the basic genome size of *S. maritima* (616Mb) and the mean size of an *Oryza sativa* gene (2.7kb; 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 16.1kb based on the fact that we expect 38,229 genes in the basic genome of *Spartina maritima* (616Mb). 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 9.9kb (IRGSP, 2005). *Musa acuminata* is predicted to have one gene every 14.3kb (D'Hont *et al*., 2012) and *A. thaliana* has one gene every 4.5kb (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 to 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, 730Mb for *S. bicolor* and 616Mb 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 50MYA; 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 Chlorioideae and Panicoideae diverged about 34.6 to 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 or 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 to 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 9 among chromosomes and 7 between chromosomes) appeared to have experienced rearrangements between genera *Spartina* and *Sorghum*. 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.

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. *Biological Journal of the Linnean Society* **82**: 475-484.

Ainouche ML, Baumel A, Salmon A, Yannic G (2004b). Hybridization, polyploidy and speciation in *Spartina* (Poaceae). *New Phytologist* **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). *Biological 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 PS, Soltis DE (eds) *Polyploidy and Genome Evolution*, Springer Berlin Heidelberg: Berlin, Heidelberg, 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 Research* **13**: 753-763.

Alberts J, Price M, Kania M (1990). Metal concentrations in tissues of *Spartina alterniflora* (Loisel) and sediments of georgia salt marshes. *Estuarine coastal and shelf science* **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. *Biological Invasions* **6**: 221-231.

Baisakh N, Subudhi PK, Varadwaj P (2008). Primary responses to salt stress in a halophyte, smooth cordgrass (*Spartina alterniflora* Loisel.). *Functional & Integrative 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). *Molecular Ecology* **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). *Molecular Phylogenetics and Evolution* **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). *Molecular Biology and Evolution* **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* x *neyrautii* re-examined. *Plant Systematics and Evolution* **237**: 87-97.

Bellot S (2010). *Evolution du génome chloroplastique chez Spartina (Poaceae, Chloridoideae)*. Université de Montpellier 2, Rapport de Master Biologie Ecologie Evolution.

Bennetzen JL (2005). Transposable elements, gene creation and genome rearrangement in flowering plants. *Current Opinion in Genetics and Development* **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). **12**: 104.

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. *Science of The Total Environment* **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. *Marine Pollution Bulletin* **56**: 2037 - 2042.

Castellanos E, Figueroa M, Davy A (1994). Nucleation and facilitation in salt-marsh succession - interactions between *Spartina maritima* and *Arthrocnemum perenne*. *Journal of Ecology* **82**: 239-248.

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. *Diversity and Distributions* **16**: 547–558.

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. *Journal of Ecology* **88**: 801–812.

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. *Molecular Genetics and Genomics* **281**: 273-288.

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. *Journal of Molecular Evolution* **58**: 424-441.

Cheung F, Town C (2007). A BAC end view of the Musa acuminata genome. *BMC Plant Biology* **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. *Current Biology* **18**: 37–43.

Christin P-A, Petitpierre B, Salamin N, Büchi L, Besnard G (2009). Evolution of C4 Phosphoenolpyruvate Carboxykinase in grasses, from genotype to phenotype. *Molecular Biology and Evolution* **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*. *Theoretical and Applied Genetics* **114**: 321-332.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **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*. doi.org/10.1038/hdy.2012.76

Ferris C, King RA, Gray AJ (1997). Molecular evidence for the maternal parentage in the hybrid origin of *Spartina anglica* C.E. Hubbard. *Molecular Ecology* **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). *Molecular phylogenetics and evolution* **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. Molecular Ecology 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. *The Plant Genome Journal* **3**: 69.

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 Cellular and Developmental Biology 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 Research* **36**: 3420 -3435.

Grass Phylogeny Working Group II (2012). New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. *New Phytologist* **193**: 304–312.

Groves H, Groves J (1880). *Spartina* x *townsendii* Nobis. *Report of the Botanical Society and exchange club of the British Isles* **1** : 37.

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. *Environmental Science and Technology* **35**: 3863-3868.

Han Y, Korban SS (2008). An overview of the apple genome through BAC end sequence analysis. *Plant Molecular Biology* **67**: 581-588.

Hilu KW, Alice LA (2001). A phylogeny of Chloridoideae (Poaceae) based on matK sequences. *Systematic Botany* **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 Biology* **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. *Functional and Integrative 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. *Cytogenetic and Genome Research* **110**: 462 - 467.

Kejnovsky E, Hawkins JS, Feschotte C (2012). Plant transposable elements: biology and evolution. In: *Plant Genome Diversity Volume 1: Plant genomes, their residents and their evolutionary dynamics*, Jonathan F. Wendel, Johann Greilhuber, Jaroslav Dolezel & Ilia J. Leitch: 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. *Tropical Plant Biology* **2**: 51–62.

Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, *et al.* (2009). Circos: An information aesthetic for comparative genomics. **19**: 1639–1645.

Lee RW (2003). Physiological adaptations of the invasive cordgrass *Spartina anglica* to reducing sediments: rhizome metabolic gas fluxes and enhanced O-2 and H2S transport. *Marine Biology* **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. *Molecular Ecology* **11**: 2453–2465.

Luo M, Wing RA (2003). An Improved Method for Plant BAC Library Construction. In: *Plant Functional Genomics*, Humana Press: New Jersey Vol 236, pp 3-20.

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). *Marine Pollution Bulletin* **38**: 49-51.

Manninen I, Schulman AH (1993). *BARE-1*, a *Copia*-like retroelement in barley (*Hordeum vulgare* L.). *Plant Molecular Biology* **22**: 829 - 846.

Marchant CJ (1967). Evolution in *Spartina* (Gramineae). I. The history and morphology of the genus in Britain. *Journal of the Linnean Society (Botany)* **60**: 1-24.

Marchant C, Goodman P (1969). *Spartina maritima* (Curtis) Fernald. *Journal of Ecology* **57**: 287-302

Ma J, Devos KM, Bennetzen JL (2004). Analyses of LTR-Retrotransposon structures reveal recent and rapid genomic dna loss in Rice. *Genome Research* **14**: 860-869.

Mobberley DG (1956). Taxonomy and distribution of the genus *Spartina*. *Iowa State College Journal of Science* **30**: 471–574.

Morgante M, Hanafey M, Powell W (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genetics* **30**: 194-200.

Mulholland MM, Otte ML (2000). Effects of varying sulphate and nitrogen supply on DMSP and glycine betaine levels in *Spartina anglica*. *Journal of Sea Research* **43**: 199 - 207.

Osoegawa K, Vessere GM, Shu CL, Hoskins RA, Abad JP, Pablos B de, *et al.* (2007). BAC clones generated from sheared DNA. *Genomics* **89**: 291 - 299.

Ouyang S, Bell R (2004). The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Research* **32**: 360D–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 Phytologist* **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 PM, Romaschenko K, Johnson G (2010). A phylogeny and classification of the Muhlenbergiinae (Poaceae: Chloridoideae: Cynodonteae) based on plastid and nuclear DNA sequences. *American Journal of Botany* **97**: 1532 –1554.

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 Research* **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. *Nature Communications* **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**: 217-217.

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. *Molecular Biotechnology* **51**: 18-26.

Raybould AF, Gray AJ, Lawrence MJ, Marshall DF (1991). The evolution of *Spartina anglica* CE HUBBARD (graminae) - Origin and genetic variability. *Biological Journal of the Linnean Society* **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. *Annals of Botany* **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. *The 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. *Proceedings of the National Academy of Sciences* **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. *Theoretical Applied Genetics* **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.). *Theoretical and Applied Genetics* **106**: 411–422.

Timmers RA, Strik DPBTB, Hamelers HVM, Buisman CJN (2010). Long-term performance of a plant microbial fuel cell with *Spartina anglica*. *Applied Microbiology and Biotechnology* **86**: 973-981.

Vitte C, Bennetzen JL (2006). Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proceedings of the National Academy of Sciences* **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 Molecular Biology* **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.

Yu J-K, Sun Q, Rota ML, Edwards H, Tefera H, Sorrells ME (2006). Expressed sequence tag analysis in tef *Eragrostis tef* (Zucc) Trotter. *Génome* **49**: 365-372.

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.

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. *American Journal of Botany* **99**: 193 –208.

Zhang D, Ayele M, Tefera H, Nguyen HT (2001). RFLP linkage map of the Ethiopian cereal tef : *Eragrostis tef* (Zucc) Trotter. *TAG Theoretical and Applied Genetics* **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.

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	*Spartina maritima*	699 (1.72% of BES)
(Nb of hits)	*Sorghum bicolor*	683 (1.68% of BES)
	Oryza sativa indica	668 (1.64% of BES)
Mitochondrion matches	*Sorghum bicolor*	175 (0.43% of BES)
(Nb of hits)	*Oryza sativa indica*	91 (0.22% of BES)

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**

Table 3 Distribution of Simple Sequence Repeats in *Spartina maritima* BESs.

Type of repeats	Number	Type of repeats	Number
A/T	1,193	**AGGC/CCTG**	1
C/G	1,417	**AGGG/CCCT**	2
AC/GT	172	**CCCG/CGGG**	1
AG/CT	459	**AAAAG/CTTTT**	2
AT/AT	268	**AAACC/GGTTT**	1
CG/CG	27	**AACAC/GTGTT**	1
AAC/GTT	138	**AACAG/CTGTT**	1
AAG/CTT	234	**ACAGC/CTGTG**	1
AAT/ATT	21	**ACAGT/ACTGT**	1
ACC/GGT	41	**AGAGG/CCTCT**	1
ACG/CGT	22	**AAAAAG/CTTTTT**	2
ACT/AGT	9	**AAACCC/GGGTTT**	1
AGC/CTG	42	**AACAAG/CTTGTT**	2
AGG/CCT	53	**AACAAT/ATTGTT**	1
ATC/ATG	59	**AACATC/ATGTTG**	1
CCG/CGG	69	**AACGGC/CCGTTG**	1
AAAC/GTTT	1	**AAGACG/CGTCTT**	1
AAAG/CTTT	9	**AAGAGG/CCTCTT**	5
AAAT/ATTT	3	**ACAGAG/CTCTGT**	1
AACC/GGTT	2	**ACATAT/ATATGT**	1
ACAG/CTGT	1	**ACCAGC/CTGGTG**	1
ACAT/ATGT	1	**ACGAGC/CGTGCT**	1
ACGC/CGTG	2	**AGAGGG/CCCTCT**	1
ACTC/AGTG	1	**AGCGAT/ATCGCT**	1
AGCC/CTGG	4	**AGGATG/ATCCTC**	1
AGCG/CGCT	2	**ATCGCC/ATGGCG**	1
AGCT/AGCT	1		
TOTAL			**4,285 SSRs (64,643bp)**

Table 4 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

Table 5 BlastN hits and comparative genomics between *Spartina maritima* BESs (39 910 masked for repeats) and the *Arabidospis thaliana*, *Brachypodium distachyon*, *Oryza sativa* and *Sorghum bicolor* genomes.

			Paired BESs			
				Co-localized on the same chromosome		
	N° of hits (% of BESs)	Single BESs	Localized on different chromosomes	Distance inf. 15kb or sup. 250kb	Distance comprised between 15kb and 250kb	
					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

Figure Legends

Figure 1: Analyses conducted on the BAC-end Sequences.

Figure 2: Phylogenetic tree of *Ty3-Gypsy* elements based on Reverse Transcriptase sequence alignments of *Spartina maritima* repeats (red branches) and the Repbase (black branches).

Figure 3: Phylogenetic tree of *Ty1-Copia* elements based on Reverse Transcriptase sequence alignments of *Spartina maritima* repeats (red branches) and the Repbase (black branches).

Figure 4: Frequency of BESs showing similarity to other sequences in the same dataset.

Figure 5: Classification of GO Annotations, (A) for biological process and (B) molecular function.

Figure 6: Summary of *Spartina maritima* BES annotation analyses.

Figure 7: Distance frequencies between paired BESs.

Figure 8: BES sequences mapped to the *Sorghum* genome. The 10 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.

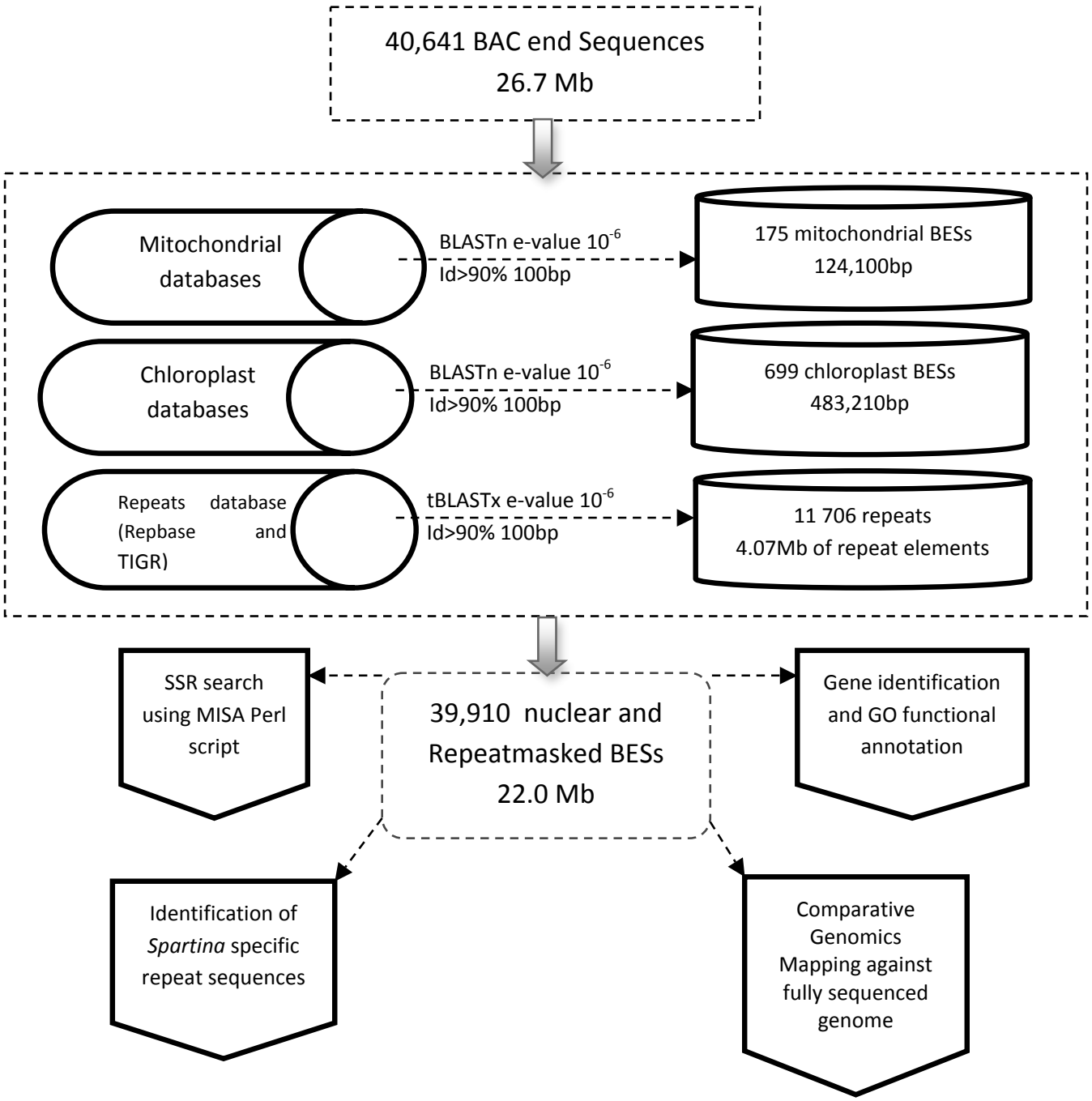
40,641 BAC end Sequences
26.7 Mb
Mitochondrial databases
BLASTn e-value 10-6
Id>90% 100bp
175 mitochondrial BESs
124,100bp
Chloroplast databases
BLASTn e-value 10-6
Id>90% 100bp
699 chloroplast BESs
483,210bp
Repeats database (Repbase and TIGR)
tBLASTx e-value 10-6
Id>90% 100bp
11 706 repeats
4.07Mb of repeat elements
SSR search using MISA Perl script
39,910 nuclear and Repeatmasked BESs
22.0 Mb
Gene identification and GO functional annotation
Identification of Spartina specific repeat sequences
Comparative Genomics Mapping against fully sequenced genome

