Reference Transcriptomes and Detection of Duplicated Copies in Hexaploid and Allododecaploid Spartina Species (Poaceae)

Julie Boutte, Julie Ferreira de Carvalho, Mathieu Rousseau-Gueutin, Julie Poulain, Corinne da Silva, Patrick Wincker, Malika Ainouche, Armel Salmon

GBE

Genome Biology and Evolution

SMBE

Reference Transcriptomes and Detection of Duplicated Copies in Hexaploid and Allododecaploid *Spartina* Species (Poaceae)

Julien Boutte[1], Julie Ferreira de Carvalho[1], Mathieu Rousseau-Gueutin[1,2], Julie Poulain[3], Corinne Da Silva[3], Patrick Wincker[3], Malika Ainouche[1], and Armel Salmon[1,*]

[1]UMR CNRS 6553 Ecobio, OSUR (Observatoire des Sciences de l'Univers de Rennes), University of Rennes 1, Rennes Cedex, France

[2]UMR Institut de Génétique, Environnement et Protection des Plantes, Institut National de la Recherche Agronomique, Le Rheu Cedex, France

[3]Genoscope, 2 rue Gaston Crémieux, Evry, France

*Corresponding author: E-mail: armel.salmon@univ-rennes1.fr.

Accepted: August 20, 2016

Abstract

In this study, we report the assembly and annotation of five reference transcriptomes for the European hexaploid *Spartina* species (*S. maritima*, *S. alterniflora* and their homoploid hybrids *S. x townsendii* and *S. x neyrautii*) and the allododecaploid invasive species *S. anglica*. These transcriptomes were constructed from various leaf and root cDNA libraries that were sequenced using both Roche-454 and Illumina technologies. Considering the high ploidy levels of the *Spartina* genomes under study, and considering the absence of diploid reference genome and the need of an appropriate analytical strategy, we developed generic bioinformatics tools to (1) detect different haplotypes of each gene within each species and (2) assign a parental origin to haplotypes detected in the hexaploid hybrids and the neo-allopolyploid. The approach described here allows the detection of putative homeologs from sets of short reads. Synonymous substitution rate (K_S) comparisons between haplotypes from the hexaploid species revealed the presence of one K_S peak (likely resulting from the tetraploid duplication event). The procedure developed in this study can be applied for future differential gene expression or genomics experiments to study the fate of duplicated genes in the invasive allododecaploid *S. anglica*.

Key words: homeo-SNPs, transcriptome *de novo* assembly, K_S distribution, polyploidy, haplotyping, paralogs–orthologs–homeologs.

Introduction

Polyploidy (resulting from whole genome duplication) appears to be a major feature of eukaryote evolution (Otto and Whitton 2000; Van de Peer et al. 2009; Mable et al. 2011). Several examples of polyploidization are reported in animals, as in amphibians (Gregory and Mable 2005) or in Teleostei fishes (Mable 2004), but this phenomenon is particularly prominent in plants. In this latter kingdom, this recurrent process (Soltis et al. 2009) has contributed to speciation, phenotypic innovation and adaptation (Leitch and Leitch 2008). Polyploidy provides the raw genomic material for natural or artificial (e.g., domestication) selection (Wendel 2000). Most of our understanding of the consequences of polyploidy derives from relatively recent polyploids, which include a large proportion of crops such as wheats, cotton, oilseed rape, tobacco or coffee (Leitch and Leitch 2008). Of particular interest are the natural and recent polyploids that have been described in Asteraceae (e.g., *Tragopogon*, Malinska et al. 2011 and *Senecio*, Abbott et al. 2008), Brassicaceae (e.g., *Cardamine*, Marhold et al. 2009), Phrymaceae (e.g., *Mimulus*, Vallejo-Marin 2012) or Poaceae (e.g., *Spartina*, Ainouche et al. 2004). These recently formed species can be compared with their actual parents and represent excellent model systems to understand the immediate consequences of hybridization and genome duplication in natural populations.

The rapidly accumulating genomic data has documented various older genome duplication events (paleopolyploidy) in eukaryotes (and most particularly plant) genomes (Blanc and Wolfe 2004; Van de Peer et al. 2009; Jiao et al. 2011). Modern plant genomes appear then shaped by recurrent rounds of polyploidization and fractionation/diploidization processes (reviewed in Wendel et al. 2016). Duplicated genes may undergo various evolutionary fates, including differential gene retention during the fractionation/diploidization

process (Langham et al. 2004), homoeologous recombination and gene conversion (Udall 2005; Nicolas et al. 2007; Salmon et al. 2009; Flagel et al. 2012; Chalhoub et al. 2014; Page et al. 2016), or reprogramming of duplicated gene expression (Adams et al. 2003; Flagel et al. 2008; Combes et al. 2013; Yoo et al. 2013). Transcription of duplicated genes in a polyploid may reflect parental additivity, mimic the level of one parent (parental transcriptome dominance) or be transgressive (over expression or under expression) compared with either parent (Grover et al. 2012).

Distinguishing genes duplicated by polyploidy is then critical to understand the evolutionary history of plant species and to explore the short- and long-term evolution of duplicated genomes. In polyploids, allelic diversity (at orthologous loci) needs to be distinguished from homoeologous divergence (reflecting divergence between the parents and subsequent evolution after allopolyploid formation), and paralogs (resulting from individual gene duplication or paleopolyploidization; Glover et al. 2016).

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

In this study, we aim at reconstructing reference transcriptomes from NGS data sets and detecting the various expected duplicated gene copies in the polyploid genus *Spartina* Schreb. (Poaceae, subfamily Chloridoideae), for two hexaploid species, their two independently formed F1 hybrids and a neo-allododecaploid species. The genus *Spartina* is characterized by recurrent interspecific hybridization and genome duplication events that resulted in various ploidy levels ranging from tetraploid to dodecaploid, with a basic chromosome number $x = 10$ (Ainouche et al. 2012). Hybridization and polyploidy had major impacts on diversification, and important ecological consequences in salt-marsh communities regarding the formation of invasive species (Ainouche et al. 2008; Strong and Ayres 2013). The history of the genus is now well-documented. *Spartina* represents a monophyletic lineage, embedded in the paraphyletic *Sporobolus* genus (Peterson et al. 2014). No diploid *Spartina* species are known among the 15 perennial species described by Mobberley (1956), which suggests that *Spartina* most likely emerged from an already polyploid common ancestor. Diploid species are reported in *Sporobolus* lineages which diverged from *Spartina* sometimes 14–20 Ma (Rousseau-Gueutin et al. 2015). *Spartina* has evolved in two lineages: a tetraploid clade and a hexaploid clade (Baumel, Ainouche, Bayer, et al. 2002), which divergence was estimated as dating back to 6–10 Ma from chloroplast genome sequences (Rousseau-Gueutin et al. 2015). Of particular interest are the hexaploid *Spartina alterniflora* Loisel. ($2n = 6x = 62$; growing on the East-American coast) and *Spartina maritima* (Curtis) Fern. ($2n = 6x = 60$, growing on the European/African Atlantic coast) which have naturally hybridized in Europe following the introduction of the American species during the 19th century. Two sterile F1 hybrids were formed, *S. alterniflora* as female parent in both hybridization events (Ferris et al. 1997; Baumel et al. 2001): *Spartina x townsendii* ($2n = 6x = 62$) in Southampton (England; Foucaud 1897) and *Spartina x neyrautii* ($2n = 6x = 62$; Marchant 1963) in Hendaye (Southwest France). Genome duplication of *S. x townsendii* (after 1890) resulted in a new allododecaploid species, *Spartina anglica* C.E. Hubbard, $2n = 120$, 122, 124 (Marchant 1968). The expansion of this fertile and invasive species that rapidly colonized Western Europe and several continents (e.g., Australia and China) has important ecological consequences. *S. anglica* is now a classical example of recent allopolyploid speciation, and this system is an excellent model to explore the early evolutionary changes following hybridization and genome duplication in natural populations (Ainouche et al. 2004). Considering the recent hybridization and allopolyploidization events, and the weak inter-population variation of the parental species in the hybridization sites (Baumel et al. 2001, 2003; Yannic et al. 2004), the parental species in Europe may be considered as good representatives of the actual parents. No major genetic changes were detected in the hybrids and neoallopolyploid (Baumel et al. 2001; Baumel, Ainouche, Kalendar, et al. 2002; Parisod et al. 2009), but homogenization

of parental rDNA homoeologous copies are being observed in populations of *S. anglica* (Dalibor et al. 2016). Hybridization appears to have entailed significant epigenetic changes (Salmon et al. 2005; Parisod et al. 2009). Using a single rice heterologous microarray, Chelaifa, Mahé et al. (2010) have analyzed the differential expression between the hexaploid parental species (*S. maritima* and *S. alterniflora*). Nonadditive transcriptomic parental patterns were observed in the hybrids and allopolyploid (Chelaifa, Monnier et al. 2010), including maternal expression dominance (from *S. alterniflora*) and transgressive expression. However, only global gene expression levels were analyzed and the employed technology could not allow distinguishing the contribution of each homeolog. A first reference transcriptome was recently assembled for the parental hexaploid species using several leaf and root cDNA libraries and Roche-454 pyrosequencing (Ferreira de Carvalho et al. 2013). This led to the annotation of c.a. 17,000 genes.

