Evidence of intense chromosomal shuffling during conifer evolution

Marina de Miguel, Jérôme Bartholome, François Ehrenmann, Florent Murat, Yoshinari Moriguchi, Kentaro Uchiyama, Saneyoshi S. Ueno, Yoshihiko Tsumura, Hélène Lagraulet, Nuria de Maria, et al.

Genome Biology and Evolution Advance Access published September 23, 2015 doi:10.1093/gbe/evv185

Genome Biology and Evolution

SMBE

Evidence of intense chromosomal shuffling during conifer evolution

Marina de Miguel[1,2,$], Jérôme Bartholomé[1,2,$], François Ehrenmann[1,2], Florent Murat[3], Yoshinari Moriguchi[4], Kentaro Uchiyama[5], Saneyoshi Ueno[5], Yoshihiko Tsumura[6], Hélène Lagraulet[1,2], Nuria de Maria[7,8], José-Antonio Cabezas[7,8], Maria-Teresa Cervera[7,8], Jean Marc Gion[1,2,9], Jérôme Salse[3], Christophe Plomion[1,2,*]

[1] INRA, UMR 1202 BIOGECO, 69 Route d'Arcachon,F-33610 Cestas, France

[2] Université de Bordeaux, UMR 1202 BIOGECO, F-33170 Talence, France

[3] INRA/UBP UMR 1095 GDEC 'Génétique, Diversité et Ecophysiologie des Céréales', 5 Chemin de Beaulieu, 63100 Clermont Ferrand, France

[4] Niigata University, Graduate School of Science and Technology, 8050, Igarashi 2-Nocho, Nishi-ku, Niigata 950-2181, Japan

[5] Forestry and Forest Products Research Institute, Department of Forest Genetics, Tsukuba, Ibaraki 305-8687, Japan

[6] University of Tsukuba, Faculty of Life & Environmental Sciences, 1-1-1, Tennodai, Tsukuba, Ibaraki 305-8572, Japan

[7] INIA-CIFOR, departamento de Ecologia y Genetica Forestal, 28040, Madrid, Spain

[8] INIA-UPM, Unidad mixta de Genomica y Ecofisiologia Forestal, Madrid, Spain

[9] CIRAD, UMR AGAP, F-33612 Cestas, France

[$]contributed equally

*Author for correspondence: Dr. Christophe Plomion, INRA UMR BioGeCo 1202, 69 Route d'Arcachon, 33612 CESTAS Cedex - France, +33 5-57-12-27-65, plomion@pierroton.inra.fr

Abstract

While recent advances have been gained on genome evolution in angiosperm lineages, virtually nothing is known about karyotype evolution in the other group of seed plants, the gymnosperms. Here we used high density gene-based linkage mapping to compare the karyotype structure of two families of conifers (the most abundant group of gymnosperms) separated around 290 million years ago: *Pinaceae* and *Cupressaceae*. We propose for the first time a model based on the fusion of 20 ancestral chromosomal blocks that may have shaped the modern karyotpes of *Pinaceae* (with n=12) and *Cupressaceae* (with n=11). The considerable difference in modern genome organization between these two lineages contrasts strongly with the remarkable level of synteny already reported within the *Pinaceae*. It also suggests a convergent evolutionary mechanism of chromosomal block shuffling that has shaped the genomes of the spermatophytes.

Keywords: chromosomal rearrangement, comparative mapping, *Cuppressaceae*, gymnosperm, *Pinaceae*, synteny.

Introduction

Knowledge about genome structure and evolution is a fundamental step towards understanding species adaptation and evolution. Genome evolution is based on genetic variability generated by mutation, recombination and the acquisition of new genes. New genes can be acquired by interspecific hybridization or the duplication of some or all the existing genes of an organism (Renny-Byfield & Wendel 2014; Cong *et al*. 2015). Plant genomes, unlike those of animals, have evolved through frequent, rapid chromosomal rearrangements, including whole-genome duplications (WGD) followed by nested chromosome fusions in particular (Abrouk *et al*. 2010; Salse *et al*. 2015). The sequencing of plant genomes has made it possible to construct evolutionary models for various angiosperm lineages (Murat *et al*. 2010; Salse 2012; Murat *et al*. 2015). These studies revealed that angiosperms have experienced successive common and lineage-specific WGDs, which have governed increases in genome size and shaped the genome structure and composition of extant species (Renny-Byfield & Wendel 2014). Evolutionary genome shuffling events (chromosomal fusions and fissions) have been identified during the course of angiosperm evolutionary history, making it possible to reconstruct the karyotypes of the common ancestors of eudicots, with seven protochromosomes, and of monocots, with five or seven protochromosomes (Abrouk *et al.* 2010; Salse 2012). However, we still know little about karyotype evolution in the other group of seed plants, the gymnosperms.

Conifers are the most abundant group of gymnosperms. Genome structure and evolution differ between conifers and angiosperms. Conifers have extremely large genomes (ranging from 18 to 35 gigabases) characterized by the presence of repetitive elements (Kovach *et al*. 2010; Mackay et al. 2012; Neale *et al*. 2014). These features have hindered attempts to sequence the genomes of these plants and the recently released draft genome sequences are

highly fragmented (Nystedt *et al*. 2013; Zimin *et al*. 2014; Warren *et al*. 2015). Consequently, the evolutionary mechanisms shaping the composition and structure of conifer genomes remain to be determined. One ancient WGD event is known to have occurred before the angiosperm-gymnosperm split around 350 million years ago (MYA) (Jiao *et al*. 2011). However, there is no evidence to suggest that other WGD events have occurred during the evolutionary history of conifers (Kovach *et al*. 2010; Nystedt *et al*. 2013), by contrast to what has been reported for angiosperms. Conifer genome size seems to have increased due to the accumulation of large numbers of retrotransposons (Morse *et al*. 2009; Nystedt *et al*. 2013). Polyploidy is exceptional in gymnosperms, with only two species from the *Cupressaceae* known to be polyploids, one of these species being hexaploid (*Sequoia semperviens* 2n=66) and the other tetraploid (*Juniperus chinensis* 2n=44). The haploid number of chromosomes in conifers ranges from 9 to 19, but karyotypes are highly conserved across species and genera, with most having 11 or 12 chromosomes (Wang & Ran 2014).

Pinaceae, the largest family of conifers, has been studied more thoroughly than other conifers, for ecological and economic reasons. *Pinus* and *Picea*, the main genera within *Pinaceae*, separated around 87 to 193 MYA (Morse *et al*. 2009). Comparative mapping between *Pinaceae* species has revealed high levels of interspecific and intergeneric synteny and macrocollinearity (Krutovsky *et al*. 2004; Pelgas *et al*. 2006; Pavy *et al*. 2012), suggesting a lack of chromosomal rearrangement within this family, despite the ancient nature of the divergence between some taxa. Nevertheless, the issue of the conservation of synteny across different families of conifers has not yet been addressed. Further studies of the evolution of conifer karyotypes are therefore required, to determine whether it has followed a pattern similar to that in angiosperms or more similar to that in the *Pinaceae*. In the absence of a completely contiguous reference genome in conifers, high-density comparative mapping

provides an opportunity to compare genomes from different lineages, thereby improving our understanding of conifer karyotype evolution.

In this work, we analyzed conifer karyotype evolution through comparative mapping, making use of published genetic linkage maps for two families: *Pinaceae* (n=12) and *Cupressaceae* (n=11). The aims of this study were: i) to analyze the degree of gene synteny and collinearity at the interfamily level; ii) to set a likely scenario of karyotype evolution between both families.

