# Tree Genetics & Genomes

## Comparative genome-wide analysis of repetitive DNA in the genus Populus L.

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | TGGE-D-17-00138R1 |
| **Full Title:** | Comparative genome-wide analysis of repetitive DNA in the genus Populus L. |
| **Article Type:** | Original Article |
| **Funding Information:** | Università di Pisa<br>(Plantomics)<br>Prof. Andrea Cavallini |
| **Abstract:** | Genome skimming was performed, using Illumina sequence reads, in order to obtain a detailed comparative picture of the repetitive component of the genome of Populus species. Read sets of seven Populus and two Salix species (as outgroups) were subjected to clustering using RepeatExplorer (Novak et al. 2010). The repetitive portion of the genome ranged from 33.8 in P. nigra to 46.5% in P. tremuloides. The large majority of repetitive sequences were long terminal repeat-retrotransposons. Gypsy elements were over-represented compared to Copia ones, with a mean ratio Gypsy to Copia of 6.7 : 1. Satellite DNAs showed a mean genome proportion of 2.2%. DNA transposons and ribosomal DNA showed genome proportions of 1.8 and 1.9%, respectively. The other repeats types accounted for less of 1% each. Long terminal repeat-retrotransposons were further characterized, identifying the lineage to which they belong and studying the proliferation times of each lineage in the different species. The most abundant lineage was Athila, which showed large differences among species. Concerning Copia lineages, similar transpositional profiles were observed among all the analyzed species; by contrast, differences in transpositional peaks of Gypsy lineages were found. The genome proportions of repeats were compared in the seven species and a phylogenetic tree was built, showing species separation according to the botanical section to which the species belongs, although significant differences could be found within sections, possibly related to the different geographical origin of the species. Overall, the data indicate that the repetitive component of the genome in the poplar genus is still rapidly evolving. |
| **Corresponding Author:** | Andrea Cavallini<br>Universita degli Studi di Pisa Dipartimento di Scienze Agrarie Alimentari e Agro-ambientali<br>Pisa, ITALY |
| **Corresponding Author Secondary Information:** | |
| **Corresponding Author's Institution:** | Universita degli Studi di Pisa Dipartimento di Scienze Agrarie Alimentari e Agro-ambientali |
| **Corresponding Author's Secondary Institution:** | |
| **First Author:** | Gabriele Usai |
| **First Author Secondary Information:** | |
| **Order of Authors:** | Gabriele Usai |
| | Flavia Mascagni |
| | Lucia Natali |
| | Tommaso Giordani |
| | Andrea Cavallini |
| **Order of Authors Secondary Information:** | |
| **Author Comments:** | Dear Sirs,<br>with this letter I am sending you the revised version of the article by Gabriele Usai, Flavia Mascagni, Lucia Natali, Tommaso Giordani, and myself, titled "Comparative genome-wide analysis of repetitive DNA in the genus Populus L.", to be submitted for |

| | |
|---|---|
| | publication in Tree Genetics and Genomes.<br>All reviewer's requests were addressed.<br>The submitted manuscript represents original work that is not being considered for publication, in whole or in part, in another journal, book, conference proceedings, or government publication with a substantial circulation; all previously published work cited in the manuscript has been fully acknowledged; the manuscript is one of a kind; all of the authors have contributed substantially to the manuscript and approved the final submission; no real or perceived conflicts of interest occur. |

# Comparative genome-wide analysis of repetitive DNA in the genus *Populus* L.

**Gabriele Usai*, Flavia Mascagni*, Lucia Natali, Tommaso Giordani, Andrea Cavallini**

Correspondence: Andrea Cavallini (andrea.cavallini@unipi.it)

Department of Agriculture, Food and Environment, University of Pisa, Via del Borghetto 80, I-56124 Pisa, Italy

* These authors contributed equally to this work

**Abstract** Genome skimming was performed, using Illumina sequence reads, in order to obtain a detailed comparative picture of the repetitive component of the genome of *Populus* species. Read sets of seven *Populus* and two *Salix* species (as outgroups) were subjected to clustering using RepeatExplorer (Novak et al. 2010). The repetitive portion of the genome ranged from 33.8 in *P. nigra* to 46.5% in *P. tremuloides*. The large majority of repetitive sequences were long terminal repeat-retrotransposons. *Gypsy* elements were over-represented compared to *Copia* ones, with a mean ratio *Gypsy* to *Copia* of 6.7 : 1. Satellite DNAs showed a mean genome proportion of 2.2%. DNA transposons and ribosomal DNA showed genome proportions of 1.8 and 1.9%, respectively. The other repeats types accounted for less of 1% each. Long terminal repeat-retrotransposons were further characterized, identifying the lineage to which they belong and studying the proliferation times of each lineage in the different species. The most abundant lineage was *Athila*, which showed large differences among species. Concerning *Copia* lineages, similar transpositional profiles were observed among all the analyzed species; by contrast, differences in transpositional peaks of *Gypsy* lineages were found. The genome proportions of repeats were compared in the seven species and a phylogenetic tree was built, showing species separation according to the botanical section to which the species belongs, although significant differences could be found within sections, possibly related to the different geographical origin of the species. Overall, the data indicate that the repetitive component of the genome in the poplar genus is still rapidly evolving.

**Keywords** *Populus*; LTR-retrotransposons; Repetitive DNA; Comparative retrotransposon dynamics

## Introduction

A large portion of eukaryotic genomes is made of repetitive DNA, which includes several types of sequences that can be transcribed (like transposable elements, TEs) or not (like tandem repeats). TEs are DNA sequences that are present in the nuclear genomes of all eukaryotes (Wicker et al. 2007) with the potential to move across the genome. Depending on the transposition mechanisms used, TEs can be collected into two different classes: class I (retrotransposons or retroelements, REs) and class II elements (DNA transposons). In eukaryotes, the main fraction of repeats is composed by class I elements that move through an RNA intermediate using a so-called "copy and paste" transposition mode which can lead to an increase in genome size.

These elements are variable in size: they can range from a few hundred base pairs to over 10 kb, and are composed of a coding portion flanked by two direct long terminal repeats (LTRs). Downstream of the 5′-LTR one primer-binding site is present, while one polypurine tract is located upstream of the 3′ LTR. The coding region includes ORFs necessary for the replication and the integration of the elements in the host chromosomes (Kumar and Bennetzen 1999) like Gag, a structural protein of the virus-like particles, and Pol. Pol encodes a polyprotein with protease, reverse transcriptase (RT), RNaseH, and integrase enzyme domains; in some LTR-REs an additional region, the chromodomain, is found upstream the 3'LTR. Although encoded enzymes are necessary for the transposition, non-autonomous LTR-REs can hijack enzymes produced by other LTR-REs to replicate and/or move (Wicker et al. 2007).

Plant LTR-REs can be grouped into two prominent superfamilies, *Copia* and *Gypsy* (Wicker et al. 2007), which differ for the integrase domain position in the polyprotein (Kumar and Bennetzen 1999). Superfamilies can be also divided into a number of major evolutionary lineages (Wicker and Keller 2007, Llorens et al. 2011).

The main *Gypsy* lineages are: *Athila* (Wright and Voytas 2002), *Chromovirus* (Gorinsek et al. 2004; Llorens et al. 2011) and *Ogre/TAT* (Neumann et al. 2003). On the other hand, for *Copia* superfamily the most represented lineages are *AleI/Retrofit/Hopscotch*, *AleII*, *Angela*, *Bianca*, *Ivana/Oryco*, *Maximus/SIRE* and *TAR/Tork* (Wicker and Keller 2007).

Class I also encompasses non-LTR-REs (Wicker et al. 2007), which can have protein domains similar to those of LTR-REs (Long-INterspersed-Elements, LINEs) or not (Short-INterspersed-

Elements, SINEs) but are not flanked by LTRs. In plants, they are quite rare (see for example Barghini et al. 2017).

Class II is made by DNA transposons, which use a DNA based enzymatic mode of transposition ("cut and paste") that can be encoded by the same element (in the case of autonomous elements) or by other elements (in the case of non-autonomous TEs) (Wicker et al. 2007). Generally, this method of transposition does not imply an increase in genome size.

The other large group of repetitive sequences are tandem repeats, commonly named satellite DNAs (Schmidt and Heslop-Harrison 1998). These sequences are arranged in tandem repeated units, where single copies are placed next to each other. Preferentially, they are located in specific positions of the chromosomes, like the telomeric, subtelomeric, pericentromeric, or intercalary regions (Kubis et al. 1998). Families of tandem repeats can be grouped according to the length of the individual unit and size of the array, and can have different redundancy, homology, and distribution pattern between related species of a plant genus (Wang et al. 1995). For example, in *Secale cereale*, satellites represent more than 6% of the genome (Bedbrook et al. 1980), in *Olea europaea* they account for around 30% of the genome (Barghini et al. 2014).