The aim of the present work is (1) to extend transcriptome assembly and annotations by combining both 454 pyrosequencing and Illumina sequencing technologies in five polyploid species: the hexaploid parents *S. maritima*, *S. alterniflora*, the F1 hybrids *S. x neyrautii* and *S. x townsendii*, and the allododecaploid *S. anglica* and (2) to detect duplicated gene copies in these highly redundant genomes, by developing a strategy aiming at reconstructing haplotypes with no diploid reference genome.

Materials and Methods

Sampling, cDNA Preparation and Sequencing

This study focused on five *Spartina* species: the two hexaploid parents *S. maritima* and *S. alterniflora*, the sterile F1 hybrids *S. x townsendii* and *S. x neyrautii* and the allododecaploid species *S. anglica*. *S. x townsendii* was collected in Hythe (Hampshire, England). Samples from *S. x neyrautii* were collected in Hendaye (Pyrénées Atlantiques, France) and *S. anglica* was sampled in Roscoff and l'Anse de Goulven (Finistère, France). RNAs were extracted from leaves and roots, from plants grown in same conditions in the greenhouse as indicated in Ferreira de Carvalho *et al.* (2013).

Roche-454 data were sequenced at the Genoscope Platform (Evry, France) and at the Environmental and Functional Genomics Platform of the University of Rennes 1 (Biogenouest, OSUR, France). Both normalized (two libraries for *S. maritima* only) and nonnormalized (two libraries for each species) data were pyrosequenced to enhance the number of assembled contigs as previously published (GenBank accession: SRP015701 and SRP015702; Ferreira de Carvalho et al. 2013). Roche-454 data of the hybrids and the allopolyploid were obtained using the same protocol as used by Ferreira de Carvalho *et al.* (2013).

Illumina libraries were prepared from cDNAs of the same samples as those used for the 454 pyrosequencing for each five species and Illumina (Hi-Seq 2000) sequencing and read-quality trimming (Phred score = 20) were performed at the Genoscope Platform (Evry, France). The number of cleaned reads obtained for each species is indicated in table 1. This project has been deposited at Genbank under the accession SRP081066 and at https://spartina-genomics.univ-rennes1.fr/.

Strategy for Assembling Roche-454 and Illumina Reads

For each species we independently assembled Roche-454 and Illumina data using the most reliable approaches (fig. 1): (1) Roche-454 reads were first assembled using the GS *de novo* assembler Software v.2.6, Roche (ml = 80 bp; mi = 90%; Margulies et al. 2005); (2) the Trinity algorithm (Grabherr et al. 2011) commonly recommended for Illumina RNA-seq assemblies (Clarke et al. 2013; Liu et al. 2013; Chopra et al. 2014) was used for assembling Illumina reads with the following parameters: k-mer size of 25 and minimum contig length of 48; (3) Roche-454 and Illumina separately assembled contigs with a length higher than (or equal to) 100 bp (to avoid the formation of chimeric contigs), were then co-assembled using the Newbler software (ml = 40 bp; mi = 90%).

The different contigs obtained after the co-assembly step and Roche-454 contigs and Illumina contigs which were not considered during the co-assembly step (with a length ranging from 40 to 100 bp) were post-processed by deleting redundant contigs and self-blasted in order to maximize the length of overlapping contigs. Contigs overlapping on 50 bp or more with an identity percentage $\geq$90% were then assembled using custom python scripts. Redundancy of the contigs was checked again using a SELFBLAST (minimum length: 40 bp and minimum identity percent: 90%).

Functional Annotation

Functional annotations were made following Ferreira de Carvalho *et al.* (2013) and using the Pfam software to detect annotated protein domains from alignments to protein families databases and using a profile Hidden Markov Model (HMM; Finn et al. 2014). All the contigs were analyzed using BLASTn and tBLASTx algorithms (*e*-value threshold of 10^{-5}; Altschul et al. 1997) against a home-built CDS database including *Oryza sativa*, *Setaria italica*, *Brachypodium distachyon*, *Sorghum bicolor* (www.phytozome.net) and *Zea mays* (concatenation of two databases downloaded on www.phytozome.net and www.plantgdb.org websites; last accessed the September 1, 2016). To obtain the homology-based functional annotation Best BLAST Hits (BBH) were selected. The Gene Ontology (GO) was analyzed using the BLAST2Go software (Conesa et al. 2005; Götz et al. 2008). GO annotations were performed using tBLASTx (*e*-value threshold of 10^{-5}) on the different assembled contigs against the *Arabidopsis thaliana* database (TAIR website, www.arabidopsis.org; *e*-value hit filter of 10^{-6} and a cutoff of 55 which corresponding to the maximum

Table 1

Sequencing Statistics Using Roche-454 and Illumina Data and Number of Reads Used for the Analysis (Illumina Cleaned Reads Length = 108 bp)

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

[a]Roche-454 data of *S. maritima* contain normalized (average read length = 576.73 ± 156.86 bp) and nonnormalized (average read length = 314.14 ± 147.02 bp) cDNA libraries.