Results and Discussion

We carried out a literature review, to select high-density gene-based linkage maps for conifers for which sequence information is publicly available. We included 18 genetic maps (Table 1) for six different species from two botanical families — *Pinaceae* (n=12) and *Cupressaceae* (n=11) — in this study. We made use of the high degree of synteny and macrocollinearity within *Pinaceae* (Krutovsky *et al*. 2004; Pavy *et al*. 2012) to establish a gene-based composite map for this botanical family. A stepwise strategy, from species to family level, was used to maximize the number of mapped markers common to different maps, thereby maximizing the number of anchor markers for the construction of the composite map for *Pinaceae* (Figure 1a). We began by constructing a composite linkage map for *Pinus pinaster* from 14 maps (Table 1). We then generated two genus-level composite maps: i) for *Pinus* sp., by combining the *P. pinaster* composite map generated in this study with a published map for *Pinus taeda* (Eckert *et al*. 2010); ii) for *Picea* sp., based on published maps for *Picea abies* (Lind *et al*. 2014), *Picea glauca* and *Picea mariana* (Pavy *et al*. 2012). The composite maps for *Pinus* sp. and *Picea* sp. were then merged into a unified composite map for *Pinaceae*. This composite map for *Pinaceae* was then compared with a published gene-based map for *Cryptomeria*

japonica (Moriguchi *et al*. 2012), a representative member of *Cupressaceae*. The strategy used to combine and compare the genetic maps is illustrated in Figures 1a and S1.

The *P. pinaster* composite map comprised 3,491 unigenes of the Pine V3 Unigene set (Canales *et al*. 2014) as well as 182 AFLP or SAMPL markers (Table 2, Figure S2). There were 3,639 unique markers, 60% of which were present on at least two component maps. Overall, we found high degrees of synteny and collinearity between all the *P. pinaster* component maps and the composite map (Figure S3). The mean proportion of markers non-collinear (inversion greater than 5 cM) between the composite map and a component map was 1.3%. The large number of markers common to different component linkage maps and the high levels of collinearity observed, increased the degree of certainty concerning the relative positions of the mapped unigenes. The *P. pinaster* composite map contributed the largest number of mapped markers for construction of the composite map for *Pinaceae* (Tables 1 and 2).

The composite map for *Pinaceae* contained 6,912 mapped markers over 2,094.9 cM (Table 2, Figure S4). As SNPs mapped in this composite map were identified from a variety of transcritptomic assays in diverse species, we considered as different gene loci only those that had an homolog in Pine V3, the gene catalog used as reference. Following this criterion, at least 5,927 different unigenes were mapped in the *Pinaceae* composite map (Tables 2 and S1). On average, 42 unigenes per LG were common to the *Pinus* sp. and *Picea* sp. maps, identifying a total of 513 othologous unigenes between both species (Figure 2, Table S1). Only 5.9% of unigenes were non-syntenic and 2.8% presented an inversion of more than 15 cM (Table S1). These markers were identified and removed from the *Pinaceae* composite map. As expected, there was a high degree of synteny and collinearity between members of the *Pinaceae*, providing support for the strategy followed in this study. A small fraction of mapped unigenes in the *Pinaceae* composite map may be originated by paralogy as revealed

by the 44 unigenes mapped in more than one LG (Table S2). Finally, the *P. pinaster*, *Pinus* and *Pinaceae* composite maps generated in this study are the densest linkage maps ever produced for conifers at species, genus and family levels, respectively.

Sequence comparisons between unigenes mapped on the *Pinaceae* and *C. japonica* genetic maps resulted in the identification of 257 and 229 homologous loci depending on the e-value cut-off applied (Figure 1b). Homologous unigenes were identified for all LGs whatever the threshold considered, from 17 on LG4 to 28 on LG3 for the *Pinaceae* map and from 13 on LG8 to 35 on LG3 for the *C. japonica* map (for an e-value cut-off of 1 e^{-30}). Linkage maps were aligned using homologous unigenes as anchor points. The alignment of both genetic maps enabled the identification of common genomic regions. Dotplot representation for map alignment (Figure S5) showed a threshold of four shared unigenes between both maps suitable for othologous LG block determination. A more stringent criterion for ortholog selection was additionally tested, which consisted in a minimum of six shared unigenes between two regions to be orthologs. A total of 12 to 20 othologous LG blocks were identified depending on the threshold used (Figure 1b). However, orthologous LG blocks covered the complete set of chromosomes of the analyzed species only when a threshold of four shared markers was used (Figure 1b). Consequently, this threshold was considered the most appropriate in view of the level of resolution of available genetic maps, and further discussion of the results is based on this threshold. The use of a threshold of four homolgous unigenes to consider an orthologous LG block resulted in the identification of 143 orthologous loci (i.e. 55.6% of the homologous markers) for an e-value cut-off lower than 1 e^{-30} and 124 (i.e. 54.2% of the homologous markers) for an e-value cut-off of 1 e^{-35}. Thus, the use of a more stringent selection criterion for the identification of homologous sequences did not decrease significantly the proportion of identified orthologous unigenes.

As a result, we found that each of the LGs on the *Pinaceae* map corresponded to one or two different LG blocks on the *Cupressaceae* map, and that each LG on the *Cupressaceae* map corresponded to one to three LG blocks on the *Pinaceae* map (Figure 3a and S5). Each pair of orthologous LG blocks determined an ancestral contiguous region (CAR). Most CARs were identified whatever the e-value threshold applied with the exception of CARs #14 and #19 (Table 3) that could not be confirmed using the most stringent criterion. Mean number of unigenes per CAR was six and seven depending of the e-value cut-off applied. The number of orthologous unigenes per CAR was slightly reduced in eight CARs for the most stringent e-value cut-off, but the size of CARs was maintained with the exception of two CARs that were reduced by 20.1 and 48.1 cM, respectively (Table 3). Thus, the number and size of identified CARs was consistent under the two different thresholds tested for homolog identification. Therefore, our results suggests the existence of 18 to 20 CARs that may have shaped the 12 chromosomes of modern *Pinaceae* species and the 11 chromosomes of modern *Cupressaceae* species through a different number fusions (Figure 3b). Taking the *Pinaceae* map as the reference and inspecting the 20 proposed CARs (e-value threshold of 1 e^{-30}), seven *C. japonica* LGs displayed crossed CARs. Taking the non-crossing CARS as a measurement of collinearity, 40% of the CARs identified were considered to be collinear. Besides, high levels of collinearity were found within CARs, with only 18.1% of orthologous markers presenting an inversion of more than 15 cM within a CAR (Figure S5). These results suggests an intense shuffling of orthologous LG blocks during the evolution of *Pinaceae* and Cupressaceae, but a higher conservation of gene order within these blocks.

Previous comparative genomics studies in *Pinaceae* have reported high levels of synteny and collinearity for genes (Chagné *et al*. 2003; Krutovsky *et al*. 2006; Pelgas *et al*. 2006; Pavy *et al*. 2012). The conservative genome macrostructure among *Pinaceae* species has been interpreted as evidence that genome rearrangement events are rare in conifers (Diaz-Sala *et al*.

2013; Nystedt *et al.* 2013). The results presented here revise this view of conifer genome evolution, which was inferred essentially from comparisons of *Pinaceae* species. Our findings also support a new hypothesis that substantial chromosome rearrangements have occurred between families. Molecular phylogenetic studies support the splitting of conifers into two groups: *Pinaceae* and *Coniferales II*, corresponding to all conifer families other than *Pinaceae* (Bowe *et al.* 2000; Gugerli *et al.* 2001; Lu *et al.* 2014). The observed chromosomal rearrangements may have generated a reproductive barrier separating the two lineages around 290 million years ago (Burleigh *et al*. 2012). On the other hand, *Pinus* and *Picea* display remarkable levels of synteny and collinearity despite their ancient divergence, confirming the exceptionally high degree of genome structure conservation within *Pinaceae*. According to Gernandt *et al*. (2011), conifers (*Coniferales*) can be grouped into six different families: *Pinaceae*, *Podocarpaceae*, *Araucariaceae*, *Sciadopityaceae*, *Taxaceae* and *Cupressaceae*. Comparative genomics studies with representatives of other conifer families are crucial, to determine whether the lack of genome rearrangement observed in *Pinaceae* is a feature common to other conifers. The adaptive radiation of some *Cupressaceae* species dates from the Oligocene (23-33 MYA), but the first fossil record of *C. japonica* dates from 55-65 MYA (Yang *et al*. 2012). The study of genome structure in other species from *Cupressaceae* with a shorter life history could provide new insight into the mechanisms and patterns of genome evolution in conifers.

