University of Pisa

PhD Course in Molecular Biotechnologies
XXIV Cycle (2009-2011)

High-throughput sequencing for the analysis of genomic DNA and gene expression in *Populus* spp.

SUPERVISOR:

Prof. Andrea Cavallini

CANDIDATE:

DOTT. Rosa Maria Cossu

XXIV Cycle (2009-2011)

To my Family

Summary

The genus *Populus* is an important crop and a model system to understand molecular processes of growth, development, and responses to environmental stimuli in trees. Moreover the entire genome of *Populus trichocarpa* was sequenced.

The aim of this research was studying genomic variation and evolution in the poplar genus, and the effects of such variations in producing heterosis in two interspecific hybrids between *Populus deltoides* and *P. nigra*.

Heterosis, intended as the superior performance of hybrid progeny compared to their inbred parents, has been one of the driving forces in poplar breeding. The two interspecific hybrids used in our experiments exhibit different levels of heterosis, i.e., their productivity is for one genotype much larger than that of parents and, for the other genotype, is similar to that of parents. The molecular bases of heterosis are still to be fully clarified, though it appears that variations in intergenic regions can have a role in the heterotic phenotype. Hence, we studied the extent of variation in the repetitive component of the genome (especially retrotransposons) and its possible consequences on gene and allelic expression.

During this research, bioinformatic and genomic analyses were performed aiming i) to characterize the repetitive component of the poplar genome, by the isolation and characterization of LTR-retrotransposons in the *P. trichocarpa* genome, and the production of a database of such elements; moreover the previously undescribed structure of poplar centromeres was evaluated by means of NGS techniques; ii) to analyze poplar genome repetitive component and its expression, studying, by Illumina RNAseq, the transcription of previously isolated LTR-retrotransposons, in control and drought stressed plants; iii) to study the poplar transcriptome, also in relation to drought and, for an indirect evaluation of cis-regulatory sequence variation in the poplar hybrid, to the differential expression between alleles in genes expressed in control and drought stress.

Concerning LTR-retrotransposons, we observed a relatively recent burst of retrotransposons activity, though counterbalanced by high levels of DNA loss. A huge fraction of retrotransposons belong to unknown superfamilies, i.e. they are non-autonomous retrotransposons because lacking coding capacity. These elements are especially expressed in poplars. We also individuated two distinct centromeric repeats, that occur in all three analysed poplar species.

Gene expression was analysed mapping RNAseq data to the complete poplar transcriptome, and a reference expression dataset was established. In several instances, the two alleles in a hybrid are flanked by different DNA sequences, affecting tissue specificity or temporal regulation of expression of genes. We found allele specific expression in many of 200 randomly chosen genes in different stress conditions. This suggests a differential role for the two alleles during hybrid growth and in its interaction with the environment. It is possible that the functional diversity of the two parental alleles in the hybrid may have an impact on hybrid performance through allelic complementation.

Table of contents

Introduction

Genome structure and repetitive DNA in eukaryotes

All eukaryotes possess a nuclear genome that is divided into two or more linear DNA molecules. Beside nuclear genome, they also have smaller, usually circular, mitochondrial genomes. The plants and other photosynthetic organisms are characterized by the presence of a third genome, located in the chloroplasts.

The basic physical structures of nuclear genomes are similar among all eukaryotes, but one feature is very different in various organisms: the genome size. The size ranges from less than 10 Mb, to 100,000 Mb and it is not related to the complexity of the organism. So the simplest eukaryotes do not have the smallest genomes, and the higher eukaryotes do not have the larger genomes.

The lack of correlation between the complexity of an organism and the size of its genome is the so-called C-value paradox.

In the early ‘70s, the discovery of non-coding DNA revealed that the genome size does not reflect gene number in eukaryotes. Most of their DNA is non-coding and therefore does not consist of genes.

Non-coding DNA plays an important role on driving the compactness or the enlargement of a genome. Now it is clear that families of repeats have sustained a massive proliferation in the genomes of certain species. However the host controls their copy number by epigenetic silencing in order to prevent genomic overload.

The bulk of intergenic DNA, the part of the genome that lie between genes, is made of repeated sequences.

These can be divided into two categories: tandemly repeated DNA, whose repeat units are placed next to each other in an array, and interspersed repeats, whose individual repeat units are distributed apparently random around the genome.

Tandemly repeated DNA

Tandemly repeated DNA is a common feature of eukaryotic genomes. This type of repeats form the so-called satellite DNA because DNA fragments containing tandemly repeated sequences tend to produce a second or ‘satellite’ bands when genomic DNA is separated on a density gradient.

In eukaryotic DNA, satellites are made up of long series of tandem repeats. A single genome can contain several different types of satellite DNA, each one made with a different repeat unit (from < 5 to > 200 bp in length).

Some satellite DNA is dispersed within the genome, however, most of it is located in the centromeres. Here, satellite repeats play a structural role as binding sites for one or more of the centromeric proteins. The centromere DNA is the last region of the chromosome to be replicated. It apparently lacks sequences acting as origins of replication to delay its replication until the end of the cell cycle. The repetitive nature of centromeric DNA may guarantee that such replication origins are absent.

Two other types of tandemly repeated DNA are classified as ‘satellite’ DNA, despite they not appear in satellite bands on density gradients. These are minisatellites and microsatellites. The minisatellites form clusters up to 20 kb in length, with repeat units up to 25 bp; microsatellite form clusters less than 150 bp, with repeat units around 13 bp or less.

Minisatellite DNA has an important function in DNA replication and some of its clusters land near the ends of chromosomes.

The telomeric DNA of most plant chromosomes comprises different copies of the motif of 7-bp 5’-TTTAGGG-3’. The telomere ends with a protective nucleoprotein cap made with DNA-binding proteins associated with telomeric microsatellite. This preserve chromosome ends from being joined by double-strand break-repair mechanisms (Lamb et al. 2012).

During the past, microsatellites were only considered as genetic markers because of the extensive length polymorphisms. Now we know that some of them act as cis-regulatory elements which can be recognized by transcription factors (Iglesias et al. 2004).

Interspersed repeats

The most abundant interspersed repeats are known as transposable elements (TEs). They have the ability to move from one position in the genome to another by transposition. TEs are present in all kingdoms. In plants with large genome size they are the major constituents of the genome, representing around 80% of the total genomic DNA (Morgante 2005; Bennetzen 2005).

TEs can be classified on the basis of the mechanism of transposition: RNA transposons move via reverse transcription of an RNA intermediate; DNA transposons move directly without an RNA intermediate.

RNA transposons

The transposition that involves an RNA intermediate is called retrotransposition. First an RNA copy of the transposon is synthesized by transcription. Then the RNA transcript is copied into DNA. This conversion of RNA to DNA, requires a reverse transcriptase enzyme. Finally the DNA copy of the transposon integrates into the genome, into the same chromosome occupied by the original unit, or into a different chromosome.

RNA transposons or retroelements can be divided into five orders: Long Terminal Repeats (LTR) retrotransposons, Dictyostelium intermediate repeat sequence-like elements (DIRS), Penelope-like elements (PLEs), long interspersed nuclear elements (LINE) and short interspersed nuclear elements (SINE).

The LTR retrotransposon length ranges from a few hundred base pairs up to 25 kb. The LTRs at either end range from a few hundred base pairs to more than 5 kb, and start with 5’-TG-3’ and end with 5’-CA-3’. The LTRs regulate transcription and play a role in the transposition process. The two main superfamilies are *Gypsy* and *Copia*.

The DIRS-like components have either inverted terminal repeats flanking the internal region or the split direct repeats (SDR). Members of this order are present in different species like green algae, animals and fungi (Goodwin 2004).

The PLEs are found in genome of different eukaryotes, including protists, fungi, plants and animals. Members of this order have LTR-like sequences with inverted or direct orientation (Arkhipova 2006).

The LINEs have been detected in all eukaryotic kingdoms. The members of this order do not have LTRs and can reach several kilobases in length. These predominate over the LTR retrotransposons in many animals, for instance the L1 family is about 20% of the human genome. Instead, in plants they seem to be less abundant compared with LTR retrotransposons.

The SINE is a non-autonomous order, SINEs are small (80–500 bp), and they rely on LINEs for trans-acting transposition functions. The best known SINE is the Alu element, which has a copy number of 500,000 in the human genome (Wicker et al. 2007).

DNA transposons

DNA transposons do not require an RNA intermediate to transpose. They are ancient and are found in almost all eukaryotes, present in lower number than retrotransposons. On the contrary, in prokaryotes they are more redundant than RNA transposons.

They have been divided into two subclasses, depending on two distinct transposition mechanisms.

Transposition in the Subclass 1 involves an excision of the element and the re-integration at a new site (conservative transposition). This subclass includes the order TIR, characterized by their terminal inverted repeats (TIRs) of variable length; and the order Crypton which lack TIRs, but seem to generate the Target Sites Duplication (TSDs) as a result of recombination and integration.

The transposition of Subclass 2 elements involves a copy of the element and integration at a new site (replicative transposition). Subclass 2 includes the order Helitron, which appear to replicate via a rolling-circle mechanism and do not generate TSDs, found mainly in plants, and the order Maverick (also known as Polintons), which are long (10–20 kb) and are bordered by long TIRs, found in different eukaryotes, but not in plants (Wicker et al. 2007).

Activation and repression of retroelements

The TEs are potentially highly mutagenic, that’s why the host has evolved epigenetic mechanism to control their proliferation. These mechanisms of silencing at the transcriptional and post-transcriptional levels, include post-transcriptional silencing of TEs by RNAi and RNAi-mediated chromatin modifications (Slotkin and Martienssen 2007).

In the post-transcriptional silencing of TEs by RNAi, dsRNA is cleaved by members of the dicer endonuclease family into 21–30-nt small interfering RNAs (siRNAs) that guide RNA- degrading complexes to a complementary transcript. Argonaute proteins constitute the catalytic component of the siRNA-guided transcript-cleavage complex (RISC). Mutations in both argonaute- and dicer-family proteins cause the reactivation of TEs.

RNAi can also lead to chromatin modifications, that suppress TE transcription, by modifications of histone tails, DNA methylation and alterations in chromatin packing and condensation. Methylation of lysine 9 in histones H3 that are associated with TEs represses transcription. Also DNA methylation on cytosine residues is an important signal that represses TE transcription.Several proteins implicated in chromatin packaging and condensation are also involved in TE silencing. For example in plants, SWI/SNF chromatin-remodelling proteins

are required for TE silencing.

The majority of transposons are inactive, methylated, and targeted by siRNAs. However, a variety of conditions are able to reverse transposon silencing. For example, the retrotransposons Tnt1 and Tto in the Solanaceae can be activated by biotic (i.e. fungi and bacteria) and abiotic stress conditions (Grandbastien et al. 1997, Grandbastien et al 2005).

Hashida et al. (2003) have found that low temperatures can lead to activation and demethylation of a DNA transposon, Tam3, in Antirrhinum, and this shift can be reversed by raising the temperature.

Also hybridization seems to have an important effect on transposon activity in some species (Otto and Whitton 2000). For example independent hybridization events between two sunflower species gave rise to three new hybrid species. Both parental and hybrids are diploid but the hybrid genomes are at least 50% larger than that of the parental species (Baack et al. 2005), and the majority of this size difference can be ascribed to the massive amplification of a single class of retrotransposons (Ungerer et al. 2006). The hybrid species are all adapted to particularly arid environments, and transposon activity is suggested to have facilitated that adaptation (Noor and Chang 2006).

Grandbastien et al. (2005) suggest that the activation of transposons may represent a programmatic and regulated response to stress, however, transposons reactivation under stress conditions may simply be a temporary shift in the equilibrium between transposons and their hosts (Lisch 2009).

Structural and regulatory roles of retrotransposons

The retrotransposons have structural and/or functional roles in centromeres, telomeres, and other heterochromatic chromosomal regions. For example, in telomeric regions, they have a role in fighting the shortening of chromosome ends (Pardue et al. 1997).

The presence of retrotransposon fragments within the regulatory regions of many plant genes indicates that they are involved in specific gene regulation. LTR retrotransposons carry a promoter within each LTR, and their insertions may cause new gene expression patterns (Flavell et al. 1994, Wessler et al. 1995). Some of them have also a function as transcriptional silencers, downregulating transcription of the enclosing genes.

In humans retroelements can modulate the transcription, the splicing of pre-mRNA and may contribute to a diversity of alternatively spliced RNAs (van de Lagemaat et al. 2006). Promoters of intronic retrotransposons may drive transcription of RNAs that are complementary to gene introns and/or exons. Some of them possess bidirectional promoters (Domansky et al. 2000, Matlik et al. 2006), and even downstream insertions of these elements relative to genes may result in the production of an antisense RNA. These complementary RNAs may alter functional host gene expression. Recently has been proved that, in the mouse, a SINE retroelement can works as insulator sequences that distinguish blocks of active and transcriptionally silent chromatin (Gogvadze and Buzdin 2009).

Retrotransposons shape genome structure

The retroelements must develop strategies for contrasting the tendency of the host to keep them under restraint. Retroelement cDNA insertion impacts on the host’s genetic material, that’s why they are target for regulatory control. Replication of REs depends on selecting a favourable chromosomal site for integration of their genomic DNA, by targeting distinctive chromosomal regions (Bushman 2003).

For example, the LTR retrotransposons of yeast are associated with domains of heterochromatin or sites bound by particular transcriptional complexes such as RNA polymerase III (Chalker and Sandmeyer 1992, Zou et al.1996). These regions are typically gene poor and may enable yeast retrotransposons to replicate without causing their host undue damage (Boeke and Devine 1998). Non-uniform chromosomal distributions are observed in other organisms as well. In Arabidopsis thaliana and Drosophila melanogaster, many retroelements are clustered in pericentromeric heterochromatin (Arabidopsis Genome Initiative 2000, Adams et al. 2000).

Beyond the yeast model, it is not known whether retroelements generally seek the spots for integration (Peterson-Burch et al. 2004).

However, it can be observed that *gypsy* and *copia* elements have a preferential localization. The first prefer mostly the centromeric and pericentromeric regions, as for example in the genus Beta (Gindullis et al. 2001) and in cereals (Presting et al. 1998, Li et al. 2004, Liu et al. 2008), the second one being rare or absent around the centromeres (Heslop-Harrison et al. 1997, Pich and Schubert 1998), for instance in the genus Helianthus, they were localized mostly in the telomeres (Santini et al. 2002, Stanton et al. 2009).

Retrotransposons affect the regulatory machinery of organisms

Activation of REs can be considered as the main source of genetic variability within a plant species. It is generally presumed that the genomes of individuals belonging to a single species do not differ in the degree to which genes remain on corresponding orders (colinearity) over time.

Instead, the intraspecific comparisons between genic and intergenic regions have revealed that, for example in maize and in barley, conservation is mainly restricted to the genic regions. Large rearrangements occurred, including gene duplications and intergenic regions divergence, through the movement of retroelements (Fu and Dooner 2002, Song and Messing 2003, Brunner et al. 2005a, Scherrer et al. 2005). Brunner et al. (2005a) have compared sequences from different inbreds at the same locus in maize. This comparison have shown that most of the non shared sequences consist of LTR-retrotransposons and other mobile elements. The differences in RE content between lines could have arisen by retrotransposition, leading to insertions, or by recombinational events, leading to deletions (Devos et al. 2002). Large variability was found in the composition, in the length of intergenic regions, and in the gene space, where several genes were missing (Fu and Dooner 2002).

Such a lack of colinearity can have many biological implications, for example non-shared sequences are excluded from recombination events.

Fu and Dooner (2002) proposed that complementation of non-shared genes could be one of the factors contributing to heterosis, a phenomenon of the superior performance of hybrid progeny in comparison with their inbred parents (Shull, 1908). The majority of these non shared genic sequences appear to be novel sequence gain (duplication events) in one inbred relative to the other. Novel functions may arise as gene fragments, or full-length genes, are copied to new chromosomal locations and potentially acquire novel expression patterns. In addition, the mechanism of transposition appears to have the capacity to perform exon shuffling and create new ORFs at a relatively high frequency (Brunner et al. 2005b, Lai et al. 2005).

Also differences in the repetitive fraction can affect heterosis. The insertion of TEs in cis-regulatory regions modulates gene products, and can be an important genetic component for quantitative trait variation (Tanksley 1993; Doebley and Lukens 1998; Mackay 2001; Buckler and Thornsberry 2002) generating novel phenotypes while preserving existing functions (Wray et al. 2003). The functional architecture of cis-regulatory regions consists of short and often redundant transcription factor binding sites, that are interspersed within apparently non functional regions. Promoters and other cis-regulatory regions form a protein/DNA complex with trans-regulatory proteins (transcription factors), thereby promoting integrative control of expression.

In several instances, conserved and active alleles in the two inbreds used to produce a hybrid, are flanked by different DNA, for example, by nonconserved retrotransposons inserted nearby (Brunner et al. 2005a). Such different repetitive sequence environments may affect tissue specificity or temporal regulation of expression of genes and have been proposed as one of the causes of heterotic complementation (Birchler et al. 2003, Song and Messing 2003) according to the overdominance theory (Crow 1948).