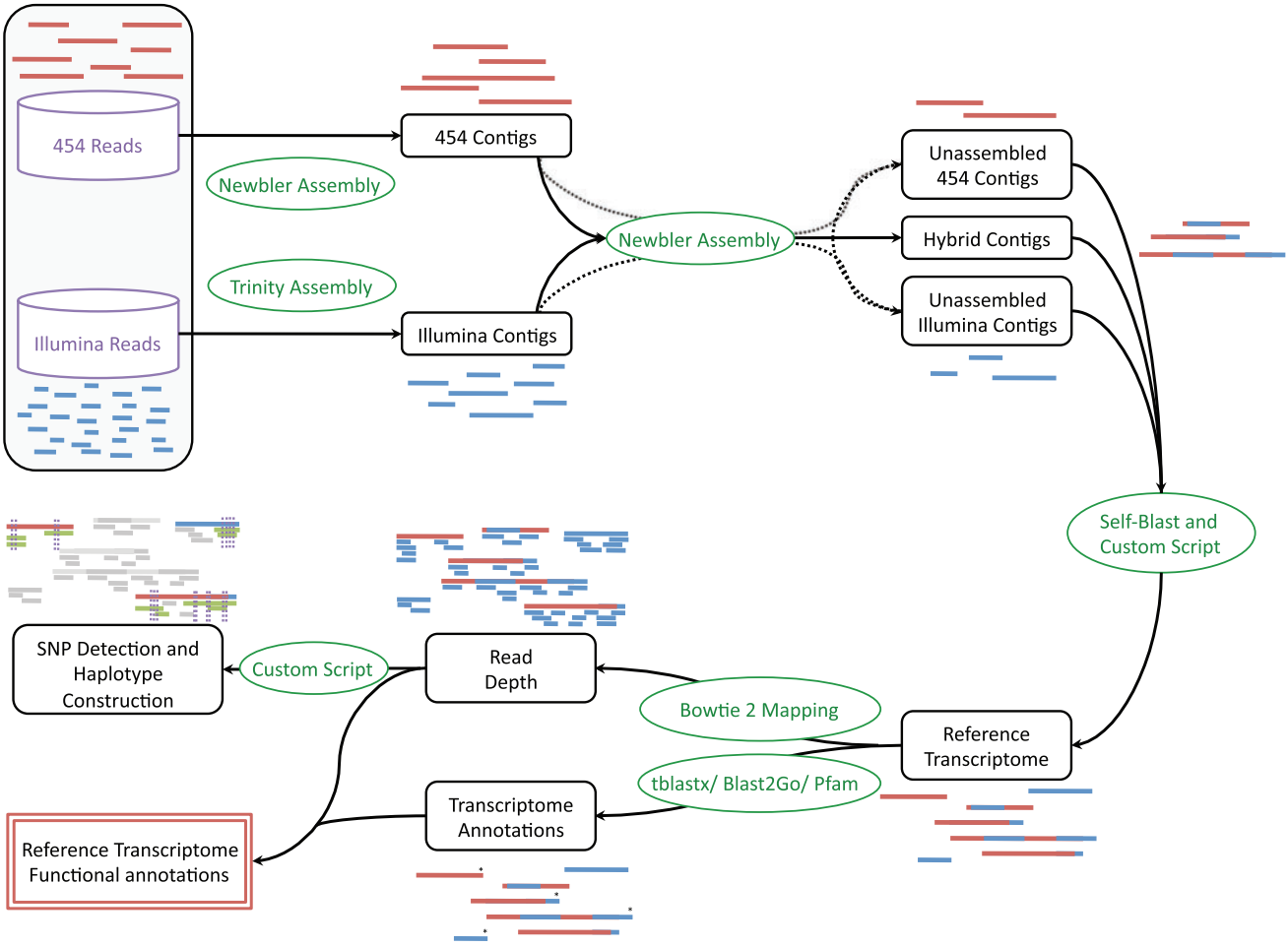

Fig. 1.—Strategy developed for assemblies and haplotype detection. First, Roche-454 and Illumina reads were assembled separately using Newbler and Trinity, respectively. The contigs obtained were co-assembled using Newbler. The length of contigs was enhanced and redundant contigs were removed using custom scripts. Reads Mapping were done with Bowtie 2 and the different contigs were annotated using three complementary methods: tBLASTx, BLAST2Go and Pfam. Polymorphisms and haplotypes were detected and constructed using mapping data.

similarity). Pfam 27.0 database was used to enrich the protein domain annotations (six reading frames tested by contig; PfamB option); Pfam results were filtered by significant hits and *e*-value $\leq 10^{-3}$ (Finn et al. 2014). Estimation of the number of exons and unigenes (transcripts from the same locus) in each *Spartina* species contigs was performed using BLASTn (≥70% of identity, ≥60 pb of overlap) against the rice genome (GFF files downloaded from www.phytozome.net).

SNP Detection and Haplotype Assembly Using Illumina Data

For each species, the Illumina reads data set was mapped on the previously built reference contigs using Bowtie 2, v2.0 (Langmead and Salzberg 2012). The parameters used were "score-min: G, 52, 8" for the natural logarithmic function $f(x) = 52 + 8 \times \ln(x)$, where x correspond to the read length. Using these parameters, all reads (with a length of 80–120 bp) presenting at least 87.06–90.30% of identity were mapped to the reference contig. The output file ".SAM" created by Bowtie 2 during the mapping step was converted to a ".PILEUP" format using the Samtools software suite (Li et al. 2009). We detected for each contig the different SNPs or Single Nucleotide Polymorphisms (minimum read depth = 30; SNP detection threshold = 2, corresponding to nucleotides that are not present more than 2/100 times per position), using custom python script. These parameters were chosen to remove potential sequencing errors in Illumina reads (<0.1%) and to avoid the use of false positives SNPs in haplotype construction (Oliphant et al. 2002).

Within each alignment of homologous reads, the different haplotypes were assembled from the ".PILEUP" file and the previously detected SNPs. To construct these haplotypes, we first identified the different reads split in the ".PILEUP" file. The next step consisted in detecting the different haplotypes using each window with a minimum length of 240 nucleotides and containing at least two SNPs. Reads that were included in this window were used to detect and to assemble the different haplotypes using the same method as that developed by Boutte *et al.* (2016) for Roche-454 data (fig. 2). Pairwise comparisons of the reads previously assembled were then performed before creating a new haplotype, by assembling them if the two compared reads present the same SNPs and if no alternative assembly (creating another haplotype) was possible. This method has the advantage of not creating chimeric haplotypes (when two or more choices are possible, the program does not assemble reads) but creates many haplotypes (cascade phenomenon). To avoid this problem, we counted the maximum number of haplotypes by sliding windows (see fig. 2D for description).

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

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

Parental Haplotype Assignation

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

K_A/K_S Tests and Molecular Dating of Duplicate Gene Divergences

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

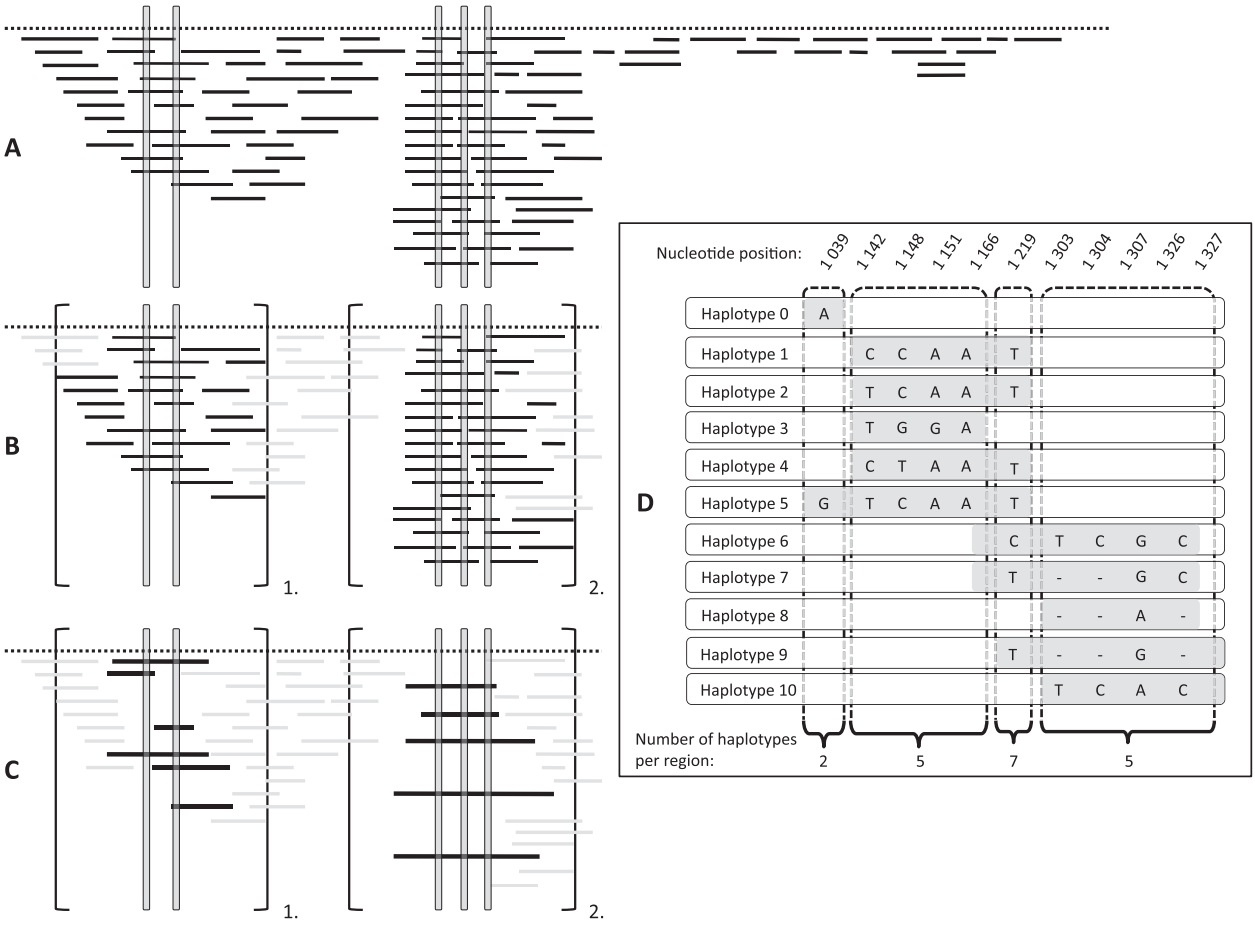

Fig. 2.—Description of the method and windows used to construct the different haplotypes. (*A*) For each contig (dotted line), the developed program detects the different SNPs (gray boxes) using mapping data. (*B*) For the windows created (1 and 2), only reads (black lines) entirely included in the windows are selected. (*C*) Detection of the different haplotypes in each window (thick black lines) using the method previously developed by Boutte et al. (2016). (*D*) Example of "the maximum number of haplotypes by window" and "cascade phenomenon" leading to the detection of multiple haplotypes. Using reads mapped to a contig of *S. maritima* annotated as nucleotidyl transferase localized in the cytoplasm (GO annotations: 0009058, 0016740, 0016779, 0005737, 0005623), 11 SNPs are detected and a total of 11 haplotypes are constructed. If the maximum number of haplotypes for this gene is seven, the hypothetical minimum number of haplotypes for this gene should be five (haplotypes seven and nine might correspond to haplotypes one, two, four or five). Because several choices are possible, the detected haplotypes are not assembled, illustrating the "cascade phenomenon".