Figure 1

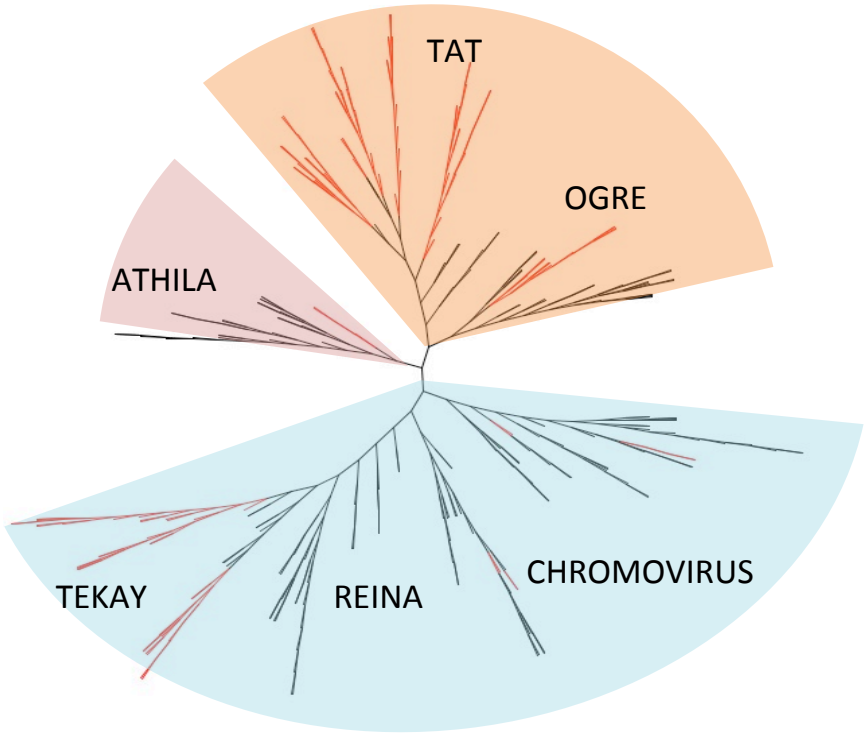

Figure 2

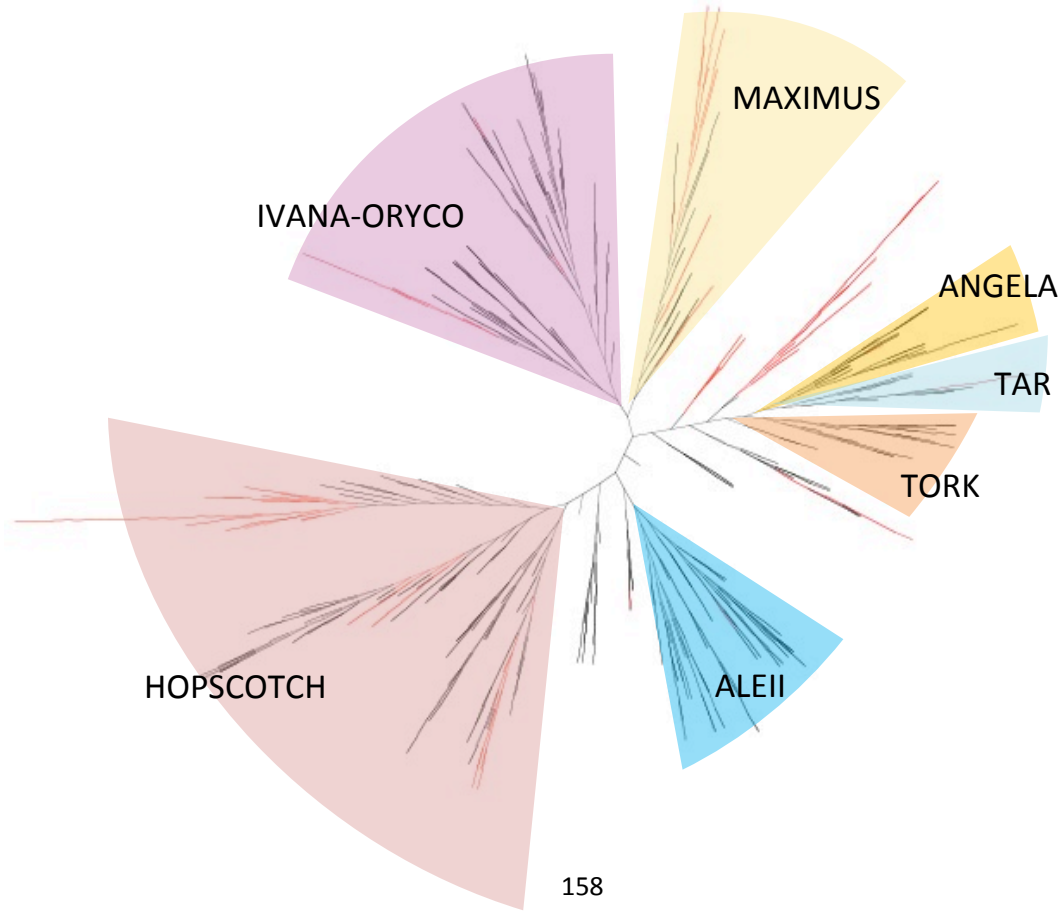

Figure 3

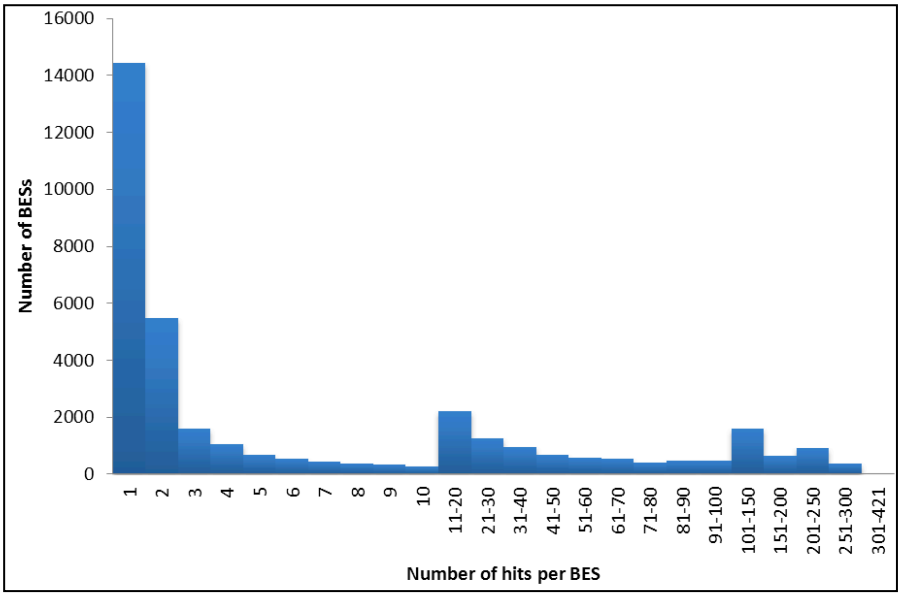

Figure 4

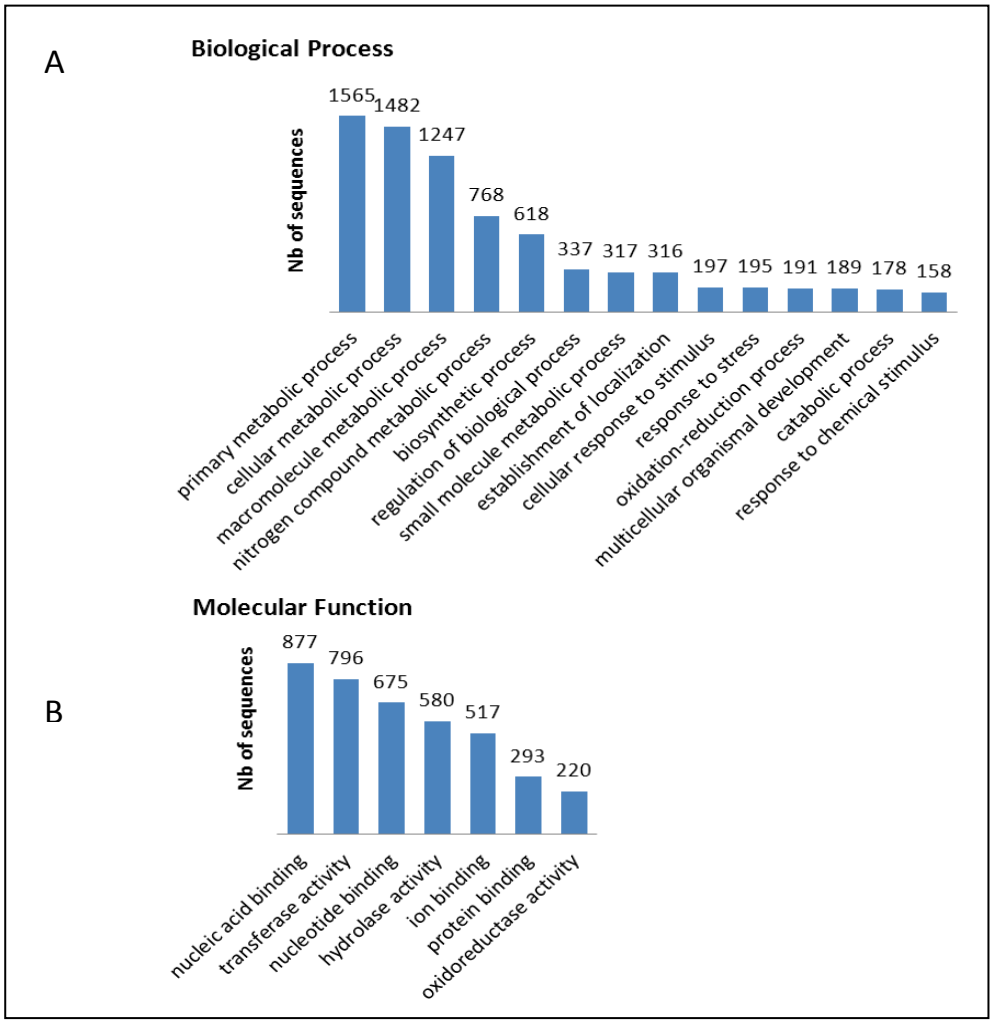

Figure 5

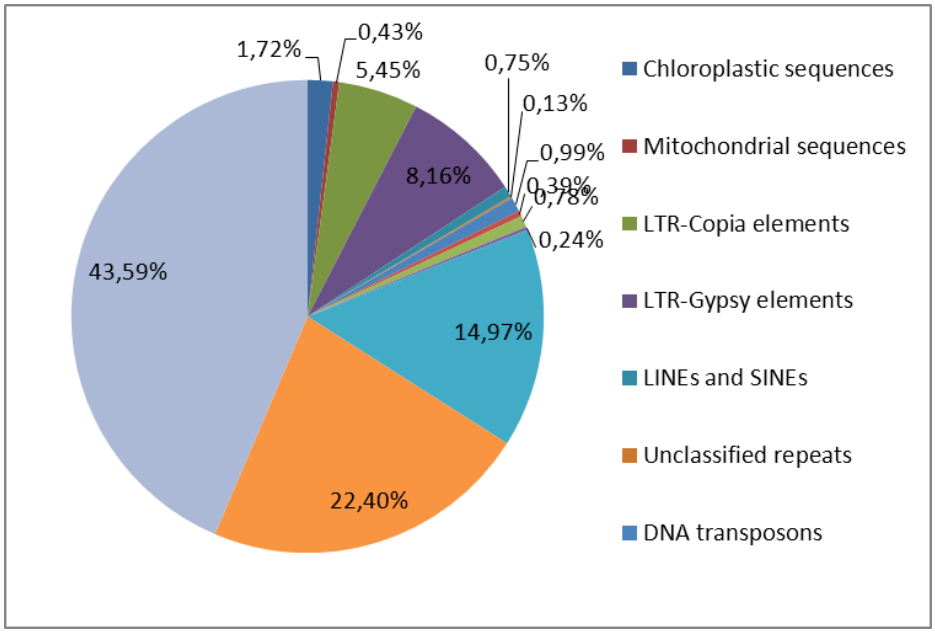

Figure 6

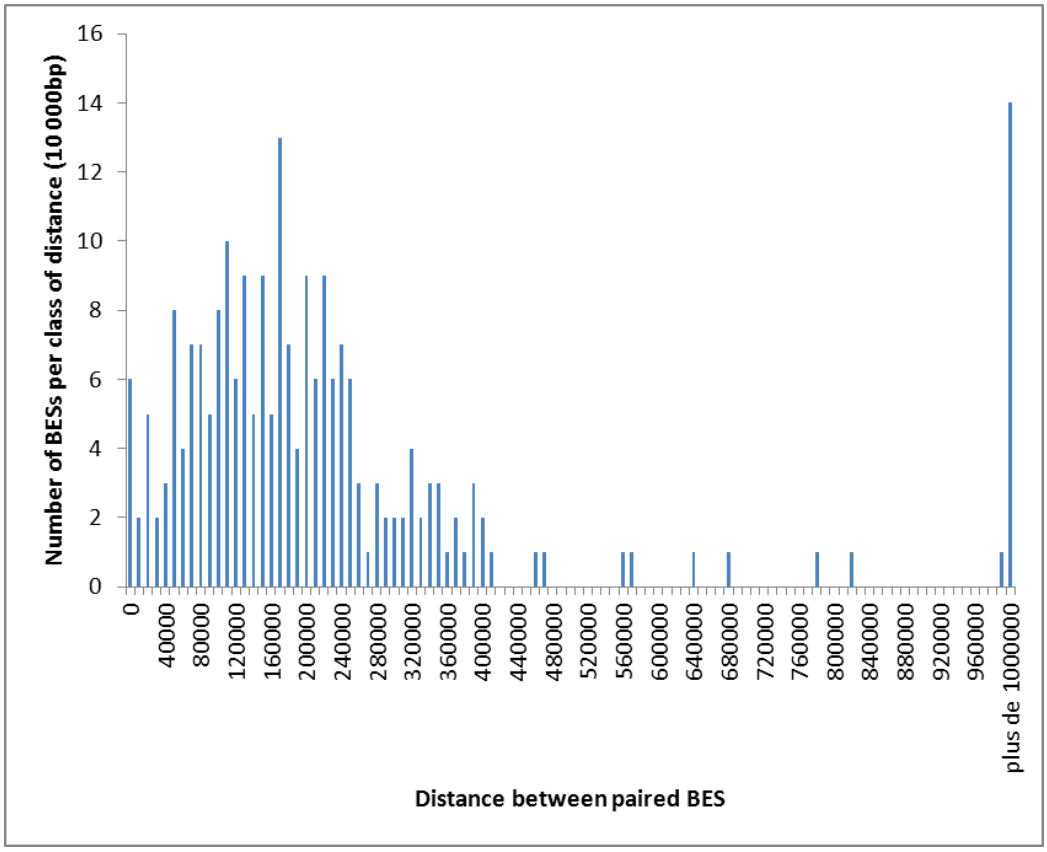

Figure 7

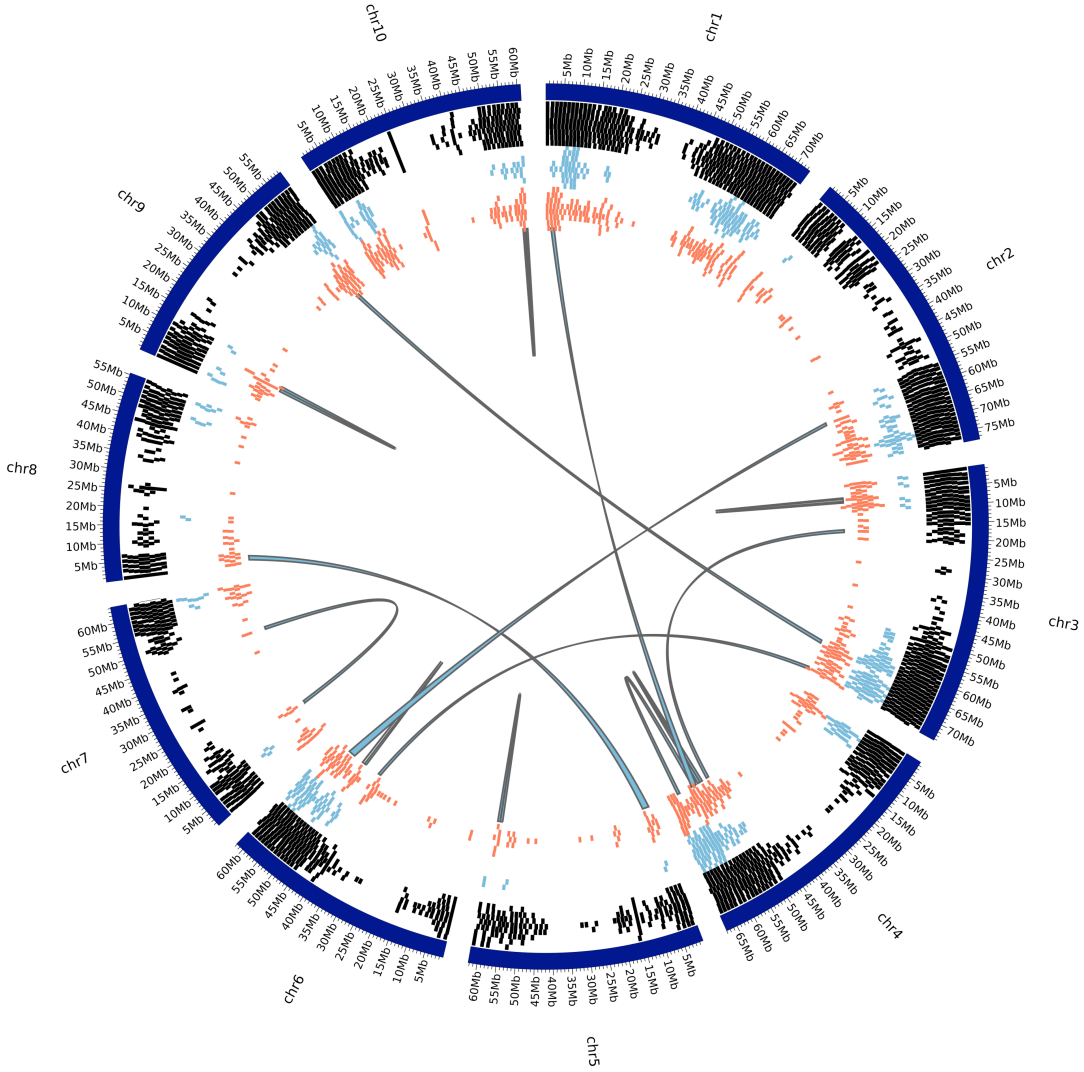

Figure 8

Partie B. Analyse des séquences répétées du génome de *Spartina maritima*

Introduction et démarche suivie

Les séquences répétées forment une part importante du génome des eucaryotes (Bennetzen, 2005) et contribuent à leur dynamique évolutive (Leitch & Leitch, 2008 ; Wicker *et al.*, 2011). Les variations de taille du génome chez les angiospermes, dont la valeur C (« Constante » représentant l'équivalent n d'un lot de chromosomes) peut être multiplié par 2000, sont causées essentiellement par les duplications et l'accumulation de séquences répétées en tandem et d'éléments transposables (Bennetzen, 2005 ; Vitte & Panaud, 2005 ; Leitch *et al.*, 2005). Il est maintenant admis qu'au cours de l'évolution, les génomes peuvent alternativement subir une expansion ou, au contraire, une contraction, dont l'histoire peut être précisée grâce aux analyses phylogénétiques (*e.g*. Jakob *et al.*, 2004 ; Hawkins *et al.*, 2009 ; Hu *et al.*, 2010). Chez les Monocotylédones (Leitch *et al.*, 2010), et notamment les Poacées (Caetano-Anolles, 2005), de grandes disparités de la valeur C sont connues. Les tailles les plus importantes du génome de base haploïde se retrouvent dans la sous-famille des Pooideae et la tribu des Triticeae, ce qui rend notamment le séquençage des espèces cultivées comme l'orge (2n=2x=14, 5,1 Gb ; The International Barley Genome Sequencing Consortium, 2012) ou le blé (*Triticum aestivum* 2n=6x=42, 17 Gb, Choulet *et al.*, 2010 ; Brenchley *et al.*, 2012) particulièrement difficile en raison de l'abondance des éléments répétés qui compliquent les tâches d'assemblage des séquences. Chez les Panicoideae, un exemple classique d'amplification massive de rétrotransposons est représenté par le maïs (2n=2x=20, 2300 Mb, Schnable *et al.*, 2009) depuis sa divergence il y a environ 12 MA d'avec le sorgho (2n=2x=20, 730 Mb, Paterson *et al.*, 2009).

Chez les Chloridoideae, l'augmentation de la taille du génome semble surtout résulter de la polyploïdie, les espèces diploïdes ayant un génome de taille relativement modeste comparées aux autres Poacées diploïdes (Caetano-Anolles, 2005). Compte tenu du niveau de ploïdie (6x) de *S. maritima*, et de la valeur 2C=3,7 – 3,8 pg mesurée par cytométrie en flux (Fortuné *et al.*, 2008), on pourrait estimer la taille de son génome de base (x) à un peu plus de 610 Mb. Chez les polyploïdes (en particulier ceux d'origine hybride), la question d'un réveil de l'activité des éléments transposables sous l'effet du « stress génomique » a été très tôt posée (McClintock, 1984). Les travaux réalisés à ce jour montrent une

transposition plutôt modérée suite à la spéciation allopolyploïde (pour revue Parisod *et al.*, 2010). La plupart des événements d'amplification notés au sein des sous-génomes homéologues des allopolyploïdes s'avèrent être intervenus antérieurement à la spéciation. De la même manière, Hu *et al.* (2010) montrent que les sous-génomes A et D du coton allotétraploïde présentent des éléments amplifiés plus anciennement chez leur parents diploïdes, sans prolifération après l'allopolyploïdisation. Chez les Spartines, cette question a été explorée à l'aide de différentes méthodes de 'transposon display' comme l'IRAP (Inter-retrotransposon Amplified Polymorphism), la REMAP (Retrotransposon-Microsatellite Amplified Polymorphism) par Baumel *et al.* (2002) ou la SSAP (Sequence Specific Amplification Polymorphisms) par Parisod *et al.* (2009). Ces études ont toutes montré une absence de transposition massive suite à l'hybridation et à la duplication du génome, avec toutefois des effets différentiels de restructuration aux sites d'insertion des éléments testés chez les deux hybrides *S.* x *neyrautii* et *S.* x *townsendii* (perte de fragments d'origine maternelle plus prononcée chez *S.* x *townsendii*, Parisod *et al*., 2009). En revanche, d'importantes altérations de la méthylation dans les régions flanquant ces éléments ont pu être détectés par l'utilisation d'enzymes de restriction sensibles à la méthylation (méthode MSTD ou Methyl-Sensitive Transposon Display) (Parisod *et al.*, 2009). Au niveau transcriptomique, Chelaifa (2010) a testé la méthode SSAP sur ADNc en ciblant les 3 éléments (deux éléments de classe I : *Cassandra* et *Wis-like*, et un élément de classe II : *Ins2*) testés précédemment en SSAP et MSTD par Parisod *et al*. (2009). Les résultats indiquent des changements plus ou moins importants suite à l'allopolyploïdie selon les amorces d'éléments testés, et en accord avec la tendance globale des changements transcriptomiques évalués par microarrays (Chelaifa *et al.*, 2010b) : profils d'expressions non-additifs des hybrides, avec dominance du parent maternel *S. alterniflora*.