The n=12 karyotype is considered the most primitive in the *Pinaceae* family, based on chromosome morphometrics (Nkongolo *et al*. 2012). However, we were unable to reconstruct the karyotype of the common ancestor of *Pinaceae* and *Cupressaceae* in this study due to the lack of a suitable outgroup species phylogenetically close to conifers and with a well assembled genome. The candidate species best matching these criteria is the basal angiosperm *Amborella trichopoda* (Amborella Genome Project 2013). A comparison between basal

angiosperms and conifers should open up promising perspectives for the construction of a model of karyotype evolution. Comparative genomics and phylogenetic studies based on genome-wide comparisons with conifers will be crucial to bridge the gaps in our understanding of the evolution of plant genomes from cryptogams to flowering plants.

Conclusion

The results reported here take us a step beyond the "stasis" already described for the *Pinaceae*, opening up new avenues of research into the evolution of conifer genomes. We propose a possible scenario for conifer genome evolution, based on the fusion of chromosomal blocks, serving as a prelude to the modern karyotype configuration in *Pinaceae* and *C. japonica*. However, further improvements in our knowledge of basal angiosperms and gymnosperms will be required, to reconstruct the karyotype of the common ancestor of seed plants.

Materials and Methods

Description of the genetic linkage maps used in this study:

The following terms were used to describe the different kinds of genetic maps used in this study, as suggested by Hudson *et al*. (2012): i) sex-averaged map: a consensus map for both parents of a pedigree; ii) consensus map: an integrated map based on segregation data from individual component maps; iii) composite map: an integrated map of different individual component maps built by a marker-merging method; iv) component map: each of the maps used in the construction of a composite map. The graphics and the representations of genetic maps were produced with R 3.1.0 (R Core Team, 2014).

Pinus pinaster

We used 14 base maps generated from seven different crosses to generate a composite genetic linkage map for *P. pinaster*. The first six maps were obtained from three controlled crosses (pedigrees #1, #2 and #3 in Figure S1) between three different genotypes: Corsica × Landes (C×L), Morocco × Landes (M×L) and Corsica ×Morocco (C×M). In total, 106, 117 and 94 full-sibs were genotyped with the 9k SNP-array described by Plomion *et al*. (2015), for C×L, M×L and C×M, respectively. The regression mapping algorithm of JoinMap 4.1 (Van Ooijen, 2011) was used to produce two maps for each parental genotype (one per cross), according to a two-way pseudo-testcross mapping strategy (Grattapaglia & Sederoff 1994), using test-cross markers (i.e. segregating in a 1:1 Mendelian ratio) only. The genetic maps were then combined into sex-averaged maps (Corsica, Landes and Morocco, see Figure S1) with the function "combine groups for map integration" of JoinMap 4.1. More details on the construction of the maps can be found in Lagraulet (2015).

Four other maps from two different mapping populations were also used. The first population was a three-generation inbred pedigree consisting of an F2 population (#4 in Figure S1) resulting from the selfing of an inter-provenance tree (Landes x Corsica). The second population was a three-generation outbred pedigree (G2, #5) resulting from a controlled cross of two intra-provenance hybrid trees (Landes x Landes). The construction of these maps was described by Plomion *et al*. (2015) for the F2 population and Chancerel *et al.* (2013) for the G2 population. For the F2 population, two different sets of individuals were used to generate two maps (F2_O and F2_N) with the RECORD algorithm (Van Os *et al.* 2005). For the G2 population, one map for each parent (G2M and G2F) was produced with the regression mapping algorithm of JoinMap 4.1 (Van Ooijen, 2011). The F2_O, G2M and G2F maps included different marker types: amplified fragment length polymorphism — AFLP, single sequence repeat — SSR, expressed sequence tag — EST and SNPs from different arrays (Chancerel *et al*. 2011, 2013), whereas the F2_N map contained only SNPs from the 9k SNP-

array (Plomion *et al*., 2015). We made use of the large number of common markers and the high level of collinearity between the two F2 maps to construct a composite map (referred to as F2C by Plomion *et al*. 2015).

The last four maps were generated from two different F1 crosses: C14×C15 (#6 in Figure S1) and Gal1056×Oria6 (#7). From the initial parental maps of the C14×C15 mapping population described by de Miguel *et al*.(2012), we mapped an additional set of 980 SNPs from the 12k SNP-array described by Chancerel *et al*. (2011) and 273 SNPs from a 1,536 SNP-array (Saez-Laguna *et al*. unpublished) here, to increase the number of anchor markers common to other maps. The parental maps of the Gal1056×Oria6 population used by de Miguel *et al*. (2014) were reconstructed in this study with the most informative individuals. For both pedigrees, we used the maximum likelihood mapping algorithm of JoinMap 4.1 to generate individual genetic maps and sex-averaged maps. The four maps from the C14×C15 and Gal1056×Oria6 crosses contained different types of molecular markers (SSRs, ESTs, selective amplification of microsatellite polymorphic loci — SAMPLs and SNPs).

For all maps, genetic distances in centimorgans (cM) were calculated with the Kosambi mapping function (Kosambi 1943).

Other Pinaceae

We carried out a literature review to identify previously published high-density gene-based linkage maps for members of the *Pinaceae* family, for which sequence information was publicly available. Only four studies satisfied these criteria. Eckert *et al*. (2010) provided a sex-averaged linkage map for a two-generation outbred pedigree based on SNPs for *Pinus taeda* (accession# TG091, http://dendrome.ucdavis.edu/cmap/). The map provided by Pavy *et al*. (2012) is the densest genetic map published to date for *Picea*. This map was a consensus of the white spruce (*Picea glauca*) and black spruce (*Picea mariana*) linkage maps. The white

spruce pedigree was an F_1 full-sib family, whereas the black spruce pedigree was a backcross representing the hybridization species complex *Picea mariana* x *Picea rubens*. *Pica glauca* and *Picea mariana* linkage maps were constructed with the regression algorithm in JoinMap 3.0 software. Lind *et al*. (2014) provided the most saturated and gene-rich map to date for *Picea abies*. This map was a sex-averaged map of the two parents of an F_1 controlled cross and was also constructed with the regression algorithm of JoinMap 3.0. A detailed list of mapping features for each the component maps included in this study is available in Table 1.

Cryptomeria japonica

A high-density linkage map for *C. japonica* was incorporated into this study, as a representative species from the *Cupressaceae* family (Moriguchi *et al*. 2012). This map was constructed from an F_1 full-sib family (Table 1), with the regression mapping algorithm implemented in JoinMap v 3.0.

Identification of homologous genes within *Pinaceae*:

The 17 maps described above were mostly constructed with SNP markers (100% of the markers for *Picea* sp., 98% for *P. pinaster*, and 90% for *P. taeda*). The flanking sequences of each SNP marker were compared with the most recent Unigene sets available for each species, to obtain the sequence of unigenes containing the mapped SNPs: Canales *et al.* (2014) for *P. pinaster*, Rigault *et al*. (2011) for *Picea glauca*, *Picea mariana* and *Picea abies* and Lorenz *et al*. (2012) for *P. taeda*. This comparison was carried out with the *blastn* tool (the BLAST 1 step in Figure 1a). Unigenes with a percentage identity exceeding 95% with mapped SNP flanking sequences were retained for the next step.