Some experimental data can be reported on the effect of allele variation on gene expression at transcription level in plants. For example, the phenotypic variation controlled by the tomato fruit weight fw2.2 gene is regulated by variation in transcription, either in the level or timing of expression rather than from protein coding differences between alleles (Cong et al. 2002). In maize, the high level of allelic variants is probably related to the high level of observed allelic expression variation (Guo et al. 2004). Results from this study showed a general allelic expression variation at the accumulated transcript level in maize hybrids and allelic variation in responding to environmental stresses.

However, there are several other potential mechanisms through which genomic variation could combine to produce a heterotic phenotype. Each of these mechanisms could occur at a subset of genes, and the combination of effects will result in heterosis (Springer and Stupar, 2007). For example, altered protein–protein interactions, novel epigenetics states, siRNAs, or altered hormone levels can profoundly alter phenotypes (Birchler et al. 2003, Osborn et al. 2003). The rate and types of allelic variation obser ved in maize inbreds, suggest that allelic variants of a large number of loci act through partial to complete dominance to provide favourable complementation resulting in superior hybrid phenotypes. The contributions of epistatic interactions and overdominance to heterosis are more difficult to establish and remain enigmatic.

Sequencing technologies for the analysis of genome structure and expression

Progresses in understanding genome structure, evolution and functions reside on the availability of complete genome sequencing. Since 2000, when Arabidopsis genome sequence was completed, many other eukaryotes (including a number of higher plants) have seen their genome sequenced allowing comprehensive studies on genome components and evolutionary dynamics.

In the past, the methods used for the sequencing were the Sanger enzymatic dideoxy technique (Sanger et al. 1977) and the Maxam and Gilbert chemical degradation method (Maxam and Gilbert 1977).

Sanger method was used for sequencing the human genome. Its great limitations were: use of gels or polymers as separation media for the fluorescently labelled DNA fragments, the low number of samples which could be analysed in parallel and the difficulty of total automation of the sample preparation methods. These limitations stimulated efforts to develop techniques without gels, which allowed sequence determination on large numbers of samples (Ansorge 2009).

The general strategies for human genome sequencing included the use of transposons to create random insertions in cloned DNA (Kimmel 1997), as well as those that used multiplex sequencing strategies in combination with several detection schemes (Church et al. 1988, Smith et al. 1977). Later was used the shotgun sequencing strategy, that was proved to be the most efficient as the cost of Sanger method decreased (Gardner et al. 1981, Messing 2001).
The principal approaches for shotgun sequencing are:

- Clone-by-clone shotgun sequencing. The most commonly strategy for genome sequencing involves the shotgun sequencing of individual mapped clones (International Human Genome Sequencing Consortium 2001). This strategy follows a ‘map first, sequence second’ progression: the target DNA is first analysed by clone-based physical mapping methods, and then individual mapped clones are selected and subjected to shotgun sequencing (Green 2001).
- Whole-genome shotgun sequencing. An alternative strategy, called whole-genome shotgun sequencing, involves the assembly of sequence reads generated random, theoretically bypassing the need for a clone-based physical map. The entire genome of an organism is fragmented into pieces of defined sizes, which in turn are subcloned into suitable plasmid vectors. Sequence reads are generated from both insert ends of most subclones, which is important for dealing with the problems presented by repetitive sequences, so as to produce highly redundant sequence coverage across the genome. Computational methods are then used to assemble the sequence reads and to deduce a corresponding consensus sequence (Edwards et al. 1990). The expected physical distances separating these juxtaposed read pairs are an important factor of an accurate sequence assembly (Green 2001).

Next-generation DNA sequencing platforms

The Next Generation Sequencing (NGS) technologies allow sequence determination on large numbers of samples in parallel. So several gigabases can be sequenced in a few weeks for only a fraction of the costs of Sanger (Mardis 2008, Ansorge 2009). Another advantage of these platforms is the possibility of single DNA fragments amplification, avoiding the need for cloning of DNA fragments, and reduction of sequencing errors.

Limitations of this technology are short read lengths, non-uniform confidence in base calling in sequence reads, particularly deteriorating 3’-sequence quality and generally lower reading accuracy in homopolar stretches of identical bases. The huge amount of short reads generated by these systems require the development of softwares and more efficient computer algorithms.

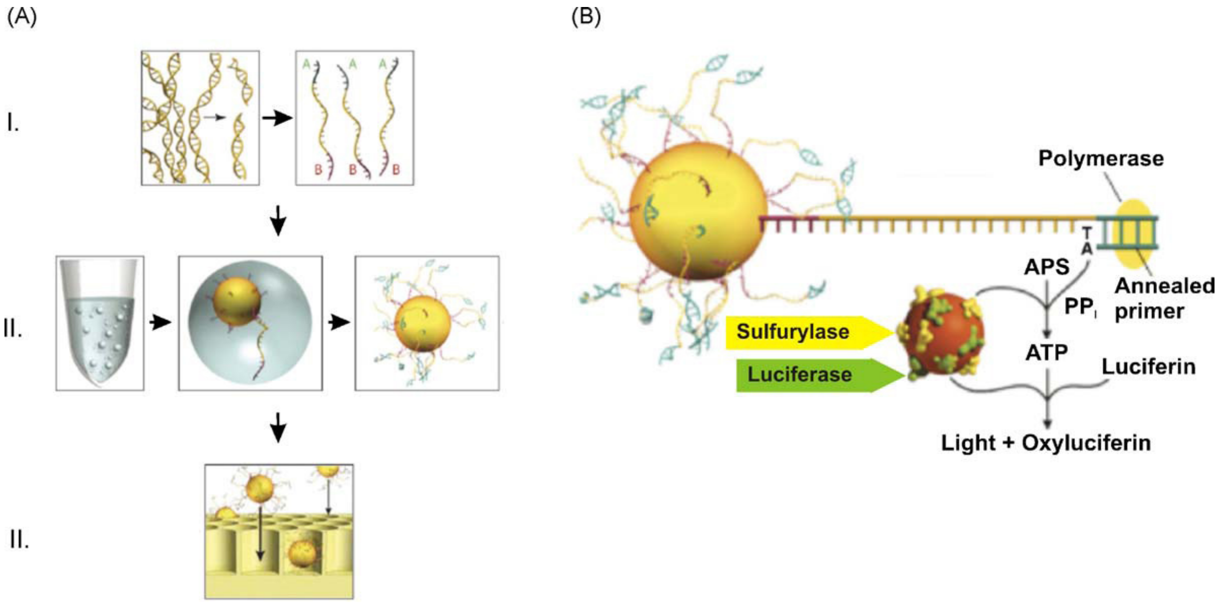


Figure 1
(A) Outline of the GS 454 DNA sequencer workflow. Library construction (I) ligates 454-specific adapters to DNA fragments (indicated as A and B) and couples amplification beads with DNA in an emulsion PCR to amplify fragments before sequencing (II). The beads are loaded into the picotiter plate (III). (B) Schematic illustration of the pyrosequencing reaction which occurs on nucleotide incorporation to report sequencing-by-synthesis. (Adapted from Ansorge 2009, http://www.454.com.)

The 454 sequencing

The 454 Genome Sequencer was introduced in 2005. In this system (Figure 1), DNA fragments are ligated with specific adapters that bind one fragment to a bead. Emulsion PCR is carried out for fragment amplification, with water droplets containing one bead and PCR reagents immersed in oil. The amplification is necessary to obtain sufficient light signal intensity for reliable detection in the sequencing-by-synthesis reaction steps. When PCR amplification cycles are completed and after denaturation, each bead is placed at the top end of a fibre in an optical fibre chip, created from glass fibre bundles. The individual glass fibres are excellent light guides, with the other end facing a sensitive CCD camera, enabling positional detection of emitted light. Each bead thus sits on an addressable position in the light guide chip, containing several hundred thousand fibres with attached beads. In the next step polymerase enzyme and primers are added to the beads, and one unlabelled nucleotide only is supplied to the reaction mixture to all beads on the chip, so that synthesis of the complementary strand can start. Incorporation of a following base by the polymerase enzyme in the growing chain releases a pyrophosphate group, which can be detected as emitted light. Knowing the identity of the nucleotide supplied in each step, the presence of a light signal indicates the next base incorporated into the sequence of the growing DNA strand. The sequencing read lengths is up to 1 kb (Ansorge 2009, http://www.454.com).

The Illumina (Solexa) Genome Analyzer

The Solexa sequencing platform was commercialised in 2006, and is based on the principle (Figure 2) of sequencing-by-synthesis chemistry. DNA fragments are ligated at both ends to adapters and, after denaturation, immobilised at one end on a solid support. The surface of the support is coated densely with the adapters and the complementary adapters. Each single-stranded fragment, immobilised at one end on the surface, creates a ‘bridge’ structure by hybridising with its free end to the complementary adapter on the surface of the support. In the mixture containing the PCR amplification reagents, the adapters on the surface act as primers for the following PCR amplification. After several PCR cycles, random clusters of about 1000 copies of single-stranded DNA fragments are created on the surface. The reaction mixture for the sequencing reactions and DNA synthesis is supplied onto the surface and contains primers, four reversible terminator nucleotides each labelled with a different fluorescent dye and the DNA polymerase. After incorporation into the DNA strand, the terminator nucleotide, as well as its position on the support surface, is detected and identified via its fluorescent dye by the CCD camera. The terminator group at the 3’-end of the base and the fluorescent dye are then removed from the base and the synthesis cycle is repeated. The maximum sequence read length is nowadays 150 nucleotides (Ansorge 2009, http://www.illumina.com/index.ilmn).

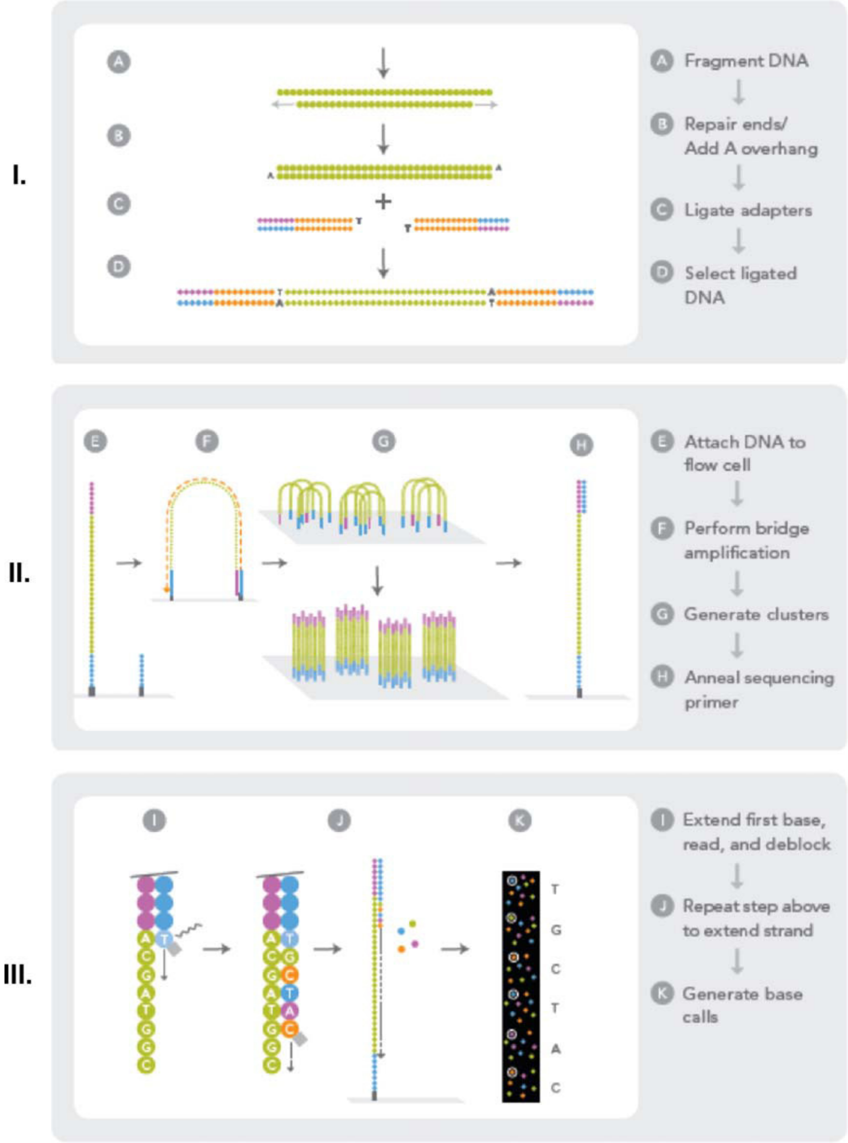

I.
A
B
C
D
A
T
A Fragment DNA
B Repair ends/ Add A overhang
C Ligate adapters
D Select ligated DNA
II.
E
F
G
H
E Attach DNA to flow cell
F Perform bridge amplification
G Generate clusters
H Anneal sequencing primer
III.
I
J
K
T
G
C
T
A
C
I Extend first base, read, and deblock
J Repeat step above to extend strand
K Generate base calls

Figure 2
Outline of the Illumina Genome Analyzer workflow. Similar fragmentation and adapter ligation steps take place (I), before applying the library onto the solid surface of a flow cell. Attached DNA fragments form 'bridge' molecules which are subsequently amplified via an isothermal amplification process, leading to a cluster of identical fragments that are subsequently denatured for sequencing primer annealing (II). Amplified DNA fragments are subjected to sequencing-bysynthesis using 30 blocked labelled nucleotides (III). (Adapted from the Genome Analyzer brochure, Ansorge 2009, http://www.illumina.com/index.ilmn).

The Applied Biosystems ABI SOLiD system

The ABI SOLiD sequencing system is a platform using chemistry based upon ligation. It was introduced in 2007. In this technique (Figure 3), DNA fragments are ligated to adapters then bound to beads. A water droplet in oil emulsion contains the amplification reagents and only one fragment bound per bead; DNA fragments on the beads are amplified by the emulsion PCR. After DNA denaturation, the beads are deposited onto a glass support surface. In a first step, a primer is hybridised to the adapter. Next, a mixture of oligonucleotide octamers is hybridised to the DNA fragments and ligation mixture added. In these octamers, the couple of fourth and fifth bases is characterised by one of four fluorescent labels at the end of the octamer. After the detection of the fluorescence from the label, bases 4 and 5 in the sequence are thus determined. The ligated octamer oligonucleotides are cleaved off after the fifth base, removing the fluorescent label, then hybridisation and ligation cycles are repeated, this time determining bases 9 and 10 in the sequence; in the subsequent cycle bases 14 and 15 are determined, and so on. The sequencing process may be continued in the same way with another primer, shorter by one base than the previous one, allowing one to determine, in the successive cycles, bases 3 and 4, 8 and 9, 13