Although repetitive DNA has long been considered "selfish", seemingly not to provide adaptive benefit to the host genome, nowadays the considerable contribution of TEs dynamics to the evolution of genomes is ascertained. For instance, LTR-REs amplification and/or deletion are key mechanisms underlying the remarkable size variation of plant genomes (Hawkins et al. 2006; Piegu et al. 2006).

Furthermore, in recent years, DNA repeats has been shown to have a primary role in different genome functions. For instance, TEs are involved in genome restructuration (Kazazian 2000), in the modification of hosts regulatory network and gene expression, and they can generate novel genes through gene fragment rearrangement (Morgante et al. 2005). Moreover, repetitive DNAs contributes to pericentromeric and intercalary heterochromatin, supplying chromatin boundary signals for heterochromatin domains, and playing a central role in higher-order physical structuring within the nucleus (Von-Sternberg and Shapiro 2005).

Gene activity appears to be strongly affected by TEs, since TEs transposition can alter the regulatory patterns of conserved coding regions resulting in the emergence of novel traits using the same repertoire of proteins and RNAs. Genetic studies conducted on different organisms (e.g. mouse and Arabidopsis) have proved the REs effect at the epigenetic level, in that they can regulate chromatin organisation and gene expression, possibly leading to phenotypic variations (Slotkin and Martienssen 2007).

Therefore, to identify and characterize the repetitive DNA is a fundamental step toward the biological and evolutionary characterization of a species. However, up to now, the study of repetitive DNA effects to genome structure and function has been restricted to organisms with a completely sequenced genome. Studies on comparative genomics of repetitive sequences within a family or a genus are still limited, especially in dicotyledons.

Although the most reliable analysis of repeated sequences can be performed when the genome (Natali et al. 2015), or at least Mbp-long sequences are available (Buti et al. 2011; Barghini et al. 2015a), next-generation sequencing (NGS) technology can be conveniently applied to identify repeats. In fact, even in species in which the genome has not been fully sequenced yet, if a repetitive sequence is present in many copies, its identification is possible thanks to the production of large numbers of random short sequences and to their assembly. In fact, the sequencing of genomic DNA at low-coverage (named “genome skimming”; Straub et al. 2012), and the subsequent clustering of sequence reads, can characterize thousands of well-represented repeats (Macas et al. 2007; Novak et al. 2010; Staton et al. 2012; Natali et al. 2013; Barghini et al. 2014, 2015b). This approach yields also detailed insights into genome evolution schemes (Leitch and Leitch 2012; Renny-Byfield et al. 2013). Poplars (i.e. species of the genus *Populus)* are among the most economically important groups of forest trees, widely exploited for the woods, besides being model organisms for biological study on trees (Stettler et al. 1996). This genus, in fact, show several peculiar quality: easy vegetative propagation, fast growth rates, and adaptability to different ecological sites (Stettler et al. 1996). These trees are largely distributed in the northern hemisphere from boreal to subtropical forests and have an important ecological role as pioneer species in riparian and boreal forests (Braatne et al. 1992).

Poplars are diploid species (2n = 38) whose genome has been estimated around 550 million pairs of bases (Tuskan et al. 2006). The exact number of species afferent to the *Populus* genus is unclear (between 22 and 85), due to some authors declaring several hybrids as their own species (Eckenwalder 1996). For instance, 29 species were grouped in six sections (Populus, Tacamahaca, Turanga, Abaso, Aigeiros, Leucoides) by Eckenwalder (1996). Currently, breeding programmes strongly rely on plant materials from the sections Populus, Aigeiros and Tacamahaca.

The afference of some taxa within such sections remains debated. For example, *P. nigra*, assigned to the Aigeiros section, shows genetic affinity to representatives of Tacamahaca.

Moreover, while the RFLP analysis of cpDNA highlighted similarity with the Populus section, RFLP patterns of nuclear rDNA suggested a possible hybrid origin of *P. nigra* (Smith and Sytsma 1990).

Phylogenies of the Salicaceae family basing on nuclear rDNA (Leskinen and Alstrom-Rapaport 1999) and chloroplast DNA sequences (Azuma et al. 2000) suggest that *Populus* is a monophyletic sister clade with *Salix*.

An outstanding feature of poplars is the occurrence of interspecific hybrids (Eckenwalder 1982; 1996). For instance, hybrids are frequently present in the contact zones of *P. angustifolia*, *P. trichocarpa*, and *P. balsamifera*, that is where species belonging to Aigeiros and Tacamahaca sections are sympatric (Brayshaw 1965). Likewise, hybridization between species of the Populus section can naturally occur (Stettler et al. 1996). A phylogeny of the *Populus* genus was reconstructed basing on nucleotide sequences of three noncoding regions of cpDNA (intergenic regions of trnT-trnL and trnL-trnF and intron of trnL) and ITS1 and ITS2 of the nuclear rDNA (Hamzeh and Dayanandan 2004). Incongruences between phylogenetic trees based on chloroplast- and nuclear-DNA sequence suggest a complex evolutionary history of this genus, and it is possible that sympatric species, even belonging to different sections, could have frequent opportunities to exchange genes (Stettler et al. 1996).

The purpose of this work was a comparative characterization of the repetitive component of the genomes of seven poplar species, belonging to the three most common and cultivated sections, Populus, Tacamahaca, and Aigeiros, to have new insights on their evolutionary relationships. The phylogenetic method we used is based on a bioinformatics estimation of different DNA repeats abundance through the analysis of reads sequenced from a small proportion of the genome. In fact, abundances of repetitive elements represents informative features for phylogenetic inference.

## Materials and methods

### Sequence data collection

The study was conducted on seven poplar species, belonging to three sections: *Populus deltoides* and *P. nigra* (section Aigeiros), *P. tremula* and *P. tremuloides* (section Populus) and *P. balsamifera*, *P. simonii* and *P. trichocarpa* (section Tacamahaca). Two willow species were used as outgroups, *Salix purpurea* and *S. suchowensis*. Illumina DNA sequences of the nine species were retrieved from the NCBI Sequence Read Archive (NCBI, Washington, USA, https://www.ncbi.nlm.nih.gov/sra). The ID codes for each sequence read set are reported in Table 1.

To analyze reads of comparable quality, all sets were checked for read quality using FastQC (Andrews 2010): then Illumina adapters and low-quality regions were removed using Trimmomatic-0.33 (Bolger et al. 2014) using the reported parameters: ILLUMINACLIP:2:30:10; LEADING:3; TRAILING:3; SLIDINGWINDOW:4:15; CROP:85; and MINLEN:85.

Organellar sequences were removed from the sequence sets by mapping against a database of chloroplast genomes of *P. balsamifera* (NCBI 675155133), *P. tremula* (768801510), *P. trichocarpa* (133712039), *S. purpurea* (772657659), and *S. suchowensis* (751371584) and mitochondrial genomes of *P. tremula* (936227452), *S. purpurea* (1011056159), and *S. suchowensis* (1002167010) using CLC-BIO GenomicWorkbench (version 7.0.4 CLC-BIO, Aahrus, Denmark), with the following parameters: Mismatch cost 1; Insertion cost 1; Deletion cost 1; Length fraction 0.7; Similarity fraction 0.7. All matching reads were considered putatively belonging to organellar genomes and removed.

## Identification of repetitive DNA

In order to perform a comparative analysis of the repetitive component of 7 species of the genus *Populus* and two species of genus *Salix*, a hybrid graph-based clustering method (RepeatExplorer, Novak et al. 2010; 2013) was applied allowing *de novo* identification of repeats and their proportion in each genome.

RepeatExplorer requires small sets of reads and produce distinct clusters of frequently connected reads, which are automatically annotated accordingly to their similarity to an internal database of repeats. A random set of 500,000 paired-end sequence reads for each species was used, to ensure that the clusters obtained were comparable.

RepeatExplorer output was further annotated performing similarity searches of the unknown clusters by RepeatMasker (developed by A.F.A. Smit, R. Hubley, and P. Green; http://www.repeatmasker.org/) against a library of *P. trichocarpa* full-length retrotransposons (Natali et al. 2015), with the following parameters: -s; -no_is; and -nolow.

## Phylogenetic trees

Contigs assembled by RepeatExplorer were also subjected to the RepeatExplorer protein domain search tool, which performed searches against its own custom-made database of protein domains (i.e., chromodomain, GAG, protease, RT, RNAse H and integrase) derived from plant mobile elements, using default parameters. The RT domains were translated and aligned to RT domains of different species using Clustal Omega (McWilliam et al. 2013). Then, phylogenetic trees were built using a neighbour joining (NJ) clustering method and multi-scale bootstrap resampling with 1,000 bootstrap replications.

Another phylogenetic analysis was based on the abundance of repeats. Pairwise comparisons between species were made for each cluster by plotting the respective genome proportion estimated using RepeatExplorer. Then, a dendrogram based on the genome proportions data of each cluster was built by using R package pvclust version 1.3-2 (Suzuki et al. 2006), which allowed the assignment of the uncertainty in hierarchical cluster analysis (10,000 bootstrap replications).