Les analyses précédemment effectuées indiquent que les régions voisines des éléments transposables sont la cible prioritaire du contrôle épigénétique en réponse au stress de l'hybridation et on peut se demander quel est leur impact sur l'expression des gènes. Toutefois, la proportion des éléments transposables par rapport aux séquences codantes dans le génome des Spartines n'est pas connue. Afin de répondre à cette question et d'identifier les familles d'éléments les plus représentées, nous avons utilisé la plateforme 454 GS FLX Titanium de chez Roche Life Science pour séquencer 96Mb de données du

génome nucléaire représentant près de 15,6% du génome de base de *S. maritima* (616Mb). Ces données ont ensuite été analysées par à une méthode de détection des séquences répétées suivant un regroupement de séquences similaires en clusters, mise au point par Novak *et al.* (2010). Les séquences présentes dans chaque cluster sont assemblées en contigs et annotées en utilisant différentes bases de données d'éléments répétés selon la procédure détaillée dans le chapitre 3.

Résultats

Le séquençage d'ADN génomique a permis d'obtenir 993 229 séquences. Après le filtrage par alignement des séquences chloroplastiques (25 568) et mitochondriales (14 417), le nombre total de séquences nucléaires analysées est de 928 170, représentant 96Mb, soit environ 15,6% du génome de base (estimé à 616 Mb) de *S. maritima*. L'identification des séquences répétées par la méthode des regroupements de séquences similaires a permis d'obtenir 67 418 clusters contenant un total de 459 097 séquences. Parmi ces clusters, nous nous sommes intéressés plus particulièrement à ceux incluant un grand nombre de séquences (susceptibles de correspondre aux éléments les plus représentés dans le génome) : les annotations ont été effectuées sur les 100 premiers clusters contenant 23 383 séquences pour le plus grand à 337 séquences pour le plus petit. Les séquences répétées représentent au total 26,71% du génome nucléaire échantillonné de *S. maritima*.

Les éléments transposables de Classe I et II

Les séquences de Classe I ou rétrotransposons sont les plus nombreuses dans le génome nucléaire échantillonné de *Spartina maritima*. Parmi elles, les rétroéléments *Gypsy* sont les plus abondants (16,19%) alors que les rétroéléments *Copia* représentent 6,83% (Figure 19). Afin d'étudier plus précisément et d'identifier les familles et sous-familles des rétrotransposons les plus abondants, des phylogénies ont été réalisées (Figure 20 et 21) à partir des domaines RT extraits des clusters de *S. maritima* alignés avec les RT annotées disponibles dans les bases de données Repbase. Trois familles d'éléments de type *Gypsy* sont présentes : la famille des Chromovirus est la plus abondante et représente 7,45% des séquences génomiques de *S. maritima* soit près de la moitié des éléments Gypsy. Les

Chromovirus sont caractérisés par la présence d'un chromodomaine (chromointégrase) et sont présent chez de nombreux eucaryotes, dont les plantes et notamment les Poaceae (Gorinsek *et al.*, 2004). Toutes les sous-familles des chromovirus (Galadriel, Tekay et Reina) sont présentes chez *S. maritima* (Figure 20). Les éléments Athila (Pélissier *et al.*, 1995) représentent ensuite 2,02%, les séquences de type Ogre/TAT (Macas & Neumann, 2007) 1,45%, et enfin, 5,26% des séquences restent non-classifiées.

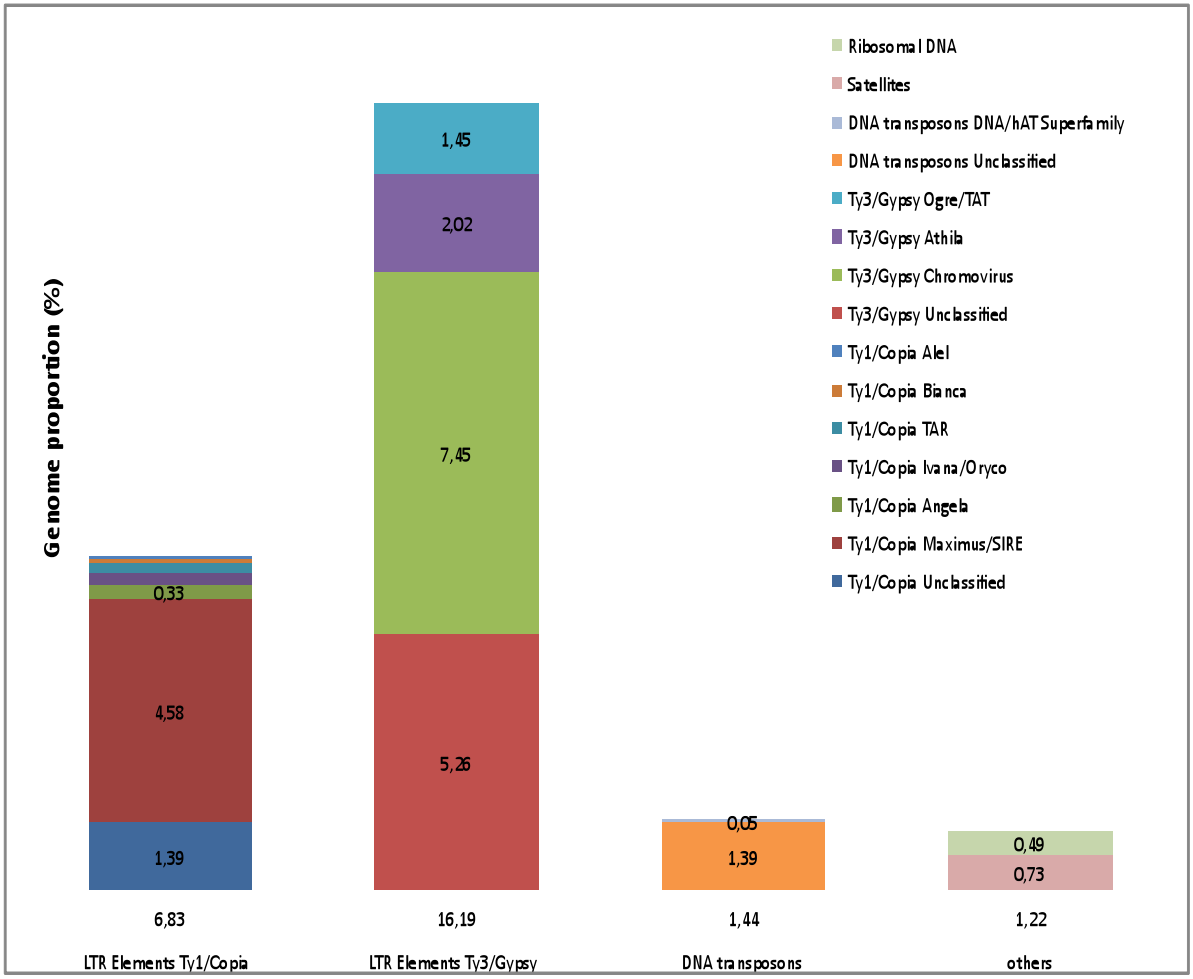

Figure 19 : Classification et proportion génomique de chaque type d'élément présent dans le génome de *Spartina maritima*.

Les rétrotransposons de type *Copia* sont moins nombreux et montrent une diversité phylogénétique moins importante que les éléments *Gypsy* (Figure 19 et 21). Dans cette super-famille, les familles Alel, Bianca, TAR, Ivan/Oryco et Angela représentent en tout

moins de 0,88% du génome de *S. maritima*. La famille Maximus/SIRE est la plus importante et compte pour 67% des éléments copia et 4,58% du génome. Sur la phylogénie des RT (Figure 21), deux lignées distinctes d'éléments Maximus sont spécifiques à *S. maritima*. De plus, une lignée à la base des familles Angela, TAR et TORK sans annotation apparaît spécifique aux séquences RT de *S. maritima* et représente 1,39% du génome échantillonné. Parmi les séquences analysées, aucune séquence correspondant à des rétrotransposons non-LTR n'a été identifiée. Les éléments de Classe II sont peu nombreux et forment 1,44% de l'échantillon (Figure 19).

La Figure 22 montre un histogramme des fréquences de moyenne de similarité entre chaque séquence du jeu de données 454 et l'ensemble des séquences présentant une région homologue à cette séquence pour *S. maritima*. Le pic principal est observé vers 89% d'identité. Autour de 98 et 99% d'identité, une faible augmentation du nombre de séquences est observée. La présence de séquences fortement similaires (98-99%) dans le jeu de donnée pourrait indiquer une expansion plus récente de ces catégories de séquences répétées ou un processus d'homogénéisation, tel que celui qui affecte les familles de gènes ribosomiques.

Autres séquences répétées

Les ADN ribosomaux (ADNr) : un seul cluster (CL15) contient 4 563 séquences assemblées codant l'ADN ribosomique. Ce cluster représente 0,49% du génome échantillonné de *S. maritima* (Figure 19). La séquence complète du gène 45S a pu être été reconstruite par les assemblages formant le contig2. Ce contig a une longueur de 8 448pb, et comprend les locus 18S-5.8S et 26S de l'ADN ribosomique bordés par les régions ETS (External Transcribed Spacer) et IGS (Intergenic Spacer).

Les microsatellites : La proportion des séquences de type satellite est de 0,73% (Figure 19). Trois clusters contiennent les 6 782 séquences assignées comme microsatellites, représentant près de 1,16Mb. Parmi les 6 287 séquences répétées identifiées, plus de 66,0% sont des monomères (A/T) et 16,3% des dimères (AT/TA). Les monomères C/G représentent 9,3% et les autres dimères (AC/GT, CG/CG et AG/CT) 8,2% des séquences. Les trinucléotides et tétranucléotides sont peu nombreux et comptent pour 0,2 et 0,03%, respectivement (Tableau 10).

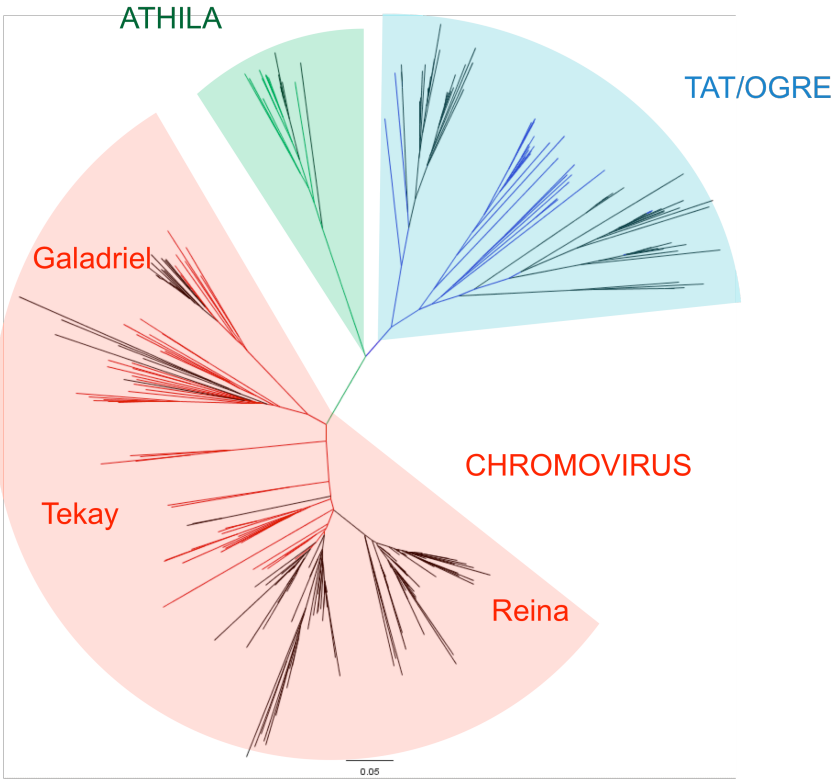

Figure 20 : Analyse phylogénétique des rétrotransposons de type *Ty3/Gypsy* présents chez *Spartina maritima* basée sur les séquences des domaines RT.

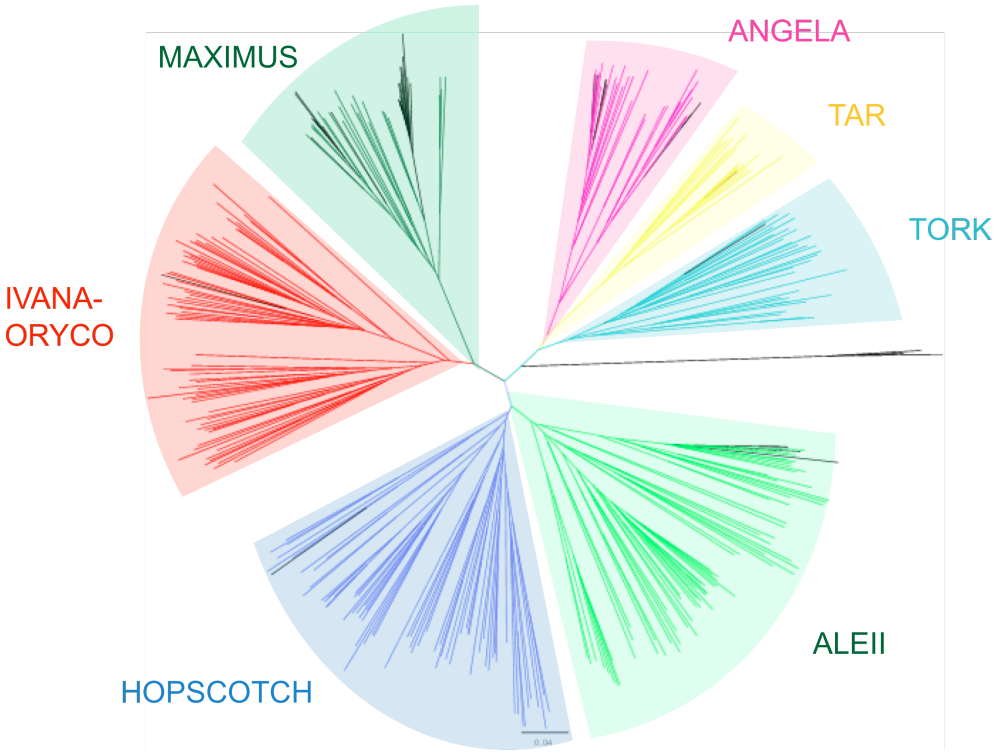

Figure 21 : Analyse phylogénétique des rétrotransposons de type *Ty1/Copia* présents chez *Spartina maritima* basée sur les séquences des domaines RT.

Figure 22 : Histogramme montrant la fréquence des moyennes de similarité entre chaque séquence de *Spartina maritima* et les séquences « homologues » (Blastn e-6 sur au moins 80pb) du jeu de données 454.

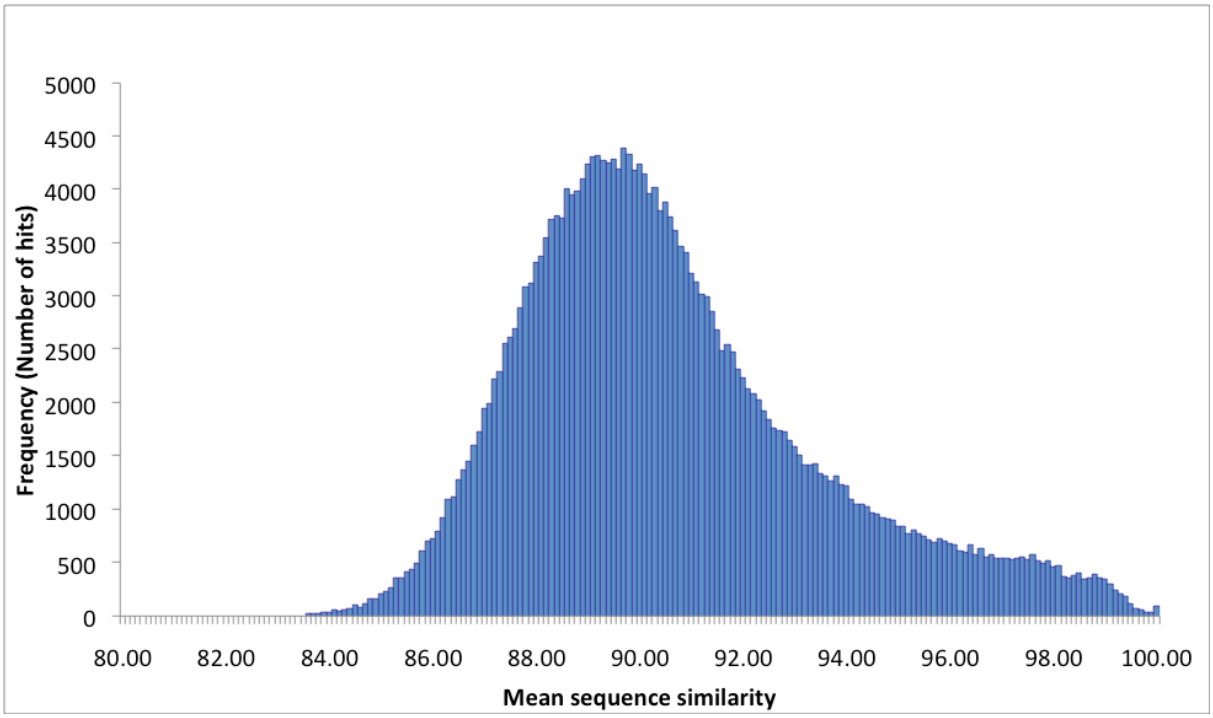

Tableau 10 : Type de microsatellites présents dans les clusters 11, 73 et 90.

Type de microsatellite	Nombre total	Pourcentage
A/T	4150	66,01%
C/G	582	9,26%
AC/GT	332	5,28%
AG/CT	149	2,37%
AT/AT	1027	16,34%
CG/CG	33	0,52%
AAC/GTT	2	0,03%
AAG/CTT	3	0,05%
AAT/ATT	1	0,02%
ACC/GGT	3	0,05%
ACG/CGT	2	0,03%
ACT/AGT	1	0,02%
ATCG/ATCG	2	0,03%

Discussion

Utilisation des NGS pour l'évaluation du compartiment répété et comparaison avec les autres Poacées

L'identification de la composition du compartiment répété a été possible grâce à la plateforme de séquençage haut-débit 454 GS FLX Titanium (Roche Life Science) qui a généré plus de 270 Mb de données sur le génome de *S. maritima* (équivalent à environ 15,6% du génome de base). La démarche de regroupement des séquences similaires en cluster par la méthode des graphes (Novak *et al*., 2010) se révèle efficace pour détecter les séquences répétées, y compris dans les cas où l'on ne dispose que d'une représentation partielle de l'ensemble du génome. Cette procédure a récemment été employée avec succès dans l'analyse de plusieurs espèces : chez le pois pour lequel une faible partie du génome (0,77% du génome de base) avait été séquencée (Macas *et al.*, 2007), le soja (Swaminathan *et al*., 2007), l'orge (Wicker *et al.*, 2009) et la banane (Hribova *et al.*, 2010). Cette procédure a également été utilisée chez des espèces polyploïdes du genre *Nicotiana* (Renny-Byfield *et al*., 2011) et de la famille des Orobanchaceae (Piednoël *et al.,* 2012). La récente publication du génome complet de *Musa acuminata* (D'Hont *et al.*, 2012) permet de tester l'efficacité de la méthode de clustering par comparaison à une détection des séquences répétées par assemblage et alignement de séquences. Ainsi, les proportions relatives d'éléments détectés sont semblables (29,43% de séquences répétées identifiées grâce à la méthode de clustering contre 32,63% dans les chromosomes assemblés de *Musa acuminata*). Néanmoins, dans chacune des catégories les proportions sont sous-évaluées avec la méthode de clustering celle-ci ne prenant en compte qu'un échantillonnage (15,7%) du génome étudié (16,94% contre 17,02% pour les éléments *Copia* ; 7,52% contre 8,28% pour les éléments *Gypsy*). Cette différence devient alors plus notable pour les éléments en faible nombre de copies et donc proportionnellement peu représentés dans les jeux partiels de séquences (1,00% contre 5,87% pour les rétroéléments non-LTR et 0,09% contre 1,42% pour les transposons ADN).