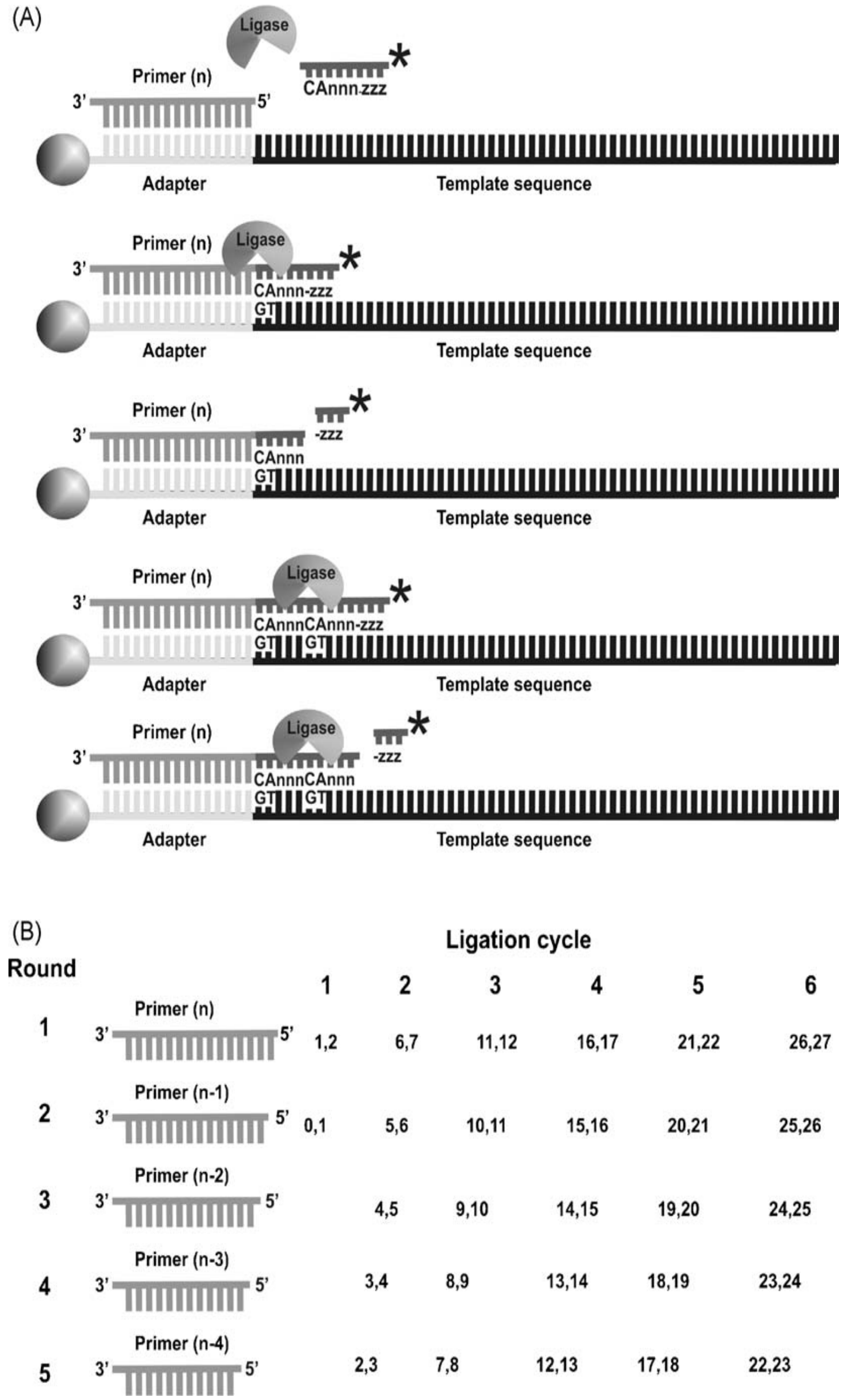


Figure 3
Sequencing-by-ligation, using the SOLiD DNA sequencing platform. (A) Primers hybridise to the P1 adapter within the library template. A set of four fluorescencelabelled di-base probes competes for ligation to the sequencing primer. These probes have partly degenerated DNA sequence (indicated by n and z) and for simplicity only one probe is shown (labelling is denoted by asterisk). Specificity of the di-base probe is achieved by interrogating the first and second base in each ligation reaction (CA in this case for the complementary strand). Following ligation, the fluorescent label is enzymatically removed together with the three last bases of the octamer. (B) Sequence determination by the SOLiD DNA sequencing platform is performed in multiple ligation cycles, using different primers, each one shorter from the previous one by a single base. The number of ligation cycles (six for this example) determines the eventual read length, whilst for each sequence tag, six rounds of primer reset occur [from primer (n) to primer (n 4)]. The dinucleotide positions on the template sequence that are interrogated each time, are depicted underneath each ligation cycle and are separated by 5-bp from the dinucleotide position interrogated in the subsequent ligation cycle. (Adapted and modified from Ansorge 2009, http://www.appliedbiosystems.com.)

and 14. The achieved sequence reading length is at present about ~50 bases. Because each base is determined with a different fluorescent label, error rate is reduced (Ansorge 2009, http://www.appliedbiosystems.com/absite/us/en/home.html).

The Helicos single-molecule sequencing device, HeliScope

The systems discussed above require the emulsion PCR amplification step of DNA fragments, to make the light signal strong enough for reliable base detection by the CCD cameras. In some instances, PCR amplification may introduce base sequence errors into the copied DNA strands, or favour certain sequences over others, thus changing the relative frequency and abundance of various DNA fragments that existed before amplification. The possibility of sequence determination directly from a single DNA molecule, without the need for PCR amplification requires a very sensitive light detection system and a physical arrangement capable of detecting and identifying light from a single dye molecule.

Helicos introduced the first commercial single-molecule DNA sequencing system in 2007. The nucleic acid fragments are hybridised to primers covalently anchored in random positions on a glass cover slip in a flow cell. The primer, the polymerase enzyme and labelled nucleotides are added to the glass support. The next base incorporated into the synthesised strand is determined by analysis of the emitted light signal, in the sequencing-by-synthesis technique (similar to Figure 2, but on only one DNA fragment, without amplification). This system also simultaneously analyses many millions of single DNA fragments simultaneously, resulting in sequence throughput in the Gigabase range. In the homopolar regions, multiple fluorophore incorporations could decrease emissions, sometimes below the level of detection; when errors did occur, most were deletions. Helicos has developed a new generation of "one-base-at-a-time" nucleotides that allows for more accurate homopolymer sequencing and lower overall error rates (Ansorge 2009, Metzker 2010, http://www.helicosbio.com/).

Applications of high-throughput DNA sequencing

Next-generation sequencing technologies are currently used for DNA resequencing, allowing the so-called personal genomics, that identifies genomic variants between individuals of one and the same species. They are being used also for de novo sequencing of genomes, either small (bacteria, viruses) or large. In this case, availability of paired-end sequences at definite and variable intervals is necessary for a correct assembly of the genome. DNA sequencing by NGS technologies is also used to investigate chromatin structure and DNA methylation patterns.

Moreover, NGS technologies offer novel, rapid ways for transcriptome-wide characterisation and profiling of mRNAs and small RNAs. For example, such technologies have allowed detailed analyses of RNA transcripts for gene expression and reliable transcript quantification.

In this field, the original serial analysis of gene expression technique (SAGE) (Velculescu et al. 1995) was limited in applications because of difficult ligation of a huge number of short DNA transcripts, subsequent cloning and Sanger sequencing. By contrast, the NGS technology allows the analysis of RNA transcripts by short sequence tags, up to 150 nt long, directly from each transcript in the sample. With this technique, transcripts are characterised through their sequence (Mortazavi et al. 2008), in contrast to the probe hybridisation employed in DNA chip techniques, with their inherent difficulties of cross hybridisation and quantification. Owing to the huge number of samples analysed simultaneously, sequence-based techniques can detect low abundance RNAs, small RNAs, or the presence of rare cells contained in the sample. Another advantage of this approach is that it does not require prior knowledge of the genome sequence.

Aim of the work

In this work we use Illumina NGS technology to study genome structure, variability and function in poplar interspecific hybrids at DNA and RNA sequencing level. The final aim of this research is contributing to clarify genomic variation and evolution in the poplar genus and the effects of such variations in producing heterotic genotypes by interspecific hybridization between *P. deltoides* and *P. nigra*.

The genus *Populus* is an important crop and a model system to understand molecular processes of growth, development, and responses to environmental stimuli in trees.

The small genome size, easiness of clonal propagation, rapid growth, ecological diversity, phylogenomic proximity to well-studied angiosperms, availability of an expressed sequence tag (EST) collection from poplar, aspen, cottonwood and their hybrids make this genus an ideal model system to understand molecular processes of growth, development and responses to environmental stimuli in trees (Sterky et al. 1998, Bradshaw et al. 2000, Taylor 2002, Kohler et al. 2003, Brunner et al. 2005, Sterky et al.

2004,

Nanjo et al. 2004). Moreover the nucleotide sequence of the entire genome of black cottonwood (*Populus trichocarpa*) was determined (Tuskan et al. 2007).

The most cultivated poplar species are *P. deltoides*, *P. nigra*, *P. trichocarpa*, as well as different interspecific hybrids. They are fast growing species, with high biomass production, and the capacity to adapt to stress conditions (Sixto et al. 2006, Dillen et al. 2007).

Heterosis has been one of the driving forces in poplar breeding (Muhle Larsen 1970). Hybrid poplars have rapid juvenile growth, high photosynthetic capacity, superior growth performance, and large woody biomass production rates that make them highly suitable for short-rotation silviculture (Ranney et al. 1987, Ceulemans et al. 1992, Barigah et al. 1994, Heilman et al. 1994).

Populus deltoides x *P. nigra* interspecific hybrids show different levels of heterosis concerning biomass production. It appears worth studying the extent of variation in the repetitive component of the genome (especially retrotransposons) and its possible consequences on gene and allelic expression. In the future, the clarification of the molecular bases of heterosis will allow a more efficient use of hybridization in crops, with favourable consequences on agriculture sustainability.

During this research, bioinformatic and genomic analyses were performed aiming to different tasks:

1) to study poplar genome and its repetitive component, we have isolated and characterized LTR-retrotransposons in the *P. trichocarpa* genome, and prepared a database of such elements, relatively unknown in poplar, until now (paper I);

2) to study poplar genome and its repetitive component, we evaluated the previously undescribed structure of poplar centromeres by means of NGS techniques (paper II);

3) to analyze poplar genome repetitive component and its expression, we have studied, by Illumina RNAseq the transcription of previously isolated LTR-retrotransposons, in control and drought stressed plants (paper III)

4) to study poplar transcriptome, we used RNAseq and determined gene expression in poplar hybrids, also in relation to drought (paper IV);

5) finally, for an indirect evaluation of cis-regulatory sequence variation in the poplar hybrid, we have analysed the occurrence of differential expression between alleles in genes expressed in control and drought stress (paper V).

References

Adams MD, Celniker SE, Holt RA et al. (2000). The genome sequence of Drosophila melanogaster. Science 287: 2185–2195.

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

Arkhipova IR (2006). Distribution and phylogeny of Penelope-like elements in eukaryotes. Systematic biology 55: 875-885.

Ansorge WJ (2009). Next-generation DNA sequencing techniques. N Biotechnol. 25: 195-203.

Baack EJ, Whitney KD, Rieseberg LH (2005). Hybridization and genome size evolution: timing and magnitude of nuclear DNA content increases in Helianthus homoploid hybrid species. New Phytol. 167: 623–630 .

Barigah TS, Saugier B, Mousseau M, Guittet J, Ceulemans R (2004). Photosynthesis, leaf area and productivity of 5 poplar clones during their establishment year. Ann. For. Sci. 51: 613-625.

Bennetzen J (2005). Transposable elements, gene creation and genome rearrangement in flowering plants. Curr. Opin. Genet. Dev. 15: 621–627.

Birchler JA, Auger DL, Riddle NC (2003). In search of a molecular basis of heterosis. Plant Cell 15: 2236–2239.

Boeke JD, Devine SE (1998). Yeast retrotransposons: finding a nice quiet neighborhood. Cell. 93: 1087–1089.

Bradshaw HD, Ceulemans R, Davis J, Stettler R (2000). Emerging model systems in plant biology: poplar (*Populus*) as a model forest tree. J. Plant Growth Regul. 19: 306-313.

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A.(2005). Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell 17: 343–360.

Brunner S, Pea G, Rafalski A (2005). Origins, genetic organization and transcription of a family of non-autonomous helitron elements in maize. Plant Journal 43: 799-810.

Buckler ES, Thornsberry JM (2002). Plant molecular diversity and applications to genomics. Curr. Opin. Plant Biol. 5: 107–111.

Bushman FD (2003). Targeting survival: Integration site selection by retroviruses and LTR-retrotransposons. Cell 115: 135–138.

Ceulemans R, Scarascia-Mugnozza G, Wiard BM (1992). Production physiology and morphology of *Populus* species and their hybrids grown under short rotation. I. Clonal comparisons of 4-year growth and phenology. Can. J. For. Res. 22: 1937-1948.

Chalker DL, Sandmeyer SB (1992). Ty3 integrates within the region of RNA polymerase III transcription initiation. Genes Dev. 6: 117–128.

Church GM, Kieffer-Higgins S (1988). Multiplex DNA sequencing. Science 240: 185–188.

Cong B, Liu J, Tanksley SD (2002). Natural alleles at a tomato fruit size quantitative trait locus differ by heterochronic regulatory mutations.

PNAS 99: 13606–13611.

Crow JF (1948). Alternative hypotheses of hybrid vigor. Genetics 33: 477–487.

Devos KM, Brown JKM, Bennetzen JL (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res. 12: 1075–1079.

Dillen SY, Marron N, Bastien C (2007). Effects of environment and progeny on biomass estimations of five hybrid poplar families grown at three contrasting sites across Europe. For. Ecol. Manage 252: 12-23.

Doebley J, Lukens L (1998). Transcriptional regulators and the evolution of plant form. Plant Cell 10: 1075–1082.

Domansky AN, Kopantzev EP, Snezhkov EV, Lebedev YB, Leib-Mosch C, Sverdlov ED (2000). Solitary HERV-K LTRs possess bi-directional promoter activity and contain a negative regulatory element in the U5 region. FEBS Lett. 472: 191–195.

Edwards A, Voss H, Rice P et al. (1990). Automated DNA sequencing of the human HPRT locus. Genomics 6: 593–608.

Flavell AJ, Pearce SR, Kumar A. (1994). Plant transposable elements and the genome. Curr. Opin. Genet. Dev. 4: 838–844.

Fu H, Dooner HK (2002). Intraspecific violation of genetic colinearity and its implications in maize. PNAS 99: 9573–9578.

Gardner RC, Howarth AJ, Hahn P, Brown-Luedi M, Shepherd RJ, Messing J (1981). The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. Nucleic Acids Res. 9: 2871–2888.

Gindullis F, Desel C, Galasso I, Schmidt T (2001). The large scale organization of the centromeric region in Beta species. Genome Res 11: 253–265.

Gogvadze E, Buzdin A (2009). Retroelements and their impact on genome evolution and functioning. Cell. Mol. Life Sci. 66: 3727–3742 .

Goodwin T, Poulter R (2004). A new group of tyrosine recombinase-encoding retrotransposons. Mol. Biol. Evol. 21: 746–759.

Grandbastien MA, Lucas H, Morel JB, Mhiri C, Vernhettes S, Casacuberta JM (1997). The expression of the tobacco Tnt1 retrotransposon is linked to plant defense responses. Genetica 100: 241–252.

Grandbastien MA, Audeon C, Bonnivard E et al. (2005). Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. Cytogenet. Genome Res. 110: 229–241.

Green ED (2008). Strategies for the systematic sequencing of complex genomes. Nature Rev. 2: 573-583.

Guo M, Rupe MA, Zinselmeier C, Habben J, Bowen BA, Smith OS (2004). Allelic variation of gene expression in maize hybrids. Plant Cell 16: 1707–1716.

Hashida SN, Kitamura K, Mikami T, Kishima Y (2003). Temperature shift coordinately changes the activity and the methylation state of transposon Tam3 in Antirrhinum majus. Plant Physiol. 132: 1207–1216.

Heilman PE, Ekuan G, Fogle C (1994). Above- and below-ground biomass and fine roots of 4-year-old hybrids of *Populus trichocarpa* × *Populus deltoides* and parental species in short-rotation culture. Can. J. For. Res. 24: 1186-1192.

Heslop-Harrison JS, Brandes A, Taketa S et al (1997). The chromosomal distribution of Ty1-*copia* group retrotransposable elements in higher plants and their implications for genome evolution. Genetica 100: 197–204.

Iglesias AR, Kindlund E, Tammi M, Wadelius C (2004). Some microsatellites may act as novel polymorphic cis-regulatory elements through transcription factor binding. Gene 341: 149-165.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Kimmel BE, Palazzolo MJ, Martin CH, Boeke JD, Devine SE. In Genome Analysis: A laboratory manual. Analyzing DNA (eds Birren B. et al.) 455–532 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1997).

Kohler A, Delaruelle C, Martin D, Encelot N, Martin F (2003). The poplar root transcriptome: analysis of 7000 expressed sequence tags. FEBS Lett. 542: 37–41.

Lai WR, Johnson MD, Kucherlapati R, Park PJ (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics 21: 3763–3770.

Lamb JC, Shakirov EV, Shippen DE. In Plant Cytogenetics. Plant Genetics and Genomics: Crops and Models, Volume 4, Part 2, 143-191 (SpringerLink Press, 2012)

Li W, Zhang P, Fellers JP, Friebe B, Gill BS (2004). Sequence composition, organization, and evolution of the core Triticeae genome. Plant J 40: 500–511.

Lisch D (2009). Epigenetic regulation of transposable elements in plants. Annu. Rev. Plant Biol. 60: 43–66.

Liu Z, Yue W, Li D et al. (2008). Structure and dynamics of retrotransposons at wheat centromeres and pericentromeres. Chromosoma 117: 445–456.

Mackay TFC (2001). Quantitative trait loci in Drosophila. Nat. Rev. Genet. 2: 11–20.

Mardis ER (2008). Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 9: 387-402.

Matlik K, Redik K, Speek M (2006). L1 antisense promoter drives tissue-specific transcription of human genes. J Biomed Biotechnol 2006: 71753

Maxam AM, Gilbert W (1977). A new method for sequencing DNA. PNAS 74: 560–564.

Metzker ML (2010). Sequencing technologies—the next generation. Nat. Rev. Genet. 11: 31-46.

Messing J (2001). The universal primers and the shotgun DNA sequencing method. Methods Mol. Biol. 167: 13–31

Morgante M (2005). Plant genome organisation and diversity: the year of the junk! Curr. Opin. Biotechnol. 17: 168–173.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5: 621–628.

Muhle Larsen C (1970). Recent advances in poplar breeding. Int. Rev. For. Res. 3: 1–67.

Nanjo T, Futamura N, Nishiguchi M, Igasaki T, Shinozaki K, Shinohara K (2004). Characterization of full-length enriched expressed sequence tags of stress-treated poplar leaves. Plant Cell Physiol 45: 1738–1748.