**LTR-retrotransposons temporal dynamics estimation**

The time course of different LTR-retrotransposon lineages in poplar species was inferred by examining distributions of pairwise divergence values for Illumina reads aligned to the RT domain encoding sequences of the different lineages, according to Piegu et al. (2006) and Ammiraju et al. (2007).

First, Illumina reads of the 7 poplar species were separately clustered using RepeatExplorer (2,000,000 paired-end reads per species). Then, assembled nucleotide sequences encoding the RT domains (at least 150 nt in length) were selected from *Gypsy* and *Copia*-related clusters of different lineages, using the protein domain search tool of RepeatExplorer. Illumina 85 nt-long reads of each species were aligned to the respective homologous RT sequences using CLC Genomics Workbench 7.0.4 with the following parameters: Mismatch cost 1; Insertion cost 1; Deletion cost 1; Length fraction 0.9; Similarity fraction 0.9. For each species, up to 100 aligned reads were collected randomly. Then, pairwise divergence values between reads were determined using MEGA version 7.0.18 (Kumar et al. 2016) under the Kimura two parameter model of sequence evolution (Kimura 1980). Peaks in the frequency distribution were interpreted as events of transposition burst, with peaks associated with lower values of divergence representing more recent proliferation events.

## Results

### Quantitative analysis of the repetitive component of *Populus* species.

The repetitive component of the genome of seven *Populus* species (and two *Salix* species, as outgroups) was studied by hybrid clustering (using RepeatExplorer, Novák et al. 2010) of nine samples of 500,000 Illumina paired-end reads (one for each species). In the hybrid clustering analysis, reads sharing sequence similarity between species cluster together, allowing the identification of repeats shared among species, and measuring the respective genome proportion in each species.

Table 2 summarizes the partitioning of read sets after hybrid clustering. Since the relatively low number of reads used by RepeatExplorer, clustered reads should belong to repetitive sequences. The mean percentage of repeated sequences in poplar species ranges from 33.8 in *P. nigra* to 46.5% in *P. tremuloides* (Table 2).

Table 3 reports the composition of the repetitive portion of each genome in terms of repeat classes, LTR-REs (*Gypsy* and *Copia*), DNA transposons, non-LTR retrotransposons (LINEs and SINEs), Helitrons, satellite DNAs, and ribosomal DNA. The genome proportions of each repeat class or superfamily derive from the sum of genome proportions of the respective individual clusters belonging to that class or superfamily of repeated sequences, considering only clusters with a genome proportion greater than 0.01%. The automatic contig annotation provided by RepeatExplorer at the end of hybrid clustering was enriched. First, using paired-end read connections, we could identify and classify unannotated interconnected clusters. Then, all clusters were subjected to a similarity search against a complete and annotated full-length LTR-RE set of *P. trichocarpa* (Natali et al. 2015). Initially, clusters with a genome proportion higher than 0.01% were 142; after excluding non-annotated clusters, low complexity or simple repeat clusters, and residual organellar DNA clusters (escaped to the removal process), a final number of 120 clusters was achieved.

LTR-REs (*Gypsy* and *Copia*) constitute the largest fraction of the annotated repetitive component of each species. The mean ratio between genome proportions of *Gypsy* and *Copia* elements was 6.7 : 1. Among minor repeat classes, the proportions of DNA transposons range from 1.5 to 2.3%, those of ribosomal DNA from 0.9 to 2.3%. Non-LTR elements and Helitrons are barely represented (less than 1%). Satellite DNAs are the most variable minor repeat class, ranging from 0.8 in *P. simonii* to 4.2% in *P. deltoides*.

It is to be noted that genome proportions of the different classes or superfamilies are quite similar among species, with the notable exception of satellite DNAs and LTR-*Gypsy* elements. *P. deltoides* and *P. tremula* show satellite DNA abundances more than 2-fold higher than those reported for all other species. On the other hand, the genome proportion of *Gypsy* elements in *P. tremuloides* is more than 2-fold higher than that of all other species.

**Characterization of LTR-REs and other repeats and phylogenetic analyses**

Poplar LTR-REs were annotated at lineage level. All *Gypsy* lineages (*Athila*, *Chromovirus*, and *Ogre*) were identified; on the contrary, only five *Copia* lineages were found, probably because of their low number in the samples of reads used for experiments. Table 4 reports the genome proportion of each identified lineage.

Apparently, the large abundance of *Gypsy* elements in *P. tremuloides* (Table 4) is related to a disproportionate percentage of *Athila* REs (14.3% of the genome in *P. tremuloides* vs. a mean of 3.5% in all other species). Large differences in percentage of *Copia* lineages are also observed among species (Table 4), although the relatively low frequencies (all lower than 1%) of such elements do not allow speculations in our experimental conditions.

The previously reported comparative analysis (Tables 3-4) was extended at sublineage level and allowed us to infer both the evolutionary trend of each LTR-RE lineage and the relationships among poplar species. In fact, separated clusters belonging to the same lineage presumably represent different sublineages according to their sequence similarity. Through hierarchical clustering analysis of LTR-RE sublineages, that was based on the genomic proportion of each cluster in each species, we identified and quantified groups of homologous clusters sharing similar abundance levels between the species (Novak et al. 2014).

The genome proportion of 120 homologous clusters belonging to the different repeat classes, superfamilies or lineages in the 7 *Populus* and 2 *Salix* species are reported in Figure 1. Each cluster represents a repeat type identified by colour. Clusters were in turn grouped according to their abundance among the 9 samples: each group represents clusters showing a similar pattern of abundance.

Nearly all repeats are shared among all the 7 poplar species (Figure 1); only some satellite DNA-related clusters resulted specific to some species. Differences in abundance evidenced 8 groups of repeat clusters with different abundance patterns. For example, group c was mainly composed of *Gypsy-Ogre* and *Copia-TAR/Tork*-related clusters having comparable abundance patterns in the analyzed species; most ribosomal DNA clusters belong to group g. *Gypsy-Athila* elements, the most abundant in poplar species, belong to only two groups, d and f. Interestingly, *Athila* clusters of group f are largely represented only in *P. tremuloides*.

Figure 1 shows the obvious separation between *Salix* and *Populus* species. From this point of view it is apparent that only a few clusters show similar abundance between the two genera: most of the ribosomal DNA clusters and some clusters belonging to *Athila*, *Ogre/TAT* and *AleI* are shared among all the species under examination, including willows. A striking difference between the two genera is the presence, only in the two willows, of three clusters representing satellite DNA sequences, with a large genome proportion, completely absent in the poplars (see group h). Figure 1 also reflects the subdivision of poplar species into the three sections analysed in these experiments. From this point of view, it is interesting to note that large differences are observed in the abundance pattern of the clusters annotated as satellite DNA. Each satellite DNA cluster corresponds to a different tandem repeat sequence. Satellite-related clusters of group f are more abundant in the Populus section (*P. tremula* and *P. tremuloides*). By contrast, satellite DNAs of the group b are especially abundant in Aigeiros section (*P. deltoides* and *P. nigra*).

The genome proportion calculated by RepeatExplorer for each cluster depends on both sequence conservation and copy number of the sequences included in the cluster. Anyway, comparisons between genomic proportions of the same clusters in different species can be used to establish relationships between the analyzed poplars.

In Figure 2, a phylogenetic tree is reported, based on the genome proportion values of all clusters, according to the method proposed by Cavallini et al. (2010) in sunflower species. The dendrogram confirms the subdivision of the genus among the three sections. The dendrogram indicates also that Tacamahaca (consisting of *P. trichocarpa*, *P. balsamifera* and *P. simonii*) and Aigeiros (*P. deltoides* and *P. nigra*) sections are phylogenetically closer to each other than the Populus section (*P. tremula* and *P. tremuloides*). This result is in agreement with the nuclear rDNA-based consensus tree reported by Hamzeh and Dayanandan (2004), in which the section Populus can be found at the base of the tree, compared to sections Tacamahaca and Aigeiros.

## Temporal dynamics of poplar LTR-REs

Predicted protein sequences of RT domains of 7 poplars and two willows were aligned, and similarity between sequences was used to build dendrograms. Figure 3 report the dendrogram obtained using *Gypsy* RTs. The dendrogram about *Copia* RT sequences is reported in Suppl. Material #1.

Figure 3 and Suppl. Material #1 show that, for both *Gypsy* and *Copia* superfamilies, the protein sequences of the analyzed RT domains are separated according to the lineage to which the elements belong rather than by plant species. This result suggests that most variation among LTR-RE lineages occurred before the separation between the *Populus* species.