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

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

where

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

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

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

Allowing the calculation of K_A and K_S:

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

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

The program outputs the length of each window, nucleotide and protein (amino acid) identities, the K_A, K_S and K_A/K_S ratios and other information for validating and/or filtering out the results. Frequency distributions of K_S values between pairs of haplotypes were performed using the R software (v. 2.13.0; R Development Core Team 2011) to detect duplication events (Blanc and Wolfe 2004).

Results

De Novo Assemblies and Functional Annotation

The number of contigs assembled from the five species ranged from 44,158 to 65,099 (table 2). Using the *O. sativa* genome and its gene annotation as a reference, 35,039 and 32,734 exons were detected in the two parental species *S. maritima* and *S. alterniflora*, respectively, and 40,365 and 34,792 in the two hybrids *S. x townsendii* and

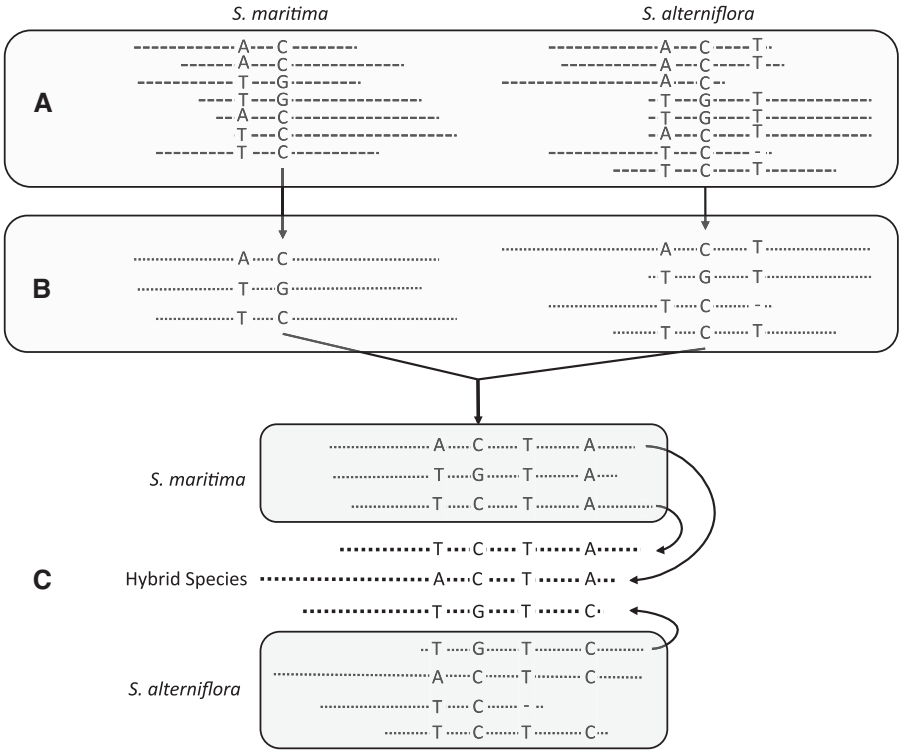

Fig. 3.—Parental haplotype assignation process. (*A*) For each read alignment, SNPs were detected and (*B*) reads assembled to obtain the different haplotypes using the developed program. (*C*) Hybrid haplotypes (detected following the procedure *A* and *B*) were aligned with parental haplotypes. Intra-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

Table 2

Summary of Assemblies' Steps and Annotations of Five *Spartina* Species

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

[a]Unigenes were detected using *Oryza sativa* genome only.

S. x neyrautii. In *S. anglica*, 35,062 were assembled. This search for exons from a reference annotated genome allowed identification of unigenes in the *Spartina* transcriptomes, ranging from 13,054 to 16,002 (which represents 26.61–32.62% of the expected number of unigenes). The number of Illumina contigs obtained using the Trinity assembler is higher in the F1 and the allopolyploid than in the parents (121,733, 110,455, 144,550 for the three hybrid species and 98,455, 76,010 for the parental species) while the number of Roche-454 contigs obtained with Newbler is higher in the parents due to a deeper sequencing and the presence of both normalized and nonnormalized libraries (table 1). The co-assembly (using Newbler) of Trinity (Illumina reads) and Newbler (Roche-454 reads) sub-assemblies formed 12,674, 9,232, 13,691, 8,768 and 11,201 Roche-454/Illumina hybrid contigs for *S. maritima*, *S. alterniflora*, *S. x townsendii*, *S. x neyrautii* and *S. anglica*, respectively, with a proportion of non co-assembled contigs ranging from 40.73% to 54.79% of the Roche-454 contigs and from 62.77% to 69.38% of the Illumina contigs. After

comparisons of annotated contigs from the five transcriptomes, a total of 37,867 different annotated genes were obtained, 15,114 genes being redundant between species. The number of annotated contigs specific to each transcriptome was determined: 4,456, 3,760, 3,528, 6,751 and 4,168 contigs were found specific to *S. maritima, S. alterniflora*, *S. x townsendii*, *S. x neyrautii* and *S. anglica*, respectively. For the five transcriptomes, the total sequence length is similar for four species (ranging from 27,548,352 to 29,841,745 bp) and lower for *S. alterniflora* (23,016,772 bp). The GC% and the N50 are similar for the five species. The high values of the N50 are explained by the presence of Roche-454 contigs in the data set (table 2). For *S. maritima*, the length of the annotated contigs ranged from 58 to 10,313 bp (742 bp on an average). For these contigs, the number of mapped reads (at 90% of identity) varied between 1 and 251,685 (643 reads per contig on an average). The average coverage of each nucleotide within contigs is estimated to be 56.51×. The length of the unannotated contigs ranged from 40 to 4,701 bp (340 bp on an average) and the number of mapped reads varied from 1 to 46,647 (146 reads per contig on an average). For these contigs, the average read depth was estimated as 37.16× (supplementary fig. S1, Supplementary Material online). The assemblies of the four other species exhibit similar contig length, read depth, for both annotated and unannotated contigs to *S. maritima* and values are available in the supplementary table S1, Supplementary Material online.

SNP Detection and Haplotype Construction

For the two parents *S. maritima* and *S. alterniflora*, we have detected a similar number of SNPs within processed alignments (3.85 SNPs for 100 bp on an average, Student's test P value > 0.05). The mean number of SNPs for 100 bp detected in the two hybrids and the allododecaploid *S. anglica* is higher than in the parents (Student's test P value < 0.001): 6.10 SNPs for 100 bp for *S. x townsendii* and a similar number of SNPs *in S. x neyrautii* and *S. anglica*, 5.32 and 5.33, respectively (Student's test P value > 0.05, fig. 4A). After detecting haplotypes from the detected SNPs, the average number of haplotypes corresponding to the maximal number of haplotypes by region was calculated. At least two haplotypes were detected in 20,085 and 13,216 contigs of the parents *S. maritima* and *S. alterniflora*. For the hybrids *S. x townsendii*, *S. x neyrautii* and the allododecaploid *S. anglica*, 24,199, 16,776 and 18,839 contigs, respectively, exhibit at least two haplotypes (table 2). A similar number of haplotypes was detected in the two parents (7.25 on an average for *S. maritima* and 6.94 on an average for *S. alterniflora*) while about twice more were detected in the other species, with 13.73, 11.63 and 11.72 haplotypes on an average in *S. x townsendii*, *S. x neyrautii* and *S. anglica*, respectively (fig. 4B). The number of SNPs for 100 bp and the maximal number of haplotypes by region is higher in *S. x townsendii* than in the parents, *S. x neyrautii* and the allopolyploid *S. anglica*. For each window, where haplotypes reconstruction was possible, we have compared the number of windows presenting a similar number of haplotypes using a Fisher's exact test. For the two parental hexaploid species (*S. maritima* and *S. alterniflora*), 43.53% and 42.38% of the windows presented between two and four haplotypes. Within the two hexaploid hybrids and the allododecaploid species *S. anglica*, a lower percentage of windows presenting two to four haplotypes were detected (18.01% in *S. townsendii*, 21.80% in *S. x neyrautii* and 21.64% in *S. anglica*). However, the number of regions presenting between 5 to 12 haplotypes is similar for the two parents, the two F1 hybrids and the allopolyploid *S. anglica* (from 43.03% in *S. x townsendii* to 48.68% in *S. x neyrautii*). The number of regions presenting more than 12 haplotypes is higher in *S. x townsendii* and *S. x neyrautii* and the allopolyploid (38.93%, 29.50% and 30.97%, respectively) compared to the two hexaploid parents (12.16% and 10.51% for *S. maritima* and *S. alterniflora*, respectively). The average length of the haplotypes reconstructed is higher than the initial length of the reads, the majority of haplotypes (≥76.44%) presents a length ranging from 150 and 450 bp for all five species (fig. 4C).