Noor MA, Chang AS (2006). Evolutionary genetics: jumping into a new species. Curr. Biol. 16: 890–892.

Osborn A J, Elledge SJ (2003). Mrc1 is a replication fork component whose phosphorylation in response to DNA replication stress activates Rad53. Genes Dev. 17:1755-1767.

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

Pardue ML, Danilevskaya ON, Traverse KL, Lowenhaupt K. (1997). Evolutionary links between telomeres and transposable elements. Genetica 100: 73–84

Peterson-Burch BD, Nettleton D, Voytas DF (2004). Genomic neighbourhoods for Arabidopsis retrotransposons: a role for targeted integration in the distribution of the Metaviridae. Genome Biol 5: R78.

Pich U, Schubert I (1998). Terminal heterochromatin and alternative telomeric sequences in Allium cepa. Chromosome Res 6: 315–321.

Presting GG, Malysheva L, Fuchs J, Schubert I (1998). A Ty3/*gypsy* retrotransposon-like sequence localized to the centromeric regions of cereal chromosomes. Plant J 16: 721–728.

Ranney JW, Wright LL, Layton PA (1987). Hardwood energy crops: the technology of intensive culture. Journal Forest 85: 17-26.

Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. PNAS 74: 5463–5467

Santini S, Cavallini A, Natali L et al (2002). Ty1/*copia*- and Ty3/*gypsy*-like DNA sequences in Helianthus species. Chromosoma 111: 192–200.

Scherrer B, Isidore E, Klein P et al. (2005). Large intraspecific haplotype variability at the Rph7 locus results from rapid and recent divergence in the barley genome. Plant Cell 17: 361–374.

Shull GH (1908). The composition of a field of maize. American Breeders Assoc. Rep. 4: 296–301.

Sixto H, Aranda I, Grau JM (2006). Assessment of salt tolerance in *Populus alba* clones using chlorophyll fluorescence. Photosynthetica 44: 169-173.

Slotkin KR, Martienssen R (2007). Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 8: 272–285.

Smith DR, Doucette-Stamm LA, Deloughery C (1997). Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics. J Bacteriol. 179: 7135–7155.

Song R, Messing J (2003). Gene expression of a gene family in maize based on noncollinear haplotypes. PNAS 100: 9055–9060.

Springer NM, Stupar RM (2007). Allelic variation and heterosis in maize: How do two halves make more than a whole? Genome Res 17: 264-275.

Sterky F, Regan S, Karlsson J et al. (1998). Gene discovery in the wood-forming tissues of poplar: analysis of 5,692 expressed sequence tags. PNAS 95: 13330–13335.

Sterky F, Bhalerao RR, Unneberg P et al. (2004). A *Populus* EST resource for plant functional genomics. PNAS 101: 13951-13956.

Stanton E, Ungerer M, Moore R (2009). The genomic organization of Ty3/*gypsy*-like retrotransposons in Helianthus (Asteraceae) homoploid hybrid species. American Journal Of Botany 96: 1646–1655.

Tanksley S (1993). Mapping polygenes. Annu. Rev. Genet. 27: 205–233.

Taylor G (2002). *Populus*: *Arabidopsis* for forestry. Do we need a model tree? Ann. Bot. (Lond.) 90: 681-689.

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

Ungerer MC, Strakosh SC, Zhen Y (2006). Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. Curr. Biol. 16: 872–873.

van de Lagemaat LN, Medstrand P, Mager DL (2006). Multiple effects govern endogenous retrovirus survival patterns in human gene introns. Genome Biol 7: R86.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995). Serial analysis of gene expression.

Science 270: 484–487

Wessler SR, Bureau TE, White SE (1995). LTR-retrotransposons and MITES, important players in the evolution of plant genomes. Curr. Opin. Genet. Dev. 5: 814– 821.

Wicker T, Sabot F, Hua-Van A et al. (2007). A uni fied classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8: 973-982.

Wray GA, Hahn MW, Abouheif E et al. (2003). The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. 29: 1377-1419.

Zou S, Ke N, Kim JM, Voytas DF (1996). The Saccharomyces retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. Genes Dev.10: 634–645.

Paper I

A computational study of the dynamics of LTR-retrotransposons in the *Populus trichocarpa* genome

Rosa Maria Cossu · Matteo Buti · Tommaso Giordani · Lucia Natali · Andrea Cavallini

Abstract Retrotransposons are a ubiquitous component of plant genomes, especially abundant in species with large genomes. *Populus trichocarpa* has a relatively small genome which was entirely sequenced, however studies focused on poplar retrotransposons dynamics are rare. With the aim to study the retrotransposon component of the poplar genome, we have scanned the complete genome sequence searching full-length LTR retrotransposons, i.e. characterised by two long terminal repeats at the 5' and 3' ends. A computational approach based on detection of conserved structural features, on building multiple alignments, and on similarity searches was used to identify 1,479 putative full-length LTR-retrotransposons. Ty1-*copia* elements were more numerous than Ty3-*gypsy*. However, many LTR-retroelements were not assigned to any superfamily because lacking of diagnostic features and non-autonomous. LTR-retrotransposon remnants were by far more numerous than full-length elements, indicating that during the evolution of poplar, large amplification of these elements was followed by DNA loss. Within superfamilies, Ty3-*gypsy* families are made of more members than Ty1-*copia* ones. Retrotransposition occurred with increasing frequency following the separation of *Populus* sections, with different waves of retrotransposition activity between Ty3-*gypsy* and Ty1-*copia* elements. Recently inserted elements appear more frequently expressed than older ones. Finally, different levels of activity of retrotransposons were observed according to their position and their density in the linkage groups. On the whole, the results support the view of retrotransposons as a community of different organisms in the genome, whose activity (both retrotransposition and DNA loss) has heavily impacted and probably continues to impact poplar genome structure and size.

Keywords: *Copia, Gypsy*, LTR-retrotransposon, poplar genome, *Populus trichocarpa*.

R.M. Cossu ·M. Buti ·T. Giordani ·L. Natali ·A. Cavallini (author for correspondence)
Dipartimento di Biologia delle Piante Agrarie,
Università di Pisa,
Via del Borghetto 80,
I-56124 Pisa, Italy
e-mail: acavalli@agr.unipi.it

Abbreviations

RE	retrotransposon
LTR-RE	LTR-retrotransposon
LTR	long terminal repeat
MY	million of years
MYA	million years ago.

Introduction

Class I transposons or retrotransposons (REs) represent the majority of the repetitive component of eukaryotic genomes. REs propagate via a "copy and paste" mechanism in which, after RE transcription, enzymes encoded by the RE synthesize dsDNA copies that are integrated back in the host genome. This mechanism resembles the replication cycle of retroviruses (Wicker et al. 2007).

REs can be separated into LTR- and non LTR-retrotransposons, depending on the presence of long terminal repeats (LTRs) flanking the coding portion at both 5'- and 3'-ends. Such repeats are identical at the time of insertion of the new element in the chromosome. They range from a few hundreds to several thousands base pairs in length. LTR-retrotransposon (LTR-RE) transcription starts in the 5'-LTR, where the TATA box usually occurs; within LTR, *cis*-regulatory motifs can be found that regulate RE transcription (Sugimoto et al. 2000). An LTR is typically delimited by two dinucleotides TG...CA, has terminal inverted repeats (TIRs) of 6 bp and is flanked by target site duplications (TSDs) of 4-6 bp. Both TIR and TSD may however be imperfect as result of mutations subsequent to LTR-RE insertion.

Internal to the 5' and 3' LTRs, respectively, are present the primer binding site (PBS) and the polypurine tract (PPT). They provide the signals for reverse transcription of RE transcripts into the cDNA that will be integrated in the genome. The PBS is complementary to a portion of a host encoded tRNA, which can act as a primer for retrotranscription (Wicker et al. 2007).

The two LTRs flank an internal portion that typically contains one or more open reading frames (ORFs) encoding the enzymes for retrotransposition (Boeke and Corces 1989; Kumar and Bennetzen 1999); *gag* (encoding a capsid protein) and *pol* (encoding aspartic proteinase, integrase, reverse transcriptase and RNaseH).

LTR-REs are subdivided into autonomous and non-autonomous elements, depending on the presence, in the internal region flanked by LTRs, of genes encoding the retrotransposition machinery.

Among autonomous LTR-REs, superfamilies Ty1-*copia* and Ty3-*gypsy* differ in the enzyme order within *pol* (Wicker et al. 2007). Both superfamilies are ubiquitous throughout the eukaryotes and have been present since the divergence of plants, animals and fungi.

Non-autonomous LTR-REs have the PBS, PPTs, and LTRs needed for transcription, replication, and integration as cDNA (Sabot and Schulman 2006) but they do not carry genes for retrotransposition and are mobilized in trans using enzymes produced by autonomous LTR-REs. Among non-autonomous LTR-REs, two main groups have been described: *T*erminal-repeat *R*etrotransposons *I*n *M*iniature (TRIMs) and *LA*rge *R*etrotransposon *D*erivatives (LARDs) (Witte et al. 2001; Kalendar et al. 2004).

Because of the error-prone nature of transcription and reverse transcription, the replicative mechanism of LTR-REs has generated different families. LTR-RE sequence heterogeneity is found in the coding, transcribed portion, and especially in the LTRs (Beguiristain et al. 2001).

The replicative activity of retrotransposons has determined the structure of eukaryotic genomes. Genome expansion by insertion of REs occurred frequently during evolution; on the other hand, retrotransposons have been the object of sequence removal - and, in part, they also have favoured DNA loss - mediated by unequal homologous recombination or by illegitimate recombination (Devos et al. 2002; Ma et al. 2004; Grover et al. 2008). The rates of both genome expansion and genome contraction processes appear to vary between species (Bennetzen et al. 2005; Vitte and Bennetzen 2006), allowing some genomes to shrink while others expand. Within a genome, for example in rice, the occurrence of illegitimate and unequal homologous recombination can be related to the gene density, being higher in coding sequences rich regions (Tian et al. 2009). Rearrangements, illegitimate and unequal homologous recombination are the processes driving DNA removal in plants by multiple mechanisms, including repair of double-strand breaks (nonhomologous end-joining) and slipstrand mispairing (Kalendar et al. 2000; Ma and Bennetzen 2004; Neumann et al. 2006; Ammiraju et al. 2007; Hawkins et al. 2008; Morse et al. 2009).

A survey of the dynamics of different RE superfamilies in eukaryotic genomes is facilitated by the availability of whole genome sequence or, at least, sequence of large portions of the genome, as BAC clones. In plants, LTR-REs have been largely surveyed in species whose genome has been entirely sequenced and in species for which the sequence of large portions of the genome are available. *Gypsy* and *Copia* superfamilies are differently represented in the genome, depending on the species, with respective ratios of 5:1 in papaya (Ming et al. 2008), 4:1 in *Sorghum* (Paterson et al. 2009), 3:1 in rice (The International Rice Genome Sequencing Project 2005), 1:2 in grapevine (The French-Italian Public Consortium for Grape Genome Characterization 2007). Maize shows a similar abundance of the two classes (Meyers et al. 2001), with *Gypsy* elements especially concentrated in gene-poor regions and *Copia* REs overrepresented in gene-rich ones (Schnable et al. 2009; Baucom et al. 2009a). Similar data are reported for other cereal species with large genomes such as wheat and barley (Vicient et al. 2005; Paux et al. 2006). Species of the *Gossypium* genus show a variable proportion of *Gypsy* versus *Copia* elements with *Gypsy* elements prevailing in species with larger genome sizes (Hawkins et al. 2006).

Recent reports have shown that retrotransposon sequences can have an impact on the expression of nearby genes (Kashkush et al. 2003) by their presence or absence in the cis-regulatory sequences of genes of the host species. Therefore, the identification and characterisation of LTR-REs is a priority in analyzing the genome of crop species.

Among species whose genome has been sequenced, poplar (*Populus trichocarpa*), grapevine and papaya are the only perennial plants and it is plausible that perennial habit affects genome dynamics in a different way from annually sexually propagated species.

In their report on poplar genome sequencing, Tuskan et al. (2006) reported that class I elements (Ty1-*copia*-like, Ty3-*gypsy*-like, LINEs, and unidentified retroelements) are the most abundant (over 5000 copies). Poplar genome is relatively small (550 Mbp) and retroelements cover approximately 176 Mbp (42% of the genome). A prevalence of *Gypsy* over *Copia* RE sequences was reported (Tuskan et al. 2006), however unidentified elements account for 120 Mbp.

Recently, a database of repetitive elements (RepPop) has been released (Zhou and Xu 2009). However, a comprehensive analysis of LTR-retrotransposon dynamics in the poplar genome is still not available (Klevebring et al. 2009). With the aim of studying the dynamics of LTR-retrotransposons in the poplar genome, we identified putative full-length retrotransposons based on the occurrence of both LTRs and established phylogenetic relationships among them according to LTR sequence similarity.

Materials and Methods

LTR-REs identification

Putative LTR-REs were identified in the sequenced genome of *P. trichocarpa* (Tuskan et al. 2006) deposited at EMBL (acc. number AARH00000000.1) using LTR FINDER software (Xu and Wang 2007). LTR-FINDER uses a suffix-array based algorithm to construct all exact matching pairs, which are extended to long highly similar pairs. Alignment boundaries are obtained adjusting the ends of LTR pair candidates using the Smith-Waterman algorithm. These boundaries are re-adjusted, based on the occurrence of typical LTR-RE features such as: i) being flanked by the dinucleotides TG and CA, at 5' and 3' ends, respectively; ii) the presence of a TSD of 4-6 bp; iii) the presence of a putative PBS, complementary to a tRNA at the end of putative 5'-LTR; iv) the occurrence of a putative polypurine tract just upstream of the 5' end of the 3' LTR. The following parameters were used: LTR sequence length from 80 to 5,000 bp, maximum distance between LTRs 20,000 bp. The sequences between two putative LTRs were subsequently analysed by BLASTX and BLASTN searches (E-value threshold 10^{-5}) against public non-redundant databases at GenBank and against REPBASE (Jurka et al. 2005). Sequences are available at the Dept. of Crop Biology of Pisa University repository website (http://www.agr.unipi.it/Sequence-Repository.358.0.html).

All sequences were masked against RepPop database (Zhou and Xu 2009) using RepeatMasker (developed by A.F.A. Smit, R. Hubley, and P. Green;(http://www.repeatmasker.org/).

LTR-REs were annotated using both structure- and homology-based methods. Relationships between LTR-REs were established according to sequence similarity between LTRs. All putative LTRs were clustered using CAP3 software (Huang and Madan 1999) using an overlap length cut off of 80% and an overlap identity cut off of 80%, following the guidelines for transposable element annotation proposed by Wicker et al. (2007).

Mutation rate estimation

Based on the estimation that separation between *Tacamahaca* and *Populus* sections (to which *P. trichocarpa* and *P. alba* belong, respectively) occurred in the Miocene between 18 and 23.3 MYA (Eckenwalder 1996), a synonymous substitution rate was calculated comparing protein coding sequences of *P. alba* (Maestrini et al. 2009) to orthologous sequences in the *P. trichocarpa* genome. Thirty-one sequences (longer than 320 bp) out of 150 available *P. alba* sequences (aligned at high similarity (> e^{-80}) with only one sequence in the *P. trichocarpa* genome) were selected for analysis. Rates of synonymous and nonsynonymous nucleotide substitution for each gene were calculated by the method of Nei and Gojobori (1986) with the Jukes–Cantor correction as implemented in the DnaSP program (Rozas and Rozas 1999). The average synonymous substitution number for 31 genes was estimated.

LTR-REs insertion time estimation

Retrotransposon insertion age was estimated comparing the 5'- and 3'-LTRs of each putative full-length retrotransposon. The two LTRs of a single retrotransposon are identical at the time of insertion because they are mostly copied from the same template. The two LTRs were aligned with ClustalX software (Thompson et al. 1994), indels were eliminated and the number of nucleotide substitutions were counted using the DnaSP program (Rozas and Rozas 1999). The insertion times of retrotransposons with both LTRs were dated using the Kimura two parameter method (K2P, Kimura 1980), calculated using DnaSP, and a synonymous substitution rate that is two-fold the one calculated for genes, according to SanMiguel et al. (1998) and to Ma and Bennetzen (2004).

LTRs copy number estimation

To estimate the number of LTR-RE remnants in the genome we have measured the number of hits obtained by BLASTN searches against *P. trichocarpa* genome at Genbank (http://blast.ncbi.nlm.nih.gov/Blast.cgi) using the LTR sequences of each putative full-length LTR-RE as queries. The occurrence of sequences with at least 80% similarity to putative LTRs in EST databases of *P. trichocarpa* was scored by BLASTN search against such databases at the same NCBI site (E-value threshold 10^{-5}).

Other sequence and statistical analyses

In other analyses, we used the TandemRepeat Finder program (Benson 1999) in conjunction with BLAST analysis against poplar genome at NCBI, to search putative centromeric repeats.

Statistical analyses were carried out using GraphPad Prism Software.