The timing of proliferation of the different LTR-RE lineages in poplar species was inferred from analyzing pairwise distances (Kimura 1980) between paralogous RT-encoding sequences belonging to elements of the same monophyletic lineages, according to the method of Piegu et al. (2006). Distances were translated into insertion dates as described by SanMiguel et al. (1996) and Piegu et al. (2006), but using a mutation rate of 4.72 x $10^{-9}$, i.e. specific to poplar and two-fold the rate calculated for synonymous substitutions in poplar gene sequences (Cossu et al. 2012), to keep into consideration that REs accumulate more mutations with time. In fact, at each insertion, the new RE copy is identical to its parental element except for mutations occurring during retrotranscription (which is error prone, Kumar and Bennetzen, 1999); then, further mutations can accumulate as time passes.

Analysis was carried out for 6 out of 8 LTR-RE lineages, i.e. those lineages for which a large number of sequences to be compared were available. This analysis enabled the identification of different retrotranspositional waves, mostly overlapping in terms of time (Fig. 4). Obviously, translation of genetic distances into insertion dates is subject to reservations, however, in our analyses we limited to compare retrotransposition waves of the same RE lineage in different species. On the other hand, a comparison between *P. trichocarpa* transpositional profiles reported in this work and transpositional profiles obtained aligning LTR pairs of full length elements (according to the method proposed by Ma and Bennetzen (2004)) showed very similar results (Cavallini, in preparation) confirming the reliability of our data.

Concerning *Copia* elements, proliferation waves are generally similar in the analyzed species (Fig. 4), with a peak of proliferation putatively dated at 10 MYA, and the exceptions of the *Ivana/Oryco* lineage in *P. balsamifera* and the *TAR/Tork* lineage in *P. simonii*, whose transposition peak can be dated at 5 MYA.

*Gypsy* RE dynamics were more complex (Fig. 4). *Athila* elements, the most abundant in the genus, show differently dated proliferation peaks at 0 (*P. tremuloides* and *P. nigra*), 5 (*P. balsamifera*), 15 (*P. deltoides*), and 10 MYA (the other species). *Chromoviruses* show a recent transposition wave in *P. balsamifera* and two proliferation peaks in *P. tremuloides* and *P. simonii*. *Ogre/TAT* proliferation profiles were more uniform among species, except *P. balsamifera*, in which these elements show a recent transposition wave.

Combined together, these results suggest an intriguing picture of species-specific increases in the abundance of REs, especially of the *Gypsy* superfamily, which in some cases should have occurred in relatively recent times, subsequent to *Populus* speciation. For example, the recent proliferation wave of *Athila* REs observed in *P. tremuloides* might be related to the accumulation of these elements in this species, in which the genome proportion of *Athila* REs is more than 3-fold that of each of the other species.

## Discussion

The repeated component of the genomes of various *Populus* species were analyzed with reference on their composition and on the existing variability among species. Actually, a detailed overall picture of the repetitive portion of the genome was obtained by using sets of low coverage Illumina sequences and applying different bioinformatics assembling, annotation and mapping methodologies. The approach of using relatively low sequencing coverage to study the repetitive component of the genome has been already used with success in other species (Natali et al. 2013; 2015; Barghini et al. 2014; 2015b; Mascagni et al. 2015), confirming the usefulness of genome skimming for this kind of sequences (Dodsworth et al. 2015).

In our analyses LTR-REs resulted the most abundant repeats in the genome of *Populus* species, as observed in all higher plants, with a mean ratio between genome proportions of *Gypsy* and *Copia* elements of 6.7 : 1. In angiosperms, different ratios between *Gypsy* and *Copia* elements abundance were observed, from 5 : 1 in papaya to 1 : 2 in grapevine (Vitte et al. 2014). Higher abundance of *Gypsy* REs compared to *Copia* ones has already been reported also in *P. trichocarpa*

(Tuskan et al. 2006) and amounted to 4.74 : 1 (Natali et al. 2015). In the present work, this ratio is 5.9 : 1, i.e. higher than that reported by Natali et al. (2015): such a discrepancy can be related to the necessarily lower number of reads used for hybrid clustering. In fact, the number of reads which can be used for hybrid clustering reduces increasing the number of species to be concurrently analysed. The use of a reduced number of reads favours the recovery of the most abundant repeats; being *Copia* families less repeated than *Gypsy*, it is possible they are under-represented in the assembled clusters.

Besides producing an accurate description of the repetitive component of the genome, our results show that, if the overall genome structure is conserved among poplars, interspecific differences occur, especially in relation to LTR-REs. Actually, large differences in abundance among LTR-RE lineages have been reported for many plant species (Du et al. 2010; Guyot et al. 2016), even at intraspecific level (see for example Mascagni et al. 2015). DNA satellites are also variable in abundance among poplar species.

Differences in the abundance of LTR-REs and DNA satellites can lead to structural variations in poplar chromosomes, as those already identified in *P. nigra*, *P. deltoides*, and *P. trichocarpa* by Pinosio et al. (2016). These authors showed the occurrence of a large number of structural variations as insertions and deletions (also at nucleotide level) and reported that variations were preferentially associated with the activity of transposable elements. Our study provides further data, which will be useful to produce and annotate a genome-wide catalogue of structural variations and to define and extend the characteristics of the poplar pan-genome.

Regarding DNA satellites, it is well known that they have often a structural role in the centromere and telomere organization (Dvořáčková et al. 2015; Lermontova et al. 2015). The different tandem repeats observed in the species of the Populus section compared to those found in species of Tacamahaca and Aigeiros sections suggests deep differences in the evolutionary processes that led to establish the present chromosomal structure in these species and will deserve future research.

As a matter of fact, interspecific differences related to repetitive DNA are much greater than the differences found in the coding gene sequences (Cossu et al. 2012; Pinosio et al. 2016). It may be speculated that such chromosomal structural differences have contributed, during evolution, to the differentiation between species, favouring their reproductive isolation. In this sense, it is interesting to note that *P. nigra* and *P. deltoides*, two species easily crossable, and which have been used to produce the most commonly cultivated interspecific hybrids (i.e. *Populus x canadensis*) are very similar in relation to the repetitive component of the genome.

Our data can also be useful in analyzing the evolution of the genus *Populus*. Poplar species are known to be widely distributed in the northern hemisphere. The species selected for analysis are typical of Northern America (*P. deltoides*, *P. balsamifera*, *P. tremuloides*, and *P. trichocarpa*), of Europe (*P. nigra*), of Eurasia (*P. tremula*) and Eastern Asia (*P. simonii*). Interestingly, no relation was found between the geographical area in which a species originated and its genome structure (i.e. the abundance and composition of the repetitive component of the genome); on the contrary, significant similarities were found between species of one and the same section, independently of the geographical area in which the species live. However, at least for some repeats, large differences in abundance were observed even within sections, suggesting that these repeats have undergone copy number variations after the separation of the species from the section progenitor, possibly related to the different environments colonized by each species. A striking example is offered by the *Athila* LTR-REs which, in *P. tremuloides* (an American Tacamahaca species) are more than three-fold than in *P. tremula* (an Eurasian species of the same section).

The overall data reported in this study evidence that *Populus* is a genus which is still rapidly evolving, especially in relation to the repetitive component of the genome, as shown also by the differences in proliferation dynamics of the various RE lineages. Actually, expression of LTR-REs has been recently reported in the interspecific hybrid *P. deltoides* x *P. nigra* (Giordani et al. 2016) and also in its parental species (Cavallini, unpublished). It may be assumed that the activity of repeated elements, besides determining genomic structural differences, also affects the coding portion of the genome, both by gene inactivation (by inserting within it) and by modifying the cis-regulatory sequences of the genes, with consequences on the plant phenotype.

**Acknowledgements** This research work was supported by Department of Agriculture, Food and Environment, University of Pisa, Italy, project Plantomics.

**Data archiving statement** All genomic DNA raw Illumina sequences used in this work are available at the NCBI Sequence Read Archive under the accession numbers SRS1328499 (*Populus deltoides*), SRS1218640 (*P. nigra*), SRS1124263 (*P. tremula*), SRS1124888 (*P. tremuloides*), SRS1115969 (*P. balsamifera*), SRS829341 (*P. simonii*), SRS844395 (*P. trichocarpa*), SRS161420 (*Salix purpurea*), and SRS1276361 (*S. suchowensis*).