La méthode de clustering permet ainsi de s'affranchir de la comparaison avec un génome séquencé de référence. Une autre méthode, que nous avons utilisée dans la détection des séquences répétées à partir des BES (Partie A) permet en plus l'identification

de séquences répétées potentiellement spécifiques au génome de *S. maritima* (*i.e.* non présentes dans les bases de données ou trop divergentes pour y être identifiées par les méthodes d'alignement utilisées) en alignant les séquences de Spartines sur elles-mêmes. Cette méthode, couplée à une détection par alignement de séquences connues a permis d'identifier une proportion d'éléments répétés de 30,45% dans le génome de *S. maritima* qui se décompose en 15,48% de séquences ayant montré une homologie dans les bases de données et 14,97% de séquences « spécifiques » de *S. maritima* (ou ayant trop divergé par rapport aux séquences des bases de données). Ainsi les résultats des deux méthodes se révèlent très proches et leur utilisation combinée permet une étude plus approfondie du compartiment répété chez *S. maritima*.

Tous les génomes de plantes analysés à ce jour présentent des quantités très variables d'éléments transposables, des éléments de Classe I et II sont toujours identifiés (Kejnovsky *et al.*, 2012). Cependant, les contributions relatives de chaque classe et de leurs sous-classes peuvent varier sensiblement d'une espèce à l'autre. Par exemple, les retrotransposons à LTR sont prédominants dans les génomes du coton et du maïs (Hawkins *et al.*, 2006 ; Vitte & Bennetzen, 2006) comparés au riz où l'on retrouve un grand nombre de transposons. Des différences sont aussi observées dans la distribution sur les chromosomes des éléments de Classe I et II. Ainsi, les régions hétérochromatiques (*e.g.* régions péricentromériques et subtélomériques) sont largement occupées par des retrotransposons à LTR (Kejnovsky *et al.*, 2006). A l'inverse, les transposons à ADN et les MITEs sont plutôt insérés dans les régions euchromatiques, à l'intérieur ou à proximité des gènes (International Rice Genome Sequencing Project, 2005). Chez *Spartina maritima*, les éléments de Classe I sont les plus abondants avec notamment un grand nombre de séquences de type *Gypsy*. En comparant ce génome avec d'autres Poaceae aux génomes complètement séquencés (*Musa acuminata,* D'Hont *et al.*, 2012, *Oryza sativa, Sorghum bicolor, Zea mais*, Paterson *et al.*, 2009) ou analysés selon la même démarche (*Musa acuminata*, Hribova *et al.,* 2010), la proportion de séquences répétées totale trouvée dans le génome de *S. maritima* se rapproche de celle de *M. acuminata* avec les rétroéléments formant la majorité des séquences répétées des deux génomes (Tableau 11). Néanmoins, dans le génome de *M. acuminata*, les éléments *Copia* sont les plus abondants et représentent plus de la moitié des rétroéléments. L'abondance des éléments *Gypsy* dans les génomes des Poaceae se retrouve

chez le riz, le sorgho et le maïs (Tableau 11). Les monocotylédones ont en effet subit une augmentation de l'activité de transposition et une accumulation des éléments à LTR par rapport aux dicotylédones (Vitte & Bennetzen, 2006). De plus, le turnover chez les monocotylédones apparaît beaucoup plus rapide avec des gains et des pertes en quelques millions d'années seulement ce qui expliquerait les expansions actuellement visibles de quelques familles de *Copia* ou de *Gypsy* (Ma *et al.*, 2004). Par exemple, la prolifération de trois familles de rétrotransposons à LTR chez *Oryza australiensis* a entrainé un doublement de la taille du génome par rapport à *O. sativa* en seulement trois millions d'années (Piegu *et al.,* 2006). De plus, Hawkins *et al.* (2006) suggèrent chez les cotons une corrélation entre la taille des génomes et le type d'éléments présents : les génomes plus larges auraient plus d'éléments *Gypsy* alors que les génomes plus petits auraient une fraction plus importante d'éléments *Copia*. Les transposons ADN ne comptent que pour une faible part du génome de *S. maritima*.

Tableau 11 : Comparaison des compartiments répétés entre *Spartina maritima* et d'autres espèces de Poaceae (données d'après Hribova *et al.*, 2010, D'Hont *et al.*, 2012 et Paterson *et al.*, 2009).

	S. maritima	*M. acuminata* (Hribova *et al.*, 2010)	*M. acuminata* (D'Hont *et al.*, 2012)	*O. sativa*	*S. bicolor*	*Z. mais*
Genome size	x=616Mb	x=523Mb	x=523Mb	x=420Mb	x=740Mb	x=2 160Mb
Ploidy level	2n=6x=60	2n=2x=22	2n=2x=22	2n=2x=24	2n=2x=20	2n=2x=20
Retrotransposons	24.05	27,55	31,3	25.78	54.52	79.44
LTR Elements Ty1/Copia	6.83	16.94	17.02	2.47	5.18	21.75
LTR Elements Ty3/Gypsy	16.19	7.52	8.28	12.03	19.00	37.73
Unclassified LTR	1.04	2.09	0.13	10.05	30.30	19.61
Non-LTR Retrotranposon	0	1	5.87	1.24	0.04	0.35
DNA transposons	1.44	0.09	1.42	13.67	7.46	2.68
Tandem repeats	1,22	1,79	NA	1.93	3.13	NA
TOTAL (%)	26.71	29,43	32,72	39.5	62.00	82.1

A l'aide du jeu de données issu de la technologie 454, nous avons aussi pu identifier 6287 motifs microsatellites (SSRs) représentant 0,73% du génome de *S. maritima.* A partir des analyses des BES précédentes (article Partie A), une proportion de 0,25% a été trouvée. Les motifs les plus répandus (les mononucléotides A/T) sont cependant les mêmes en utilisant ces deux approches. Chez *S. pectinata*, Gedye et ses collaborateurs (2010) ont identifié les SSRs présents dans les régions transcrites. Dans cette étude, 3,2% des séquences assemblées sont des SSRs qui présentent une large proportion de motifs riches en GC. Les marqueurs microsatellites, dont le nombre de répétition est variable au niveau intraspécifique, pourront être utilisés pour explorer la variation des populations de Spartines ou détecter les évènements d'hybridation dans les populations naturelles. Des études antérieures avaient pu utiliser une vingtaine de marqueurs microsatellites chez les Spartines (Blum *et al*., 2003 ; Sloop *et al*., 2006) pour analyser la structure génétique des populations de *S. alterniflora*, *S. foliosa* et leurs hybrides envahissants dans la baie de San Francisco (Sloop *et al.,* 2011).

Evolution des séquences répétées chez les polyploïdes

Spartina maritima étant une espèce hexaploïde, la proportion d'éléments répétés et la fréquence de certaines familles dans son génome reflètent les conséquences de la spéciation allopolyploïde à long terme. Chez des allopolyploïdes relativement anciens comme *Nicotiana quadrivalis* (âgé d'un million d'années) et *N. nesophila* (4,5 millions d'années), des remaniements de séquences répétées entre sous-génomes homéologues ont été montrés (Lim *et al.*, 2007). Ainsi, en moins de 5 millions d'années, les allopolyploïdes du genre *Nicotiana* ont vu leurs régions intergéniques complètement renouvelées. Charles *et al.* (2008) ont aussi montré une amplification différentielle d'éléments transposables dans les sous-génomes A et B des blés polyploïdes antérieure à l'évènement d'allotétraploïdisation ayant conduit à la formation de *Triticum turgidum*. A l'inverse, les analyses de trois diploïdes du genre *Brassica* (*B. rapa, B. oleracea, B. nigra*) et de trois allopolyploïdes (*B. napus, B. carinata, B. juncea*) n'ont révélé aucune prolifération de retrotransposons chez les allopolyploïdes (Alix & Heslop-Harrison, 2004) ni d'homogénéisation de retrotransposons entre les sous-génomes (Alix *et al.*, 2008). Certaines familles d'éléments *Gypsy* montrent des groupes de séquences proches phylogénétiquement, présentes uniquement chez *Spartina*

maritima par rapport aux espèces disponibles dans les bases de données. C'est notamment le cas de Chromovirus proches des éléments *Reina*, initialement décrits chez le maïs (Avramova *et al.*, 1996). C'est également le cas de certains éléments *Copia* de la lignée Maximus, particulièrement diversifiés chez les Triticeae mais également chez le riz (Wicker et Keller, 2007) et connus pour la grande taille de leur LTR. Tous ces clades pourraient résulter d'une amplification récente, et il serait intéressant de savoir si cette expansion est postérieure à la divergence des Spartines des autres Chloriodoideae, cette sous-famille étant remarquablement peu connue du point de vue de la nature des séquences répétées. L'évolution à moyen et long terme de lignées polyploïdes peut également s'accompagner d'amplification ou de perte d'éléments. Ce processus est observé chez *Nicotiana sylvestris*, une espèce diploïde où un grand nombre de séquences répétées montrent une similarité moyenne élevée avec les séquences proches dans le même génome (Renny-Byfield *et al.*, 2011). A l'inverse, chez le diploïde *N. tomentosiformis* et l'allotétraploïde *N. tabacum* peu de séquences montrent une similarité moyenne élevée avec d'autres séquences. Ceci a été interprété comme résultant d'une homogénéisation/expansion de séquences répétées chez *N. sylvestris* après la formation de *N. tabacum* ou une élimination de ces séquences du génome de *N. tabacum* (Renny-Byfield *et al.*, 2011).

Les tailles du génome estimées à ce jour chez les Spartines suggèrent une dynamique des séquences répétées au sein du genre. Les quantités d'ADN des espèces tétraploïdes (2n=40) évaluées par cytométrie en flux (valeur 2C) varient de 1,45 pg en moyenne chez *S. bakeri* (Fortuné *et al.*, 2008), 1,56 pg en moyenne chez *S. pectinata* (Kim *et al.*, 2012), 1,54 pg chez *S. cynusoroides*, 1,98 pg chez *S. patens*, à 2,00 pg chez *S. spartinae* (syn. *S. argentinensis*) (Baisakh *et al.*, 2009), soit un écart d'environ 30% entre *S. bakeri* et *S. spartinae*. Chez les Spartines hexaploïdes, *S. alterniflora* et *S. foliosa* (2n=62) présentent une taille de génome de 4,3 à 4,6 pg (Fortuné *et al.*, 2008, Ayres *et al.*, 2008) et *S. maritima* 3,7 à 3,8 pg (Fortuné *et al.*, 2008). Les assemblages et annotations des éléments répétés que nous avons réalisés fournissent à présent des données permettant d'explorer les causes de cette variation de la taille des génomes au cours de l'histoire du genre *Spartina*. Il sera particulièrement intéressant de savoir si la différence de taille de génome notée entre *S. alterniflora* et *S. maritima*, qui auraient divergé depuis moins de 3 millions d'années (Bellot 2010, Bellot *et al.*, en préparation) résulte simplement de l'aneuploïdie (2n=62 pour *S.*

alternifora et 2n=60 pour *S. maritima*) ou si une évolution des séquences répétées est aussi intervenue depuis leur divergence sur les côtes américaines d'une part (*S. alterniflora*), et euro-africaines (*S. maritima*) d'autre part. Les analyses phylogénétiques basées sur les séquences de la Réverse Transcriptase ont montré (tant chez les rétrotransposons de type *Gypsy*, les plus abondants, que chez les types *Copia*) l'existence de lignées qui semblent avoir été plus particulièrement amplifiées chez *Spartina maritima*. Il serait maintenant intéressant d'analyser plus particulièrement ces éléments en produisant des sondes spécifiques des séquences de Spartines de façon à explorer leur évolution au sein du genre et dans la lignée des Chloridoideae. Ces sondes pourraient être utilisées en cytogénétique moléculaire (Hybridation *in situ*) dans le but d'explorer l'importance et la distribution des familles d'éléments répétés sur les chromosomes de Spartines. Par exemple, à la suite des travaux d'identification des éléments transposables chez *Musa acuminata* à partir de séquençage massif (Roche 454), certaines familles de *Copia* identifiées par Hribova *et al*. (2010) ont été analysées par FISH (Fluorescent In situ Hybridization) et s'avèrent distribués sur tous les chromosomes, tandis que d'autres se situent plus préférentiellement en position distale. Les éléments *Gypsy* apparaissent principalement dans les régions péri-centromériques avec quelques exceptions en position distale (Hribova *et al*., 2010). Cette procédure a aussi permis de détecter des séquences satellites abondantes et spécifiques des génomes homéologues de l'allopolyploïde *T. miscellus* et ainsi de pouvoir caractériser cytogénétiquement les chromosomes parentaux et leurs remaniements chez l'allopolyploïde (Chester *et al.*, 2012).

Parmi les séquences répétées, les familles des gènes ribosomiques (et notamment celle codant l'unité 45S) sont connues pour leur évolution rapide, y compris chez les polyploïdes (Wendel *et al.*, 1995), suite à l'évolution concertée des copies paralogues et homéologues. Le contig 45S d'ADN ribosomique que nous avons assemblé a pu être analysé au laboratoire dans le but de détecter les différentes copies présentes dans le génome hexaploïde de *S. maritima*. En analysant les SNPs (Single Nucleotide Polymorphisms) partagés entre reads composant le contig 45S, Boutte (2011) a pu reconstruire différents haplotypes présents chez *S. maritima* et délimiter les zones les plus polymorphes (espaceurs des régions ETS et IGS) moins affectées par l'évolution concertée (Boutte *et al.*, *en préparation*).

En conclusion, nos analyses de l'ADN génomique à partir des séquences d'extrémités de BACs d'une part, et du jeu de données de séquences généré par le pyroséquençage Roche-454 d'autre part, nous ont permis d'avoir pour la première fois une idée de la proportion des séquences répétées dans le génome des Spartines, et de détecter les familles d'éléments qui sont fort probablement les plus représentées dans ce génome. Ces données vont faciliter les annotations ultérieures de séquences générées dans des analyses de RNA-Seq et d'annotations de BACs et de séquences génomiques en cours d'obtention dans l'équipe sur les autres espèces de Spartines. Ces séquences constituent aujourd'hui une première base de données et représentent la première banque d'éléments répétés connus chez les Spartines. Il n'y a pas à notre connaissance de base de référence disponible pour les Chloridoideae.

Chapitre 7

Conclusion et perspectives

On ne fait jamais attention à ce qui a été fait ; on ne voit que ce qui reste à faire
— Marie Curie

Lettre à son frère (18 Mar s 1894)

Au cours de cette thèse, nous avons appréhendé l'analyse et l'évolution du génome des Spartines polyploïdes en exploitant les données issues des nouvelles technologies de séquençage et mené les premières analyses de séquençage haut-débit dans un système biologique non-modèle, caractérisé par une forte redondance des génomes. Nous avons centré nos efforts sur le génome de l'espèce hexaploïde euro-africaine *Spartina maritima* et les transcriptomes des espèces hexaploïdes *S. maritima* et *S. alterniflora*, leurs hybrides F1, *S.* x *neyrautii* et *S.* x *townsendii* et l'allododécaploïde *S. anglica*. Les études ont été menées dans le contexte, récurrent dans ce genre, de la spéciation polyploïde et de l'évolution réticulée par hybridation interspécifique.

Dans la littérature, les polyploïdes montrent souvent des capacités envahissantes plus importantes que leurs parents diploïdes (te Beest *et al.,* 2012). Chez les Spartines, *S. anglica* et *S. alterniflora* ont envahi plusieurs continents et sont reconnues comme des espèces tolérantes à l'immersion dans l'eau salée, aux stress oxydants et résistantes aux pollutions par les hydrocarbures. Ces espèces possèdent une croissance et une vigueur leur conférant des avantages adaptatifs sur les marais salés. Leurs puissants systèmes racinaires et rhizomateux leur permettent d'occuper rapidement les espaces disponibles sur l'estran et d'accélérer la sédimentation, remodelant ainsi la structure physique des marais côtiers ce qui leur vaut le nom d' « ingénieurs d'écosystèmes ».

Dans un premier volet de cette thèse, nous nous sommes plus particulièrement intéressés à la caractérisation du compartiment génique du génome des Spartines et à l'évolution de l'expression de gènes d'intérêt (notamment ceux impliqués dans la tolérance aux stress abiotiques) chez ces espèces polyploïdes. Dans cette perspective, nous avons réalisé un premier transcriptome de référence des espèces hexaploïdes à partir de données de pyroséquençage (Roche 454) transcriptomique de feuilles et de racines. Les assemblages et annotations ont permis d'identifier plusieurs dizaines de gènes d'intérêts ensuite utilisés

pour l'analyse de la variation de l'expression dans les populations naturelles de Spartines hexaploïdes, leurs hybrides F1 et l'allododécaploïde

Dans un second temps, nous avons analysé les données génomiques de l'espèce hexaploïde *S. maritima* à travers les BES et des données de pyroséquençage, afin d'identifier et de caractériser les compartiments codants et non-codants chez cette espèce. Plus particulièrement, nous nous sommes intéressés aux séquences répétées et aux éléments transposables représentant une fraction du génome potentiellement dynamique chez les hybrides et l'allopolyploïde.

Enfin, les nouvelles connaissances sur les gènes et le compartiment non-codant ont permis de poser la question de l'histoire évolutive des Chloridoideae au sein des Poaceae. La sous-famille à laquelle appartiennent les Spartines est en effet peu connue et peu documentée par rapport aux autres sous-familles contenant la majeure partie des espèces de Poaceae cultivées.

L'apport des NGS à la connaissance du génome et du transcriptome d'espèces non modèles

Les récentes avancées dans le domaine du séquençage de génome couplées à la multiplication des capacités de mémoire et de calcul de l'informatique, ouvrent de nouvelles opportunités dans différents domaine de la Biologie, de la médecine personnalisée à la découverte de gènes et leur utilisation en agronomie, mais aussi en écologie évolutive et génomique comparative. Les espèces non-modèles et les gros génomes de plantes peuvent aujourd'hui être séquencés à moindre coût et depuis ces trois dernières années un nombre exponentiel de publications est sorti révélant les avancées incroyables réalisées grâce aux technologies NGS. Les applications sont nombreuses : les transcriptomes d'espèces non-modèles peuvent désormais être caractérisés (nous pouvons citer l'eucalyptus : Novaes *et al*., 2009 ; l'avocatier : Wall *et al*., 2009 ; *Artemisia annua* : Wang *et al.,* 2009) et peuvent être couplés à des études d'expression (chez *Cicer arietinum* : Molina *et al.*, 2008 et le châtaigner : Barakat *et al.*, 2009) et de découverte de gènes candidats (*e.g*. *Amaranthus tuberculatus* : Lee *et al.*, 2009 ou la canne à sucre : Bundock *et al.*, 2009). Les espèces polyploïdes relativement récentes possédant un génome plus complexe commencent aussi à

profiter de ces avancées technologiques (*e.g*. *Tragopogon miscellus* : Buggs *et al.*, 2010 et 2012, et *Nicotiana tabacum* Renny-Byfield *et al.*, 2011, Bombarely *et al.,* 2012).