P. pinaster Unigene set from Canales *et al*. (2014), Pine V3, was then used as the reference for the identification of homologous unigenes within *Pinaceae* species. A second sequence comparison (R-BLAST 2 in Figure 1a) was performed, between the mapped unigenes of each

species and the unigene sequences of Pine V3. For this interspecific comparison, a stringent reciprocal *tblastx* analysis was performed. Only sequences with a reciprocal best hit with a percentage identity exceeding 85%, an e-value below e^{-65} and an alignment of more than 200 bp were retained as homologous unigenes. Homologous unigenes between different species were considered as orthologs if they were positioned in the same LG (ie. syntenic unigenes). Identified orthologous unigenes were used as anchor markers to construct a composite linkage map for each genus (*Pinus* and *Picea*), as a preliminary step in the construction of a composite map for the *Pinaceae* family including both genera.

Construction of composite genetic linkage maps:

We used the R package *LPmerge* (Endelman & Plomion 2014) to integrate component linkage maps into a composite map without the use of segregation data. *LPmerge* assessed the goodness of fit of the composite map by calculating a root mean squared error (RMSE) per linkage group (LG), by comparing the position (in cM) of all markers on the composite map with that on the component maps. We calculated this metric for different maximum interval sizes (parameter *K* in the algorithm), ranging from one to ten. The value of *K* minimizing the mean RMSE per LG was selected for construction of the composite map. This method was used for the construction of all the composite species maps reported here. Further details about the production of each composite map are described below.

Pinus pinaster

Before integrating the 14 base genetic linkage maps into a single composite map, we established consensus maps (Figure S1) based on markers common to different accessions across pedigrees (Corsica, Landes, Morocco genotypes for pedigrees #1, #2 and #3, respectively), or accessions within pedigrees (Coca and GxO for pedigrees #6 and #7, respectively), or based on the merging of different datasets of the same pedigree (F2 for

pedigree #4). This process, designed to increase the number of markers common to component maps, was facilitated by the use of the same 12k (Chancerel *et al*. 2013) and 9k (Plomion *et al*. 2015) SNP-arrays for some pedigrees.

The SNP marker ID of each component map was replaced by the corresponding maritime pine unigene ID from Canales *et al*. (2014). This step, which was essential for the use of *LPmerge* (i.e. same marker name for orthologous markers), also made it possible to check the collinearity between maps. Thus, non-syntenic unigenes between different *P.pinaster* linkage maps were removed from the analysis with the exception of those validated for at least two other component maps (Table S2). Finally, *LPmerge* was used to create the composite map for *P. pinaster*. Given that similar numbers of genotypes were used to obtain the base maps and the high degree of synteny between base maps, each component map was assigned the same weight in *LPmerge*.

Pinaceae

The SNP marker ID of each species component map was replaced by the corresponding homologous unigene ID of *P. pinaster* (Canales *et al.* 2014). We established composite maps at genus level before integrating the four genetic maps for each species into a single composite map for *Pinaceae*. This process was designed to increase the number of markers common to the component maps for each genus (*Pinus* and *Picea*). *LPmerge* was used to build these two composite maps, following the same procedure as described for *P. pinaster*. We discarded non-syntenic unigenes, except for non-syntenic unigenes validated by at least two component maps in the *P. pinaster* composite map, from the construction of composite linkage maps. Non-collinear unigenes with inversions of more than 15 cM were also excluded from the construction of the composite linkage map. A large inversion of a group of markers was detected in LG7 of *Picea abies* (Lind *et al*. 2014), when the map for this species was

compared with that for *Picea glauca* (Pavy *et al*. 2008). *Picea abies* LG7 (renamed LG2 after comparison with the *P. pinaster* reference map) was reconstructed from genotyping data provided as supplementary material by Lind *et al*. (2014), using the same parameters described in Lind's article and the same mapping software (JoinMap v4.1). Two markers with a $-Log_{10}(p) > 1$ that produced a large number of double recombinants were excluded from this LG map. We were thus able to map 16 additional markers, and a much higher degree of synteny and collinearity was found between the homologous LGs of *Picea abies* and *Picea glauca* (Figure S3).

Comparison with *C. japonica*

The available map for *C. japonica* consisted of 77% of SNP markers (Moriguchi *et al*. 2012). Sequences of unigenes containing the mapped SNPs were retrieved from the Unigene set developed by Ueno *et al*. (2012). Then, unigene sequences from Ueno *et al*. (2012) mapped in *C. japonica* linkage map (Moriguchi *et al*. 2012) were compared to those of *P. pinaster* (Canales *et al*. 2014) mapped in the *Pinaceae* composite map using *tblastx* (BLAST 3, Figure 1a). Different e-value cutoff for homologous unigenes identification between *Pinaceae* and *C. japonica* were tested: lower than $1\ e^{-30}$ and $1\ e^{-35}$.

Selected homologs from the *Pinaceae* and *C. japonica* linkage maps were used for comparative mapping. We established orthologous blocks within linkage groups where several homologous unigenes were shared between both families. Different thresholds were also tested to consider an orthologous block within a linkage group: blocks with at least four and six shared unigenes. Each pair of orthologous chromosomal blocks determined a contiguous ancestral region (CAR) between the two families. The most parsimonious evolutionary model between *Pinaceae* and *Cupressaceae* considering the existence of the

identified CARs was proposed. Circular genetic maps used in inter-family comparative mapping were drawn with Circos software (Krzywinski *et al.* 2009).

Supplementary Material

Supplementary Figures S1-S5 and Tables S1-S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the European Union Seventh Framework Programme: Procogen [grant number 289841]; postdoctoral fellowship from Procogen [MdM]; "Conseil Général des Landes" postdoctoral fellowship [JB] and INRA [EFPA division and ACCAF metaprogram] and Région Aquitaine [grant number 20111203004] PhD fellowship [HL].

Data accessibility

All the linkage maps described here are available from the Pinus portal (a European genetic and genomic resource for *Pinus*) via the PinusMap application (https://w3.pierroton.inra.fr/PinusPortal/index.php). Accession numbers for marker sequences used in this study are available in supporting information Table S3.

Authors' contributions

JB: constructed the composite map for *Pinus pinaster* and wrote the script for the construction and representation of composite maps. MdM: constructed the *Pinaceae* composite map. MdM, FM, JS, J-MG: developed the model of karyotype evolution; FE: performed the bioinformatic analyses; YM, KU, SU, YT: provided the sequence information for *Cryptomeria japonica;* NdM, J-AC, MTC: provided unpublished SNP segregation data for two pedigrees of *P. pinaster*; HL: provided unpublished SNP segregation data for three pedigrees of *P. pinaster*; CP: designed the study and coordinated the project; MdM, JB, CP: wrote the paper. All the authors have read and approved the submitted manuscript.(

Literature cited

Abrouk M *et al.* 2010. Palaeogenomics of plants: synteny-based modelling of extinct ancestors. Trends Plant Sci. 15:479–87.

Amborella Genome Project. 2013. The Amborella genome and the evolution of flowering plants. Science. 342:1241089.

Birol I *et al.* 2013. Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics. 29:1492–7.

Bowe L, Coat G, DePamphilis C. 2000. Phylogeny of seed plants based on all three genomic compartments: Extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. Natl. Acad. Sci. 97:4092–4097.

Burleigh JG, Barbazuk WB, Davis JM, Morse AM, Soltis PS. 2012. Exploring Diversification and Genome Size Evolution in Extant Gymnosperms through Phylogenetic Synthesis. J. Bot. 2012:1–6

Canales J *et al.* 2014. De novo assembly of maritime pine transcriptome: implications for forest breeding and biotechnology. Plant Biotechnol. J. 12:286–99.