## References

Ammiraju JS, Zuccolo A, Yu Y, Song X, Piegu P, Chevalier F, Walling JG, Ma J, Talag J, Brar DS, SanMiguel PJ, Jiang N, Jackson SA, Panaud O, Wing RA (2007) Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*. Plant J 52:342–351

Andrews S (2010) A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Azuma T, Kajita T, Yokoyama J, Ohashi H (2000) Phylogenetic relationships of *Salix* based on rbcL sequence data. Am J Bot 87: 67-75

Barghini E, Natali L, Cossu RM, Giordani T, Pindo M, Cattonaro F, Scalabrin S, Velasco R, Morgante M, Cavallini A (2014) The peculiar landscape of repetitive sequences in the olive (*Olea europaea* L.) genome. Genome Biol Evol 6:776-791

Barghini E, Natali L, Giordani T, Cossu RM, Scalabrin S, Cattonaro F, Šimková H, Vrána J, Doležel J, Morgante M, Cavallini A (2015a) LTR retrotransposon dynamics in the evolution of the olive (*Olea europaea*) genome. DNA Res 22:91-100

Barghini E, Mascagni F, Natali L, Giordani T, Cavallini A (2015b) Analysis of the repetitive component and retrotransposon population in the genome of a marine angiosperm, *Posidonia oceanica* (L.) Delile. Mar Genomics 24:397–404

Barghini E, Mascagni F, Natali L, Giordani T, Cavallini A (2017) Identification and characterisation of Short Interspersed Nuclear Elements in the olive tree (*Olea europaea* L.) genome. Mol Genet Genomics 292:53-61

Bedbrook JR, Jones J, O'Dell M, Thompson RD, Flavell RB (1980) A molecular description of telomeric heterochromatin in *Secale* species. Cell 19:545–560

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Braatne JH, Hinckly TM, Stettler RF (1992) Influence of soil water supply on the physiological and morphological components of plant water balance in *Populus trichocarpa*, *Populus deltoides* and their F1 hybrids. Tree Physiol 11:325-340

Brayshaw TC (1965) Native poplars of southern Alberta and their hybrids. Can Forest Serv Publication 1109, Ottawa, Ontario, Canada

Buti M, Giordani T, Cattonaro F, Cossu RM, Pistelli L, Vukich M, Morgante M, Cavallini A, Natali L (2011) Temporal dynamics in the evolution of the sunflower genome as revealed by sequencing and annotation of three large genomic regions. Theor Appl Genet 123:779-791

Cavallini A, Natali L, Zuccolo A, Giordani T, Jurman I, Ferrillo V, Vitacolonna N, Sarri V, Cattonaro F, Ceccarelli M, Cionini PG, Morgante M (2010) Analysis of transposons and repeat composition of the sunflower (*Helianthus annuus* L.) genome. Theor Appl Genet 120:491-508

Cossu RM, Buti M, Giordani T, Natali L, Cavallini A (2012) A computational study of the dynamics of LTR retrotransposons in the *Populus trichocarpa* genome. Tree Genet Genomes 8:61–75

Dodsworth S, Chase MW, Kelly LJ, Leitch IJ, Macas J, Novák P, Piednoël M, Weiss-Schneeweiss H, Leitch AR (2015) Genomic repeat abundances contain phylogenetic signal. Syst Biol 64:112–126

Dvořáčková M, Fojtová M, Fajkus J (2015) Chromatin dynamics of plant telomeres and ribosomal genes. Plant J 83:18-37

Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, Shoemaker RC, Ma J (2010) Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J 63:584–598

Eckenwalder JE (1982) *Populus xinopia* hybr. nov. (*Salicaceae*), a natural hybrid between the native North American *P. fremontii* S. Watts and the introduced Eurasian *P. nigra* L. Madrono 29:67-78

Eckenwalder JE (1996) Systematics and evolution of *Populus*. In: Stettler RF, Bradshaw HD, Heilman PE, Hinckley TM (eds) Biology of *Populus* and its implications for management and conservation, NRC Research Press, National Research Council of Canada, Ottawa, Ontario, Canada, pp 7–32

Giordani T, Cossu RM, Mascagni F, Marroni F, Morgante M, Cavallini A, Natali L (2016) Genome-wide analysis of LTR-retrotransposon expression in leaves of Populus × canadensis water-deprived plants. Tree Genet Genomes 12:75

Gorinsek B, Gubensek F, Kordis D (2004) Evolutionary genomics of chromoviruses in eukaryotes. Mol Biol Evol 21:781–798

Guyot R, Darré T, Dupeyron M, de Kochko A, Hamon S, Couturon E, Crouzillat D, Rigoreau M, Rakotomalala JJ, Raharimalala NE, Doffou Akaffou S, Hamon P (2016) Partial sequencing reveals the transposable element composition of *Coffea* genomes and provides evidence for distinct evolutionary stories. Mol Genet Genomics 291:1979-1990

Hamzeh M, Dayanandan S (2004) Phylogeny of *Populus* (Salicaceae) based on nucleotide sequences of chloroplast trnt-trnf region and nuclear rDNA. Am J Bot 91:1398-1408

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

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

Kazazian HH (2000) L1 retrotransposons shape the mammalian genome. Science 289:1152–1153

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

Kubis S, Schmidt T, Heslop-Harrison JS (1998) Repetitive DNA elements as a major component of plant genomes. Ann Bot 82:45–55

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Annu Rev Genet 33:479-532

Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870-1874

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

Lermontova I, Sandmann M, Mascher M, Schmit AC, Chabouté ME (2015) Centromeric chromatin and its dynamics in plants. Plant J 83:4–17

Leskinen E, Alstrom-Rapaport C (1999) Molecular phylogeny of *Salicaceae* and closely related *Flacourtiaceae*: evidence from 5.8S, ITS] and ITS2 of the rDNA. Plant Syst Evol 215:209-227

Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, Aguilar-Rodríguez J, Vicente-Ripolles M, Fuster G, Bernet GP, Maumus F, Munoz-Pomer A, Sempere JM, Latorre A, Moya A (2011) The *Gypsy* database (GyDB) of mobile genetic elements: release 2.0. Nucl Acids Res 39:D70–D74

Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA 101:12404‑12410

Mascagni F, Barghini E, Giordani T, Rieseberg LH, Cavallini A, Natali L (2015) Repetitive DNA and plant domestication: variation in copy number and proximity to genes of LTR-retrotransposons among wild and cultivated sunflower (*Helianthus annuus*) genotypes. Genome Biol Evol 7:3368-3382

McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, Lopez R (2013) Analysis tool web services from the EMBL-EBI. Nucl Acids Res 41:W597-W600

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

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37:997-1002

Natali L, Cossu RM, Barghini E, Giordani T, Buti M, Mascagni F, Morgante M, Gill N, Kane NC, Rieseberg L, Cavallini A (2013) The repetitive component of the sunflower genome as revealed by different procedures for assembling next generation sequencing reads. BMC Genomics 14:686

Natali L, Cossu RM, Mascagni F, Giordani T, Cavallini A (2015) A survey of *Gypsy* and *Copia* LTR-retrotransposon superfamilies and lineages and their distinct dynamics in the *Populus trichocarpa* (L.) genome. Tree Genet Genomes 11:107

Neumann P, Požárková D, Macas J (2003) Highly abundant pea LTR-retrotransposon *Ogre* is constitutively transcribed and partially spliced. Plant Mol Biol 53:399–410

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

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a galaxy based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. Bioinformatics 29:792–793

Novák P, Hřibová E, Neumann P, Koblížková A, Doležel J, Macas J (2014) Genome-wide analysis of repeat diversity across the family *Musaceae*. PLoS One 9:e98918

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K,Brar DS, Jackson S,Wing RA, Panaud O (2006) Doubling genome size without polyploidization: dynamics of retrotransposition driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res 16:1262–1269

Pinosio S, Giacomello S, Faivre-Rampant P, Taylor G, Jorge V, Le Paslier MC, Zaina G, Bastien C, Cattonaro F, Marroni F, Morgante M (2016) Characterization of the poplar pan-genome by genome-wide identification of structural variation. Mol Biol Evol 33:2706–2719

Renny-Byfield S, Kovarik A, Kelly LJ, Macas J, Novak P, Chase MW, Nichols RA, Pancholi MR, Grandbastien MA, Leitch AR (2013) Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high copy sequences. Plant J 74:829–839

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765–768

Schmidt T, Heslop-Harrison JS (1998) Genomes, genes and junk: the large scale organization of plant chromosomes. Trends Plant Sci 3:195–199

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nature Rev Genet 8:272-285

Smith RL, Sytsma KJ (1990) Evolution of *Populus nigra* (sect. Aigeiros): introgressive hybridization and the chloroplast contribution of *Populus alba* (sect. Populus). Am J Bot 77:1176–1187

Staton SE, Bakken BE, Blackman BK, Chapman MA, Kane NC, Tang S,Ungerer MC, Knapp SJ, Rieseberg LH, Burke JM (2012) The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements. Plant J 72:142–153

Stettler RF, Bradshaw HD, Heilman PE, Hinckley TM (1996) Biology of *Populus* and its implications for management and conservation. NRC Research Press, Ottawa, Ontario, Canada

Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A (2012) Navigating the tip of the genomic iceberg: Next generation sequencing for plant systematics. Am J Bot 99:349–364

Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22:1540-1542

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Déjardin A, dePamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596-1604

Vitte C, Fustier MA, Alix K, Tenaillon MI (2014) The bright side of transposons in crop evolution. Brief Funct Genom 13:276-295