La mise en place d'un projet de séquençage haut-débit chez une espèce non-modèle polyploïde doit prendre en compte différents paramètres. Tout d'abord, il faut prendre en compte le niveau de ploïdie, identifier la taille du génome, le niveau d'hétérozygotie chez l'espèce en question et la présence d'espèces parentales diploïdes et/ou d'un ou de plusieurs génomes de référence séquencés. Suivant la question scientifique et le but de l'étude, le séquençage de plusieurs organes selon plusieurs conditions peut être nécessaire. Les techniques de normalisation évitent la sur-représentation de transcrits très abondants et permettent d'augmenter le nombre de gènes séquencés différents. Selon ces paramètres, le choix de la plateforme et de la méthode d'assemblage des données s'impose donc aux chercheurs. C'est dans ce contexte que le projet GENOSPART (Genomics of *Spartina*) a vu le jour au cours de l'année 2009. Le choix de la technologie s'est naturellement porté sur la technologie Roche-454 qui fournissait alors des séquences d'une longueur moyenne de 300pb facilitant l'assemblage *de novo* de transcriptomes par rapport aux autres technologies fournissant des fragments de séquences plus courts. De plus, la stratégie de combiner des banques de feuilles et de racines ainsi que des banques normalisées et non-normalisées s'est révélée efficace puisque près de 20 000 gènes différents ont été assemblés à l'aide des données générées pour les cinq espèces de Spartines. Ce transcriptome représente entre 71% et 86% des 28 236 gènes codants trouvés chez *Oryza sativa* (RAP2, Rice Annotation Project, 2008) et ceux trouvés chez *Setaria italica* (entre 24 000 et 29 000 ; Bennetzen *et al*., 2012). *Sorghum bicolor* possède un nombre de gènes intermédiaire, estimé à 27 640 gènes codants (v1.4 Paterson *et al*., 2009). Ainsi, en utilisant une base de données de plusieurs Poacées séquencées, plus de la moitié des gènes potentiellement présents chez les Spartines hexaploïdes sont identifiés.

Dans les perspectives à court terme de ce travail, nous pouvons envisager d'enrichir le transcriptome à partir d'autres organes à différents stades de développement (jeunes tiges et feuilles, inflorescences, rhizomes…). De plus, des études récentes ont montré l'intérêt de combiner différents types de données issues de plusieurs séquenceurs. He *et al.* (2012) ont testé plusieurs combinaisons d'assemblage de séquences Sanger, Roche-454 et Illumina chez *Phragmites australis.* La combinaison des technologies permet un assemblage de meilleure

qualité, des séquences consensus plus longues et un taux d'annotation des contigs de plus de 80%. L'acquisition récente de jeux de données transcriptomiques séquencés par la technologie Illumina (RNA-Seq) sur les 5 espèces de Spartines permettra d'améliorer la quantité et la qualité des informations sur le transcriptome de ces espèces (J. Boutte, A. Salmon et M. Ainouche, analyses en cours). Ces assemblages pourront aussi tirer profit du séquençage de la banque BAC construite récemment par notre équipe chez *Spartina maritima*.

Hétérogénéité des séquences orthologues et identification des homéologues chez les allopolyploïdes

L'analyse des copies de gènes dupliqués par polyploïdie (orthologues chez les parents diploïdes et appelés homéologues au sein du génome allopolyploïde) est d'une importance fondamentale dans les recherches sur l'évolution des polyploïdes. Leur identification permet de reconstruire l'histoire du polyploïde, et ces copies portent les traces des forces évolutives qui s'exercent à court et long terme sur les séquences et sur leur expression.

Chez les polyploïdes relativement récents pour lesquels les parents diploïdes sont encore vivants, il est possible d'utiliser les informations génomiques de ces derniers comme référence pour identifier et analyser l'évolution des sous-génomes homéologues dupliqués au sein de l'allopolyploïde. C'est notamment la démarche employée chez les Tragopogons allotétraploïdes formés il y a environ un siècle (Buggs *et al*., 2010) ou chez le coton allotétraploïde formé il y a 1 à 2 millions d'années (Udall *et al*., 2006 : Salmon *et al*., 2010 ; Flagel *et al*., 2012). Quelques analyses à ce jour ont posé la question de la présence des copies homéologues à détecter dans les assemblages de séquences issues de NGS chez les polyploïdes (*e.g*. Collins *et al*., 2008 chez *Pachycladon enysii ;* Schreiber *et al*., 2012 chez *Triticum aestivum*).

Chez les Spartines, la spéciation allopolyploïde récente (à l'origine de *S. anglica*) fait intervenir des parents déjà hexaploïdes et il n'existe pas de Spartine diploïde connue à ce jour. Dans ce contexte, la réalisation d'un transcriptome de référence chez les parents est essentielle : elle est un préalable à l'analyse ultérieure de l'hétérogénéité de séquence à

chaque locus. La procédure d'assemblage des reads 454 que nous avons utilisée pour la constitution du transcriptome de référence des espèces hexaploïdes (Chapitre 4) visait à maximiser les chances d'assembler des séquences homologues incluant les reads de séquences potentiellement homéologues, de façon à construire des contigs « consensus », tout en essayant d'éviter d'assembler des séquences plus divergentes issues de gènes paralogues (résultant de duplications individuelles de gènes). Ceci a nécessité plusieurs ajustements de la stringence des assemblages (97%, 96%, 95% et 90% d'identité), pour finalement retenir une identité de 90% sur un minimum de 100 pb. Nous ne pouvons toutefois exclure à ce niveau de stringence (et c'est le cas pour tous les assemblages), la possibilité d'assembler des paralogues récents, que seules des analyses complémentaires (validation expérimentale, analyses phylogénétiques) pourraient établir. En tout état de cause, les contigs générés peuvent à présent servir de référence pour aligner les reads correspondants et rechercher les polymorphismes (SNPs) permettant de détecter les différents haplotypes (qui contiendraient les homéologues potentiels et / ou les allèles au sein de chaque paire d'homéologues) pour un contig donné. Cette démarche fait à présent l'objet de développements bioinformatiques et validation expérimentale au laboratoire (J. Boutte et A. Salmon) initiés sur une famille de gènes ribosomiques (Boutte, 2011 ; Aliaga, 2012), et poursuivis actuellement sur jeux de données transcriptomiques 454 et Illumina (Boutte, 2012 ; Boutte *et al., en préparation*). La détection des homéologues putatifs permettra notamment de reconstruire l'histoire de la lignée des Spartines hexaploïdes et de préciser à l'échelle d'un grand nombre de gènes, le degré de rétention ou de perte des copies homéologues (3 paires de copies dupliquées attendues par locus). Fortuné *et al*. (2007) avaient montré par clonage, séquençage et analyses phylogénétiques, la rétention différentielle de copies dupliquées du gène *Waxy* chez les Spartines polyploïdes. L'analyse de 4 contigs assemblés chez *S. alterniflora* et *S. maritima* (Ferreira de Carvalho *et al.*, 2013, Chapitre 4) a montré la présence de 3 à 4 haplotypes différents par contig au sein de chaque espèce, et dans chaque cas, 2 copies nettement plus divergentes (la troisième et la quatrième étant un variant très similaire à l'une ou l'autre de ces copies divergentes). Dans l'hypothèse où ces deux copies représenteraient des homéologues (et l'une ou les deux autres des variants alléliques), ce nombre d'haplotypes dans le transcriptome pourrait s'expliquer soit par une mise sous silence de la troisième copie homéologue attendue chez un hexaploïde, soit par la perte physique ou la conversion d'une de ces copies qu'il

conviendrait de vérifier sur ADN génomique. Notons que deux copies divergentes sont observées au niveau génomique et transcriptomique, chez l'espèce hexaploïde *S. maritima* pour le gène métal tolérance protein, ce qui suggèrerait dans ce cas la possibilité de perte d'un homéologue (Chapitre 5).

Les datations moléculaires concernant l'origine des polyploïdes sont assez difficiles à établir avec certitude (Doyle & Egan, 2010) et cette question est peu documentée dans la sous-famille des Chloridoideae au sein de la tribu des Zoysieae où se placent les Spartines. En utilisant le génome chloroplastique des *S. maritima* reconstruit à parti de séquençage Roche-454 (Bellot, 2010 ; Bellot *et al.*, *en préparation*), de premières estimations à partir de plusieurs régions chloroplastiques codantes et non codantes et sur la base de calibrations connues dans la famille des Poacées, la divergence entre représentants des lignées tétraploïde et hexaploïde de Spartines a été estimée à moins de 6 millions d'années et la divergence entre les hexaploïdes *S. alterniflora* et *S. maritima* à moins de 3 millions d'années. Les travaux de la littérature suggèrent que la dynamique de perte de gènes, ou de conversion génique peut intervenir dans une période évolutive similaire (exemple : le coton allotétraploïde formé il y a 1 à 2 millions d'années, Salmon *et al*., 2010 ; Flagel *et al.*, 2012) ou même juste après la formation du polyploïde comme dans les populations naturelles de *Tragopogon miscellus* (*e.g*. Tate *et al*., 2009).

Les analyses en cours au laboratoire permettront de vérifier les hypothèses émises au vu de ces premières explorations du génome et du transcriptome des Spartines, sur un plus grand nombre de gènes. L'identification des copies homéologues pourra permettre d'analyser le taux de rétention des homéologues au sein des génomes hexaploïdes, leur histoire, et pour chaque gène, l'évaluation du niveau global de son expression, et les contributions respectives de chacun des homéologues au transcriptome. Ceci pourra ensuite être plus précisément analysé chez les hybrides F1 et l'allododécaploïde récemment formés. Des analyses sont également en cours au laboratoire, concernant le séquençage de plusieurs BACs (collaboration avec le CNRGV, Toulouse et le Génoscope, Evry) contenant des régions d'intérêt (la région ADH-1, qui a fait l'objet de plusieurs analyses comparatives chez les Poacées, *e.g*. Jannoo *et al.*, 2007, et la région contenant le gène CAD-2 codant la Cinnamoyl Alcool Deshydrogénase impliquée dans la synthèse de la lignine des parois cellulaires). J'ai contribué au screening de la banque BAC ayant permis la détection des BACs

potentiellement homéologues chez *S. maritima* et l'analyse des séquences est actuellement en cours. L'analyse comparative de ces BACs amènera également des informations sur l'évolution locale de régions homéologues.

Evolution de l'expression des gènes des Spartines polyploïdes sur les marais salés

Les variations de l'expression globale de 13 gènes en conditions naturelles ont été examinées au niveau intra- et interspécifique chez les Spartines hexaploïdes et leurs descendants hybrides et allododécaploïde. Une telle approche va refléter à la fois l'hétérogénéité des conditions stationnelles et environnementales dans lesquelles les échantillons ont été récoltés, la réponse intrinsèque de l'expression des gènes à l'environnement, au stade physiologique et de développement de l'organe récolté, et les effets de l'évolution génétique suite à la spéciation. Différents types de spéciation concernent les polyploïdes analysés : spéciation divergente entre les deux parents hexaploïdes (estimée il y a moins de 3 millions d'année comme indiqué précédemment), évolution réticulée chez les deux hybrides *S.* x *townsendii* et *S.* x *neyrautii* formés indépendamment, et effets de la duplication d'un génome hybride (spéciation allopolyploïde) chez *S. anglica*. Malgré la complexité évidente de l'interprétation des changements d'expression des gènes susceptibles d'être observés chez ces espèces sur le terrain, nous avons pris le parti d'explorer les profils d'expression d'une sélection de gènes d'intérêt dans les conditions dans lesquelles évoluent naturellement les Spartines, démarche importante dans la compréhension des capacités adaptatives de ces espèces.

Nous avons noté une faible variabilité d'expression des gènes au sein des populations analysées. Ceci pourrait, comme nous l'avons discuté au Chapitre 6, résulter de biais statistiques (variance élevée dans certaines populations, biais liés à la proportion des transcrits détectés par la PCR quantitative). On notera également la base génétique réduite connue au niveau inter-individuel dans les populations de Spartines, pour différentes raisons biologiques et/ou historiques : ces espèces pérennes ont en effet une multiplication végétative prédominante. *Spartina maritima* ne produit pas de graines dans les populations analysées, et son uniformité génétique dans les populations européennes a été soulignée par Yannic *et al*. (2004). *Spartina alterniflora* est une espèce allogame, qui bien que

possédant une diversité génétique dans sa région native sur les côtes Atlantiques américaines (Utomo *et al*., 2009) présente certainement moins de diversité introduite dans l'ouest de l'Europe où un seul haplotype chloroplastique est détecté à ce jour (Gharib, 2012). Les hybrides F1 et *S. anglica* se sont donc formés à partir de génotypes parentaux très similaires, ce qui a eu pour conséquence un fort goulot génétique à l'origine de la nouvelle espèce allododécaploïde dans laquelle une très faible diversité génétique a été notée dans sa région native (Europe) et dans les régions où elle a été introduite (Baumel *et al.*, 2001 ; 2002). Bien que peu de populations pour chaque espèce aient été testées, on a toutefois pu noter des différences significatives d'expression entre populations de la même espèce, qui reflèteraient probablement l'effet des conditions locales sur l'expression des gènes.

Au niveau interspécifique, nous avons retrouvé pour plusieurs gènes (7/13 analysés) la tendance précédemment observée en mêmes conditions expérimentales à l'aide de microarrays de riz par Chelaifa *et al.* (2010a et b), à savoir une sur-expression des gènes de *S. alterniflora* par rapport à *S. maritima*. De plus, certains gènes qui n'apparaissaient pas différentiellement exprimés en mêmes conditions de culture s'avèrent surexprimés, toujours chez *S. alterniflora*. Les profils d'expression de *S.* x *townsendii* et *S.* x *neyrautii* sont très similaires pour la majorité (12/13) des gènes examinés, et dans la plupart des cas ils sont sous-exprimés par rapport aux deux parents. En revanche, on a noté une augmentation significative des niveaux d'expression chez *S. anglica* par rapport à *S.* x *townsendii*. Il est donc intéressant de noter que nous retrouvons, en conditions naturelles, sur des sites différents où les espèces ont été échantillonnées, les effets marqués et différentiels de l'hybridation interspécifique d'une part et de la duplication du génome d'autre part, qui avaient été détectés par une approche différente, en mêmes conditions de culture expérimentale (Chelaifa *et al*., 2010a). Les différentes formes de cette évolution de l'expression des gènes par rapport à l'additivité attendue des niveaux d'expression des parents hexaploïdes (dominance d'expression parentale, sur-expression ou sous-expression) et les fonctions des gènes concernés ont été discutées au Chapitre 6 et ouvrent la voie à de nouvelles perspectives d'exploration des rôles de stress environnementaux (stress salin, pollution aux métaux lourds…) sur ces espèces.

Les transcriptomes de référence que nous avons réalisés représentent dans ce contexte une base permettant l'analyse à haut débit (RNA-Seq) de l'expression des gènes dans les

populations naturelles ; les contigs que nous avons assemblés serviront de référence d'alignement des lectures de séquences obtenues avec une plus grande profondeur par des technologies telles que celle d'Illumina. Le transcriptome de référence obtenu par la technologie 454 a également permis de concevoir une puce spécifique aux Spartines hexaploïdes (A. Salmon, non publié) qui est actuellement utilisée pour analyser les effets de de la marée noire (intervenue en 2010 dans le Golfe du Mexique le long des côtes atlantiques américaines) sur le transcriptome dans les populations naturelles de *Spartina alterniflora* (collaboration avec C. Richards, Université de Floride, USA).

Séquences répétées et dynamique du génome chez les Spartines

En utilisant les séquences génomiques (BAC End Sequences et run de 454) obtenues chez l'espèce hexaploïde européenne *S. maritima*, nous avons pour la première fois, estimé la proportion des séquences répétées dans un génome de Spartine et identifié les familles les plus représentées dans notre échantillonnage du génome. Ce travail exploratoire devra être poursuivi et ouvre de nombreuses perspectives : il reste notamment à préciser la nature des séquences répétées « spécifiques » aux Spartines (lignées se différenciant des clades d'éléments répertoriés dans les bases de données, « clusters » de séquences non annotées…), et l'analyse de leur évolution dans les différentes lignées de Spartines. Les variations de la taille des génomes de Spartines de même niveau de ploïdie rapportées à ce jour (Fortuné *et al*., 2007 ; Baisakh *et al*., 2009 ; Kim et *al*., 2012) suggèrent une dynamique des séquences répétées qu'il sera intéressant d'analyser.

Une question importante sur l'impact des éléments transposables dans le génome des Spartines concerne leur distribution par rapport aux gènes, qui pourra par exemple être précisée localement par les analyses comparatives de BAC en cours d'annotation chez *Spartina maritima*. La poursuite du séquençage à haut débit (amélioration de la profondeur de séquençage) ou les méthodes de captures de séquences cibles (comme celles des gènes dont l'expression s'est avérée rapidement altérée par l'hybridation ou la duplication du génome permettraient de mieux cerner l'environnement des gènes, donnée importante pour explorer les mécanismes épigénétiques (*e.g*. distribution de la méthylation de l'ADN) qui semblent jouer un rôle important sous l'effet de l'hybridation interspécifique chez les

Spartines (Salmon *et al*., 2005) et plus particulièrement dans les régions voisines d'éléments transposables (Parisod *et al*., 2009). L'exploration des éléments dans le transcriptome est une approche qui reste également à mener.

Génomique comparative

L'histoire évolutive des Poacées est aujourd'hui bien documentée chez les espèces pour lesquelles des ressources génomiques étaient disponibles. Néanmoins, les relations phylogénétiques entre la sous-famille des Chloridoideae et les autres Poacées restent encore à approfondir. Dans les approches comparatives présentées dans le chapitre 2, nous avons noté que très peu de données génomiques de cette sous-famille sont disponibles, accentuant la difficulté de bien intégrer les Spartines au sein des autres Poacées. L'étude de la synténie entre *S. maritima* et les génomes de Poacées phylogénétiquement proches permet d'apporter un premier éclairage sur les relations évolutives entre les Chloridoideae et les autres lignées (Panicoideae, Ehrartoideae), où seulement quelques potentiels remaniements interchromosomiques (de l'ordre de la Mb) sont observés en utilisant les données BES. La carte physique d'*Eleusine coracana* a été comparée à celle du riz. Les auteurs ont pu mettre en évidence 3 évènements de fusions afin de donner un génome au nombre chromosomique de base de x=9. Au sein des Chloridoideae, le nombre de base varie de x=7 à x=10 (comme chez les Spartines), aussi serait-il intéressant de savoir si ces fusions sont partagées entre les différentes tribus de la sous-famille ou si les génomes de chaque lignée ont subi des évènements de fusions / cassures chromosomiques différents. La sous-famille des Chloridoideae fait l'objet d'études phylogénétiques qui permettent de mieux appréhender la tribu des Zoysieae à laquelle fait partie les Spartines. Les nouvelles données moléculaires pourraient permettre de dater plus précisément les divergences entre les différentes tribus, et la divergence entre les Chloridoideae et les Panicoideae.

Au sein même de la lignée des Spartines les relations phylogénétiques, l'histoire évolutive et les évènements qui ont conduit à l'obtention d'espèces uniquement polyploïdes sont encore mal connus. Les premiers résultats sur l'hétérogénéité de séquences (sur le plan génomique et transcriptomique) chez les Spartines suggèrent la présence de deux copies homéologues qui devra être vérifiée sur un plus grand nombre de gènes et des analyses phylogénétiques

pour établir l'origine évolutive des copies détectées. Les approches permises par le séquençage massif parallèle devraient permettre d'élucider l'histoire profonde des différents clades de Spartines dans un avenir très proche.