Results

Identification and classification of REs with complete LTRs

An intact LTR retrotransposon was defined as one that contains two relatively intact LTRs and identified PPT and PBS sites and is also flanked by TSDs (Ma et al. 2004), irrespective of encoding or not enzymes for retrotransposition. Using this definition, we started our analyses searching for every sequence flanked by two highly similar sequences longer than 80 bp and with the above specified typical features.

We mined putative LTR-REs of poplar from the entire *P. trichocarpa* genome using LTR-FINDER software (Xu and Wang 2007). False positives were eliminated by careful checking each sequence separately. To estimate the frequency of false negatives we masked the sequence of chromosome I with all identified poplar LTR-REs using RepeatMasker. Then, the masked and unmasked sequences of chromosome I were analyzed by tBLASTn using two poplar sequences, a *Copia* retrotranscriptase and a *Gypsy* integrase. The unmasked chromosome I showed 172 hits for the *Copia* sequence and 88 hits for the *Gypsy* sequence; the masked chromosome I showed only one *Gypsy* sequence that revealed a retrotransposon fragment. Hence, we estimated that the number of false negatives was negligible.

On the whole, we collected 325 intact elements. Moreover, putative LTR REs with two or one of the above described three typical LTR-RE features (PPT, PBS, and TSD) were identified (1,150 and 4 elements, respectively). Hereafter, the complete set of 1,479 putative LTR-REs are referred as full-length LTR-REs. Their sequences are available at the Dept. of Crop Biology of Pisa University repository website (http://www.agr.unipi.it/Sequence-Repository.358.0.html). see also Supplemental file 1).

The collected elements were masked against repetitive sequences present in the RepPop database (Zhou and Xu 2009) using RepeatMasker. Beside the overlaps, there are significant portions unique to both sets. Forty-three per cent of bases of our dataset resulted unmasked. Moreover, 132 out of 1,479 LTR-REs resulted masked only for 0-15% of their sequence, hence can be considered as specific to our dataset.

Nearly all elements found using this approach are isolated, i.e., apparently adjacent to sequences of the host genome. In only 31 loci were we able to recognise nested elements, i.e. an element within another one. We cannot exclude the possibility that more complex nested structures are present in the poplar genome, as observed for example in maize (SanMiguel et al. 1996). However, we decided to limit our search to full-length and linear elements to analyse a homogeneous RE sample.

The recorded putative LTRs had a mean length of 566 bp, but large length variability was observed (up to 4,848 bp, standard deviation = 631.82 bp). As for full-length retrotransposons, the mean length was 7,225 bp, again with a large standard deviation (5,436 bp).

The full-length LTR-REs were compared with the GenBank nr database by BLAST analysis (E-value threshold 10^{-5}) to explore whether sequences encoding RE enzymes were present. Of 1,479 putative LTR-REs, only 595 (40.2%) were found to contain at least one of the coding domains needed for retrotransposition.

LTR-REs were first classified as belonging to Ty3-*gypsy*, Ty1-*copia*, or Unknown superfamilies according to BLAST analysis of their internal portion (i.e. between LTRs) in comparisons with GenBank and REPBASE databases. Table 1 reports the number of full-length Ty1-*copia*-like, Ty3-*gypsy*-like and Unknown LTR-REs identified in the poplar genome. Unknown putative elements are the most represented in our sample, followed by Ty1-*copia*-like and *Ty3-gypsy*-like ones.

Concerning Unknown full-length elements (855 LTR-REs), in some cases BLAST analysis showed the presence of coding sequences with similarity to non-LTR retrotransposons (34 elements), to DNA transposons (44 elements), or to helitrons (6 elements) between the putative LTRs. These elements possibly originated by insertion of such sequences in previously existing LTR-REs. In 41 cases BLAST analysis showed the occurrence of *pol* or *gag* encoding sequences, but the attribution to a superfamily was not allowed. The internal domain of other Unknown LTR-REs (730 elements) lacked strong homology to any known LTR-RE proteins.

According to Wicker et al. (2007), all elements lacking typical LTR-RE protein encoding sequences can be classified as TRIMs when they had a length less than 4 Kbp and as LARDs when longer than 4 Kbp. On the whole, elements not showing any RE enzyme coding portion, or elements containing sequences with similarity to DNA transposons or non LTR-REs and not sharing their LTR sequence with any *Copia* or *Gypsy* superfamily were classified as Unknown (Wicker et al. 2007).

Chromosome distribution of LTR-REs

Table 1 reports the number of full-length LTR-REs in the 19 linkage groups (LGs) of *P. trichocarpa*. The putative full-length REs identified in our analysis represent 3.47% of the poplar genome, i.e. a mean of one full-length retroelement every 208,141 bp. The distribution in the 19 LGs is somewhat different, from 6.29% in the LG XIX to 2.00% in the LG VI. *Copia* LTR-REs are especially frequent in the LG I. *Gypsy* LTR-REs are more frequent than *Copia* in 5 out of the 19 LGs.

In Figure 1 and Supplemental file 2 the distribution of the 1,479 LTR-REs on the 19 linkage groups of *P. trichocarpa* is reported. REs are mostly dispersed throughout the chromosomes. Unfortunately, the current *Populus* genome sequence does not annotate the centromeric regions (Klevebring et al. 2009). Moreover, a complete cytogenetic map of the poplar, based on linkage groups as determined by whole genome sequencing, is still to be established (see Islam-Faridi et al. 2009). The fact that, in some cases, *Gypsy*-like and Unknown LTR-REs are especially clustered in one chromosome position, might suggest that this is the centromere position, where *Gypsy* REs are usually very frequent (Santini et al. 2002 and references therein).

To determine if clustered LTR-REs are actually centromeric, we searched for putative centromeric satellites in the poplar genome using the TandemRepeat Finder software. We identified two types of putative centromeric repeats. The first type, whose consensus sequence is 107 bp long, should allow the identification of the centromere position in chromosomes IV, V, VIII, X, XI, XII, XIII, XIV, and XV. The second, a consensus sequence 142 bp long, should identify the centromere of chromosomes I, III, IX, XVI, XVIII and XIX (Cossu, unpublished, see Supplemental file 3). No putative centromeric repeats were found in chromosomes II, VI, VII, and XVII, probably because of underrepresentation of repetitive sequences in the currently available poplar genome sequence (Klevebring et al. 2009). It is to be noted that the 142 bp long sequence shows high similarity to a 145 bp tandem repeat sequence isolated by Rajagopal et al. (1999) in *Populus deltoides* and *P. ciliata*, that was described as putatively centromeric.

We overlapped a map track of putative centromeric repeats for each chromosome with the distribution of *Copia*, *Gypsy*, and Unknown LTR-REs along chromosomes (Figure 1 and Supplemental files 2a and 2b). In all chromosomes in which the centromere position seemed to be identified, there was a significant overlap between the putative centromeric position and the accumulation of full-length *Gypsy* LTR-REs, suggesting the association between centromeric repeats and *Gypsy* LTR-REs. It is however to be recalled that the definition of the centromere position requires biochemical and cytological validation, for example by BAC in situ hybridization (Islam-Faridi et al. 2009).

Family distribution and frequency of LTR-REs in the poplar genome

Usually, structural and sequence similarities are used for the classification of non-autonomous LTR retrotransposons into families; such a classification is used, for example, in Repbase, a database of eukaryotic repetitive and transposable elements (Jurka et al. 2005). Wicker et al. (2007) established application rules to a hierarchical transposable element classification similar to that used in Repbase and defined a family of retrotransposons as a group of REs that have high DNA sequence similarity in their coding region (if present) or internal domain, or in their LTR. Specifically, they proposed that two REs are assumed to belong to the same family if at least 80% of the aligned sequence (LTRs, or internal portion, or both) show 80% or more similarity, analyzing segments longer than 80 bp.

We classified the full-length LTR-REs into families based on their LTR sequence similarity. We used LTR sequences to classify families rather than more commonly used retrotranscriptase (RT) coding domain sequences because many nonautonomous LTR-REs lack an intact RT domain.

The set of 1,479 LTR pairs (longer than 80 bp) were compared using CAP3 algorithm, setting 80% identity of 80% LTR length, with reference to the so called 80-80-80 rule, according to Wicker et al. 2007). A schematic representation of LTR alignments of the four most redundant *Gypsy* families are reported in Figure 2 as an example. In the case of the G126 family, all 12 LTRs overlap. In the other cases in Figure 2, overlapping is not complete; some LTRs do not share their sequence with other LTRs that have been attributed to the same family. Such attribution is justified because if members A and B fulfil the 80-80-80 rule, then they should belong to the same family and, if members B and C also fulfil that rule, then also members A and C should belong to the same family, because they should share a common ancestor. Such

Table 1 Number of full-length LTR-retrotransposons in the 19 linkage groups of *P. trichocarpa*. For each linkage group, length, percentage of full-length LTR-REs (calculated as the ratio between total length of LTR-REs in a chromosome and the total length of that chromosome), full-length LTR-RE density (the mean number of bp between two LTR-REs), and the mean insertion date (MY) are reported

Linkage Group	Nr. LTR-REs	Nr. *Copia* REs	Nr. *Gypsy* REs	Nr. Unknown LTR-REs	Chromosome length (bp)	% LTR-REs	LTR-RE density	Mean insertion date
I	173	58	27	88	35,571,569	3.09%	205,616	9.3
II	92	22	11	59	24,482,572	2.47%	266,115	12.2
III	86	22	14	50	19,129,466	3.39%	222,436	9.3
IV	107	25	17	65	16,625,654	4.70%	155,380	11.3
V	65	15	14	36	17,991,592	2.24%	276,794	9.2
VI	58	15	6	37	18,519,121	2.00%	312,911	10.6
VII	43	8	6	29	12,805,987	2.17%	291,338	11.2
VIII	55	12	19	24	16,228,216	2.64%	295,058	7.9
IX	36	8	8	20	12,525,049	2.11%	347,918	9.4
X	98	24	13	61	21,101,489	3.38%	208,046	10.1
XI	84	19	21	44	15,120,528	4.58%	171,755	10.4
XII	91	12	21	58	14,142,880	4.44%	148,513	12.1
XIII	83	23	18	42	13,101,108	5.09%	157,845	9.8
XIV	59	9	16	34	14,699,529	3.06%	241,529	10.2
XV	55	17	4	34	10,599,685	4.26%	184,504	10.1
XVI	80	18	9	53	13,661,513	4.05%	170,769	10.7
XVII	45	6	11	28	6,060,117	5.07%	134,669	12.1
XVIII	73	20	17	36	13,470,992	4.61%	175,790	10.1
XIX	96	25	14	57	12,003,701	6.29%	125,039	10.8
Total	1479	358	266	855	307,840,768	3.47%	208,141	10.3

transitivity might induce errors in classification, as reported by Seberg and Petersen (2009). However, no alternatives have been proposed at present.

Based on this classification, in some cases, autonomous, defective, and non-autonomous elements could be attributed to one and the same family, even in the absence of the coding portion. In such cases, we assumed that non-autonomous and defective elements originated from autonomous elements with which they share LTR sequence.

One-hundred-twenty-six LTR-RE families were established by this method. Nine hundreds-eighty-one elements did not cluster and remained single. The mean number of full-length elements per family was 3.94. The distribution of LTR-RE families in relation to the number of components is reported in Figure 3. The vast majority of families comprise 2-3 components and only 10 families had more than 8 components. *Copia* and *Gypsy* families were also analysed separately and *Gypsy* families resulted more redundant than *Copia* ones (Figure 3). The majority of *Copia* and *Gypsy* families were specific to poplar. Analysis using RepBase showed, in four cases, similarity to *Tto1 Copia* elements of *Nicotiana tabacum*. Some *Gypsy* families were similar to *Diaspora* elements of *Asparagus officinalis*.

In another analysis, the LTR sequence of each full-length RE was compared to the whole poplar genome to measure the frequency of LTR-RE remnants containing that LTR, hence belonging to the same LTR-RE family. The LTR-RE remnants include solo-LTR and isolated LTR fragments, and REs with only one complete or fragmented LTR. The frequency of RE remnants was calculated for each LTR-RE family (126 entries) and for single LTR-REs (981 entries) keeping *Copia*, *Gypsy*, and Unknown elements separate (Table 2).

A correlation occurs between number of full-length LTR-REs and number of LTR-RE remnants (not shown); accordingly, the most numerous family (G011) showed the highest number of LTR-RE remnants in the genome. The mean number of LTR-RE remnants per family or single LTR-RE is by far higher for *Gypsy* than for *Copia* elements.

The above described correlation is especially true for *Gypsy* elements, being not significant for *Copia* REs (not shown). This should indicate that retrotransposition activity and DNA loss (by rearrangements and by homologous and illegitimate recombination) of *Gypsy* elements is more ancient than that of *Copia* elements and/or that mechanisms of DNA loss in *Gypsy* elements are more efficient (possibly because they are longer than *Copia*).

Putative insertion dates of LTR-REs

The availability of both complete LTRs allows the insertion time of a LTR-RE to be estimated. Insertion time estimates are based on the occurrence of nucleotide substitutions in the LTRs, which are supposed to be identical at the retroelement insertion time, using a nucleotide substitution rate suitable for such elements (SanMiguel et al. 1998; Ma and Bennetzen 2004). It should be noted that the calculation of insertion date by the number of mutations in sister LTRs is subjected to error because it assumes the same mutation rate in all LTR-RE sequences and all chromosome positions. However, this method appears as the most suitable to study LTR-RE dynamics.

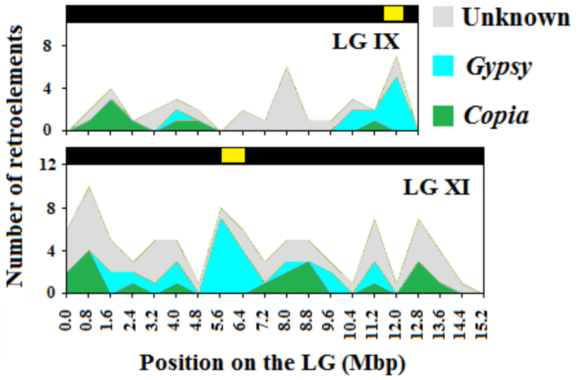


Fig. 1 Distribution of putative full-length *Gypsy*, *Copia* and Unknown LTR-REs in the linkage groups IX and XI of *P. trichocarpa*. The putative positions of centromeres, as indicated by the occurrence of centromeric repeats, are evidenced in the box in the black track map over the LTR-RE profiles

We estimated the synonymous substitution rate by comparing orthologous cDNA sequences of *P. alba* and *P. trichocarpa,* i.e. thirty-one coding sequences for a total of 18,344 bp. The mean number of synonymous substitutions per site (K_s) was 0.0483 (Table 3).

Based on the dating of fossil leaves in the second part of the Miocene, the separation between the sections Tacamahaca and Populus (to which *P. trichocarpa* and *P. alba* belong, respectively) is estimated as 18-23 MYA, i.e., a common ancestor should have existed in the early Miocene (Eckenwalder 1996, and references therein]. Recent data based on dating polyploidization events in different *Populus* species, indicates that genus speciation occurred 8-13 MYA (Sterck et al. 2005; Tuskan et al. 2006). The difference in dating *Populus* speciation was attributed to the use of substitution rates calculated in herbaceous monocots and dicots (Sterck et al. 2005), considering that the generation time of a species is known to affect its nucleotide-substitution rate (Gaut 1998) and that poplar has a much longer generation time than herbaceous species.

Assuming an average of 20.5 MY as insertion date and a K_s of 0.0483, the resulting synonymous substitution rate was 2.36 x 10^{-9} substitutions per years. It has been suggested that mutation rates for LTR-retrotransposons may be approximately twofold higher than silent site mutation rates for protein coding genes (Xu and Wang 2007). Consequently, a substitution rate per year of 4.72 x 10^{-9} was used in our calculations of LTR-RE insertion dates.

LTR pairs were compared in their sequence, excluding deletions from comparisons and the putative insertion date was calculated for each full-length LTR-RE based on the number of substituted nucleotides per site. When the whole set of usable retrotransposons was taken into account, the nucleotide distance (K) between sister LTRs showed large variation between retroelements (0 to 0.602, Kimura 2-parameter method), representing a time span of at most 124 million years. The putative mean age of analysed LTR-REs is 10.4 MY, with great variability (standard deviation = 8.9 MY). The distribution of full-length LTR-REs according to their putative insertion date is reported in Figure 4. As expected, since the most ancient LTR-REs should have accumulated the largest variations in their sequences (being not recognised by LTR-FINDER), the frequency of LTR-REs with older insertion date reduces progressively. Analysis of the insertion date profiles provides evidence for overlapping among retrotransposition waves of *Gypsy*, *Copia*, and Unknown full-length LTR-REs (Figure 4). When taking into consideration the last 20 MY (i.e. after the separation of poplar sections), peaks of retrotransposition by *Gypsy* and *Copia* elements alternate. However, it is to be considered that most full length LTR-REs were not assigned to any family. If *Gypsy* and *Copia*-related Unknown elements in this class were not distributed with nearly 1:1 ratio, different profiles would be observed.