Von-Sternberg R, Shapiro JA (2005) How repeated retroelements format genome function. Cytogenet Genome Res 110:108–116

Wang ZX, Kurata N, Saji S, Katayose Y, Minobe Y (1995) A chromosome 5-specific repetitive DNA-sequence in rice (*Oryza sativa* L.). Theor Appl Genet 90:907–913

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

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification systemfor eukaryotic transposable elements. Nature Rev Genet 8:973–982

Wright DA, Voytas DF (2002) *Athila4* of *Arabidopsis* and *Calypso* of soybean define a lineage of endogenous plant retroviruses. Genome Res 12:122–131

**Table 1** List of the poplar and willow analyzed species. For poplars, the botanical section and the geographical origin are reported. The Sequence Reads Archive (SRA) code corresponding to the sequence reads used for the analysis is reported for each species

| Species | Section | Origin | SRA |
|---|---|---|---|
| *P. deltoides* | Aigeiros | Eastern North America | SRS1328499 |
| *P. nigra* | Aigeiros | Europe | SRS1218640 |
| *P. tremula* | Populus | Europe, Northern Asia | SRS1124263 |
| *P. tremuloides* | Populus | North America | SRS1124888 |
| *P. balsamifera* | Tacamahaca | Northern North America | SRS1115969 |
| *P. simonii* | Tacamahaca | Northeast Asia | SRS829341 |
| *P. trichocarpa* | Tacamahaca | Western North America | SRS844395 |
| *S. purpurea* | | | SRS161420 |
| *S. suchowensis* | | | SRS1276361 |

**Table 2** Sequence reads used for hybrid clustering by RepeatExplorer. The Sequence Reads Archive (SRA) code is reported for each set of Illumina reads

| Species | Total reads | Analysed reads | Clustered reads | % | Single reads | % |
|---|---|---|---|---|---|---|
| *P. deltoides* | 500,000 | 166,524 | 63,786 | 38.3 | 102,738 | 61.7 |
| *P. nigra* | 500,000 | 166,966 | 56,383 | 33.8 | 110,583 | 66.2 |
| *P. tremula* | 500,000 | 167,570 | 67,133 | 40.1 | 100,437 | 59.9 |
| *P. tremuloides* | 500,000 | 146,648 | 68,232 | 46.5 | 78,416 | 53.5 |
| *P. balsamifera* | 500,000 | 167,650 | 60,153 | 35.9 | 107,497 | 64.1 |
| *P. simonii* | 500,000 | 167,650 | 63,372 | 37.8 | 104,278 | 62.2 |
| *P. trichocarpa* | 500,000 | 167,348 | 56,860 | 34.0 | 110,488 | 66.0 |
| *S. purpurea* | 500,000 | 168,378 | 65,063 | 38.6 | 103,315 | 61.4 |
| *S. suchowensis* | 500,000 | 167,230 | 69,914 | 41.8 | 97,316 | 58.2 |
| Mean | | 165,107 | 63,433 | 38.5 | 101,674 | 61.4 |

**Table 3** Genome percentages of different repeat classes and superfamilies in poplar species. Total repeats refer to the proportion of reads clustered by RepeatExplorer. Percentages of different repeat classes and superfamilies are calculated on annotated clusters with a genome proportion higher than 0.01%

| Species | Total Repeats | LTR-*Gypsy* | LTR-*Copia* | DNA Trans-posons | Non-LTR | Heli-tron | Satel-lite DNA | rDNA |
|---|---|---|---|---|---|---|---|---|
| *P. deltoides* | 38.3 | 8.3 | 1.2 | 1.9 | 0.8 | 0.2 | 4.2 | 2.3 |
| *P. nigra* | 33.8 | 7.6 | 1.5 | 1.7 | 0.8 | 0.4 | 2.1 | 2.2 |
| *P. tremula* | 40.1 | 9.0 | 1.2 | 1.5 | 0.4 | 0.2 | 4.1 | 1.4 |
| *P. tremuloides* | 46.5 | 19.6 | 1.9 | 1.7 | 0.4 | 0.4 | 1.4 | 0.9 |
| *P. balsamifera* | 35.9 | 9.6 | 1.6 | 1.9 | 0.5 | 0.3 | 1.7 | 1.9 |
| *P. simonii* | 37.8 | 8.9 | 1.5 | 1.7 | 0.5 | 0.2 | 0.8 | 2.0 |
| *P. trichocarpa* | 34 | 7.7 | 1.3 | 2.3 | 0.6 | 0.5 | 1.3 | 2.3 |
| Mean | 38.1 | 10.1 | 1.5 | 1.8 | 0.6 | 0.3 | 2.2 | 1.9 |

**Table 4** Genome percentages of different LTR-RE lineages in poplar species. Percentages of lineages are calculated on annotated clusters with a genome proportion higher than 0.01%

| | *Gypsy* | | | | *Copia* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Species | Total | *Athila* | *Chrom-ovirus* | *Ogre / Tat* | Total | *AleI/ Retrofit* | *AleII* | *Ivana / Oryco* | *Maximus / SIRE* | *TAR/ Tork* |
| *P. deltoides* | 8.3 | 4.0 | 2.5 | 1.7 | 1.2 | 0.3 | 0.4 | 0.1 | 0.0006 | 0.6 |
| *P. nigra* | 7.6 | 3.5 | 1.9 | 2.2 | 1.5 | 0.5 | 0.2 | 0.1 | 0.002 | 0.6 |
| *P. tremula* | 9.0 | 3.4 | 2.6 | 3.0 | 1.2 | 0.5 | 0.2 | 0.2 | 0.002 | 0.4 |
| *P. tremuloides* | 19.6 | 14.3 | 2.2 | 3.2 | 1.9 | 0.6 | 0.2 | 0.3 | 0.090 | 0.7 |
| *P. balsamifera* | 9.6 | 3.3 | 2.6 | 3.8 | 1.6 | 0.3 | 0.2 | 0.2 | 0.002 | 0.8 |
| *P. simonii* | 8.9 | 3.5 | 2.3 | 3.1 | 1.5 | 0.3 | 0.5 | 0.2 | 0.002 | 0.6 |
| *P. trichocarpa* | 7.7 | 3.3 | 2.3 | 2.0 | 1.3 | 0.6 | 0.2 | 0.1 | 0.002 | 0.4 |
| Mean | | 5.0 | 2.4 | 2.7 | | 0.4 | 0.3 | 0.2 | 0.014 | 0.6 |

## Legends for figures

**Fig. 1** Representation of cluster abundances in the genome of 7 *Populus* and 2 *Salix* species. The size of the rectangle is proportional to the genome proportion of a cluster for each species. The colour of the rectangles corresponds to the repeat class, superfamily or lineage. Bar plot in the top row shows the size of the clusters as number of reads in the comparative analysis. Upper lines label groups of clusters as assessed by a hierarchical clustering of the results