Bibliographie

A

Abbott RJ, Lowe AJ (2004). Origins, establishment and evolution of new polyploid species: *Senecio cambrensis* and *S. eboracensis* in the British Isles. *Biological Journal of the Linnean Society* **82**: 467-474.

Abrouk M, Murat F, Pont C, Messing J, Jackson S, Faraut T, *et al.* (2010). Palaeogenomics of plants: Synteny-based modelling of extinct ancestors. *Trends in Plant Science* **15**: 479–487.

Adams KL, Cronn R, Percifield R, Wendel JF (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences* **100**: 4649-4654.

Adams KL, Percifield R, Wendel JF (2004). Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics* **168**: 2217 - 2226.

Adams KL, Wendel JF (2004). Exploring the genomic mysteries of polyploidy in cotton. *Biological Journal of the Linnean Society* **82**: 573–581.

Adams KL, Wendel JF (2005). Novel patterns of gene expression in polyploid plants. *Trends in Genetics* **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. *Biological Journal of the Linnean Society* **82**: 475-484.

Ainouche ML, Baumel A, Salmon A, Yannic G (2004b). Hybridisation, polyploidy and speciation in *Spartina* Schreb. (Poaceae). *New Phytologist* **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). *Biological Invasions* **11**: 1159–1173.

Ainouche M, Chelaifa H, Ferreira de Carvalho J, Bellot S, Ainouche A, Salmon A (2012). Polyploid Evolution in *Spartina*: Dealing with Highly Redundant Hybrid Genomes. In: Soltis PS, Soltis DE (eds) *Polyploidy and Genome Evolution*, Springer Berlin Heidelberg: Berlin, Heidelberg, pp 225-243.

Ainouche ML, Jenczewski E (2010). Focus on polyploidy. *New Phytologist* **186**: 1-4.

Albertin W, Brabant P, Catrice O, Eber F, Jenczewski E, Chevre AM, *et al.* (2005). Autopolyploidy in cabbage (*Brassica oleracea* L.) does not alter significantly the proteomes of green tissues. *Proteomics* **5**: 2131-2139.

Albertin W, Balliau T, Brabant P, Chevre A-M, Eber F, Malosse C, *et al.* (2006). Numerous and rapid nonstochastic modifications of gene products in newly synthesized *Brassica napus* allotetraploids. *Genetics* **173**: 1101-1113.

Aliaga B (2012). *Reconstruction et assemblage des copies d'ADN ribosomique (ADNr 45S) chez les Spartines polyploïdes*. Rapport de Master 1 Biologie Intégrée (Molécules, Population et Développement Durable), Université de Perpignan – Via Domitia.

Alix K, Heslop-harrison JS (pat) (2004). The diversity of retroelements in diploid and allotetraploid *Brassica* species. *Plant Molecular Biology* **54**: 895–909.

Alix K, Joets J, Ryder CD, Moore J, Barker GC, Bailey JP, *et al.* (2008). The CACTA transposon *Bot1* played a major role in *Brassica* genome divergence and gene proliferation. *The Plant Journal* **56**: 1030–1044.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.

An SQ, Gu BH, Zhou CF, Wang ZS, Deng ZF, Zhi YB, *et al.* (2007). *Spartina* invasion in China: implications for invasive species management and future research. *Weed Research* **47**: 183–191.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796-815.

Arrivault S, Senger T, Krämer U (2006). The *Arabidopsis* metal tolerance protein AtMTP3 maintains metal homeostasis by mediating Zn exclusion from the shoot under Fe deficiency and Zn oversupply. *The Plant Journal* **46**: 861–879.

Avramova Z, Tikhonov A, SanMiguel P, Jin Y-K, Liu C, Woo S-S, *et al.* (1996). Gene identification in a complex chromosomal continuum by local genomic cross-referencing. *The Plant Journal* **10**: 1163–1168.

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. *Biological Invasions* **6**: 221–231.

Ayres DR, Zaremba K, Sloop CM, Strong DR (2008). Sexual reproduction of cordgrass hybrids (*Spartina foliosa* x *alterniflora*) invading tidal marshes in San Francisco Bay. *Diversity and Distributions* **14**: 187–195.

B

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 Science* **170**: 1141–1149.

Baisakh N, Subudhi PK, Varadwaj P (2008). Primary responses to salt stress in a halophyte, smooth cordgrass (*Spartina alterniflora* Loisel.). *Functional & Integrative Genomics* **8**: 287–300.

Baisakh N, Subudhi PK, Arumuganathan K, Parco AP, Harrison SA, Knott CA, *et al.* (2009). Development and interspecific transferability of genic microsatellite markers in *Spartina* spp. with different genome size. *Aquatic Botany* **91**: 262 – 266.

Baisakh N, RamanaRao MV, Rajasekaran K, Subudhi P, Janda J, Galbraith D, *et al.* (2012). Enhanced salt stress tolerance of rice plants expressing a vacuolar H+ -ATPase subunit c1 (SaVHAc1) gene from the halophyte grass Spartina alterniflora Löisel. *Plant Biotechnology Journal* **10**: 453–464.

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 Biology* **9**: 51.

Barrientos A, Barros MH, Valnot I, Rötig A, Rustin P, Tzagoloff A (2002). Cytochrome oxidase in health and disease. *Gene* **286**: 53 – 63.

Baumel A, Ainouche ML, Levasseur JE (2001). Molecular investigations in populations of *Spartina anglica* C.E. Hubbard (Poaceae) invading coastal Brittany (France). *Molecular Ecology* **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). *Molecular Phylogenetics and Evolution* **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). *Molecular Biology and Evolution* **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* x *neyrautii* re-examined. *Plant Systematics and Evolution* **237**: 87–97.

Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K, *et al.* (2011). Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**: 245–249.

te Beest M, Le Roux JJ, Richardson DM, Brysting AK, Suda J, Kubešová M, *et al.* (2012). The more the better? The role of polyploidy in facilitating plant invasions. *Annals of Botany* **109**: 19 -45.

Bellot S (2010). *Evolution du génome chloroplastique chez Spartina (Poaceae, Chloridoideae)*. Université de Montpellier 2, Rapport de Master Biologie Ecologie Evolution.

Bennetzen JL (2005). Transposable elements, gene creation and genome rearrangement in flowering plants. *Current Opinion in Genetics Development* **15**: 621 – 627.

Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, *et al.* (2012). Reference genome sequence of the model plant *Setaria*. *Nature Biotechnology* **30**: 555–561.

Bérard A, Le Paslier MC, Dardevet M, Exbrayat-Vinson F, Bonnin I, Cenci A, *et al.* (2009). High-throughput single nucleotide polymorphism genotyping in wheat (*Triticum* spp.). *Plant Biotechnology Journal* **7**: 364–374.

Binzel ML (1995). NACL-induced accumulation of tonoplast and plasma-membrane H+ ATPase message in tomato. *Physiologia Plantarum* **94**: 722–728.

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.

Blanc G, Wolfe KH (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**: 1667 – 1678.

Blum MJ, Sloop CM, Ayres DR, Strong DR (2003). Characterization of microsatellite loci in *Spartina* species (Poaceae). *Molecular Ecology Notes* **4**: 39–42.

Bombarely A, Edwards KD, Sanchez-Tamburrino J, Mueller LA (2012). Deciphering the complex leaf transcriptome of the allotetraploid species *Nicotiana tabacum*: A phylogenomic perspective. *BMC Genomics* **13**: 406.

Bourdau P (2012). *Polyploïdie et variation de l'expression génique dans les populations naturelles de Spartines envahissant les marais salés*. Université de Rennes 1.

Boutte J (2011). *Détection automatisée des copies dupliquées de gènes ribosomiques chez une espèce hexaploïde (Spartina maritima), à partir de données de pyroséquençage*. Rapport de Master 1 MSB (Modélisation de Systèmes biologiques), Université de Rennes 1.

Boutte J (2012). *Evolution des génomes homéologues chez les Spartines polyploïdes : Assemblages de séquences transcriptomiques (Roche-454 et Illumina) et détection de copies dupliquées*. Rapport de Master 2 MSB (Modélisation de Systèmes biologiques), Université de Rennes 1.

Brautigam A, Mullick T, Schliesky S, Weber APM (2011). Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C3 and C4 species. *Journal of Experimental Botany* **62**: 3093–3102.

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, *et al.* (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**: 705–710.

Buggs RJ, 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. *Molecular Ecology* **19**: 132 – 146.

Buggs RJA, Chamala S, Wu W, Tate JA, Schnable PS, Soltis DE, *et al.* (2012). Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Current Biology* **22**: 248 - 252.

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 biotechnology journal* **7**: 347–354.

Bundock PC, Casu RE, Henry RJ (2012). Enrichment of genomic DNA for polymorphism detection in a non-model highly polyploid crop plant. *Plant Biotechnology Journal* **10**: 657–667.

Buschiazzo E, Gemmell NJ (2006). The rise, fall and renaissance of microsatellites in eukaryotic genomes. *BioEssays* **28**: 1040–1050.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, *et al.* (2008). ALLPATHS: *De novo* assembly of whole-genome shotgun microreads. *Genome Research* **18**: 810–820.

C

Caetano-Anollés G (2005). Evolution of Genome Size in the Grasses. *Crop Science* **45**: 1809.

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. *Marine Pollution Bulletin* **56**: 2037 – 2042.

Campos JA, Herrera M, Biurrun I, Loidi J (2004). The role of alien plants in the natural coastal vegetation in central-northern Spain. *Biodiversity and Conservation* **13**: 2275–2293.

Chagué V, Just J, Mestiri I, Balzergue S, Tanguy A-M, Huneau C, *et al.* (2010). Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. *New Phytologist* **187**: 1181-1194.

Chang P, Dilkes B, McMahon M, Comai L, Nuzhdin S (2010). Homoeolog-specific retention and use in allotetraploid *Arabidopsis suecica* depends on parent of origin and network partners. *Genome Biology* **11**: R125.

Chantret N, Salse J, Sabot F, Rahman S, Bellec A, Laubin B, *et al.* (2005). Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *The Plant Cell Online* **17**: 1033 -1045.

Chao S, Sharp PJ, Worland AJ, Warham EJ, Koebner RMD, Gale MD (1989). RFLP-based genetic maps of wheat homoeologous group 7 chromosomes. *Theoretical and Applied Genetics* **78**: 495–504.

Charles M, Belcram H, Just J, Huneau C, Viollet A, Couloux A, *et al.* (2008). Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. *Genetics* **180**: 1071 –1086.

Chaudhary B, Flagel L, Stupar RM, Udall JA, Verma N, Springer NM, *et al.* (2009). Reciprocal silencing, transcriptional bias and functional divergence of komeologs in polyploid cotton (*Gossypium*). *Genetics* **182**: 503 - 517.

Check Hayden E (2012). Nanopore genome sequencer makes its debut. *Nature*. doi:10.1038/nature.2012.10051

Chelaifa H, Mahé F, Ainouche M (2010a). Transcriptome divergence between the hexaploid salt-marsh sister species *Spartina maritima* and *Spartina alterniflora* (Poaceae). *Molecular Ecology* **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 Phytologist* **186**: 161–174.

Chelaifa H (2010). *Spéciation allopolyploïde et dynamique fonctionnelle du génome chez les Spartines*. Manuscrit de thèse, Université de Rennes 1.

Chen H, Lai Z, Shi J, Xiao Y, Chen Z, Xu X (2010). Roles of arabidopsis WRKY18, WRKY40 and WRKY60 transcription factors in plant responses to abscisic acid and abiotic stress. *BMC Plant Biology* **10**: 281.

Chen M, Ha M, Lackey E, Wang J, Chen Z (2008). RNAi of met1 reduces DNA methylation and induces genome-specific changes in gene expression and centromeric small RNA accumulation in *Arabidopsis* allopolyploids. *Genetics* **178**: 1845 - 1858.

Chen ZJ, Wang J, Tian L, Lee H-S, Wang JJ, Chen M, *et al.* (2004). The development of an *Arabidopsis* model system for genome-wide analysis of polyploidy effects. *Biological Journal of the Linnean Society* **82**: 689-700.

Chen ZJ (2007). Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annual Review of Plant Biology* **58**: 377-406.

Chester M, Gallagher JP, Symonds VV, Cruz da Silva AV, Mavrodiev EV, Leitch AR, *et al.* (2012). Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proceedings of the National Academy of Sciences* **109**: 1176 –1181.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WEG, Wetter T, *et al.* (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research* **14**: 1147 – 1159.

Chevreux B, Wetter T, Suhai S (1999). Genome sequence assembly using trace signals and additional sequence information. *Comput Sci Biol: Proceedings of the German Conference on Bioinformatics (GCB)* **99**: 45 – 56.

Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, *et al.* (2010). Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *The Plant Cell Online* **22**: 1686 –1701.

Christin PA, Besnard G, Samaritani E, Duvall MR, Hodkinson TR, Savolainen V, *et al.* (2008). Oligocene CO2 decline promoted C4 photosynthesis in grasses. *Current Biology* **18**: 37–43.

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.

Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology* **4**: 265–270.

Collins LJ, Biggs PJ, Voelckel C, Joly S (2008). An approach to transcriptome analysis of non-model organisms using short-read sequences. *Genome informatics International Conference on Genome Informatics* **21**: 3–14.

Comai L (2000). Genetic and epigenetic interactions in allopolyploid plants. *Plant Molecular Biology* **43**: 387-399.

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.

Consortium TIBGS (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**: 711–716.

Cottet M, de Montaudouin X, Blanchet H, 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**: 629–640.

Cui LY, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, *et al.* (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Research* **16**: 738–749.

Cusimano N, Sousa A, Renner SS (2012). Maximum likelihood inference implies a high, not a low, ancestral haploid chromosome number in *Araceae*, with a critique of the bias introduced by "x." *Annals of Botany* **109**: 681 -692.

D

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.

Dai N, Schaffer A, Petreikov M, Shahak Y, Giller Y, Ratner K, *et al.* (1999). Overexpression of arabidopsis hexokinase in tomato plants inhibits growth, reduces photosynthesis, and induces rapid senescence. *The Plant Cell Online* **11**: 1253 –1266.

Deragon JM, Casacuberta JM, Panaud O (2008). Plant Transposable Elements. In: Volff J-N (ed) *Genome Dynamics*, KARGER: Basel, pp 69–82.

Deschamps S, Campbell MA (2009). Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Molecular Breeding* **25**: 553–570.

Devos KM, Millan T, Gale MD (1993a). Comparative RFLP maps of the homoeologous group-2 chromosomes of wheat, rye and barley. *Theoretical and Applied Genetics* **85**: 784-792.

Devos KM, Atkinson MD, Chinoy CN, Francis HA, Harcourt RL, Koebner RMD, *et al.* (1993b). Chromosomal rearrangements in the rye genome relative to that of wheat. *Theoretical and Applied Genetics* **85**: 673-680.

Devos KM (2005). Updating the "Crop Circle." *Current Opinion in Plant Biology* **8**: 155 – 162.

Devos KM (2010). Grass genome organization and evolution. *Current Opinion in Plant Biology* **13**: 139 – 145.

Dida MM, Srinivasachary, Ramakrishnan S, Bennetzen JL, Gale MD, Devos KM (2006). The genetic map of finger millet, *Eleusine coracana*. *Theoretical and Applied Genetics* **114**: 321–332.

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, *et al.* (2008). Evolutionary genetics of genome merger and doubling in plants. *Annual Review of Genetics* **42**: 443–461.

Doyle JJ, Egan AN (2010). Dating the origins of polyploidy events. *New Phytologist* **186**: 73-85.

Dubcovsky J, Luo MC, Zhong GY, Bransteitter R, Desai A, Kilian A, *et al.* (1996). Genetic map of diploid wheat, *Triticum monococcum* l., and its comparison with maps of *Hordeum vulgare* L. *Genetics* **143**: 983–999.

E

Earl D, Bradnam K, St. John J, Darling A, Lin D, Fass J, *et al.* (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research* **21**: 2224–2241.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792 –1797.

Edwards D, Wilcox S, Barrero RA, Fleury D, Cavanagh CR, Forrest KL, *et al.* (2012). Bread matters: a national initiative to profile the genetic diversity of Australian wheat. *Plant Biotechnology Journal* **10**: 703–708.

Edwards EJ, Osborne CP, Stromberg CAE, Smith SA, Consortium CG, Bond WJ, *et al.* (2010). The origins of C4 grasslands: integrating evolutionary and ecosystem science. *Science* **328**: 587–591.

Edwards EJ, Smith SA (2010). Phylogenetic analyses reveal the shady history of C4 grasses. *Proceedings of the National Academy of Sciences* **107**: 2532 –2537.

Egan AN, Schlueter J, Spooner DM (2012). Applications of next-generation sequencing in plant biology. *American Journal of Botany* **99**: 175 –185.

Ekblom R, Galindo J (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**: 1–15.

F

Feldman M, Levy A, Chalhoub B, Kashkush K (2012). Genomic plasticity in polyploid wheat. In: Soltis PS, Soltis DE (eds) *Polyploidy and Genome Evolution*, Springer Berlin Heidelberg, pp 109-135.

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*. doi: 10.1038/hdy.2012.76

Ferris C, King RA, Gray AJ (1997). Molecular evidence for the maternal parentage in the hybrid origin of *Spartina anglica*. *Molecular Ecology* **6**: 185–187.

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

Flagel LE, Wendel JF (2009). Gene duplication and evolutionary novelty in plants. *New Phytologist* **183**: 557–564.

Flagel LE, Wendel JF (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytologist* **186**: 184 - 193.

Flagel LE, Blackman BK. (2012). The first ten years of plant genome sequencing and prospects for the next decade. In: *Plant Genome Diversity Volume 1: Plant genomes, their residents and their evolutionary dynamics*, Jonathan F. Wendel, Johann Greilhuber, Jaroslav Dolezel & Ilia J. Leitch: Vienna, pp 1-16.

Flagel L, Wendel J, Udall J (2012). Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. *BMC Genomics* **13**: 302.

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). *Molecular phylogenetics and evolution* **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. *Molecular Ecology* **17**: 4304-4316.

Foucaud (1897). Un *Spartina* inédit. *Annales de la Société de Sciences naturelles Charente Inférieure* **32**: 220–222.

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.

Freeling M (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology* **60**: 433–453.

G

Gabriel S, Ziaugra L, Tabbaa D (2009). SNP Genotyping Using the Sequenom MassARRAY iPLEX Platform. In: Haines JL, Korf BR, Morton CC, Seidman CE, Seidman JG, Smith DR (eds) *Current Protocols in Human Genetics*, John Wiley & Sons, Inc.: Hoboken, NJ, USA, p .

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.

Gaeta RT, Pires JC (2010). Homoeologous recombination in allopolyploids: the polyploid ratchet. *New Phytologist* **186**: 18-28.

Gale MD, Devos KM (1998). Plant comparative genetics after 10 years. *Science* **282**: 656 –659.

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 research* **18**: 53.

Gaut BS (2002). Evolutionary dynamics of grass genomes. *New Phytologist* **154**: 15–28.

Gaut BS, Doebley JF (1997). DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences* **94**: 6809 –6814.

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. *The Plant Genome Journal* **3**: 69.

Gholami M, Bekele WA, Schondelmaier J, Snowdon RJ (2012). A tailed PCR procedure for cost-effective, two-order multiplex sequencing of candidate genes in polyploid plants. *Plant Biotechnology Journal* **10**: 635–645.

Giegé P, Sweetlove LJ, Cognat V, Leaver CJ (2005). Coordination of nuclear and mitochondrial genome expression during mitochondrial biogenesis in *Arabidopsis*. *The Plant Cell Online* **17**: 1497 – 1512.

Gong L, Salmon A, Yoo M-J, Grupp KK, Wang Z, Paterson AH, *et al.* (2012). The cytonuclear dimension of allopolyploid evolution: an example from cotton using rubisco. *Molecular Biology and Evolution* **29**: 3023 -3036.