Chagné D *et al.* 2003. Comparative genome and QTL mapping between maritime and loblolly pines. Mol. Breed. 12:185–195.

Chancerel E *et al.* 2011. Development and implementation of a highly-multiplexed SNP array for genetic mapping in maritime pine and comparative mapping with loblolly pine. BMC Genomics. 12:368.

Chancerel E *et al.* 2013. High-density linkage mapping in a pine tree reveals a genomic region associated with inbreeding depression and provides clues to the extent and distribution of meiotic recombination. BMC Biol. 11:50.

Cong Q, Borek D, Otwinowski Z, Grishin N V. 2015. Tiger Swallowtail Genome Reveals Mechanisms for Speciation and Caterpillar Chemical Defense. Cell Rep. 10:910–919. Diaz-Sala C *et al.* 2013. The uniqueness of conifers. In: From Plant Genomics to Plant Biotechnology. Poltronieri, P, Burbulis, N, & Fogher, C, editors. Woodhead Publishing: Cambridge pp. 67–96.

Eckert AJ *et al.* 2010. Patterns of population structure and environmental associations to aridity across the range of loblolly pine (Pinus taeda L., Pinaceae). Genetics. 185:969–82. doi: 10.1534/genetics.110.115543.

Endelman JB, Plomion C. 2014. LPmerge: an R package for merging genetic maps by linear programming. Bioinformatics. 30:1623–4.

Gernandt D., Willyard A, Syring J., Liston A. 2011. The Conifers (Pinophyta). In: Genetics, Genomics and Breeding of Conifers. Plomion, C, Bousquet, J, & Kole, C, editors. Science Publishers: St. Helier, Jersey, British Channel Islands pp. 1–39.

Grattapaglia D, Sederoff R. 1994. Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: Mapping strategy and RAPD markers. Genetics. 137:1121–1137.

Gugerli F *et al.* 2001. The evolutionary split of Pinaceae from other conifers: evidence from an intron loss and a multigene phylogeny. Mol. Phylogenet. Evol. 21:167–175.

Hudson CJ *et al.* 2012. A reference linkage map for Eucalyptus. BMC Genomics. 13:240.

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

Kosambi DD. 1943. The estimation of map distances from recombination values. Ann. Eugen. 12:172–175.

Kovach A *et al.* 2010. The Pinus taeda genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics. 11:420.

Krutovsky K V., Elsik CG, Matvienko M, Kozik A, Neale DB. 2006. Conserved ortholog sets in forest trees. Tree Genet. Genomes. 3:61–70.

Krutovsky K V., Troggio M, Brown GR, Jermstad KD, Neale DB. 2004. Comparative mapping in the Pinaceae. Genetics. 168:447–461.

Krzywinski M *et al.* 2009. Circos: An information aesthetic for comparative genomics. Genome Res. 19:1639–1645.

Lagraulet H. 2015. Plasticité phénotypique et architecture génétique de la croissance et de la densité du bois du pin maritime (Pinus pinaster Ait.). Université de Bordeaux, France.

Lind M *et al.* 2014. A Picea abies linkage map based on SNP markers identifies QTLs for four aspects of resistance to Heterobasidion parviporum infection. PLoS One. 9:e101049.

Lorenz WW *et al.* 2012. Conifer DBMagic: a database housing multiple de novo transcriptome assemblies for 12 diverse conifer species. Tree Genet. Genomes. 8:1477–1485.

Lu Y, Ran J, Guo D, Yang Z, Wang X. 2014. Phylogeny and Divergence Times of Gymnosperms Inferred from Single-Copy Nuclear Genes. PLoS One. 9:e107679.

De Miguel M *et al.* 2012. Annotated genetic linkage maps of Pinus pinaster Ait. from a Central Spain population using microsatellite and gene based markers. BMC Genomics. 13:527.

De Miguel M *et al.* 2014. Genetic control of functional traits related to photosynthesis and water use efficiency in Pinus pinaster Ait. drought response: integration of genome annotation, allele association and QTL detection for candidate gene identification. BMC Genomics. 15:464.

Mackay J *et al.* 2012. Towards decoding the conifer giga-genome. Plant Mol. Biol. 80:555–69.

Moriguchi Y *et al.* 2012. The construction of a high-density linkage map for identifying SNP markers that are tightly linked to a nuclear-recessive major gene for male sterility in Cryptomeria japonica D. Don. BMC Genomics. 13:95.

Morse AM *et al.* 2009. Evolution of genome size and complexity in Pinus. PLoS One. 4:1–11.

Murat F *et al.* 2010. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. Genome Res. 20:1545–1557.

Murat F *et al.* 2015. Karyotype and Gene Order Evolution from Reconstructed Extinct Ancestors Highlights Contrasts in Genome Plasticity of Modern Rosid Crops. Genome Biol. Evol. 7:735–749.

Neale DB *et al.* 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 15:R59.

Nkongolo KK, Mehes-Smith M, Gustafson P. 2012. Karyotype evolution in the Pinaceae: implication with molecular phylogeny. Genome. 55:735–753.

Nystedt B *et al.* 2013. The Norway spruce genome sequence and conifer genome evolution. Nature. 497:579–84.

Van Os H, Stam P, Visser RGF, Van Eck HJ. 2005. RECORD: A novel method for ordering loci on a genetic linkage map. Theor. Appl. Genet. 112:30–40.

Pavy N *et al.* 2012. A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. BMC Biol. 10:84.

Pavy N *et al.* 2008. Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. BMC Genomics. 9:21.

Pelgas B *et al.* 2006. Comparative genome mapping among Picea glauca, P. mariana x P. rubens and P. abies, and correspondence with other Pinaceae. Theor. Appl. Genet. 113:1371–93.

Plomion, C *et al.* 2015 High-density SNP assay development for genetic analysis in maritime pine (*Pinus pinaster*). Mol Ecol Res *(in press)*.

Renny-Byfield S, Wendel JF. 2014. Doubling down on genomes: Polyploidy and crop plants. Am. J. Bot. 101:1–15.

R Core Team. 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria: URL http://www.R-project.org/.

Rigault P *et al.* 2011. A white spruce gene catalog for conifer genome analyses. Plant Physiol. 157:14–28.

Salse J. 2012. In silico archeogenomics unveils modern plant genome organisation, regulation and evolution. Curr. Opin. Plant Biol. 15:122–30.

Salse J *et al.* 2015. Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. Proc. Natl. Acad. Sci. U. S. A. 106:14908–14913.

Ueno S *et al.* 2012. A second generation framework for the analysis of microsatellites in expressed sequence tags and the development of EST-SSR markers for a conifer, Cryptomeria japonica. BMC Genomics. 13:136.

Van Ooijen JW 2011. JoinMap 4.1, Software for the Calculation of Genetic Linkage Maps in Experimental Populations. Wageningen, Netherlands: Kyazma BV.

Wang X-Q, Ran J-H. 2014. Evolution and biogeography of gymnosperms. Mol. Phylogenet. Evol. 75:24–40.

Warren RL *et al*. 2015. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. Plant J. 83:189-212

Yang Z-Y, Ran J-H, Wang X-Q. 2012. Three genome-based phylogeny of Cupressaceae s.l.: Further evidence for the evolution of gymnosperms and Southern Hemisphere biogeography. Mol. Phylogenet. Evol. 64:452–470.

Zimin A *et al.* 2014. Sequencing and assembly of the 22-gb loblolly pine genome. Genetics. 196:875–90.

Tables

Table 1 Characteristics of the genetic linkage maps used in this study: i) to establish composite maps for *P. pinaster* (from 14 individual maps), and for the *Pinaceae* family (combining *P. pinaster*, *P. taeda*, *P. glauca*, *P. mariana* and *P abies* linkage maps), and ii) to compare the *Pinaceae* composite map with the map of one representative (*C. japonica*) of the *Cupressaceae* family.