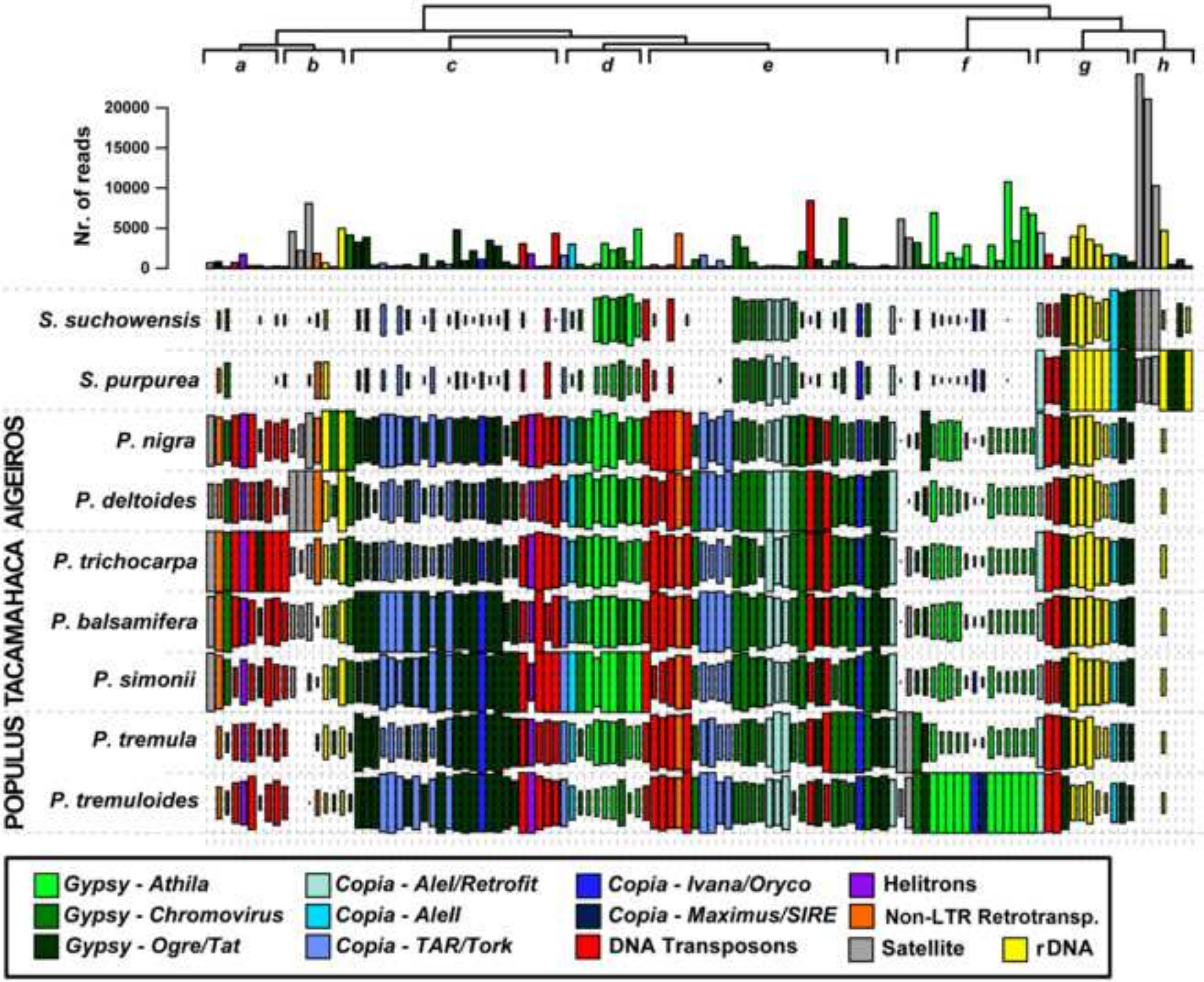
**Fig. 2** Dendrogram obtained by a hierarchical clustering analysis based on genome proportion data of clusters, as obtained by hybrid clustering using RepeatExplorer, in different poplar and willow species. Asterisks indicate multiscale bootstrap resampling (only values > 50% are given). The bar represents the genetic distance

**Fig. 3** Distance tree of *Gypsy* RT domains of 7 poplar and two willow species subjected to NJ analysis. Bootstrap values higher than 0.6 are shown with asterisk. Stars indicate RT domains of species different from poplars and willows, i.e. outgroups. The lineage to which RTs belong is reported

**Fig. 4** Timing of retrotranspositional activity of 6 LTR-RE lineages in 7 poplar species, based on the pairwise comparisons of Illumina reads matching RT-encoding sequences. The y axis reports the product of the percentage of pairwise comparisons for the average coverage of the RT sequence in each species