Parental Haplotype Assignation

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

For the three species, the number of haplotypes assigned to the parental species *S. maritima* is similar, ranging from 49.80% to 50.71% and represents the majority of the copy assigned. The number of haplotypes assigned to *S. alterniflora* is ranging from 40.69% to 41.18% and similar for the three species. The number of unassigned copies for the three data

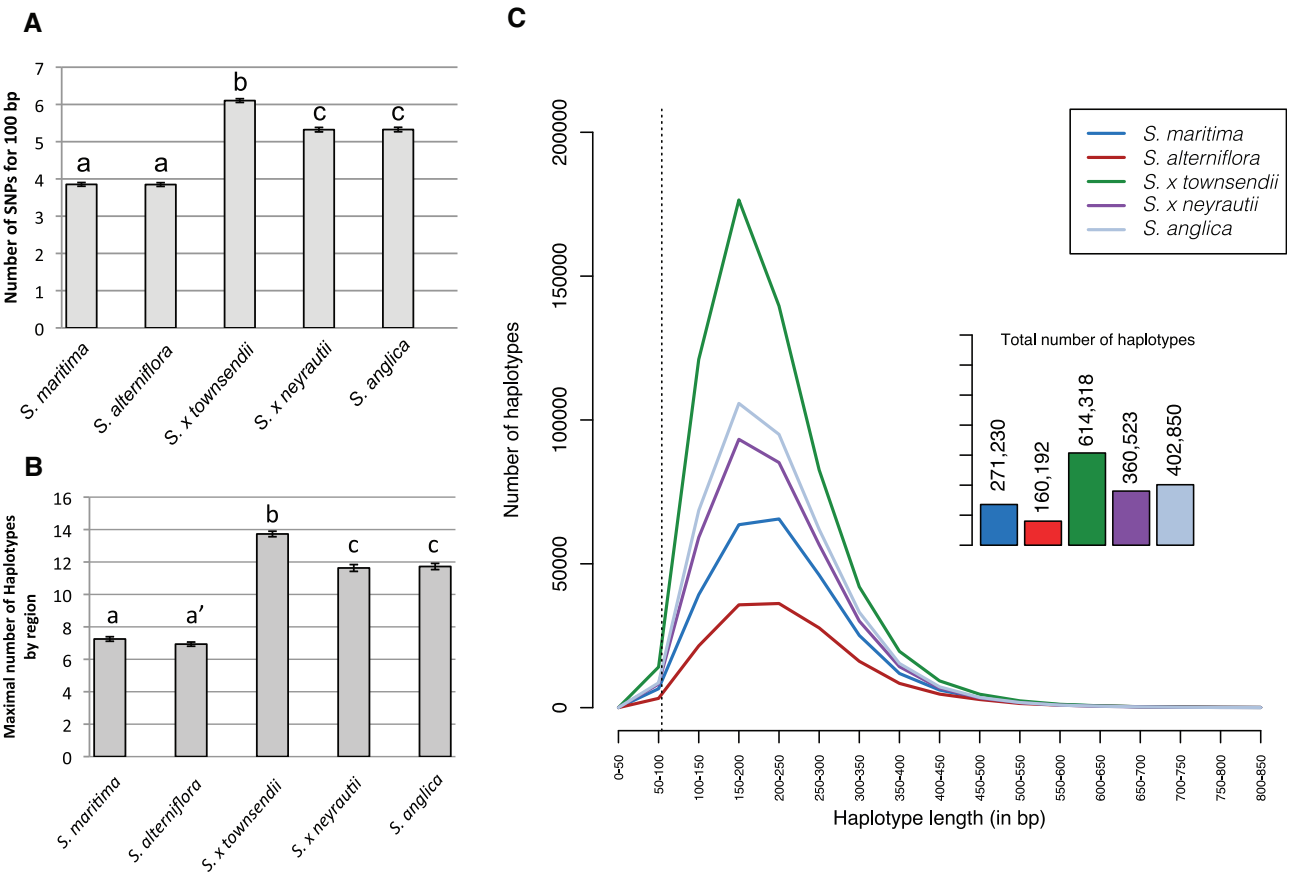

Fig. 4.—Graphical representation of (*A*) the number of SNPs for 100 bp and (*B*) the maximum number of haplotypes by windows for each species studied (Student's test *P* value > 0.05 and *P* value < 0.001). Errors bars indicate confidence interval to 95%. (*C*) Distribution of the haplotype length (in bp) reconstructed for the five species studied. Dotted vertical bar represent the length of the Illumina reads.

Table 3

Identification of the Parental Origin of the Haplotypes in the Hybrids, Using 271,230 and 160,192 Haplotypes of *S. maritima* and *S. alterniflora*, Respectively

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

Unassigned haplotypes correspond to those where the parental haplotypes are not found or to haplotypes where the two parental copies are similar.

set is <10% (similar for the three species, between 8.60% and 9.30%; table 3).

K_A/K_S and Molecular Dating of Duplicate Genes

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

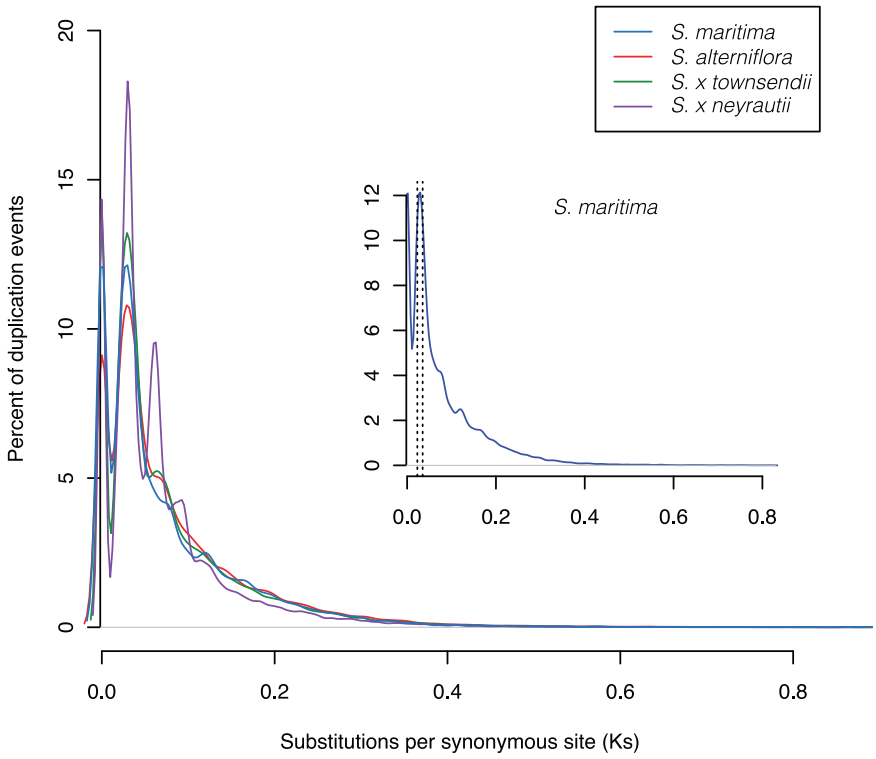

Fig. 5.—K_S distribution for the two parents *S. maritima*, *S. alterniflora*, and the two F1 hybrids *S. x townsendii* and *S. x neyrautii*. Dotted vertical bars represent the estimations of the duplication event (0.023–0.035).