Species	Pedigree name	Linkage map ID	Number of individuals	Number of markers	Length (cM)	Mean marker interval (cM)	Reference
Pinus pinaster	C×L	C	106	574	1,488	2.8	Lagraulet 2015
		L		826	1,863	2.3	
	M×L	M	117	627	1,658	2.8	
		L		920	1,953	2.2	
	C×M	C	94	728	1,886	2.6	
		M		630	1,619	2.6	
	F2	F2_O	69	1,481	1,688	0.98	Plomion *et al.*, 2015
		F2_N	92	2,052	1,993	1.17	
	G2	G2M	83	619	1,425	2.3	Chancerel *et al.* 2013
		G2F		562	1,445	2.57	
	C14×C15	C14	63	812	1,714	2.1	extended from de Miguel *et al.* 2012
		C15		923	1,577	1.71	
	Gal1056xOria6	Gal1056	69	666	1,426	2.14	modified from de Miguel *et al.* 2014
		Oria6		755	1,296	1.72	
Pinus taeda	qtl	Ptaeda	172	1,815	1,899	1.1	Eckert *et al.* 2010
Picea glauca	D, P	Pglauca	500, 260	2,270	2,083	1.1	Pavy *et al.* 2012
Picea mariana	9920002		283				
Picea abies	S21K7622162 x S21K7621678	Pabies	247	686	1,889	2.8	Lind *et al.* 2014
Cryptomeria japonica	YI	Cjaponica	150	1,262	1,405	1.1	Moriguchi *et al.* 2012

Table 2 Details of the composite genetic linkage maps generated in this study.

	P. pinaster	***Pinus***	***Picea***	***Pinaceae***
Nb of LGs	12	12	12	12
Size (cM)	1,721.7	1,943	2,032.9	2,094.9
Nb of markers	3,673	5,195	2,325	6,912
Nb of markers corresponding to PineV3[a] unigenes	3,491	4,639	1,940	5,971
Nb of unique unigenes	3,457	4,605	1,931	5,927
Nb of unique positions	1,819	2,336	2,006	3,077
Mean marker interval (cM)	0.47	0.39	0.88	0.30
Mean unique position interval (cM)	0.96	0.93	1.02	0.68

[a]from Canales *et al*. (2014).

Table 3 Number of unigenes and size (cM) of identified orthologous LG blocks (CARs) at two e-value cut-offs for homologous unigenes identification. Changes in the number of unigenes or size of CARs following the most stringent threshold are indicated in bold.

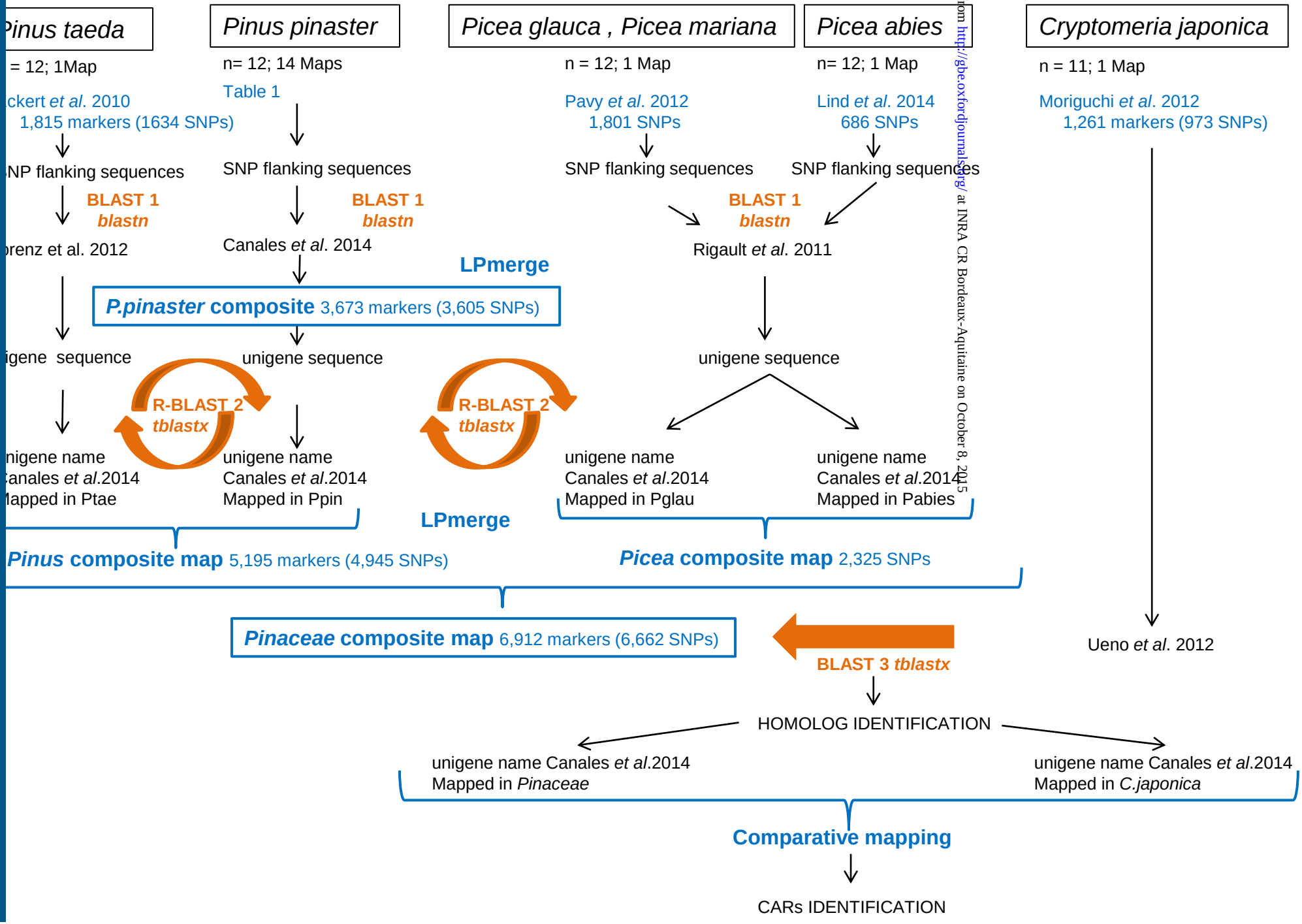
	e-30		e-35	
CAR	Nb unigenes	cM (*Pinaceae*)	Nb unigenes	cM (*Pinaceae*)
1	9	10.9-155.8	9	10.9-155.8
2	10	11.9-84.9	**9**	11.9-84.9
3	12	0-146.5	**11**	0-146.5
4	7	65.3-147.5	7	65.3-147.5
5	5	27.6-69	**4**	27.6-69
6	5	13.5-123	5	13.5-123
7	6	12.2-70.8	6	12.2-70.8
8	7	37.4-157.9	7	37.4-157.9
9	8	42.9-83.7	**7**	42.9-83.7
10	6	77.2-146.4	**4**	**125.3-146.4**
11	4	27.7-64.7	4	27.7-64.7
12	4	116.6-161.8	4	116.6-161.8
13	6	30.9-69.9	**5**	30.9-69.9
14	5	25.9-74	**0**	/
15	8	59.8-132	8	59.8-132
16	10	5-163.7	**9**	5-163.7
17	6	6.7-41.2	6	6.7-41.2
18	9	41.9-124.7	**8**	**62-124.7**
19	4	39.5-62.6	**0**	/
20	12	14.9-163.3	**11**	14.9-163.3