Figure 1

Click here to download Figure Usai_Figure 1.tif



Click here to download Figure Usai_Figure 2.tif

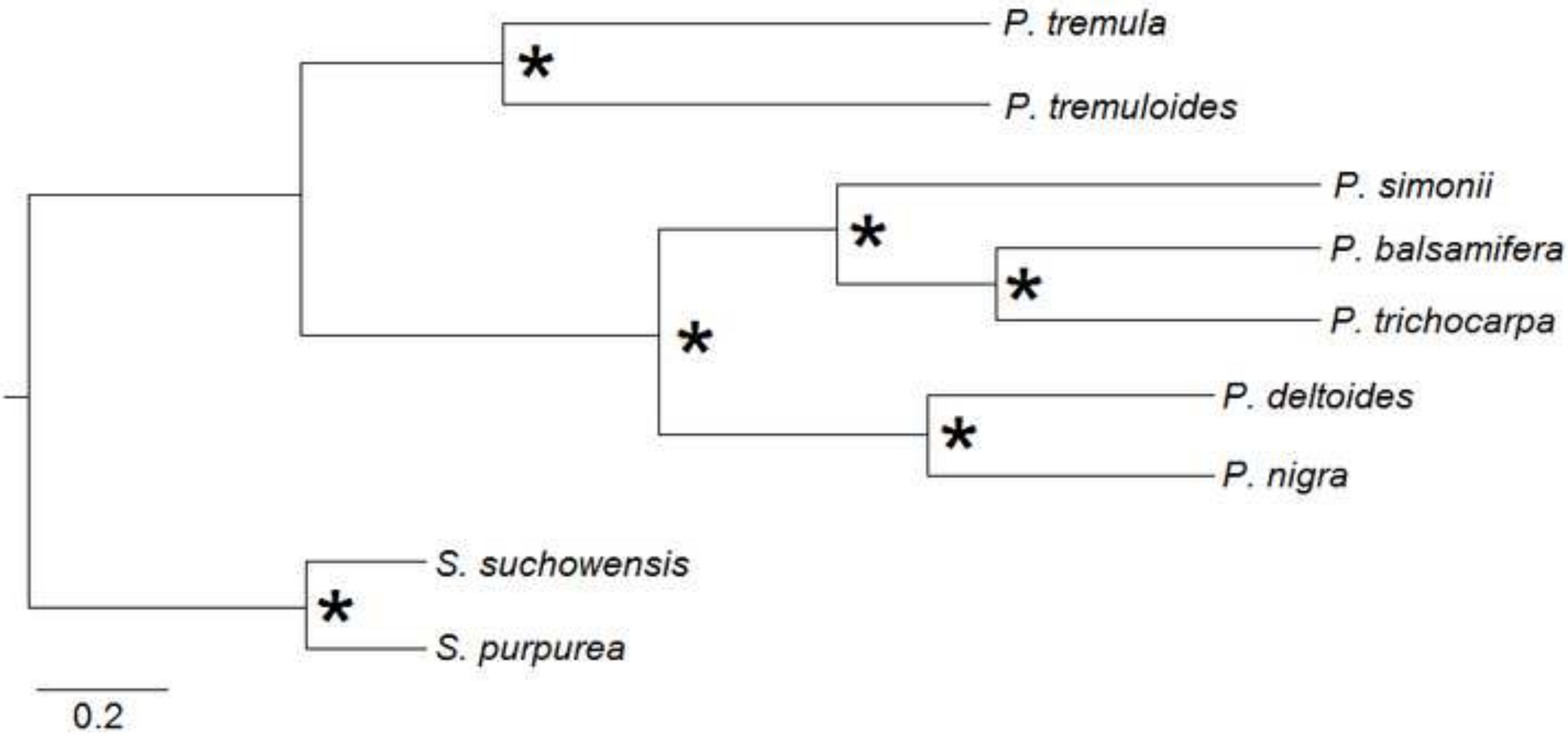


Figure 3

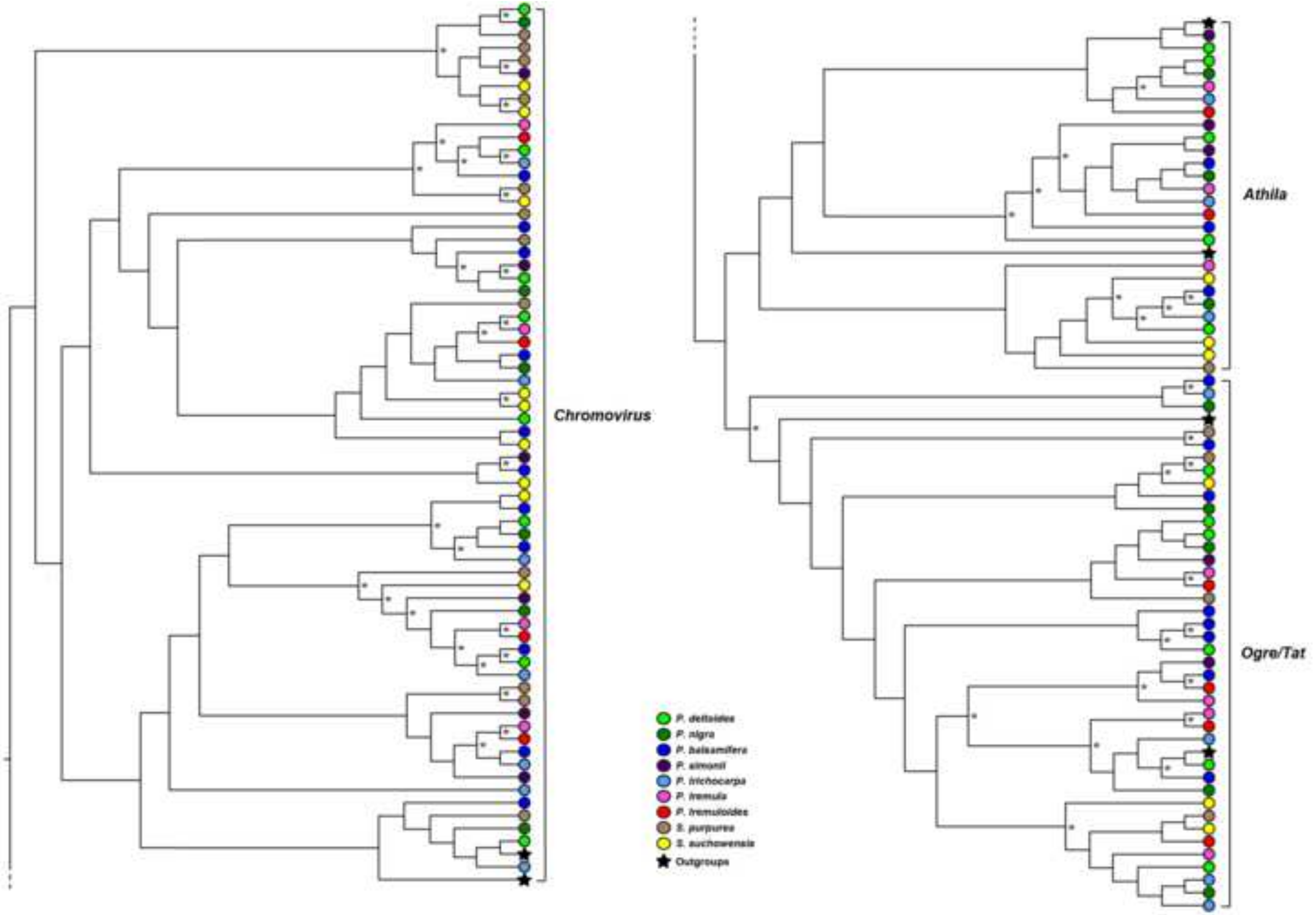
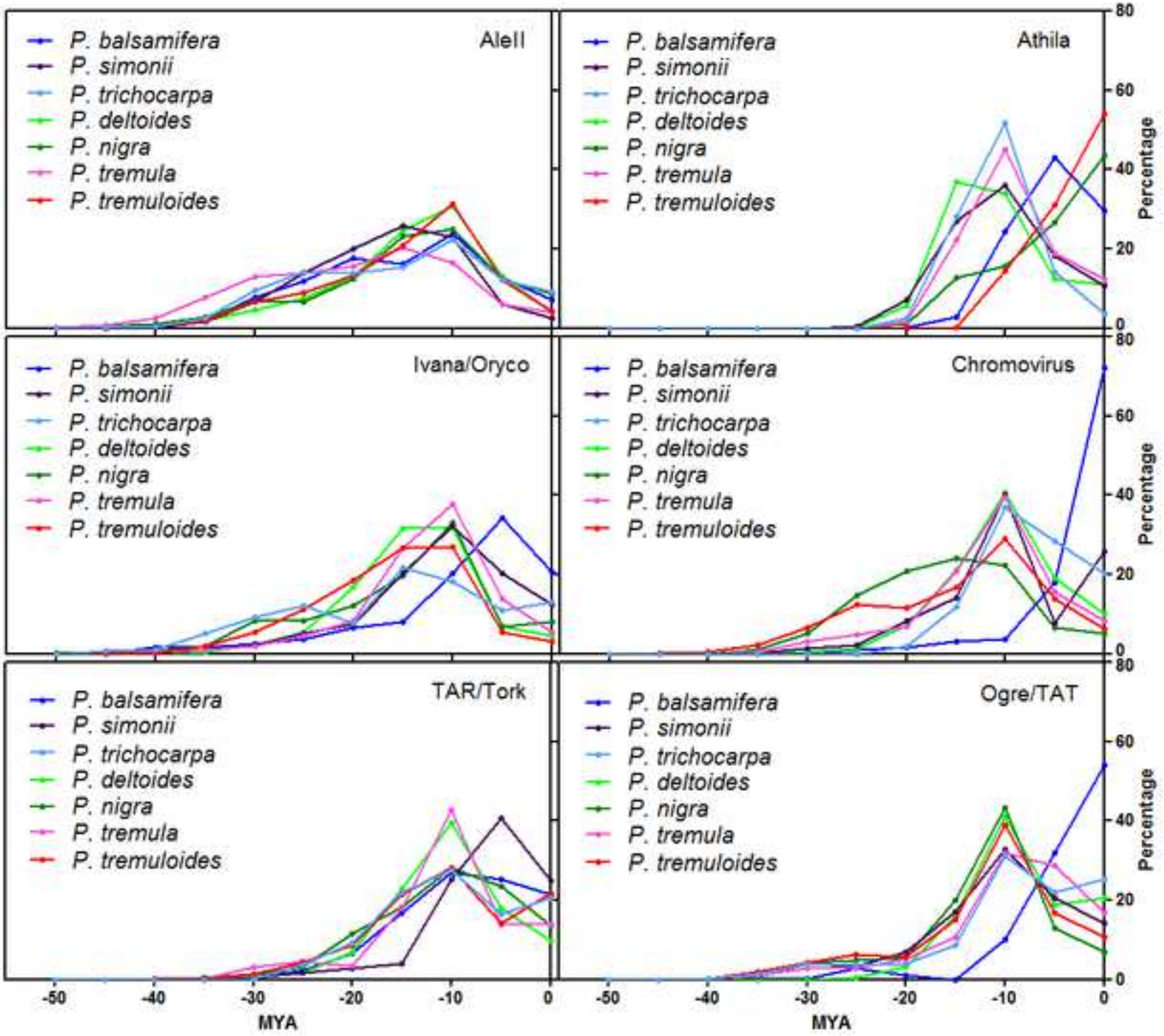

Chromovirus
Athila
Ogre/Tat
P. deltoides
P. nigra
P. balsamifera
P. simonii
P. trichocarpa
P. tremula
P. tremuloides
S. purpurea
S. suchowensis
Outgroups

Figure 4


P. balsamifera
P. simonii
P. trichocarpa
P. deltoides
P. nigra
P. tremula
P. tremuloides
AleII
Athila
Ivana/Oryco
Chromovirus
TAR/Tork
Ogre/TAT
Percentage
80
60
40
20
0
-50
-40
-30
-20
-10
0
MYA

Supplementary Material

TGGE-D-17-00138
Comparative genome-wide analysis of repetitive DNA in the genus Populus L.

G. Usai, F. Mascagni, L. Natali, T. Giordani, A. Cavallini

Response to reviewers

We are returning the revised version of the manuscript titled "Comparative genome-wide analysis of repetitive DNA in the genus Populus L."(TGGE-D-17-00138). In the revised manuscript, changes are evidenced in red. In the paragraphs that follow, we describe our responses to reviewer’s comments. Comments of the reviewers are copied from the e-mail provided by the Editor. Our responses are in red and follow these comments.

Reviewer #1: This is clearly written manuscript addressing a question of composition and evolution of repetitive genome fractions in a group of related plant species. Thanks to the availability of genome skimming NGS data the authors performed comparative repeat analysis using RepeatExplorer pipeline in a set of seven poplar and two willow species employed as outgroups. Overall repeat composition as well as its differences between the species were elucidated. I have only minor questions/comments to the manuscript:

1./ Table 2 - Comparative clustering analysis was performed using 500,000 reads from each species (column "Total reads"); I suppose the column "Analysed reads" gives numbers of reads that were actually analysed due to limited hardware resources of the RepeatExplorer server. As expected, the numbers of randomly sampled reads are very similar in all species except for P. tremuloides where only 146,648 reads were analysed (contrary to ~ 167,000 in other species). Why ?
RE: The actual number of reads that can be processed depends on the number of similarity hits they produce because all read overlaps must be loaded into the server memory during the graph-based clustering step. In the case of P. tremuloides, some sequences, whose repeated units are especially repeated in the genome (i.e. Athila retrotransposons), might have determined a more rapid saturation of the server memory than for the other species and, consequently, the use of a minor number of reads.

2./ Page 8, l. 3-5: RepeatExplorer protein domain search tools does not use RepBase but it uses its own custom-made database of transposon protein domains.
RE: It is correct, change made.

3./ Page 8, l. 50-55: Up to 100 reads were used for examination of pairwise divergence of RT domains. Is it statistically sufficient number to provide reliably estimates ? Why there were not more reads used instead ?
RE: 100 different (completely overlapping) reads represent 100 different retrotransposons belonging to a lineage and in our opinion are a reliable number to estimate a transpositional profile. In other experiments, we compared P. trichocarpa transpositional profiles (obtained by measuring RT pairwise divergence) with transpositional profiles obtained using other methods for measuring retrotransposon insertion ages and found a very good correspondence. This comparison is the object of another publication we are preparing. We added this comment in the Results section as "unpublished data".

4./ Some parts of the text in Results would be more suitable for Discussion, for example p. 10, l. 11-25.
RE: We agree, change made.

5./ Could you discuss how observed differences in repeat proportions and composition correlate with genome size variation in Populus ?
RE: No, we couldn't. The genome size is not available for all analysed species.