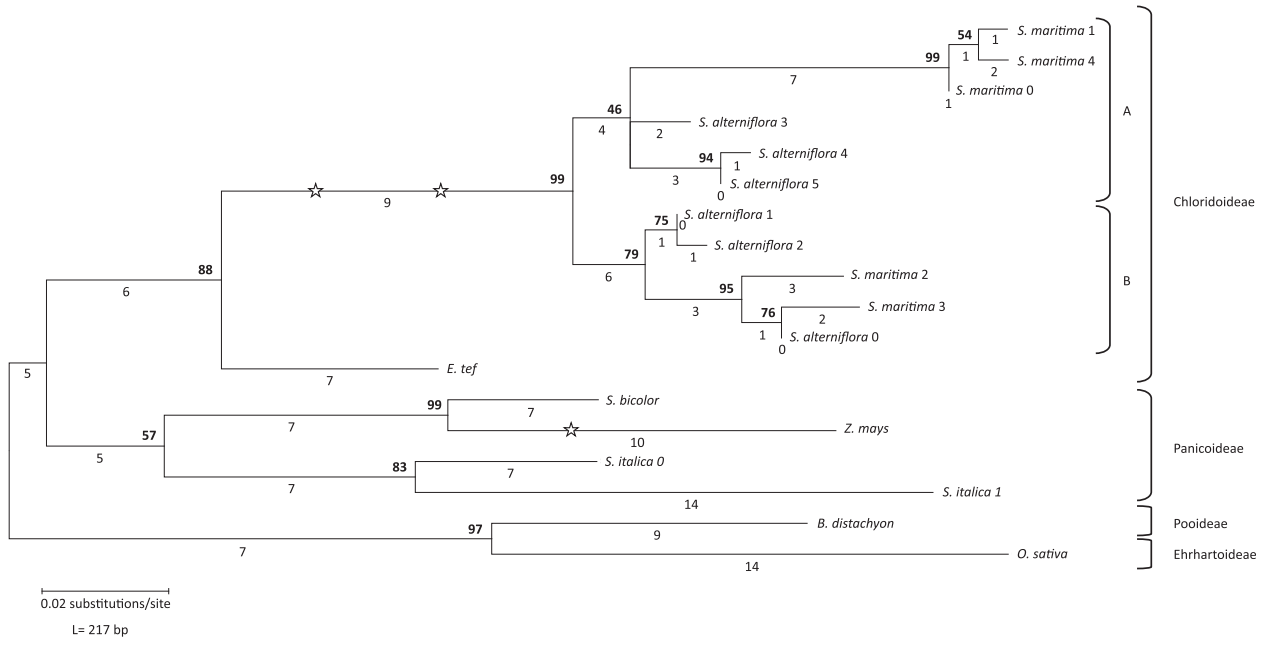

Fig. 6.—Phylogenetic analysis of the PPR gene (Pentatricopeptide repeat superfamily protein, GO annotations: 0003674, 0008150, 0005739) with Maximum Likelihood method (K2 + G model). The numbers of substitutions indicated under the branches were obtained from a Maximum Parsimony analysis which generated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

the two parents, *S. maritima* and *S. alterniflora* and the two F1 hybrids one peak (0.023–0.035) is observed.

Phylogenetic Analysis of the Haplotypes Detected in the PPR Gene

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

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

Discussion

In this study, we report the assembly and annotation of five reference transcriptomes for the European hexaploid *Spartina* species (*S. maritima*, *S. alterniflora* and their homoploid hybrids *S. x townsendii* and *S. x neyrautii*) and the allododecaploid invasive species *S. anglica.* The use of a deep sequencing technology significantly enhanced the previously assembled and published reference transcriptomes built for the hexaploid parental species (Ferreira de Carvalho et al. 2013) with 60,644 and 44,158 contigs against 25,239 and 14,317 for *S. maritima* and *S. alterniflora*, respectively, and up to 30% more functionally annotated contigs. We also provide here the first reference transcriptomes of the two hybrids and the allododecaploid species *S. anglica*. As the redundant nature of *Spartina* genomes and transcriptomes due to their high ploidy levels and their hybrid origin was challenging, we developed generic bioinformatics tools to (1) detect different haplotypes of each gene within these species and (2) assign a parental origin to haplotypes detected in the hybrids and the allopolyploid. The approach described here allows the detection of putative homeologs from sets of short reads and can be applied for future differential gene expression or genomics experiments to study the fate of duplicated genes in the allododecaploid *S. anglica*

Spartina Transcriptomics

Before the NGS revolution, *Spartina* transcriptomic resources were restricted to few EST sequences available on NCBI databases (Baisakh et al. 2008; Chelaifa, Mahé et al. 2010). Whole genome expression experiments were designed using heterologous rice microarrays to demonstrate differential expression in similar growing conditions of *S. maritima* and *S. alterniflora* (Chelaifa, Mahé et al. 2010) and the relative effects of hybridization and genome duplication on nonadditive expression in *S. x neyrautii*, *S. x townsendii* and *S. anglica* (Chelaifa, Monnier et al. 2010). Besides the nonspecies specific design of the array of this approach that was limiting the number of transcript detected, the measured signals included all putative homeologs, disabling the study of each duplicated gene. NGS technologies were then first used to build reference transcriptomes of the tetraploid species *Spartina pectinata*, using Roche-454 data (Gedye et al. 2010) and for *S. maritima* and *S. alterniflora* (Ferreira de Carvalho et al. 2013). In our study, we used a combination of Roche-454 and Illumina deep sequencing reads data sets to improve the reference transcriptomes of the two parents *S. maritima* and *S. alterniflora*, and to assemble the first reference transcriptomes for the two hybrids *S. x townsendii*, *S. x neyrautii* and the allododecaploid *S. anglica*. We assembled independently the Roche-454 (with Newbler) and Illumina read data sets (with Trinity) before co-aligning them (with Newbler and custom scripts to enhance assemblies by self-BLAST). The Newbler software is commonly used for Roche-454 data (Margulies et al. 2005) and showed positive results on similar data sets (Ferreira de Carvalho et al. 2013). The choice of the Trinity software is based on the results of several studies (Clarke et al. 2013; Liu et al. 2013) and comparative tests on our data set. The hybrid assembly strategy for Roche-454 and Illumina contigs showed good results in several studies (Sirota-Madi et al. 2010; Barthelson et al. 2011; Jiang et al. 2011) using assemblers such as Mira or Abyss. The length of Roche-454 contigs obtained in the first step of our assembly process motivated the choice of the Newbler assembler. The large number of Illumina contigs obtained by the Trinity software (76,010–121,733 contigs) is explained by the presence of several similar copies (identity $\geq$ 90%) and were automatically re-assembled with Newbler in the co-assembly step. The parameters used for the different assemblies (90%) are consistent with the literature (Franchini et al. 2011; Ferreira de Carvalho et al. 2013; Liu et al. 2013) and fitted in order to get consensus sequences of all putative homeologs for each species. The functional annotations were made using a method similar to that used by Ferreira de Carvalho and collaborators

(tBLASTx, BLAST2Go approach; Ferreira de Carvalho et al. 2013) and using the Pfam software used in annotation pipeline such as TRAPID (Van Bel et al. 2013). The number of annotated contigs represents 36.50–43.57% of the total number of contig, the unannotated contigs have a lower average length, but also a lower average read depth and are reconstructed using a limited number of reads compared with annotated contigs (see supplementary fig. S1 and table S1, Supplementary Material online, for details); they also are shorter (40–200 bp). The annotated contigs of the two parents *S. maritima* and *S. alterniflora* have an average length of 741.75 and 761.11 bp, respectively, similar to contigs assembled by Ferreira de Carvalho and collaborators (2013) who reported an average length included between 415 and 759 (617 and 415 bp for *S. maritima* and 759 bp for *S. alterniflora*). The average length of the unannotated contigs corresponds to 339.56 and 336.00 bp for *S. maritima* and *S. alterniflora*, respectively (these results are similar for the other species). To validate the contig constructed and to detect the different SNPs and haplotypes, we have mapped (to 90% of identity) between 34.22% and 57.84% among different species. Contigs have an average read depth included to 25.12×–90.52×. These values are similar to the study of Franchini *et al.* (2011) where 10,635,178 Illumina paired-end reads (42.10%) have been used (36.6× of average read depth) to construct the transcriptome of an abalone species.