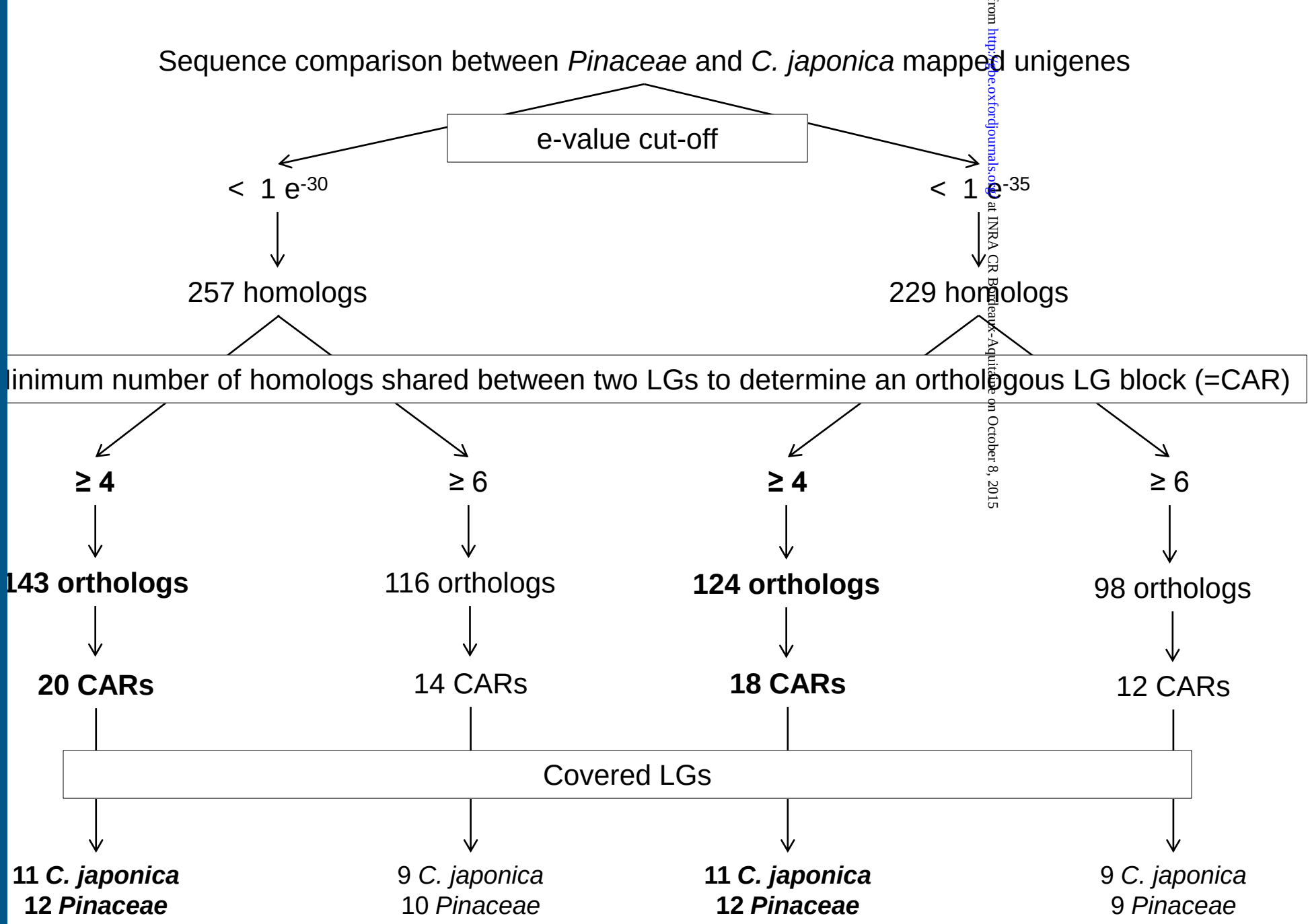
Figures

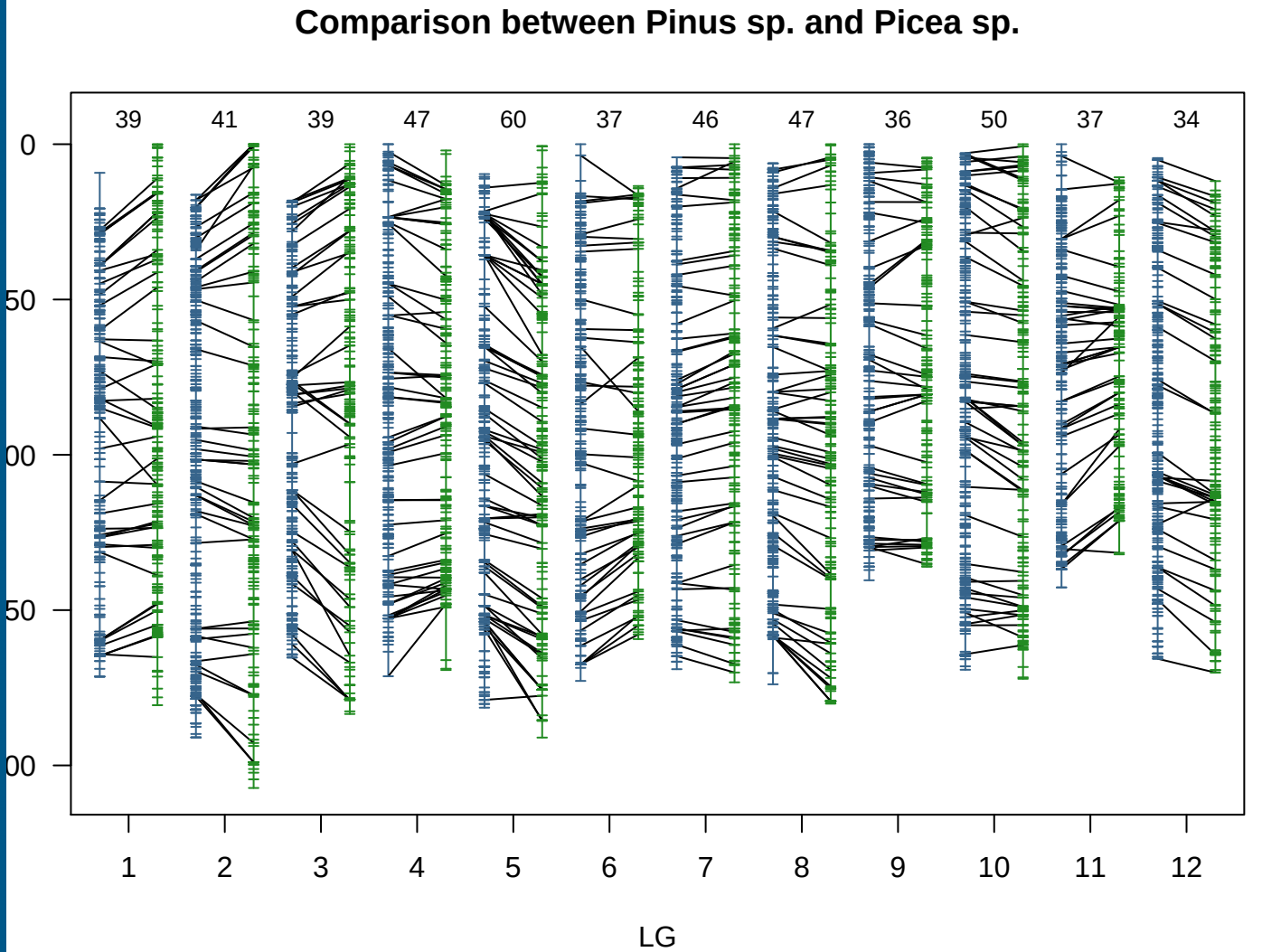
Fig. 1. Flowchart of the comparative analysis between *Pinaceae* and *C. japonica* a) Bioinformatic analysis developed for homologous genes identification. **b)** Results of the comparative analysis between *Pinaceae* and *C. japonica* testing different confidence thresholds applied at two different steps: sequence comparison for homolog unigne identification and comparative gene position for orthologs identification. Pathways that allowed the identification of orthologs covering the full set of chromosomes from *C. japonica* and *Pinaceae* are marked in bold.