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. *Gautam Sarath Publications*.

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 Research* **36**: 3420 – 3435.

Gouy M, Guindon S, Gascuel O (2010). SeaView Version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* **27**: 221 – 224.

GPWG II (2012). New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. *New Phytologist* **193**: 304–312.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**: 644 – 652.

Grandbastien M-A, Casacuberta JM (2013). *Plant transposable elements: Impact on genome structure and function*. Grandbastien & Casacuberta (eds), Springer Topics in Current Genetics Vol.24, 330p.

Griffin PC, Robin C, Hoffmann AA (2011). A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to *Poa* grasses. *BMC Biology* **9**: 19.

Gross BL, Rieseberg LH (2005). The ecological genetics of homoploid hybrid speciation. *The Journal of heredity* **96**: 241–252.

Grover CE, Gallagher JP, Szadkowski EP, Yoo MJ, Flagel LE, Wendel JF (2012a). Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist* **196**: 966-971.

Grover CE, Salmon A, Wendel JF (2012b). Targeted sequence capture as a powerful tool for evolutionary analysis. *American Journal of Botany* **99**: 312 –319.

Groves H, Groves J (1880). *Spartina* x *townsendii* Nobis. *Report of the Botanical Society and exchange club of the British Isles* **1** : 37.

H

Harada E, Kim J-A, Meyer AJ, Hell R, Clemens S, Choi Y-E (2010). Expression profiling of tobacco leaf trichomes identifies genes for biotic and abiotic stresses. *Plant and Cell Physiology* **51**: 1627 –1637.

Harrington GN, Bush DR (2003). The bifunctional role of hexokinase in metabolism and glucose signaling. *The Plant Cell Online* **15**: 2493 –2496.

Ha M, Lu J, Tian L, Ramachandran V, Kasschau KD, Chapman EJ, *et al.* (2009a). Small RNAs serve as a genetic buffer against genomic shock in *Arabidopsis* interspecific hybrids and allopolyploids. *Proceedings of the National Academy of Sciences* **106**: 17835–17840.

Ha M, Kim E-D, Chen ZJ (2009b). Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proceedings of the National Academy of Sciences* **106**: 2295-2300.

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Research* **16**: 1252 – 1261.

Hawkins JS, Proulx SR, Rapp RA, Wendel JF (2009). Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proceedings of the National Academy of Sciences* **106**: 17811 –17816.

He R, Kim M-J, Nelson W, Balbuena TS, Kim R, Kramer R, *et al.* (2012). Next-generation sequencing-based transcriptomic and proteomic analysis of the common reed, *Phragmites australis* (Poaceae), reveals genes involved in invasiveness and rhizome specificity. *American Journal of Botany* **99**: 232 - 247.

Hegarty MJ, Barker GL, Wilson ID, Abbott RJ, Edwards KJ, Hiscock SJ (2006). Transcriptome shock after interspecific hybridization in *Senecio* is ameliorated by genome duplication. *Current Biology* **16**: 1652–1659.

Hegarty MJ, Barker GL, Brennan AC, Edwards KJ, Abbott RJ, Hiscock SI (2009). Extreme changes to gene expression associated with homoploid hybrid speciation. *Molecular Ecology* **18**: 877–889.

Hegarty MJ, Batstone T, Barker GL, Edwards KJ, Abbott RJ, Hiscock SJ (2011). Nonadditive changes to cytosine methylation as a consequence of hybridization and genome duplication in *Senecio* (Asteraceae). *Molecular Ecology* **20**: 105–113.

Henson J, Tischler G, Ning Z (2012). Next-generation sequencing and large genome assemblies. *Pharmacogenomics* **13**: 901–915.

Hervé M (2012). *RVAideMemoire: Diverse basic statistical and graphical functions*.

Hilu KW, Alice LA (2001). A phylogeny of Chloridoideae (Poaceae) based on matK sequences. *Systematic Botany* **26**: 386–405.

Hirst M, Marra MA (2010). Next generation sequencing based approaches to epigenomics. *Briefings in Functional Genomics* **9**: 455 –465.

Hovav R, Udall JA, Chaudhary B, Rapp R, Flagel L, Wendel JF (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**: 6191 –6195.

Hřibová E, Neumann P, Matsumoto T, Roux N, Macas J, Doležel J (2010). Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. *BMC Plant Biology* **10**: 204.

Huang X, Madan A (1999). CAP3: A DNA Sequence Assembly Program. *Genome Research* **9**: 868–877.

Hubbard JCE (1968). *Grasses*. Penguin Books London.

Hu G, Hawkins JS, Grover CE, Wendel JF (2010). The history and disposition of transposable elements in polyploid *Gossypium*. *Genome* **53**: 599–607.

Hu G, Houston NL, Pathak D, Schmidt L, Thelen JJ, Wendel JF (2011). Genomically biased accumulation of seed storage proteins in allopolyploid cotton. *Genetics* **189**: 1103 -1115.

I

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. *American Journal of Botany* **99**: 383 –396.

Imelfort M, Duran C, Batley J, Edwards D (2009). Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnology Journal* **7**: 312–317.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature* **436**: 793–800.

Iuchi S, Suzuki H, Kim Y-C, Iuchi A, Kuromori T, Ueguchi-Tanaka M, *et al.* (2007). Multiple loss-of-function of *Arabidopsis* gibberellin receptor AtGID1s completely shuts down a gibberellin signal. *The Plant Journal* **50**: 958–966.

J

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, *et al.* (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463-467.

Jakob SS, Meister A, Blattner FR (2004). The considerable genome size variation of *Hordeum* Species (Poaceae) is linked to phylogeny, life form, ecology, and speciation rates. *Molecular Biology and Evolution* **21**: 860 –869.

Jannoo N, Grivet L, Chantret N, Garsmeur O, Glaszmann JC, Arruda P, *et al.* (2007). Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. *The Plant Journal* **50**: 574–585.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, *et al.* (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**: 97-100.

Jukes TH, Cantor CR (1969). Evolution of protein molecules. In: *Mammalian protein metabolism, III*, Munro H.N.: New York, pp 21–132.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic Genome Research* **110**: 462 – 467.

K

Kantar M, Akpınar BA, Valárik M, Lucas SJ, Doležel J, Hernández P, *et al.* (2012). Subgenomic analysis of microRNAs in polyploid wheat. *Functional & Integrative Genomics* **12**: 465–479.

Kashkush K, Feldman M, Levy AA (2003). Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nature Genetics* **33**: 102-106.

Kejnovsky E, Hawkins JS, Feschotte C (2012). Plant transposable elements: biology and evolution. In: *Plant Genome Diversity Volume 1: Plant genomes, their residents and their evolutionary dynamics*, Jonathan F. Wendel, Johann Greilhuber, Jaroslav Dolezel & Ilia J. Leitch: Vienna, pp 17-34.

Kejnovsky E, Kubat Z, Macas J, Hobza R, Mracek J, Vyskot B (2006). Retand: a novel family of *gypsy*-like retrotransposons harboring an amplified tandem repeat. *Molecular Genetics and Genomics* **276**: 254–263.

Keller B, Feuillet C (2000). Colinearity and gene density in grass genomes. *Trends in Plant Science* **5**: 246 – 251.

Kellogg EA (2001). Evolutionary History of the Grasses. *Plant Physiology* **125**: 1198 –1205.

Kenan-Eichler M, Leshkowitz D, Tal L, Noor E, Melamed-Bessudo C, Feldman M, *et al.* (2011). Wheat hybridization and polyploidization results in deregulation of small RNAs. *Genetics* **188**: 263–272.

Kim C, Tang H, Paterson AH (2009). Duplication and divergence of grass genomes: integrating the Chloridoids. *Tropical Plant Biology* **2**: 51–62.

Kim E-D, Chen ZJ (2011). Unstable transcripts in *Arabidopsis* allotetraploids are associated with nonadditive gene expression in response to abiotic and biotic stresses. *PLoS ONE* **6**: e24251.

Kim S, Rayburn AL, Lee DK (2010). Genome Size and Chromosome Analyses in Prairie Cordgrass. *Crop Science* **50**: 2277.

Kim S, Rayburn AL, Parrish A, Lee DK (2012). Cytogeographic Distribution and Genome Size Variation in Prairie Cordgrass (*Spartina pectinata* Bosc ex Link). *Plant Molecular Biology Reporter* **30**: 1073–1079.

Kim Y, Morris MD (1996). Ultrafast high resolution separation of large DNA fragments by pulsed-field capillary electrophoresis. *Electrophoresis* **17**: 152–160.

Kircher M, Kelso J (2010). High-throughput DNA sequencing - concepts and limitations. *BioEssays* **32**: 524–536.

Koh J, Chen S, Zhu N, Yu F, Soltis PS, Soltis DE (2012). Comparative proteomics of the recently and recurrently formed natural allopolyploid *Tragopogon mirus* (Asteraceae) and its parents. *New Phytologist* **196**: 292-305.

Kolpakov R, Bana G, Kucherov G (2003). Mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research* **31**: 3672–3678.

Kovarik A, Dadejova M, Lim YK, Chase MW, Clarkson JJ, Knapp S, *et al.* (2008). Evolution of rDNA in *Nicotiana* allopolyploids: A potential link between rdna homogenization and epigenetics. *Annals of Botany* **101**: 815-823.

Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, *et al.* (2009). Circos: An information aesthetic for comparative genomics. *Genome Research* **19**: 1639–1645.

Kumar S, Blaxter M (2010). Comparing *de novo* assemblers for 454 transcriptome data. *BMC genomics* **11**: 571.

L

Lackey E, Ng DW-K, Chen ZJ (2010). RNAi-mediated down-regulation of DCL1 and AGO1 induces developmental changes in resynthesized *Arabidopsis* allotetraploids. *New Phytologist* **186**: 207-215.

Lai K, Duran C, Berkman PJ, Lorenc MT, Stiller J, Manoli S, *et al.* (2012). Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnology Journal* **10**: 743–749.

Lai Z, Gross BL, Zou Y, Andrews J, Rieseberg LH (2006). Microarray analysis reveals differential gene expression in hybrid sunflower species. *Molecular Ecology* **15**: 1213–1227.

Lamb C, Dixon RA (1997). The oxidative burst in plant disease resistance. *Annual Review of Plant Physiology and Plant Molecular Biology* **48**: 251–275.

Landegren U, Kaiser R, Sanders J, Hood L (1988). A ligase-mediated gene detection technique. *Science* **241**: 1077–1080.

Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M (2004). Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* **166**: 935–945.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**: R25.

Lee RW (2003). Physiological adaptations of the invasive cordgrass *Spartina anglica* to reducing sediments: rhizome metabolic gas fluxes and enhanced O-2 and H2S transport. *Marine Biology* **143**: 9–15.

Lee RM, Thimmapuram J, Thinglum KA, Gong G, Hernandez AG, Wright CL, *et al.* (2009). Sampling the waterhemp (*Amaranthus tuberculatus*) genome using pyrosequencing technology. *Weed Science* **57**: 463-469.

Lehr A, Kirsch M, Viereck R, Schiemann J, Rausch T (1999). cDNA and genomic cloning of sugar beet V-type H+- ATPase subunit A and c isoforms: evidence for coordinate expression during plant development and coordinate induction in response to high salinity. *Plant Molecular Biology* **39**: 463–475.

Leitch AR, Leitch IJ (2012). Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytologist* **194**: 629-646.

Leitch AR (2004). Biological relevance of polyploidy: ecology to genomics. In: Leitch, AR, Soltis, DE, Soltis, PS, Leitch, IJ et al.: London Vol 82.

Leitch AR, Leitch IJ (2008). Genomic plasticity and the diversity of polyploid plants. *Science* **320**: 481 – 483.

Leitch IJ, Soltis DE, Soltis PS, Bennett MD (2005). Evolution of DNA amounts across land plants (Embryophyta). *Annals of Botany* **95**: 207 –217.

Liao C, Peng R, Luo Y, Zhou X, Wu X, Fang C, *et al.* (2008). Altered ecosystem carbon and nitrogen cycles by plant invasion: a meta-analysis. *New Phytologist* **177**: 706–714.

Lim KY, Kovarik A, Matyasek R, Chase MW, Clarkson JJ, Grandbastien MA, *et al.* (2007). Sequence of events leading to near-complete genome turnover in allopolyploid *Nicotiana* within five million years. *New Phytologist* **175**: 756–763.

Lim KY, Matyasek R, Lichtenstein CP, Leitch AR (2000). Molecular cytogenetic analyses and phylogenetic studies in the *Nicotiana* section *Tomentosae*. *Chromosoma* **109**: 245-258.

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

Li B, Liao C-hang, Zhang X-dong, Chen H-li, Wang Q, Chen Z-yi, *et al.* (2009b). *Spartina alterniflora* invasions in the Yangtze River estuary, China: An overview of current status and ecosystem effects. *Ecological Engineering* **35**: 511 – 520.

Li X, Zhu J, Hu F, Ge S, Ye M, Xiang H, *et al.* (2012). Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics* **13**: 300.

Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology* **11**: 2453–2465.

Liu B, Vega JM, Feldman M (1998). Rapid genomic changes in newly synthesized amphiploids of *Triticum* and *Aegilops*: Changes in low-copy coding DNA sequences. *Genome* **41**: 535 - 542.

Liu Q, Triplett JK, Wen J, Peterson PM (2011). Allotetraploid origin and divergence in *Eleusine* (Chloridoideae, Poaceae): evidence from low-copy nuclear gene phylogenies and a plastid gene chronogram. *Annals of Botany* **108**: 1287–1298.

Lukens LN, Pires JC, Leon E, Vogelzang R, Oslach L, Osborn T (2006). Patterns of sequence loss and cytosine methylation within a population of newly resynthesized *Brassica napus* allopolyploids. *Plant Physiology* **140**: 336-348.

Lukens L, Quijada P, Udall J, Pires JC, E S, Osborn TC (2004). Genome redundancy and plasticity within ancient and recent Brassica crop species. *Biological Journal of the Linnean Society* **82**: 665–674.

Luo M, Wing RA (2003). An improved method for plant BAC library construction. In: *Plant Functional Genomics*, Humana Press: New Jersey Vol 236, pp 3–20.

Lynch M, Conery JS (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151 – 1155.

Lynch M, Force A (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**: 459 - 473.

M

Macas J, Neumann P (2007). Ogre elements - A distinct group of plant Ty3/*gypsy*-like retrotransposons. *Gene* **390**: 108 – 116.

Macas J, Neumann P, Navratilova A (2007). Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics* **8**: 427.

Macas J, Kejnovský E, Neumann P, Novák P, Koblížková A, Vyskot B (2011). Next Generation Sequencing-based analysis of repetitive DNA in the model dioceous plant *Silene latifolia*. *PLos ONE* **6**: e27335.

Madlung A, Masuelli RW, Watson B, Reynolds SH, Davison J, Comai L (2002). Remodeling of DNA methylation and phenotypic and transcriptional changes in synthetic *Arabidopsis* allotetraploids. *Plant Physiology* **129**: 733–746.

Madlung A, Tyagi AP, Watson B, Jiang H, Kagochi T, Doerge RW, *et al.* (2005). Genomic changes in synthetic *Arabidopsis* polyploids. *Plant Journal* **41**: 221-230.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, *et al.* (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods* **7**: 111–118.

Marchant CJ (1963). Corrected chromosome numbers for *Spartina* x *townsendii* and its parent species. *Nature* **199**: 929.

Marchant CJ (1967). Evolution in *Spartina* (Gramineae). I. The history and morphology of the genus in Britain. *Botanical Journal of the Linnean Society (Botany)* **60**: 1-24.

Marchant CJ (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.

Marchant C, Goodman P (1969). *Spartina maritima* (Curtis) Fernald. *Journal of Ecology* **57**: 287-302

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.

Marmagne A, Brabant P, Thiellement H, Alix K (2010). Analysis of gene expression in resynthesized *Brassica napus* allotetraploids: transcriptional changes do not explain differential protein regulation. *New Phytologist* **186**: 216-227.

Martin JA, Wang Z (2011). Next-generation transcriptome assembly. *Nature Review Genetics* **12**: 671–682.

Ma J, Devos KM, Bennetzen JL (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic dna loss in rice. *Genome Research* **14**: 860–869.

McCarthy FM, Gresham CR, Buza TJ, Chouvarine P, Pillai LR, Kumar R, *et al.* (2011). AgBase: supporting functional modeling in agricultural organisms. *Nucleic Acids Research* **39**: 497–506.

McClintock B (1984). The significance of responses of the genome to challenge. *Science* **226**: 792-801.

Mestiri I, Chagué V, Tanguy A-M, Huneau C, Huteau V, Belcram H, *et al.* (2010). Newly synthesized wheat allohexaploids display progenitor-dependent meiotic stability and aneuploidy but structural genomic additivity. *New Phytologist* **186**: 86-101.

Metzker ML (2010). Sequencing technologies - the next generation. *Nature Review Genetics* **11**: 31 – 46.

Miller JR, Koren S, Sutton G (2010). Assembly algorithms for next-generation sequencing data. *Genomics* **95**: 315 – 327.

Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, *et al.* (2010). Tablet—next generation sequence assembly visualization. *Bioinformatics* **26**: 401 –402.

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, *et al.* (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**: 991-996.

Mobberley DG (1956). Taxonomy and distribution of the genus *Spartina*. *Iowa State College Journal of Science* **30**: 471–574.

Molina C, Rotter B, Horres R, Udupa S, Besser B, Bellarmino L, *et al.* (2008). SuperSAGE: the drought stress-responsive transcriptome of chickpea roots. *BMC Genomics* **9**: 553.

Moore B, Zhou L, Rolland F, Hall Q, Cheng W-H, Liu Y-X, *et al.* (2003). Role of the *Arabidopsis* glucose sensor hxk1 in nutrient, light, and hormonal signaling. *Science* **300**: 332 –336.

Morozova O, Hirst M, Marra MA (2009). Applications of new sequencing technologies for transcriptome analysis. *Annual Review of Genomics and Human Genetics* **10**: 135–151.

Mullen JL, Weinig C, Hangarter RP (2006). Shade avoidance and the regulation of leaf inclination in *Arabidopsis*. *Plant, Cell & Environment* **29**: 1099–1106.

Myers EW (1995). Toward Simplifying and Accurately Formulating Fragment Assembly. *Journal of Computational Biology* **2**: 275—290.

N

Nicolas SD, Mignon GL, Eber F, Coriton O, Monod H, Clouet V, *et al.* (2007). Homeologous recombination plays a major role in chromosome rearrangements that occur during meiosis of *Brassica napus* haploids. *Genetics* **175**: 487 -503.

Nieto Feliner G, Rossello JA (2012). Concerted evolution of multigene families and homoeologous recombination. In: *Plant Genome Diversity Volume 1: Plant genomes, their residents and their evolutionary dynamics*, Jonathan F. Wendel, Johann Greilhuber, Jaroslav Dolezel & Ilia J. Leitch: Vienna, pp 171-194.

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.

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

O

Ohno S (1970). Evolution by Gene Duplication. *Springer-verlag, New York.*

Okada M, Lanzatella C, Tobias CM (2011). Single-locus EST-SSR markers for characterization of population genetic diversity and structure across ploidy levels in switchgrass (*Panicum virgatum* L.). *Genetic Resources and Crop Evolution* **58**: 919–931.

Orlando L, Ginolhac A, Raghavan M, Vilstrup J, Rasmussen M, Magnussen K, *et al.* (2011). True single-molecule DNA sequencing of a pleistocene horse bone. *Genome Research* **21**: 1705–1719.

Osborne CP, Freckleton RP (2009). Ecological selection pressures for C4 photosynthesis in the grasses. *Proceedings of the Royal Society B: Biological Sciences* **276**: 1753–1760.