The mean insertion dates of the most numerous

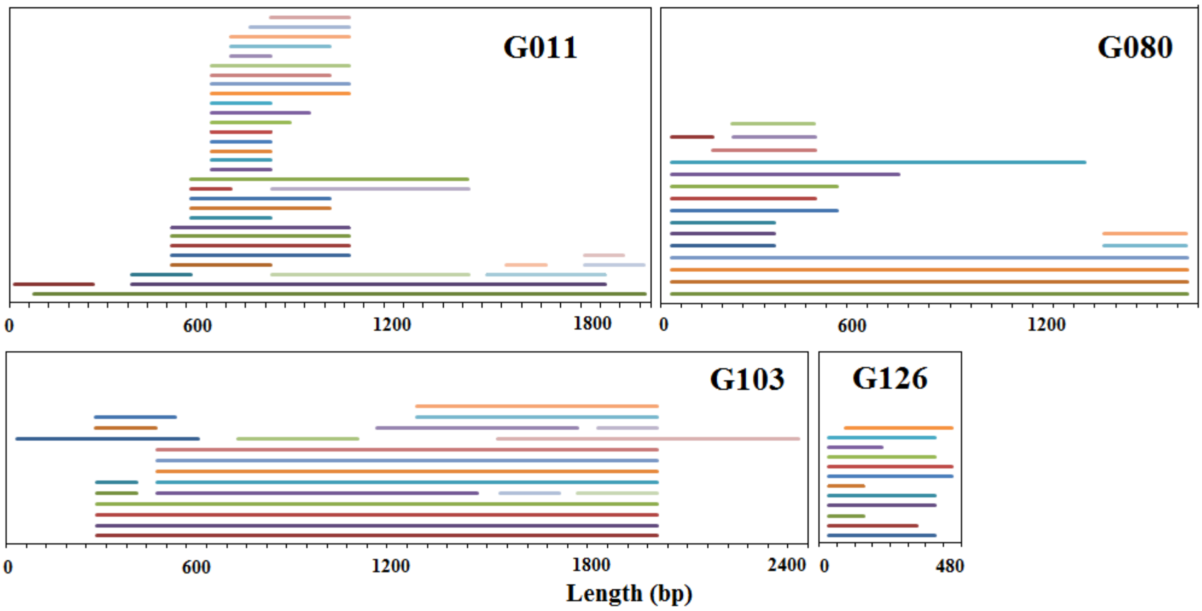


Fig. 2 Schematic representation of overlapping of LTR sequences (horizontal bars) in the four most repeated *Gypsy* families (G011, G080, G103, G126)

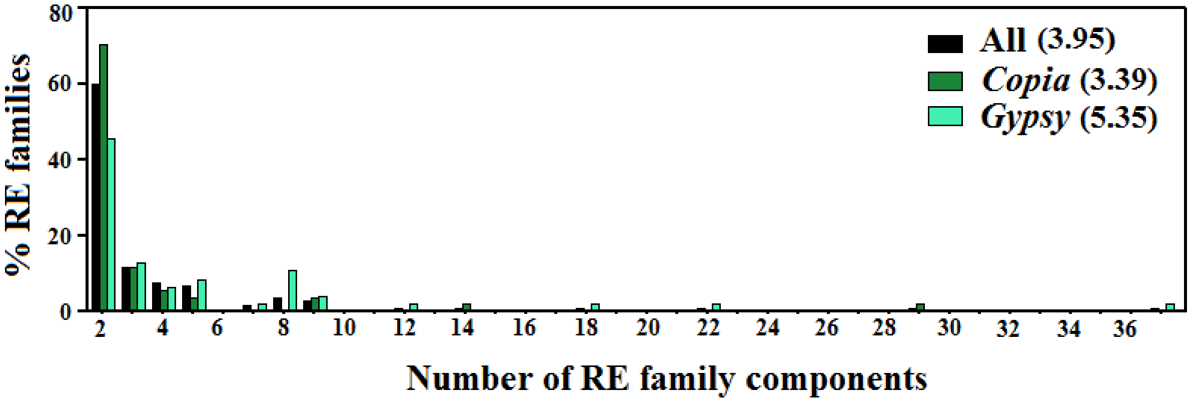


Fig. 3 Size distribution of LTR-RE families obtained using the CAP3 assembler. The histogram depicts the percentage of LTR-RE families (Y axis) containing a specified number of full-length LTR-REs (X axis)

Gypsy (6) and *Copia* (4) families (with number of full-length LTR-REs ≥ 9) show that different families underwent amplification in different time spans (Figure 5), as indicated also by one way ANOVA (Table 4).

The profiles of LTR-RE insertion age along the 19 linkage groups are reported in Figure 6 and Supplemental file 4. Comparisons between the profiles and the mean insertion age of each LG or of the entire genome suggest that retrotransposition occurred at different times in the different chromosomes and chromosome positions (see for example LGX), or that mutation rate changes according to chromosome positions. Actually, the concentration of older elements in pericentromeric regions might reflect the suppressed recombination in these areas (Tian et al. 2009).

Transcriptional activity of LTR-REs

The transcriptional activity of LTR-REs of our sample was computationally evaluated by BLASTN searches of putative LTR sequences against the available EST databases of *P. trichocarpa*. Such evaluation represents just a qualitative indication of RE activity, and it should be confirmed by RT-PCR experiments. The available EST collection includes 139,007 sequences from terminal vegetative buds (two libraries), young and mature leaves, along with green shoot tips (one library) phloem and cambium (one library), outer xylem (three libraries) (Ralph et al. 2006) and 17,727 sequences from male catkins, female catkins and floral buds (Sterky et al. 2004). We are conscious that RE-related EST might result from DNA contamination of the EST library, mostly because of the repetitiveness of RE sequences in the genomes. Moreover, finding ESTs with similarity to LTR sequences could be also related to the expression of siRNA: it has been shown that, in young leaves of poplar, the majority of 24 nt short RNA correspond to LTR elements (Klevebring et al. 2009). However, as we found numerous EST matches to LTR-RE sequences, this should be a strong indication that those elements are expressed. We established a threshold of 5 EST matches to consider a LTR-RE as transcriptionally active. The distribution of full-length *Copia*, *Gypsy*, and Unknown LTR-REs according to their expression and insertion date is reported in Table 5. Actually, for the vast majority (1188/1479) of LTR-REs, no match to EST sequences was found. The percentages of active full-length LTR-REs (with number of EST matches > 5) range from 3.91 (for *Copia* REs) to 11.65 (for *Gypsy* REs). *Gypsy* REs are apparently more active than *Copia* ones. Though variations are not significant, there is a tendency for completely inactive full-length LTR-REs (showing no EST matches) to be older than the mean of their superfamily, indicating that transcriptional activity is maintained mostly by young LTR-REs and ancient elements are repressed.

We also related RE transcriptional activity to the frequency of RE remnants for each family. Low copy number families are generally more expressed than highly redundant ones (not shown). This result confirms data in the literature that low copy number REs are the most active (Meyers et al. 2001; Yamazaki et al. 2001).

Relationship between RE density and activity

To study the effect of LTR-RE density on LTR-RE activity, we established two subsets of full-length LTR-REs; the first subset, called clustered LTR-REs, contained the elements found in 400,000 bp long regions in which at least 10 full-length LTR-REs are present; the second subset, called dispersed elements, contained the elements found in 1 million bp long regions, in which only one full-length element is present. A descriptive statistics of these two subsets compared to the entire sample of poplar LTR-REs is reported in Table 6. It is to be noted that the two subgroups are placed in opposition to the data of the entire set; LTRs of dispersed elements are less represented in the genome; these elements show lower transcriptional activity and are putatively younger than the entire full-length LTR-RE population. On the contrary, LTR of clustered elements are more common in the genome and these elements are more transcribed and older than the mean of the

Table 2 Number of full-length LTR-RE families and of single full-length LTR-REs (i.e., not belonging to any family) and mean number of LTR-RE remnants with similarity to LTRs per family and per single element of *Copia*, *Gypsy*, and Unknown LTR-RE superfamilies

Superfamily	Number of LTR-RE families	Mean number of LTR-RE remnants per family	Number of single LTR-REs	Mean number of LTR-RE remnants per single LTR-RE
Copia	51	95.14	226	28.23
Gypsy	46	774.72	123	104.46
Unknown	29	352.76	632	17.90
Total	126	398.88	981	31.15

Table 3 Length (L), number of synonymous (^{S}) and nonsynonymous (or non coding, ^{A}) sites, number of synonymous and non synonymous (or non coding) substitutions per site (Ks and Ka, respectively) in 33 orthologous gene sequences of *P. trichocarpa* and *P. alba*. For each gene sequence, the identification code in *P. trichocarpa* and in *P. alba* (Maestrini et al. 2009) and the putative function is reported

ID code in *P. trichocarpa*	ID code in *P. alba*	Putative function[a]	L	S	A	K_s	K_a
eugene3.00440183	B3/H1	Unknown	401	132.33	268.67	0.0549	0.0113
estExt_fgenesh1_pm_v1.C_LG_III0004	B5/H3	Enoyl-ACP reductase	767	378.83	388.17	0.0133	0.0026
fgenesh1_pg.C_LG_V000487	B3/C8	Ca++/calmodulin kinase	398	128.33	269.67	0.0483	0.0000
estExt_fgenesh1_pm_v1.C_LG_XI0014	B3/F3	Dehydration responsive	379	124.67	254.33	0.0455	0.0099
eugene3.00012771	B1/C3	C2 domain-containing	718	279.75	438.25	0.0576	0.0386
estExt_fgenesh1_pg_v1.C_LG_VI0517	B3/D5	MIP1	651	196.17	454.83	0.0419	0.0066
estExt_fgenesh1_pm_v1.C_LG_V0518	B1/B5	Purple acid phosphatase	639	185	454	0.0445	0.0022
eugene3.00090981	B3/D3	Unknown	861	242.83	618.17	0.0424	0.0247
fgenesh1_pg.C_scaffold_129000034	B3/E4	Timing of CAB	784	281.17	502.83	0.0630	0.0181
fgenesh1_pg.C_LG_VII000308	L2/B11	GRP1 cell wall	424	193.08	230.92	0.0537	0.0446
eugene3.00170186	L2/C2	Ubiquitin-associated	710	240.58	469.42	0.1070	0.0172
estExt_fgenesh1_pm_v1.C_LG_II0684	L2/E1	Ubiquitin-protein ligase	430	183.5	246.5	0.0221	0.0000
eugene3.00400367	L1/C1	E3 ubiquitin ligase	396	148.17	247.83	0.0136	0.0040
fgenesh1_pg.C_LG_I001051	L3/E9	Acetyl-CoA carboxylase	620	328.17	291.83	0.0154	0.0104
estExt_fgenesh1_pg_v1.C_LG_VII0605	L3/D1	RNA pol II subunit	353	97.25	255.75	0.0104	0.0039
estExt_fgenesh1_kg_v1.C_LG_X0113	L4/H3	Ethylene responsive	387	138.17	248.83	0.0842	0.0000
estExt_Genewise1_v1.C_LG_XV2114	B3/G4	Oxidoreductase	624	142.33	481.67	0.0816	0.0000
grail3.0021011101	B4/B8	Unknown	720	168.33	551.67	0.0716	0.0119
eugene3.00081670	B4/C1	Ankyrin	564	123.17	440.83	0.0082	0.0137
estExt_Genewise1_v1.C_LG_III0385	B4/E6	Vacuolar invertase	339	75.5	263.5	0.0840	0.0231
fgenesh1_pg.C_LG_VIII001653	B4/H3	Cellulase	429	95.17	333.83	0.0213	0.0151
estExt_fgenesh1_pg_v1.C_LG_XI1305	B4/H4	UDP-D-xyl 4-epimerase	426	106.33	319.67	0.0288	0.0094
eugene3.00150320	B1/C2	Protein kinase	372	89.17	282.83	0.0344	0.0107
grail3.0006033201	B1/E4	B-box zinc finger	384	85.25	298.75	0.0870	0.0342
fgenesh1_pg.C_LG_X000373	B1/G7	Kinesin-related	816	185	631	0.0503	0.0144
eugene3.00070342	B1/H1	D123-like	726	168	558	0.0751	0.0090
estExt_Genewise1_v1.C_LG_XIII3457	B1/G4	NADH dehydrogenase	528	116.33	411.67	0.0262	0.0073
estExt_fgenesh1_pm_v1.C_LG_VIII0327	B1/G6	Phosphoglucomutase	639	153.5	485.5	0.1010	0.0093
estExt_fgenesh1_pm_v1.C_290015	L4/B2	Iron transporter	549	129.92	419.08	0.0685	0.0354
estExt_Genewise1_v1.C_LG_VI0164	L2/F4	ABI3-interacting	525	113.08	411.92	0.0089	0.0098
fgenesh1_pg.C_LG_X000444	L1/E12	CoF420 hydrogenase	477	115.33	361.67	0.0633	0.0111
grail3.0001095801	L3/C11	Unknown	732	180.83	551.17	0.0282	0.0128
grail3.0057014501	L1/G9	Hydrolase	576	132.58	443.42	0.0387	0.0160
Mean						0.0483	0.0133

[a]determined by evaluating top BLASTX hits in Genbank database

whole full-length LTR-RE population. The observed different transcriptional activities between the two subsets might suggest that silencing is more efficient when a LTR-RE is dispersed. Concerning the putative insertion age, dispersed elements show more similar sister LTRs, therefore they should be younger than clustered ones.

Discussion

We have analysed poplar LTR-retrotransposons based on sister LTRs identification. By this approach, only putative full-length retroelements, i.e. with two very similar LTRs, are scored. On the whole, we have isolated 1,479 full-length LTR-REs, of which 132 were identified for the first time, being absent in the existing database of poplar repeated sequences, RepPop (Zhou and Xu 2009) and so adding new retroelements to those already available.

Our data show that *Copia* full-length retroelements are more common than *Gypsy* ones (Table 1). However, *Gypsy* RE remnants were much more common in the genome than *Copia* ones (Table 2).

Our analysis also showed that the majority of full-length LTR-REs of poplar are of unknown nature, without any apparent coding sequence. Some unknown elements are to be classified as LARDs or TRIMs. To account for the origin of LARDS, it has been proposed that they are the product of transduction of a genomic sequence from the host genome, flanked by two solo LTRs. Alternatively, LARDs may have originated from the virus-like particle by co-encapsulation of a mRNA of the autonomous element with a mRNA of any host gene, followed by strand exchange between the two during the reverse transcription step (Jiang et al. 2002). LARDS and TRIMS could also have originated by rearrangements, deletions and/or illegitimate recombination of old functional elements, both *Gypsy* and *Copia*. Some of the LARDs identified in our analyses have probably maintained the capacity to retrotranspose, as indicated by the presence of families with genetically uniform LTRs (Table 2), by the putative very recent insertion dates of some of them (Table 5) and by the occurrence of such sequences in EST libraries (Table 5). Examples of recently inserted nonautonomous LTR-REs are known in other plant species, such as *Glycine max* (Wawrzynski et al. 2008).

The occurrence of retrotransposon families in poplar was established according to sequence similarity of their LTRs (Wicker et al. 2007). The number of full-length LTR-REs per family is generally low. *Gypsy* families contain more members than *Copia* ones (Figure 3). No family is made of a large number of elements; only 10 families show more than 8 LTR-REs. Prevalence of small LTR-RE families has been observed also in medium to large sized genome angiosperms as maize (Schnable et al. 2009) and sunflower (Cavallini et al. 2010).

Our data show a direct relationship between the number of full-length LTR-REs of a family and the number of LTR-RE remnants of that family in the genome. For instance, the LTR sequence of the largest family, G011, made of 37 full-length elements, shows high similarity with 3,754 sequences in the genome, indicating that this family has been active in ancient times and the vast majority of components of this family are now LTR-RE remnants. This aspect is generally true for poplar full-length LTR-retrotransposons (Table 2). The equilibrium between enlargement of the genome by retrotransposition and RE DNA loss affects the genome size of a species (Devos et al. 2002; Ma et al. 2004; Grover et al. 2008). Our data suggest that, in poplar, a small sized genome species, the equilibrium between retrotransposition activity and loss of DNA is biased towards DNA loss and that, probably, many REs have been active also in ancient times.