Fig. 2. Comparison between the composite linkage maps for *Pinus* sp. and *Picea* sp. The *Pinus* sp. composite map is shown in blue and the *Picea* sp. composite map is shown in green. Orthologous markers are linked by black lines connecting the two maps. The number of orthologous markers is indicated at the top of each linkage group.

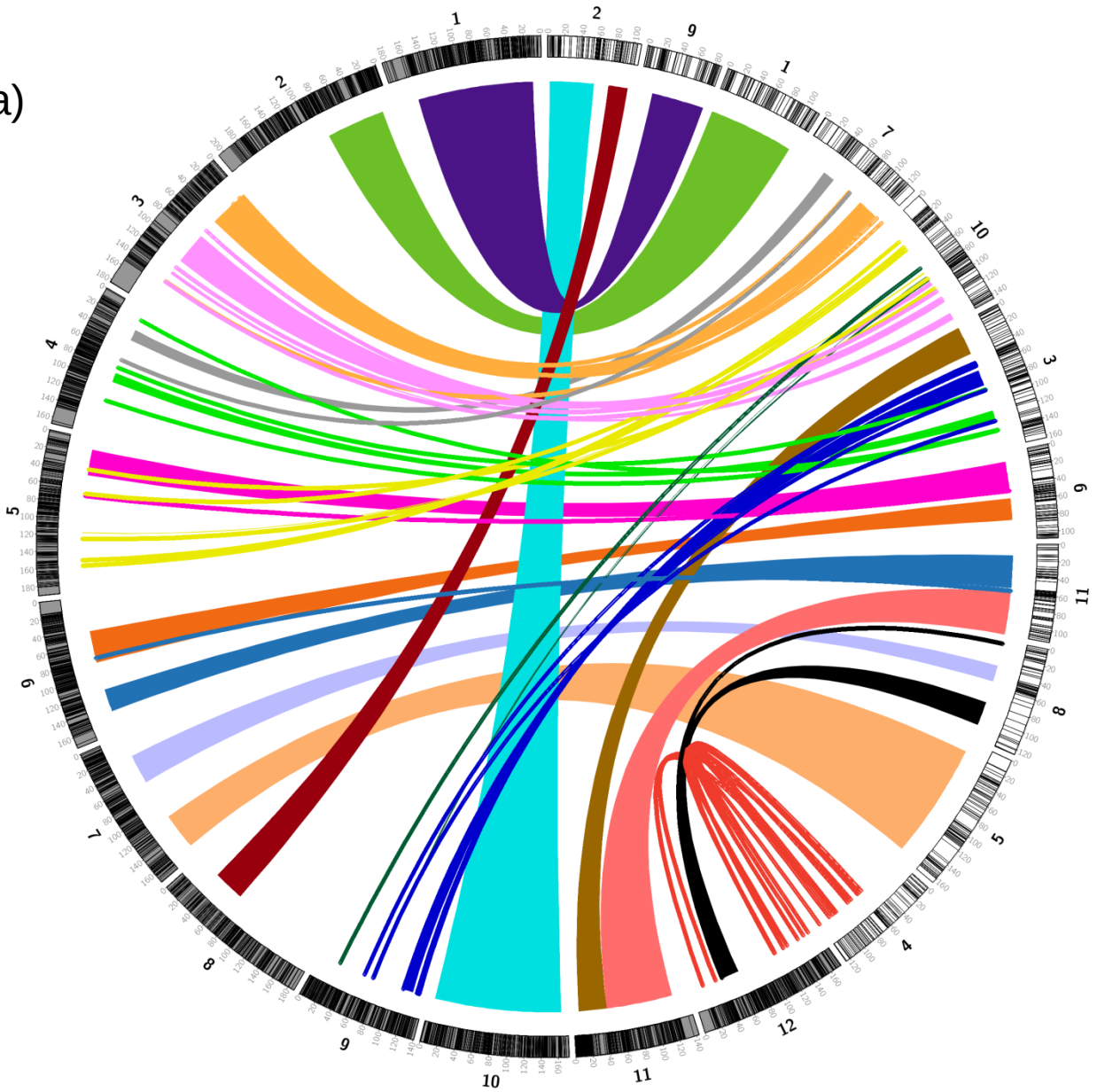
Fig. 3. *Pinaceae* – *Cupressaceae* comparative mapping. Results for an e-value cut-off of e^{-30} for homolog unigene identification and a threshold of at least four homologs shared between the two maps to determine an orthologous LG block. **a**) Positions of the 143 orthologous unigenes mapped for representative species of both *Pinaceae* and *Cupressaceae*. Orthologous LG blocks are indicated by color-coded ribbons connecting the *Pinaceae* (in gray) and *Cupressaceae* (in white) linkage groups (LG). LG number and genetic distance in cM are indicated outside the circle. *Pinaceae* LGs are ordered from 1 to 12 and *C. japonica* LGs are ordered to facilitate graphical visualization. **b**) Representation of the more parsimonious model of evolution of the identified orthologous LG blocks mapped on *Pinaceae* and *C. japonica*. Each orthologous LG block determine a contiguous ancestral region (CAR). Identified CARs are numbered from 1 to 20.



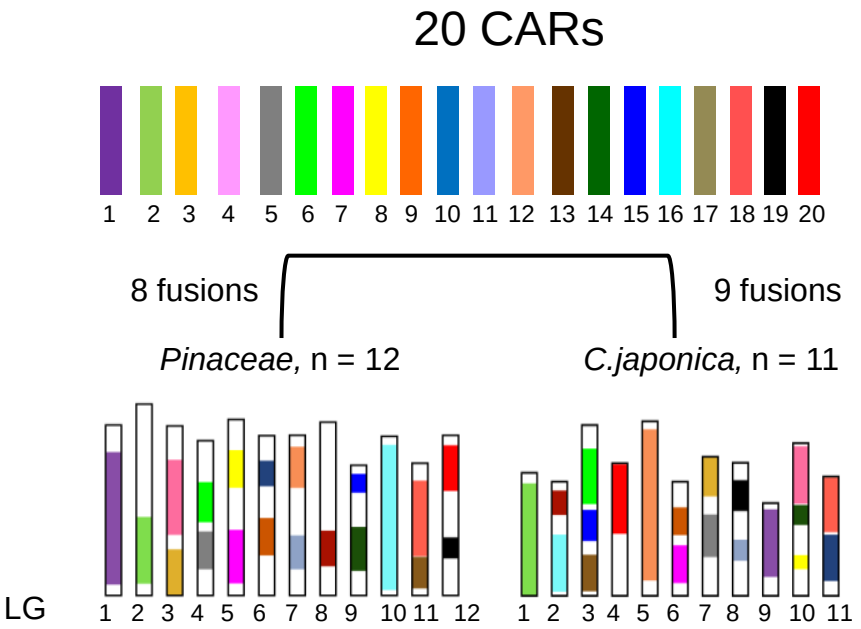



Comparison between Pinus sp. and Picea sp.
39
41
39
47
60
37
46
47
36
50
37
34
0
50
00
50
00
1
2
3
4
5
6
7
8
9
10
11
12
LG

a)



b)


20 CARs
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
8 fusions
9 fusions
Pinaceae, n = 12
C.japonica, n = 11
LG
1 2 3 4 5 6 7 8 9 10 11 12
1 2 3 4 5 6 7 8 9 10 11