Osborn T, Pires J, Birchler J, Auger D, Chen Z, Lee H-S, *et al.* (2003). Understanding mechanisms of novel gene expression in polyploids. *Trends in Genetics* **19**: 141 - 147.

Ouyang S, Bell R (2004). The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Research* **32**: 360D–363.

Ozkan H, Levy AA, Feldman M (2001). Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group. *Plant Cell* **13**: 1735–1747.

Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, *et al.* (2009). Direct RNA sequencing. *Nature* **461**: 814–818.

P

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 Phytologist* **184**: 1003-1015.

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, *et al.* (2010). Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytologist* **186**: 37-45.

Paterson AH, Bowers JE, Chapman BA (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**: 9903 –9908.

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.

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, *et al.* (2012). Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**: 423-427.

Pearson WR, Lipman DJ (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America* **85**: 2444–2448.

Pélissier T, Tutois S, Deragon JM, Tourmente S, Genestier S, Picard G (1995). Athila, a new retroelement from *Arabidopsis thaliana*. *Plant Molecular Biology* **29**: 441 – 452.

Peterson PM, Romaschenko K, Johnson G (2010). A phylogeny and classification of the Muhlenbergiinae (Poaceae: Chloridoideae: Cynodonteae) based on plastid and nuclear DNA sequences. *American Journal of Botany* **97**: 1532 –1554.

Petit M, Guidat C, Daniel J, Denis E, Montoriol E, Bui QT, *et al.* (2010). Mobilization of retrotransposons in synthetic allotetraploid tobacco. *New Phytologist* **186**: 135-147.

Piednoël M, Aberer AJ, Schneeweiss GM, Macas J, Novak P, Gundlach H, *et al.* (2012). Next-Generation Sequencing reveals the impact of repetitive DNA across phylogenetically closely related genomes of Orobanchaceae. *Molecular Biology and Evolution* **29**: 3601–3611.

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 Research* **16**: 1262 – 1269.

Pignatta D, Dilkes BP, Yoo S-Y, Henry IM, Madlung A, Doerge RW, *et al.* (2010). Differential sensitivity of the *Arabidopsis thaliana* transcriptome and enhancers to the effects of genome doubling. *New Phytologist* **186**: 194-206.

Pires JC, Zhao JW, Schranz ME, Leon EJ, Quijada PA, Lukens LN, *et al.* (2004). Flowering time divergence and genomic rearrangements in resynthesized *Brassica* polyploids (Brassicaceae). *Biological Journal of the Linnean Society* **82**: 675-688.

Prasad V, Strömberg CAE, Alimohammadian H, Sahni A (2005). Dinosaur coprolites and the early evolution of grasses and grazers. *Science* **310**: 1177 –1180.

Q

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 and Soil* **352**: 157–171.

R

R (2005). *Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0*.

RamanaRao MV, Weindorf D, Breitenbeck G, Baisakh N (2012). Differential expression of the transcripts of *Spartina alterniflora* Loisel (smooth cordgrass) induced in response to petroleum hydrocarbon. *Molecular Biotechnology* **51**: 18–26.

Ramsey J, Schemske DW (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual Review of Ecology And Systematics* **29**: 467-501.

Rapp R, Udall J, Wendel J (2009). Genomic expression dominance in allopolyploids. *BMC Biology* **7**: 18.

Rausch T, Kirsch M, Löw R, Lehr A, Viereck R, Zhigang A (1996). Salt stress responses of higher plants: The role of proton pumps and Na+/H+-antiporters. *Journal of Plant Physiology* **148**: 425 – 433.

Raybould AF, Gray AJ, Hornby DD (2000). Evolution and current status of the salt marshes grass, *Spartina anglica* in the Solent. In: *Solent science - a review*, Collins M. & Ansell K.: Amsterdam, pp 299–302.

Raybould AF, Gray AJ, Lawrence MJ, Marshall DF (1991). The evolution of *Spartina anglica* CE HUBBARD (graminae) - Origin and genetic variability. *Biological Journal of the Linnean Society* **43**: 111–126.

Reboreda R, Caçador I (2008). Enzymatic activity in the rhizosphere of *Spartina maritima*: Potential contribution for phytoremediation of metals. *Marine Environmental Research* **65**: 77 – 84.

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. *Annals of Botany* **105**: 527–533.

Renny-Byfield S, Chester M, Kovařík A, Le Comber SC, Grandbastien M-A, Deloger M, *et al.* (2011). Next Generation Sequencing reveals genome downsizing in *Allotetraploid Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. *Molecular Biology and Evolution* **28**: 2843 –2854.

Rice Annotation Project (2008). The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Research* **36**: D1028–D1033.

Richards CL, Rosas U, Banta J, Bhambhra N, Purugganan MD (2012). Genome-wide patterns of *Arabidopsis* gene expression in nature (G Gibson, Ed.). *PLoS Genetics* **8**: e1002662.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, *et al.* (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**: 348–352.

Rozen S, Skaletsky H (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology (Clifton, N.J.)* **132**: 365–386.

S

Sage RF (2004). The evolution of C4 photosynthesis. *New Phytologist* **161**: 341–370.

Salmon A, Ainouche ML, Wendel JF (2005). Genetic and epigenetic consequences of recent hybridization and polyploidy in *Spartina* (Poaceae). *Molecular Ecology* **14**: 1163-1175.

Salmon A, Flagel L, Ying B, Udall JA, Wendel JF (2010). Homoeologous nonreciprocal recombination in polyploid cotton. *New Phytologist* **186**: 123-134.

Salmon A, Udall JA, Jeddeloh JA, Wendel J (2012). Targeted capture of homoeologous coding and noncoding sequence in polyploid cotton. *G3: Genes\Genomes\Genetics* **2**: 921 –930.

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. *The Plant Cell Online* **20**: 11.

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, *et al.* (2011). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research* **22**(3):557-567.

Sanita di Toppi L, Gabbrielli R (1999). Response to cadmium in higher plants. *Environmental and Experimental Botany* **41**: 105–130.

Sarret G, Harada E, Choi Y-E, Isaure M-P, Geoffroy N, Fakra S, *et al.* (2006). Trichomes of tobacco excrete zinc as zinc-substituted calcium carbonate and other zinc-containing compounds. *Plant Physiology* **141**: 1021 –1034.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, *et al.* (2010). Genome sequence of the palaeopolyploid soybean. *Nature* **463**: 178-183.

Schnable JC, Freeling M, Lyons E (2012). Genome-wide analysis of syntenic gene deletion in the Grasses. *Genome Biology and Evolution* **4**: 265 –277.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, *et al.* (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science* **326**: 1112 -1115.

Schreiber A, Hayden M, Forrest K, Kong S, Langridge P, Baumann U (2012). Transcriptome-scale homoeolog-specific transcript assemblies of bread wheat. *BMC Genomics* **13**: 492.

Shcheglov AS, Zhulidov PA, Bogdanova EA, Shagin DA (2007). Normalization of cDNA libraries. In: *Nucleic Axids Hybridization*, Springer. A. Buzdin & S. Lukyanov, pp 97–124.

Shi X, Ng DW-K, Zhang C, Comai L, Ye W, Jeffrey Chen Z (2012). *Cis*- and *trans*-regulatory divergence between progenitor species determines gene-expression novelty in *Arabidopsis* allopolyploids. *Nature Communications* **3**: 950.

Shivaprasad PV, Dunn RM, Santos BA, Bassett A, Baulcombe DC (2011). Extraordinary transgressive phenotypes of hybrid tomato are influenced by epigenetics and small silencing RNAs. *EMBO Journal* **31**: 257-266.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009). ABySS: a parallel assembler for short read sequence data. *Genome Research* **19**: 1117 – 1123.

Sloop CM, Ayres DR, Strong DR (2011). Spatial and temporal genetic structure in a hybrid cordgrass invasion. *Heredity* **106**: 547–556.

Sloop CM, McGray HG, Blum MJ, Strong DR (2006). Characterization of 24 additional microsatellite loci in *Spartina* species (Poaceae). *Conservation Genetics* **6**: 1049–1052.

Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, *et al.* (2008). Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Research* **18**: 1638–1642.

Smit A, Hubley R, Green P (1996-2010). *RepeatMasker Open-3.0*. http://www.repeatmasker.org

Soltis DE, Soltis PS, Pires JC, Kovarik A, Tate JA, Mavrodiev E (2004). Recent and recurrent polyploidy in *Tragopogon* (Asteraceae): cytogenetic, genomic and genetic comparisons. *Biological Journal of the Linnean Society* **82**: 485-501.

Soltis DE, Buggs RJ, Barbazuk B, Chamala S, Chester M, Gallagher JP, *et al.* (2012). The early stages of polyploidy: Rapide and repeated evolution in *Tragopogon*. In: *Polyploidy and Genome Evolution*, Soltis & Soltis: Berlin, Heidelberg, pp 271–292.

Song K, Lu P, Tang K, Osborn TC (1995). Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. *Proceedings of the National Academy of Sciences* **92**: 7719–7723.

Sonnhammer ELL, Durbin R (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**: 1 – 10.

Srinivasachary, Dida MM, Gale MD, Devos KM (2007). Comparative analyses reveal high levels of conserved colinearity between the finger millet and rice genomes. *Theoretical and Applied Genetics* **115**: 489–499.

Stebbins GL (1950). Variation and evolution in plants. In: Colombia University Press.

Strickler SR, Bombarely A, Mueller LA (2012). Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *American Journal of Botany* **99**: 257 –266.

Subudhi PK, Baisakh N (2011). *Spartina alterniflora* Loisel., a halophyte grass model to dissect salt stress tolerance. *In Vitro Cellular & Developmental Biology - 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.

Swaminathan K, Varala K, Hudson ME (2007). Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *BMC Genomics* **8**: 132.

Swigoňová Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, *et al.* (2004). Close split of *Sorghum* and *Maize* genome progenitors. *Genome Research* **14**: 1916–1923.

Swofford DL (2003). *PAUP: Phylogenetic analysis using parsimony (and other methods). Version 4.* Sinauer Associates, Sunderland, Massachusetts.

Szadkowski E, Eber F, Huteau V, Lode M, Huneau C, Belcram H, *et al.* (2010). The first meiosis of resynthesized *Brassica napus*, a genome blender. *New Phytologist* **186**: 102-112.

T

Tang H, Bowers JE, Wang X, Paterson AH (2010). Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proceedings of the National Academy of Sciences* **107**: 472 –477.

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 Biology* **9**: 80.

Thiebaut F, Grativol C, Carnavale-Bottino M, Rojas C, Tanurdzic M, Farinelli L, *et al.* (2012). Computational identification and analysis of novel sugarcane microRNAs. *BMC Genomics* **13**: 290.

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.). *Theoretical and Applied Genetics* **106**: 411–422.

Thomas BC, Pedersen B, Freeling M (2006). Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research* **16**: 934 – 946.

Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z (1999). Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 7409–7414.

Treangen TJ, Salzberg SL (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13**(1): 36-46.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, *et al.* (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596-1604.

U

Udall JA, Swanson JM, Nettleton D, Percifield RJ, Wendel JF (2006). A novel approach for characterizing expression levels of genes duplicated by polyploidy. *Genetics* **173**: 1823 – 1827.

Utomo HS, Wenefrida I, Materne MD, Harrison SA (2009). Genetic diversity and population genetic structure of saltmarsh *Spartina alterniflora* from four coastal Louisiana basins. *Aquatic Botany* **90**: 30-36.

V

Vandenbussche F, Vriezen WH, Smalle J, Laarhoven LJJ, Harren FJM, Van Der Straeten D (2003). Ethylene and auxin control the *Arabidopsis* response to decreased light intensity. *Plant Physiology* **133**: 517 –527.

Van de Peer Y, Maere S, Meyer A (2009). The evolutionary significance of ancient genome duplications. *Nature Review Genetics* **10**: 725-732.

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, *et al.* (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**: e1326.

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, *et al.* (2010). The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nature Genetics* **42**: 833-839.

Vitte C, Panaud O (2005). LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenetic and Genome Research* **110**: 91–107.

Vitte C, Bennetzen JL (2006). Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proceedings of the National Academy of Sciences* **103**: 17638 –17643.

Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, *et al.* (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**: 763-768.

W

Wall PK, Leebens-Mack J, Chanderbali A, Barakat A, Wolcott E, Liang H, *et al.* (2009). Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics* **10**: 347.

Wang J, Lee JJ, Tian L, Lee HS, Chen M, Rao S, *et al.* (2005). Methods for genome-wide analysis of gene expression changes in polyploids. *Methods in Enzymology* **395**: 570-596.

Wang J, Tian L, Lee HS, Wei NE, Jiang H, Watson B, *et al.* (2006). Genomewide nonadditive gene regulation in *Arabidopsis* allotetraploids. *Genetics* **172**: 507-517.

Wang W, Wang Y, Zhang Q, Qi Y, Guo D (2009). Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics* **10**: 465.

Wang K, Wang Z, Li F, Ye W, Wang J, Song G, *et al.* (2012). The draft genome of a diploid cotton *Gossypium raimondii*. *Nature Genetics* **44**: 1098-1103.

Watanabe S, Nakagawa A, Izumi S, Shimada H, Sakamoto A (2010). RNA interference-mediated suppression of xanthine dehydrogenase reveals the role of purine metabolism in drought tolerance in *Arabidopsis*. *FEBS Letters* **584**: 1181 – 1186.

Wei F, Coe E, Nelson W, Bharti AK, Engler F, Butler E, *et al.* (2007). Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genetics* **3** (7) e123.

Welchen E, Chan RL, Gonzalez DH (2002). Metabolic regulation of genes encoding cytochrome c and cytochrome c oxidase subunit Vb in *Arabidopsis*. *Plant, Cell & Environment* **25**: 1605–1615.

Wendel JF, Schnabel A, Seelanan T (1995). Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proceedings of the National Academy of Sciences* **92**: 280 -284.

Wendel JF (2000). Genome evolution in polyploids. *Plant Molecular Biology* **42**: 225-249.

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.

Wicker T, Keller B (2007). Genome-wide comparative analysis of *Copia* retrotransposons in *Triticeae*, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *Copia* families. *Genome Research* **17**: 1072 – 1081.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, *et al.* (2007). A unified classification system for eukaryotic transposable elements. *Nature Review Genetics* **8**: 973–982.

Wicker T, Narechania A, Sabot F, Stein J, Vu GTH, Graner A, *et al.* (2008). Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC Genomics* **9**: 518.

Wicker T, Taudien S, Houben A, Keller B, Graner A, Platzer M, *et al.* (2009). A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *The Plant Journal* **59**: 712–722.

Wicker T, Mayer KFX, Gundlach H, Martis M, Steuernagel B, Scholz U, *et al.* (2011). Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *The Plant Cell Online* **23**: 1706 –1718.

Winfield MO, Wilkinson PA, Allen AM, Barker GLA, Coghill JA, Burridge A, *et al.* (2012). Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnology Journal* **10**: 733–742.

Winston F, Chaleff DT, Valent B, Fink GR (1984). Mutations affecting Ty-mediated expression of the *HIS4* gene of *Saccharomyces cerevisiae*. *Genetics* **107**: 179 – 197.

Winston F (2001). Control of eukaryotic transcription elongation. *Genome Biology* **2**: 1-3.

Y

Yaakov B, Kashkush K (2011a). Methylation, transcription, and rearrangements of transposable elements in synthetic allopolyploids. *International Journal of Plant Genomics* **2011**: 1-7.

Yaakov B, Kashkush K (2011b). Massive alterations of the methylation patterns around DNA transposons in the first four generations of a newly formed wheat allohexaploid. *Genome* **54**: 42-49.

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.

Yan Y-E, Wang H, Feng Y-H (2005). Alterations of placental cytochrome P450 1A1 and P-glycoprotein in tobacco-induced intrauterine growth retardation in rats. *Acta Pharmacologica Sinica* **26**: 1387–1394.

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.

Z

Zagursky R, McCormick R (1990). DNA sequencing report - DNA sequencing separations in capillary gels on a modified commercial DNA sequencing instrument. *Biotechniques* **9**: 74-79.

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. *American Journal of Botany* **99**: 193 –208.

Zerbino DR, Birney E (2008). Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research* **18**: 821–829.

Zhang D, Ayele M, Tefera H, Nguyen HT (2001). RFLP linkage map of the Ethiopian cereal tef (*Eragrostis tef* (Zucc)). *Theoretical and Applied Genetics* **102**: 957–964.

Zhao C, Hanada A, Yamaguchi S, Kamiya Y, Beers EP (2011). The *Arabidopsis Myb* genes *MYR1* and *MYR2* are redundant negative regulators of flowering time under decreased light intensity. *The Plant Journal* **66**: 502–515.

Zhu H, Senalik D, McCown BH, Zeldin EL, Speers J, Hyman J, *et al.* (2011). Mining and validation of pyrosequenced simple sequence repeats (SSRs) from American cranberry (*Vaccinium macrocarpon* Ait.). *Theoretical and Applied Genetics* **124**: 87–96.

Evolution du génome des Spartines polyploïdes envahissant les marais salés: apport des nouvelles techniques de séquençage haut-débit

Les Spartines jouent un rôle écologique majeur sur les marais salés. Elles représentent un excellent modèle pour appréhender les conséquences écologiques de la spéciation par hybridation et polyploïdie dans le contexte d'invasion biologique. On s'intéresse plus particulièrement, à l'hybridation récente entre une espèce hexaploïde d'origine américaine *Spartina alterniflora* et une espèce hexaploïde européenne *S. maritima* ayant donnés deux hybrides F1 (*S.* x *townsendii* et *S.* x *neyrautii*) et la nouvelle espèce envahissante allododécaploïde (*S. anglica*). Les nouvelles technologies de séquençage haut-débit facilitent l'exploration de ces génomes peu connus. L'assemblage et l'annotation d'un transcriptome de référence ont permis d'annoter 16 753 gènes chez les spartines hexaploïdes et d'identifier des gènes d'intérêts écologique et évolutif. Une sélection de ces gènes a ensuite été analysée à travers une étude d'expression par PCR quantitative sur les populations naturelles des 5 espèces du complexe. Les résultats ont permis de mettre en évidence une expression homogène intra-populations mais une grande variabilité entre les espèces. L'analyse du génome des Spartines a ciblé prioritairement le développement de ressources génomiques concernant l'espèce *S. maritima* pour l'analyse des compartiments codant et répété à l'aide de séquençage d'une banque BAC et d'un run de pyroséquençage d'ADN génomique. Les analyses ont permis d'évaluer une proportion d'éléments répétés représentant près de 30% du génome. Les données générées ont alors été comparées avec les génomes séquencés phylogénétiquement proches et ont permis de premières comparaisons entre les spartines et les autres Poaceae.

Genome evolution of polyploid *Spartina* species invading salt-marshes: Contribution of Next-generation Sequencing technologies

Spartina species play an important ecological role on salt marshes. They represent an excellent system to study the ecological consequences of hybrid and polyploid speciation in biological invasion contexts. In this study, we examined the effects of hybridization between the hexaploid American-native species *Spartina alterniflora* and the European species *S. maritima*, that gave rise to two F1 hybrids (*S.* x *townsendii* in England et *S.* x *neyrautii* in France) and the new invasive allododecaploid species (*S. anglica*). Next-generation sequencing technologies offer new perspectives to explore these previously poorly known genomes. The assembly of a reference transcriptome (from 454 Roche pyrosequencing) allowed annotation of 16,753 genes in hexaploid *Spartina* and identification of ecologically and evolutionary important genes. Expression levels of a subset of these genes were analyzed by quantitative PCR in *Spartina* natural populations. The results indicate intrapopulation homogenous expression but extreme variability between species. The European *S. maritima* beneficiated from genomic resource development through a BAC library and one pyrosequencing run. Our analyses estimated the relative proportions of repetitive sequences as about 30% and have identified the main transposable element families Data generated were also compared to closely related sequenced species and provided the first insights into the evolution of *Spartina* genomes in the Poaceae family.