Analysis of sister LTR similarity indicates that, in

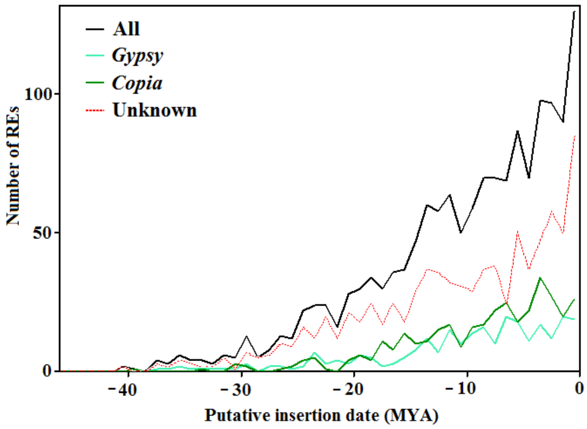


Fig. 4 Distributions of *Copia*, *Gypsy*, and Unknown full-length LTR-REs according to their estimated insertion ages (MYA) in the last 45 MY

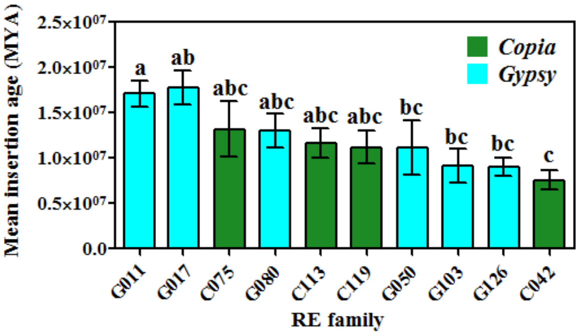


Fig. 5 Mean estimated insertion ages (MYA) of full-length LTR-REs belonging to the most numerous *Gypsy* and *Copia* families (number of full-length LTR-REs ≥ 9). Families with the same letter are not significantly different at the 5% level according to Tukey's test

Table 4 One-way ANOVA for estimated insertion age of full-length LTR-retrotransposons belonging to the 10 most numerous families (number of LTR-retrotransposons > 9)

Source of variation	SS	degrees of freedom	MS	F	P
Between families	2.18×10^{15}	9	2.42×10^{14}	4.22	0.0114
Within families	9.08×10^{15}	158	5.74×10^{13}		
Total	1.13×10^{16}	167			

poplar, both *Gypsy* and *Copia* REs have been active in the same period. Nearly all the identified full-length elements appear to be mobilised in a time span of 40 MY (Figure 4). It is conceivable that more ancient REs are no more recognizable because of accumulation of variability between sister LTRs.

The mean insertion date of poplar *Copia* full-length REs is lower than that of *Gypsy* ones (9.301 vs. 10.259 MY, Table 5). The insertion date profiles indicate that, after separation of poplar sections, *Copia* and *Gypsy* REs have both been active, but with different time courses. It can also be observed that different *Copia* and *Gypsy* families show different mean insertion times (Figure 5, Table 4). Similar results have been reported in other species, in which retrotransposon superfamilies are subjected to different amplification histories during the evolution of the host; for instance, in wheat, *Copia* and *Gypsy* superfamilies are differently represented in the A and B genome (Charles et al. 2008). Another example of different amplification histories among LTR-RE families was reported for *Copia* elements of *Vitis vinifera* (Moisy et al. 2008).

Concerning LTR-RE activity, a search for LTR sequences in EST databases of *P. trichocarpa* showed that only a small number of families appear to be transcriptionally active, often composed of one or at most two full-length elements. Generally, ancient full-length LTR-REs are inactive or less active than young ones, probably because of the accumulation of mutations determining premature stop codons in the coding portion of the LTR-RE, as observed in rice (Baucom et al. 2009b). Moreover, there is also a strong control of retrotransposon activity by the host species; it has been established that retrotransposons are especially silenced by siRNA (Lisch 2009). It is plausible that the large number of LTR-RE fragments spread throughout the poplar genome can produce siRNAs that silence related retroelements. Many 24-nt small RNAs associated to LTRs have been recently discovered in the poplar (Klevebring et al. 2009).

LTR-REs are present in poplar chromosomes at different densities. No loci are found with more than sixteen full-length REs inserted therein. Non significant variations are observed for mean insertion age between chromosomes, though such values range from 7.9 to 12.2 MY (Table 1). Within chromosomes, large regions are found in which the mean insertion age of full-length retrotransposons are either higher, or lower than the mean insertion age of LTR-REs in the whole chromosome (Fig. 6). Not only have LTR-REs inserted in different positions at different ages, but their retrotransposition activity appears to be somehow specific to their position in the chromosome (Table 6). In fact, LTR-REs inserted in regions with high full-length elements density belong to families whose LTR is largely represented in the genome (the number of LTR-RE remnants containing single LTRs or LTR fragments related to those elements is higher than the general mean), a feature related to the past activity of a LTR-RE family. On the other hand, dispersed full-length LTR-REs belong to families with lower numbers of related remnants than the general mean, i.e. with low past activity. Also a parameter indicating present activity (LTR-RE transcription) shows a difference

Table 5 Number of *P. trichocarpa* EST matches to LTRs of *Copia*, *Gypsy* and Unknown poplar full-length LTR-REs. The mean insertion dates for differently expressed LTR-RE groups are reported

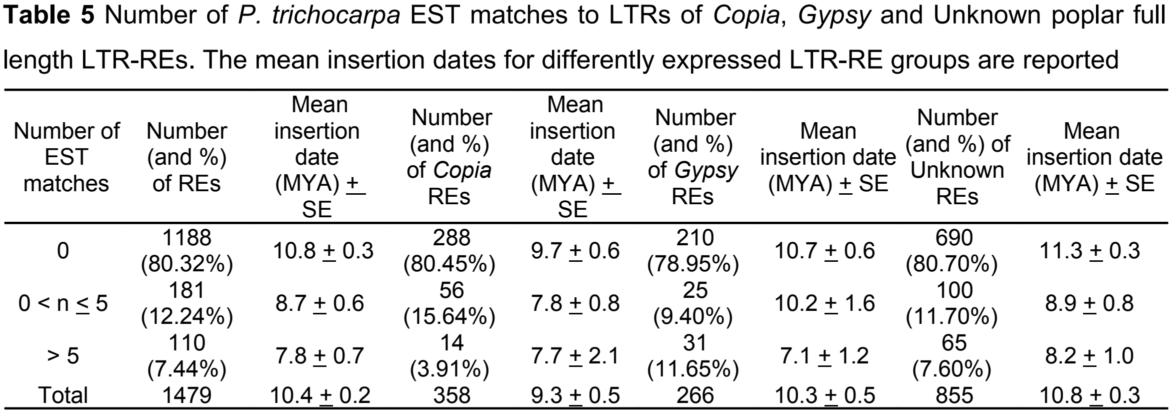


Table 5 Number of *P. trichocarpa* EST matches to LTRs of *Copia*, *Gypsy* and Unknown poplar full length LTR-REs. The mean insertion dates for differently expressed LTR-RE groups are reported

Number of EST matches	Number (and %) of REs	Mean insertion date (MYA) ± SE	Number (and %) of *Copia* REs	Mean insertion date (MYA) ± SE	Number (and %) of *Gypsy* REs	Mean insertion date (MYA) ± SE	Number (and %) of Unknown REs	Mean insertion date (MYA) ± SE
0	1188 (80.32%)	10.8 ± 0.3	288 (80.45%)	9.7 ± 0.6	210 (78.95%)	10.7 ± 0.6	690 (80.70%)	11.3 ± 0.3
0 < n ≤ 5	181 (12.24%)	8.7 ± 0.6	56 (15.64%)	7.8 ± 0.8	25 (9.40%)	10.2 ± 1.6	100 (11.70%)	8.9 ± 0.8
> 5	110 (7.44%)	7.8 ± 0.7	14 (3.91%)	7.7 ± 2.1	31 (11.65%)	7.1 ± 1.2	65 (7.60%)	8.2 ± 1.0
Total	1479	10.4 ± 0.2	358	9.3 ± 0.5	266	10.3 ± 0.5	855	10.8 ± 0.3

between clustered (higher than the general mean) and dispersed elements (lower than the general mean).

Dispersed elements seem also younger than clustered ones because of a higher similarity of sister LTRs. This result could however be explained hypothesising that the mutation rate of LTR-REs is higher in clustered than in dispersed elements. In fact, clustered elements are found in regions with a low number of predicted genes, on the contrary, dispersed elements lie in gene-rich regions, that are probably preserved from retrotransposition and, in general, from mutations; in this sense the higher identity shown by sister LTRs of dispersed elements should depend more on the region in which the element is found and less on the insertion age of the retrotransposon. Such a conclusion should support the hypothesis of the existence of different mutation rates in different kinds of transposon sequences or in different chromosome positions (Zuccolo et al. 2010) and would also indicate that insertion ages measured on sequence dissimilarity between LTR pairs are to be taken with caution.

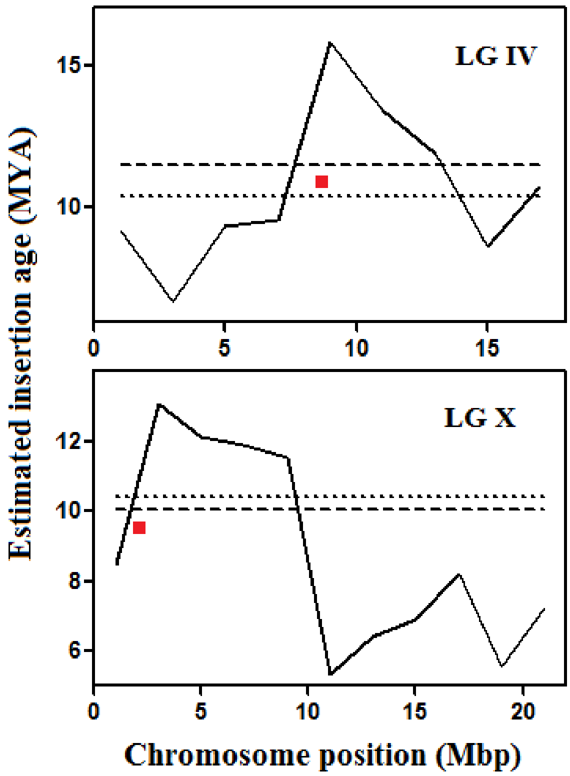


Fig. 6 Distributions of estimated insertion ages (MYA) of full-length LTR-REs along the poplar linkage groups IV and X. For each linkage group two horizontal lines are reported, representing the mean of all full-length LTR-REs in the genome (…………..) and in each linkage group (- - - - - - - - -). The box represents the putative position of the centromere as indicated by the occurrence in that position of centromeric repeats

Our analyses show the relationships between sequence characteristics, estimated age of LTR retrotransposons and their transcriptional activity in poplar LTR-REs. They are similar to those observed in other plant species, and support the theory of a ''life-history'' common to all LTR-REs, that includes birth through transposition, followed by silencing and then death by both random mutation and possibly deletion from the genome (Baucom et al. 2009b). However, we observed that different superfamilies and families are subjected to transposition in different time spans and show different transcription levels suggesting that if dynamics are similar, the factors inducing such dynamics might be different in different families and possibly related to the “ecosystem” in which the REs interact and compete, as proposed by Le Rouzic et al. (2007). In this sense, according to Venner et al. (2009), we suggest that poplar REs are a community of different organisms in the genome, with RE superfamilies, that can be described as species, and with “subspecies” characterised by different LTR sequence, activity, and evolution history.

Acknowledgements Research work supported by PRIN-MIUR, Italy, Project “Verso la delucidazione delle basi molecolari dell’eterosi nelle piante coltivate: variazione cis-regolatoria ed espressione genica in ibridi di pioppo”. Thanks are due to Dr.

Table 6 Number of *Copia*, *Gypsy*, and Unknown full-length LTR-REs, mean number of LTR-RE remnants, of ESTs and mean insertion age of clustered (> 10 elements within 400,000 bp) or dispersed (one LTR-RE within 1 million bp, with at least 300,000 bp between two adjacent elements) full-length LTR-REs. The general values obtained for all full-length LTR-REs are reported for comparison

LTR-RE positions	Number of *Copia* REs	Number of *Gypsy* REs	Number of Unknown LTR-REs	Total	Number of gene models[a]	Number of LTR-RE remnants (mean ± SE)	Number of ESTs (mean ± SE)	Insertion age in MY (mean ± SE)
Clustered	9	14	32	55	34.0 ± 2.2	920 ± 157	4.9 ± 2.5	13.8 ± 1.6
Single	13	4	51	68	92.3 ± 2.5	114 ± 58	1.9 ± 0.8	8.1 ± 0.9
General	462	508	540	1,492		287 ± 19	1.8 ± 0.2	10.4 ± 0.2

[a] number of genes (per 1 Mbp) predicted by Genewise, Fgenesh, GrailEXP6 and Eugene and selected by JGI annotation pipeline (http://genome.jgi-psf.org)

Andrea Zuccolo (Arizona Genomics Institute, USA) and Dr. John A. Walsh (Warwick University, UK) for their critical reading of the manuscript.

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

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, SanMiguel PJ, Bennetzen JL (2009a) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet 5:e1000732. doi:10.1371/journal.pgen.1000732.

Baucom RS, Estill JC, Leebens-Mack J, Bennetzen JL (2009b) Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Res 19:243-254

Beguiristain T, Grandbastien MA, Puigdomenech P, Casacuberta JM (2001) Three Tnt1 subfamilies show different stress-associated patterns of expression in tobacco. Consequences for retrotransposon control and evolution in plants. Plant Physiol 127:212–221

Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot 95:127–132

Benson G (1999) Tandem Repeat Finder: a program to analyze DNA sequences. Nucl Acids Res 27:573-580

Boeke JD, Corces VG (1989) Transcription and reverse transcription of retrotransposons. Ann Rev Microbiol 43:403-434

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

Charles M, Belcram H, Just J, Huneau C, Viollet A, Couloux A, Segurens B, Carter M, Huteau V, Coriton O, Appels R, Samain S, Chalhoub B (2008) Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics 180:1071–1086

Devos KM, Brown JKM, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res 12:1075–1079

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

The French-Italian Public Consortium for Grape Genome Characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463-467

Gaut BS (1998) Molecular clocks and nucleotide substitution rates in higher plants. In: Hecht MK, Macintyre RJ, Clegg MT (eds) Evolutionary Biology, vol 30. Plenum Press, New York, pp 93–120

Grover C, Hawkins J, Wendel J (2008) Phylogenetic insights into the paceand pattern of plant genome size evolution. In: Volff JN (ed) Plant Genomes. Genome Dynamics, vol 4. Karger, Basel, pp 57-68

Hawkins JS, Kim HR, 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

Hawkins JS, Hu G, Rapp RA, Grafenberg JL, Wendel JF (2008) Phylogenetic determination of the pace of transposable element proliferation in plants: *Copia* and LINE-like elements in *Gossypium*. Genome 51:11-18

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9:868-877

Islam-Faridi MN, Nelson CD, DiFazio SP, Gunter LE, Tuskan GA (2009) Cytogenetic analysis of *Populus trichocarpa* – Ribosomal DNA, Telomere Repeat Sequence, and Marker-selected BACs. Cytogenet Genome Res 125:74–80

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

Jiang N, Jordan IK, Wessler SR (2002) Dasheng and RIRE2. A nonautonomous long terminal repeat element and its putative autonomous partner in the rice genome. Plant Physiol 130:1697–1705

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

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH (2000) Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc Natl Acad Sci USA 97:6603-6607

Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoyb A, Schulman AH (2004) Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics

166:1437-1450

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

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

Klevebring D, Street NR, Fahlgren N, Kasschau KD, Carrington JC, Lundeberg J, Jansson S (2009) Genome-wide profiling of *Populus* small RNAs. BMC Genomics 10:620

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

Le Rouzic A, Dupas S, Capy P (2007) Genome ecosystem and transposable elements species. Gene 390:214–220

Lisch D (2009) Epigenetic regulation of transposable elements in plants. Annu Rev Plant Biol 60:43-66

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

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

Maestrini P, Cavallini A, Rizzo M, Giordani T, Bernardi R, Durante M, Natali L (2009) Isolation and expression analysis of low temperature-induced genes in white poplar (*Populus alba*). J Plant Physiol 166:1544-1556

Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11:1660-1676

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Albert H, Suzuki JY, Tripathi S, Moore PH, Gonsalves D (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452:991-997

Moisy C, Garrison KE, Meredith CP, Pelsy F (2008) Characterization of ten novel Ty1/*Copia*-like retrotransposon families of the grapevine genome. BMC Genomics 9:469

Morse AM, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, Garcia SA, Kubisiak TL, Amerson HV, Carlson JE, Nelson CD, Davis JM (2009) Evolution of genome size and complexity in Pinus. PLoS One 4:e4332

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418-426

Neumann P, Koblizkova A, Navratilova A, Macas J (2006) Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. Genetics 173:1047-1056

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev I, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Rahman M, Ware D, Westhoff P, Mayer KFX, Messing M, Rokhsar DS (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551-556.