Haplotype Detection

Several studies have focused on the detection of different copies in polyploid genomes such as cotton, coffee, strawberry or even the paleopolyploid soybean genome (Flagel et al. 2008; Salmon et al. 2009; Ilut et al. 2012; Combes et al. 2013; Tennessen et al. 2014). Nevertheless, the strategies developed in these studies can only be applied on species with known diploid parents. Detection of the different copies in *Spartina* hexaploid species using Roche-454 data was previously restricted to a few genes (Ferreira de Carvalho et al. 2013) and was recently automated for rDNA gene copies in *S. maritima* (Boutte et al. 2016). Our study reports here the automated detection of haplotypes at a whole transcriptome scale enabling us to identify the parental origin of the hybrid and polyploid haplotypes.

In our study, the number of haplotypes detected in the investigated *Spartina* species is correlated with the number of SNPs detected and the number of copies expected. For the parents, we have detected around seven haplotypes by windows and 12–14 for the hybrids and the allododecaploid species. These values are higher than the number of homoeologous copies expected (three pairs for the hexaploids parents and for the hybrids and six pairs for the allopolyploid) suggesting co-alignements and detection of either paralogs or alleles. A previous study focusing on a few targeted genes demonstrated the detection of four haplotypes in the parental species with Roche-454 data that indicated the presence of two homoeologous copies (Ferreira de Carvalho et al. 2013). A study realized on Waxy genes using cloning and Sanger sequencing indicates the presence of one homoeologous copy in *S. maritima* and three copies in *S. alterniflora* (Fortune et al. 2007). Furthermore, the nonadditive gene expression in polyploid species could lead to a lower number of detected copies (Yoo et al. 2014). The phylogeny of the different haplotypes for a Pentatricopeptide repeat (PPR) superfamily protein gene indicates the presence of two divergent homoeologous copies and additional alleles (two to three per copy in *S. maritima* and three in *S. alterniflora*). Prevalence of reticulate evolution in *Spartina* and previous gene topologies (e.g., *Waxy* gene, Fortune et al. 2007) suggested an allopolyploid origin of the hexaploid ancestor to *S. maritima* and *S. alterniflora*; however, we cannot rule out a possible allo-auto hexaploid origin, which would result in divergent homeologs and additional related homologous alleles.

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

Frequency distributions of K_S values between pairs of haplotypes (Blanc and Wolfe 2004) exhibited one peak common to the four hexaploid species (0.023–0.035): *S. maritima*, *S. alterniflora*, *S. x townsendii* and *S. x neyrautii*. Simulation of an allododecaploid species using parental haplotypes provided similar results (supplementary fig. S2A, Supplementary Material online). The different peaks observed in the two hybrids *S. x townsendii* and *S. x neyrautii* are due to the presence of haplotypes from both parents. The hybrid simulation using parental mapping reads process confirmed these results (supplementary fig. S2B, Supplementary Material online). The second peak observed in the hybrid *S. x neyrautii* corresponds to the sum of the divergence between homeologous copies present in hexaploid *Spartina* species and of the divergence

between *S. maritima* and *S. alterniflora*. The peak observed in all species indicates a burst of the number of duplicated genes resulting from whole genome duplication, which is most likely related to the tetraploidy event in *Spartina*. In agreement to this hypothesis, a similar peak was also found in two tetraploid species (*Spartina versicolor* Fabre and *Spartina bakeri* Merr.; unpublished data). The absence of an additional peak in these latter hexaploid species as expected when considering two allopolyploidization events scenario for hexaploids (Fortune et al. 2007) may be explained by different hypotheses. It may result from the presence of two combined peaks, suggesting that the tetraploid and hexaploid clades formed within a short evolutionary time. Another explanation to the absence of this second peak may be related to the mapping parameters used in this study (>88% of identity). These parameters may be too stringent for identifying an additional and very divergent duplicated copy in the hexaploid species. The presence of a single peak is in accordance with the identification of only two homoeologous copies within genomic or transcriptomic data of hexaploid species (Ferreira de Carvalho et al. 2013; C. Charron, personnal communication). Alternatively, the hexaploid *Spartina* species might have formed by the merging of low divergent genomes (e.g., auto-allopolyploidy). Further analyses are needed to explore these hypotheses.

Conclusion

In conclusion, five new *Spartina* reference transcriptomes were assembled by combining two NGS technologies (Roche-454 and Illumina). After transcriptomic assembly and annotation of the different contigs, SNP detection allowed reconstructing different haplotypes, which could correspond to paralogous, homoeologous and even allelic copies. About seven haplotypes were most frequently detected for the hexaploid parents and 12–14 haplotypes were observed in the two hybrids and the allododecaploid; their parental origins were assigned. K_S distribution peak indicate one duplication event in *Spartina* species common to tetraploid and hexaploid species, which indicates an allotetraploid formation of this monophyletic lineage. Origin of the hexaploid *Spartina* clade is not yet resolved using the K_S method.

The *Spartina* reference transcriptomes constructed may provide useful informations to explore gene expression in the context of *Spartina* ecology, such as genes implicated in responses to abiotic stresses (salt and oxidative stress or to heavy metal stress for example), biotic interactions (Gray and Benham 1990) and in the context of allopolyploid speciation. It is now possible to study the different transcription levels of the detected copies in different natural or experimental conditions; this opens new perspectives for studying duplicate gene expression evolution in the context of the adaptive success of *S. anglica*.

Supplementary Material

Supplementary table S1 and figures S1 and S2 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

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, Iowa State University (Ames) and the Partner University Funds (to M.A., A.S., and J.B.). Sequencing was supported by Genoscope funds (GENOSPART Project). We would like to thank the members of POLY-BNF group: A. M. Chèvre, D. Lavenier, P. Peterlongo and C. Lemaitre for interaction about Illumina assemblies. The analyses benefited from the Molecular Ecology (UMR CNRS 6553 Ecobio) and Genouest (Bioinformatics) facilities. J.B. benefited from a PhD scholarship from the University of Rennes 1 (France). We thank two anonymous reviewers for helpful comments on the article.

Literature Cited

Abbott RJ, et al. 2008. Recent hybrid origin and invasion of the British Isles by a self-incompatible species, Oxford ragwort (*Senecio squalidus* L., Asteraceae). Biol Invasions 11:1145–1158.

Adams KL, Cronn R, Percifield R, Wendel JF. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc Natl Acad Sci. 100: 4649–4654.

Ainouche ML, Baumel A, Salmon A. 2004. *Spartina anglica* CE Hubbard: a natural model system for analysing early evolutionary changes that affect allopolyploid genomes. Biol J Linn Soc. 82:475–484.

Ainouche ML, et al. 2008. Hybridization, polyploidy and invasion: lessons from *Spartina* (Poaceae). Biol Invasions 11:1159–1173.

Ainouche ML, et al. 2012. Polyploid evolution in *Spartina*: dealing with highly redundant hybrid genomes. In: Soltis PS, Soltis DE, editors. Polyploidy and genome evolution: dealing with highly redundant hybrid genomes. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 225–243.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

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.

Barthelson R, McFarlin AJ, Rounsley SD, Young S. 2011. Plantagora: modeling whole genome sequencing and assembly of plant genomes. Pellegrini, M, editor. PLoS One 6:e28436.

Baumel A, Ainouche M, Kalendar R, Schulman AH. 2002. 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. 2002. 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* CE 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.

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

Boutte J, et al. 2016. Haplotype detection from next-generation sequencing in high-ploidy-level species: 45S rDNA gene copies in the hexaploid *Spartina maritima.* G3 (Bethesda, Md.) 6:29–40.

Buggs RJA, et al. 2012. Next-generation sequencing and genome evolution in allopolyploids. Am J Bot. 99:372–382.

Chalhoub B, et al. 2014. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science 345:950–953.

Chelaifa H, Mahé F, Ainouche M. 2010. Transcriptome divergence between the hexaploid salt-marsh sister species *Spartina maritima* and *Spartina alterniflora* (Poaceae): transcriptome divergence in *Spartina*. Mol Ecol. 19:2050–2063.

Chelaifa H, Monnier A, Ainouche M. 2010. Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species *Spartina* × *townsendii* and *Spartina anglica* (Poaceae). New Phytol. 186:161–174.

Chopra R, et al. 2014. Comparisons of *de novo* transcriptome assemblers in diploid and polyploid species using peanut (Arachis spp.) RNA-seq data. PloS One 9:e115055.

Clarke K, Yang Y, Marsh R, Xie L, Zhang KK. 2013. Comparative analysis of *de novo* transcriptome assembly. Sci China Life Sci. 56:156–162.

Clevenger JP, Ozias-Akins P. 2015. SWEEP: a tool for filtering high-quality SNPs in polyploid crops. G3 (Bethesda, Md.) 5:1797–1803.

Combes M-C, Dereeper A, Severac D, Bertrand B, Lashermes P. 2013. Contribution of subgenomes to the transcriptome and their intertwined regulation in the allopolyploid *Coffea arabica* grown at contrasted temperatures. New Phytol. 200:251–260.

Conesa A, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676.

Dalibor H, et al. 2016. Persistence, dispersal and genetic evolution of recently formed *Spartina* homoploid hybrids and allopolyploids in Southern England. Biol Invasions 18:2137–2151.

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

Ferreira de Carvalho J, et al. 2013. Transcriptome *de novo* assembly from next-generation sequencing and comparative analyses in the hexaploid salt marsh species *Spartina maritima* and *Spartina alterniflora* (Poaceae). Heredity 110:181–193.

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

Finn RD, et al. 2014. Pfam: the protein families database. Nucleic Acids Res. 42:D290–D301.

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

Flagel LE, Wendel JF, Udall JA. 2012. Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. BMC Genomics 13:302.

Fortune PM, et al. 2007. Evolutionary dynamics of Waxy and the origin of hexaploid *Spartina* species (Poaceae). Mol Phylogenet Evol. 43:1040–1055.

Foucaud J. 1897. Un *Spartina* inédit. Ann. Société Sci. Nat. Charante Inférieure. 32:220–222.

Franchini P, Van der Merwe M, Roodt-Wilding R. 2011. Transcriptome characterization of the South African abalone *Haliotis midae* using sequencing-by-synthesis. BMC Res Notes 4:59.

Gedye K, et al. 2010. Investigation of the transcriptome of Prairie cord grass, a new cellulosic biomass crop. Plant Genome J. 3:69.

Glover NM, Redestig H, Dessimoz C. 2016. Homoeologs: what are they and how do we infer them? Trends Plant Sci. 21:609–621.

Götz S, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36:3420–3435.

Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29:644–652.

Gray AJ. PEM, Benham 1990. *Spartina anglica* – a research review. London: HMSO.

Gregory TR, Mable BK. 2005. Polyploidy in animals. In: Gregory TR, editors. The evolution of the genome. San Diego, California, USA: Elsevier. p. 428–517.

Grover CE, et al. 2012. Homeolog expression bias and expression level dominance in allopolyploids. New Phytol. 196:966–971.

Higgins J, Magusin A, Trick M, Fraser F, Bancroft I. 2012. Use of mRNA-seq to discriminate contributions to the transcriptome from the constituent genomes of the polyploid crop species *Brassica napus*. BMC Genomics 13:247.

Ilut DC, 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.

Jiang Y, et al. 2011. A pilot study for channel catfish whole genome sequencing and *de novo* assembly. BMC Genomics 12:629.

Jiao Y, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 16:111–120.

Langham RJ, et al. 2004. Genomic duplication, fractionation and the origin of regulatory novelty. Genetics 166:935–945.

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

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

Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol. 2:150–174.

Liu S, Li W, Wu Y, Chen C, Lei J. 2013. *De novo* transcriptome assembly in Chili Pepper (*Capsicum frutescens*) to identify genes involved in the biosynthesis of *Capsaicinoids Schönbach*, C, editor. PLoS One 8:e48156.

Mable BK. 2004. 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. Biol J Linn Soc. 82:453–466.

Mable BK, Alexandrou MA, Taylor MI. 2011. Genome duplication in amphibians and fish: an extended synthesis: polyploidy in amphibians and fish. J Zool. 284:151–182.

Malinska H, et al. 2011. Ribosomal RNA genes evolution in Tragopogon: a story of new and old world allotetraploids and the synthetic lines. Taxon 60:348.

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

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

Margulies M, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.

Marhold K, et al. 2009. Cytotype diversity and genome size variation in eastern Asian polyploid *Cardamine* (Brassicaceae) species. Ann Bot. 105:249–264.

Milne I, et al. 2009. Tablet–next generation sequence assembly visualization. Bioinformatics 26:401–402.

Mobberley DG. 1956. Taxonomy and distribution of the genus *Spartina*. Iowa State Coll. J Sci. 30:471–574.

Nicolas SD, et al. 2007. Homeologous recombination plays a major role in chromosome rearrangements that occur during meiosis of *Brassica napus* haploids. Genetics 175:487–503.

Oliphant A, Barker DL, Stuelpnagel JR, Chee MS. 2002. BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. BioTechniques Suppl:56–58:60–61.

Otto SP, Whitton J. 2000. Polyploidy incidence and evolution. Annu Rev Genet. 34:401–437.

Page JT, et al. 2016. DNA sequence evolution and rare homoeologous conversion in tetraploid cotton. PLoS Genet. 12(5):e1006012.

Page JT, Gingle AR, Udall JA. 2013. PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. G3 (Bethesda, Md.) 3:517–525.

Page JT, Huynh MD, et al. 2013. Insights into the evolution of cotton diploids and polyploids from whole-genome re-sequencing. G3 (Bethesda, Md.) 3:1809–1818.

Page JT, Udall JA. 2015. Methods for mapping and categorization of DNA sequence reads from allopolyploid organisms. BMC Genet. 16(Suppl 2):S4.

Parisod C, et al. 2009. Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in *Spartina*. New Phytol. 184:1003–1015.

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

Peterson PM, Romaschenko K, Arrieta YH, Saarela JM. 2014. A molecular phylogeny and new subgeneric classification of *Sporobolus* (Poaceae: Chloridoideae: Sporobolinae). Taxon 63:1212–1243.

R Development Core Team. 2011. R: a language and environment for statistical computing. Austria: Vienna. Available from: http://www.R-project.org/.

Rousseau-Gueutin M, et al. 2015. The chloroplast genome of the hexaploid *Spartina maritima* (Poaceae, Chloridoideae): comparative analyses and molecular dating. Mol Phylogenet Evol. 93:7–5.

Salmon A, Ainouche ML. 2015. Next generation sequencing and the challenge of deciphering evolution of recent and highly polyploid genomes. Germany: Koeltz Scientific Books. [cited 2015 Sep 27] Available from: www.iapt-taxon.org.

Salmon A, Ainouche ML, Wendel JF. 2005. Genetic and epigenetic consequences of recent hybridization and polyploidy in *Spartina* (Poaceae): genetic and epigenetic changes in *Spartina*. Mol Ecol. 14:1163–1175.

Salmon A, Flagel L, Ying B, Udall JA, Wendel JF. 2009. Homoeologous nonreciprocal recombination in polyploid cotton. New Phytol. 186:123–134.

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

Sirota-Madi A, et al. 2010. Genome sequence of the pattern forming *Paenibacillus vortex* bacterium reveals potential for thriving in complex environments. BMC Genomics 11:710.

Soltis DE, et al. 2009. Polyploidy and angiosperm diversification. Am J Bot. 96:336–348.

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

Tamura K, et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28:2731–2739.

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

Udall JA. 2005. Detection of chromosomal rearrangements derived from homeologous recombination in four mapping populations of *Brassica napus* L. Genetics 169:967–979.

Udall JA, et al. 2006. A global assembly of cotton ESTs. Genome Res. 16:441–450.

Vallejo-Marin M. 2012. *Mimulus peregrinus* (Phrymaceae): a new British allopolyploid species. PhytoKeys 14:1–14.

Van Bel M, et al. 2013. TRAPID: an efficient online tool for the functional and comparative analysis of *de novo* RNA-Seq transcriptomes. Genome Biol. 14:R134.

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

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191.

Wendel JF. 2000. Genome evolution in polyploids. Plant Mol Biol. 42:225–249.

Wendel JF, Jackson SA, Meyers BC, Wing RA. 2016. Evolution of plant genome architecture. Genome Biol. 17:37.

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 MJ, Szadkowski E, Wendel JF. 2013. Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity 110:171–180.

Yoo M-J, Liu X, Pires JC, Soltis PS, Soltis DE. 2014. Nonadditive gene expression in polyploids. Annu Rev Genet. 48:485–517.

Associate editor: Ellen Pritham