Paux E, Roger D, Badaeva E, Gay G, Bernard M, Sourdille P, Feuillet C (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J 48:463–474

Rajagopal J, Das S, Khurana DK, Srivastava PS, Lakshmikumaran M (1999) Molecular characterization and distribution of a 145-bp tandem repeat family in the genus *Populus*. Genome 42:909–918

Ralph S, Oddy C, Cooper D, Yueh H, Jancsik S, Kolosova N, Philippe RN, Aeschliman D, White R, Huber D, Ritland CE, Benoit F, Rigby T, Nantel A, Butterfield YSN, Kirkpatrick R, Chun E, Liu J, Palmquist D, Wynhoven B, Stott J, Yang G, Barber S, Holt RA, Siddiqui A, Jones SJM, Marra MA, Ellis BE, Douglas CJ, Ritland K, Bohlmann J (2006) Genomics of hybrid poplar (*Populus trichocarpa* × *deltoides*) interacting with forest tent caterpillars (*Malacosoma disstria*): normalized and full-length cDNA libraries, expressed sequence tags, and a cDNA microarray for the study of insect-induced defences in poplar. Mol Ecol 15:1275-1297

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174-175

Sabot F, Schulman AH (2006) Parasitism and the retrotransposon life cycle in plants: A hitchhiker's guide to the genome. Heredity 97:381-388

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

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nature Genet 20:43-45

Santini S, Cavallini A, Natali L, Minelli S, Maggini F, Cionini PG (2002) Ty1/*Copia*- and Ty3/*Gypsy*-like DNA sequences in *Helianthus* species. Chromosoma 111:192–200

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:doi:1126/science.1178534

Seberg O, Petersen G (2009) A unified classification system for eukaryotic transposable elements should reflect their phylogeny. Nature Rev Genet 10:276

Sterck L, Rombauts S, Jansson S, Sterky F, Rouzé P, Van de Peer Y (2005) EST data suggest that poplar is an ancient polyploid. New Phytol 167:165-170

Sterky F, Bhalerao RR, Unneberg P, Segerman B, Nilsson P, Brunner AM, Charbonnel-Campaa L, Lindvall JJ, Tandre K, Strauss SH, Sundberg B, Gustafsson P, Uhlen M, Bhalerao RP, Nilsson O, Sandberg G, Karlsson J, Lundeberg J, Jansson S (2004) A *Populus* EST resource for plant functional genomics. Proc Natl Acad Sci USA 101:13951–13956

Sugimoto K, Takeda S, Hirochika H (2000) MYB-related transcription factor NtMYB2 induced by wounding and elicitors is a regulator of the tobacco retrotransposon tto1 and defense-related genes. Plant Cell 12:2511-2528

Thompson JD, Desmond G, Gibson H, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 22:4673-4680

Tian Z, Rizzon C, Du JC, Zhu L, Bennetzen JL, Jackson SA, Gaut BS, Ma J (2009) Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res 19:2221-2230

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

Venner S, Feschotte C, Biemont C (2009) Dynamics of transposable elements: towards a community ecology of the genome. Trends Genet 25:317-323

Vicient CM, Kalendar R, Schulman AH (2005) Variability, recombination, and mosaic evolution of the barley BARE-1 retrotransposon. J Mol Evol 61:275–291

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA 103:17638-17643

Wawrzynski A, Ashfield T, Chen NWG, Mammadov J, Nguyen A, Podicheti R, Cannon SB, Thareau V, Ameline-Torregrosa C, Cannon E, Chacko B, Couloux A, Dalwani A, Denny R, Deshpande S, Egan AN, Glover N, Howell S, Ilut D, Lai H, Martin del Campo S, Metcalf M, O'Bleness M, Pfeil BE, Ratnaparkhe MB, Samain S, Sanders I, Ségurens B, Sévignac M, Sherman-Broyles S, Tucker DM, Yi J, Doyle JJ, Geffroy V, Roe BA, Saghai Maroof MA, Young NA, Innes RW (2008) Replication of nonautonomous retroelements in soybean appears to be both recent and common. Plant Physiol 148:1760–1771

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 system for eukaryotic transposable elements. Nature Rev Genet 8:973-982

Witte CP, Le QH, Bureau T, Kumar A (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci USA 98:13778-13783

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucl Acids Res 35:W265-W268

Yamazaki M, Tsugawa H, Miyao A, Yano M, Wu J, Yamamoto S, Matsumoto T, Sasaki T, Hirochika H (2001) The rice retrotransposon *Tos17* prefers low-copy-number sequences as integration targets. Mol Genet Genomics 265:336-344

Zhou F, Xu Y (2009) *RepPop*: a database for repetitive elements in *Populus trichocarpa*. BMC Genomics 10:14

Zuccolo A, Sebastian A, Yu Y, Jackson S, Rounsley S, Billheimer D, Wing RA (2010) Assessing the extent of substitution rate variation of retrotransposon long terminal repeat sequences in *Oryza sativa* and *Oryza glaberrima*. Rice 3:242-250

Paper II

Putative centromeric sequences in three poplar species

Abstract. The evolution of centromere structure in plants is far to be clarified. It is generally composed of tandem repeated sequences and retrotransposons, mainly of the *Gypsy* superfamily. Though the genome of *Populus trichocarpa* has been entirely sequenced, the structure of centromeres has received little attention. We searched for putative centromeric satellites in the poplar genome using the TandemRepeat Finder software and identified two types of putative centromeric repeats. The first type, whose consensus sequence is 107 bp long, should allow the identification of the centromere position in 9 over 19 chromosomes forming the haploid complement. The second, a consensus sequence 142 bp long, should identify the centromere of 6 chromosomes. No putative centromeric repeats were found in the remaining 4 chromosomes. In all chromosomes in which the centromere position seemed to be identified, there was a significant overlap between the putative centromeric position and the accumulation of full-length *Gypsy* retrotransposons. The presence of two different centromeric repeats in two groups of chromosomes should be related to an ancient interspecific hybridization occurred during *P. trichocarpa* evolution. Clustering of sequences belonging to different chromosomes showed a clear differentiation among chromosomes, especially for the 107 bp repeat. We also performed Illumina sequencing of genomic DNA of two poplar species, *P. deltoides* and *P. nigra*, mapping Illumina reads of these two species to the two *P. trichocarpa* putative centromeric repeats. Such repeats occur also in these two species, at different redundancy. Sequence clustering showed that putative centromeric repeats have evolved also after poplar species differentiation.

Introduction

The centromere is a highly differentiated and extremely important structure of eukaryote chromosomes. The centromeres are responsible for sister chromatid cohesion and for normal chromosomal segregation during mitosis and meiosis, which are essential for development and cellular proliferation in all organisms. These functions are conserved across species, but the DNA components in centromeres differ greatly. Satellite DNA and retrotransposons are the most abundant DNA elements found in plant centromere regions (Jiang et al. 2003). Centromeric repeats often extend over several hundreds of thousands or millions of base pairs. The characterized satellite repeats are mainly composed of 150–180 bp tandem repeat motifs (Thompson et al. 1996, Round et al. 1997, Ananiev et al. 1998, Cheng et al. 2002, Nagaki et al. 2003, Kulikova et al. 2004, Lim et al. 2005, Birchler et al. 2011). Although the repeat length is similar between taxa, their sequence composition can be very different, even between closely related species (Malik and Henikoff 2002, Jiang et al. 2003, Lamb et al. 2004, Henikoff and Dalal 2005). There is also often a particular type of *Gypsy* retrotransposon that is present in centromeres but this is less well defined (Birchler et al. 2011).

The diversity of both satellite and retrotransposon sequences at the centromeres of different species is in sharp contrast to the protein components of the kinetochore that are highly conserved across species (Karpen and Allshire 1997, Henikoff et al. 2001). For example, the centromere satellite sequences of oat and maize are quite distinct but the oat proteins can function on the maize sequences as demonstrated by the existence of oat–maize addition lines (Jin et al. 2004).

This dichotomy is referred to as the centromere paradox (Henikoff et al. 2001). The rapid evolution of the centromeric sequences in most species has not been explained, although the centromere drive model has been put forward as one possibility as favouring reproductive isolation and consequent species differentiation (Henikoff et al. 2001).

In recent years, identification and characterization of centromeres have been achieved in several plant species whose genome has been completely sequenced. For example, the centromeres of Arabidopsis and rice contain 178 and 155 bp long tandem repeats arranged in blocks of 2.8 to 4.0 Mb and 0.06 to 1.9 Mb, respectively (Kumekawa et al. 2000, 2001; Cheng et al. 2002; Hosouchi et al. 2002). Repeat arrays are flanked by pericentromeric repetitive sequences and retrotransposons. In maize centromere the tandem satellite repeats CentC (156 pb) are interspersed with centromeric retrotransposons (Wolfgruber et al. 2009).

However, the repetitive nature of centromeric DNA can determine difficulties in sequence assembling, especially when using new generation sequencing methods, for which sequenced fragments are short (Yin et al. 2011).

Information on poplar centromeres is still lacking. Despite the whole genome of *Populus trichocarpa* has been sequenced (Tuskan et al. 2006), the current *Populus* genome sequence does not annotate the centromeric regions (Klevebring et al. 2009). Moreover, a complete cytogenetic map of the poplar, based on linkage groups as determined by whole genome sequencing, is still to be established (see Islam-Faridi et al. 2009). Cossu et al. (2011) surveyed the *P. trichocarpa* genome searching tandem repeats and reported two types of putative centromeric repeats. One of these repeats shows high similarity to a tandem repeat sequence isolated by Rajagopal et al. (1999) in *P. deltoides* and *P. ciliata*, which was described as putatively centromeric.

In this paper, we report a detailed bioinformatic characterization of the two putative centromeric repeats identified by Cossu et al. (2011) in *P. trichocarpa*. Moreover we use Illumina sequencing to isolate a number of similar repeats in *P. deltoides* and *P. nigra* and perform a comparative analysis of the three poplar species.

Materials and Methods

The TandemRepeat Finder program (Benson 1999) in conjunction with BLAST analysis was used to search putative centromeric repeats in the sequenced genome of *P. trichocarpa* (Tuskan et al. 2006) deposited at EMBL (accession number AARH00000000.1).

Sequences were identified and isolated all over the genome when sharing at least 80% of the length and 80% of sequence similarity (Cossu et al. 2011). Two putative centromeric sequences were identified, C107 (107 bp long) and C142 (142 bp).

For the isolation of similar putative centromeric sequences in *P. deltoides* and *P. nigra*, genomic DNA was extracted from leaflets of single plants (0.5 g fresh weight) as described by Doyle and Doyle (1989). Genomic libraries were prepared from 5 *μ*g of genomic DNA from *P. deltoides* or *P. nigra* leaves using the Illumina PE DNA Sample Prep kit according to the manufacturer. After spin column extraction and quantification, libraries were loaded on Cluster Station to create CSMA (clonal single molecular array) and sequenced at ultra-high throughput on the Illumina's Genome Analyzer IIx platform to produce 75- or 100-bp reads.

Illumina DNA reads were assembled using CLC Bio Workbench 4.9 software. The putative centromeric sequences were identified on the resulting contigs by local BLAST using the consensus sequence of the two putative centromeric repeats of *P. trichocarpa* as queries. Sequences were selected when sharing at least 80% of the length and 80% of sequence similarity.

To determine redundancy, *P. deltoides* and *P. nigra* Illumina reads were mapped to the two putative centromeric sequences of *P. trichocarpa*, using CLC Bio Workbench 4.9 software under the following parameters: Similarity = 0.7, Length fraction = 0.7, Insertion cost = 1, Deletion cost = 1, Mismatch cost = 1. To obtain comparable redundancy data in *P. trichocarpa*, the entire sequenced genome was splitted three times into 75mers, with 25 nt overlapping, obtaining a 3x coverage. These “artificial” reads were then used for mapping to the two putative centromeric sequences.

Relationships among centromeric repeats were investigated by the neighbor-joining (NJ) method (distance algorithm after Kimura), using the PHYLIP program package Version 3.572 (Felsenstein 1989): sequences were selected based on their similarity to the consensus C107 and C142, then, after DNA sequence alignment, trees were generated using DNADIST and NEIGHBOR programs, using default options. Strict consensus trees were obtained from the available trees using the CONSENSE program and visualized using TREEVIEW (Page 1996).

Results

Putative centromeric repeats of P. trichocarpa

Two putative centromeric sequences were found surveying the *P. trichocarpa* genome using TandemRepeat Finder. The first type, whose consensus sequence is 107 bp long (hereafter called C107), should allow the identification of the centromere position in chromosomes IV, V, VIII, X, XI, XII, XIII, XIV, and XV. The second, a consensus sequence 142 bp long (hereafter called C142), should identify the centromere of chromosomes I, III, IX, XVI, XVIII, and XIX. No putative centromeric repeats were found in chromosomes II, VI, VII, and XVII. The 142-bp long sequence shows high similarity to a putatively centromeric 145-bp tandem repeat sequence isolated by Rajagopal et al. (1999) in *Populus deltoides* and *Populus ciliata*. Both C142 and C107 are also found in all chromosomes as rare singlets.

C142 or C107 repeats are usually arranged tandemly, with different numbers of repeats, along *P. trichocarpa* chromosomes. Clusters of repeats are separated by non coding DNA, retrotransposons, and retrotransposon fragments. C142 and C107 repeats colocalize with *P. trichocarpa* centromeric retrotransposons identified by Neumann et al. (2011) in the available poplar genome sequence (data not shown).

A map track of putative centromeric repeats for each chromosome and the distribution of *Copia*, *Gypsy*, and unknown LTR-REs along chromosomes is reported in Fig. 1 (see also Cossu

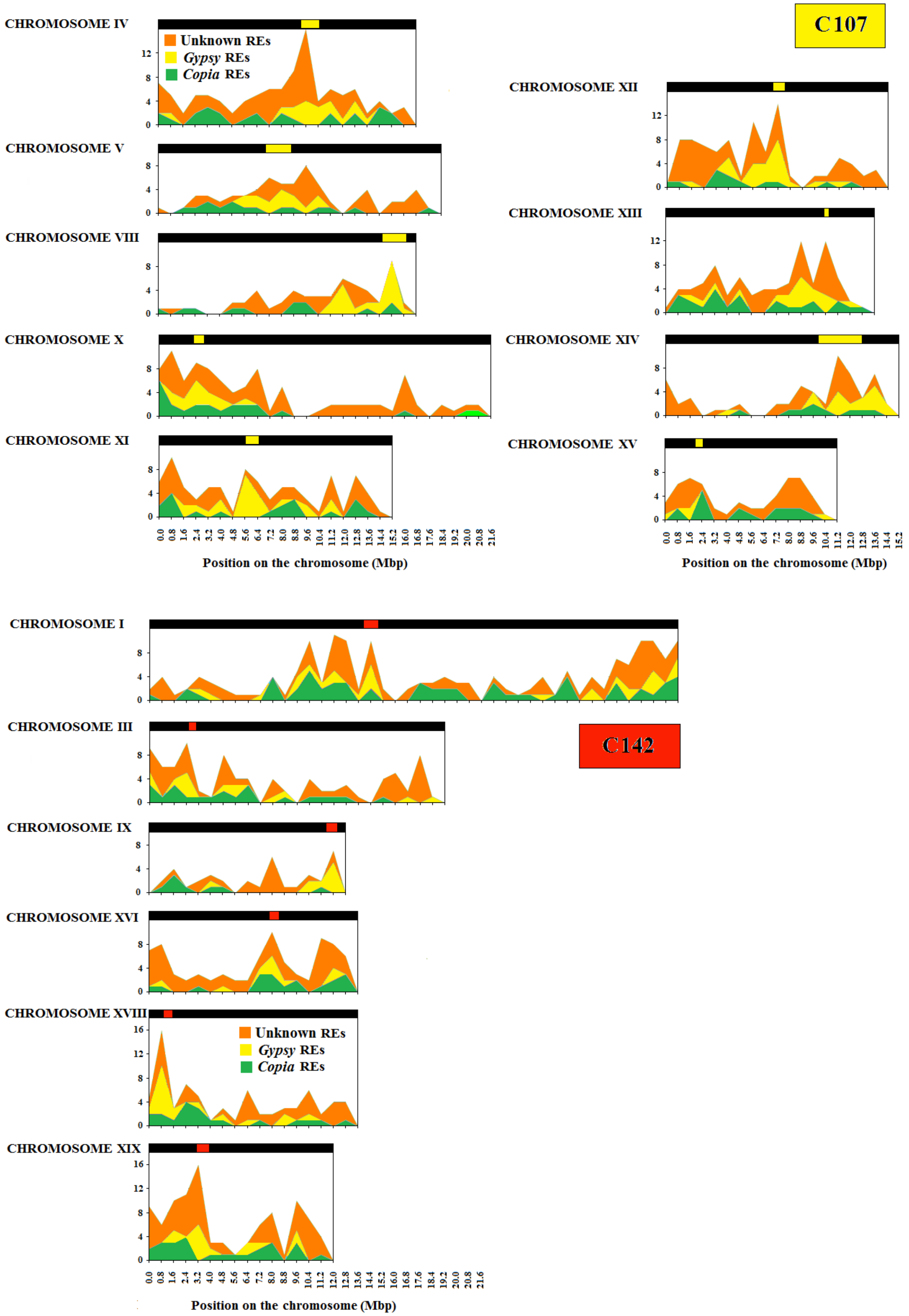


Fig. 1. Distribution of putative centromeric repeats and retrotransposons along poplar chromosomes. The positions of putative centromeric repeats, are evidenced in yellow (for C107, on the top) and in red (for C142, on the bottom) in the black track map over the retrotransposon profiles.

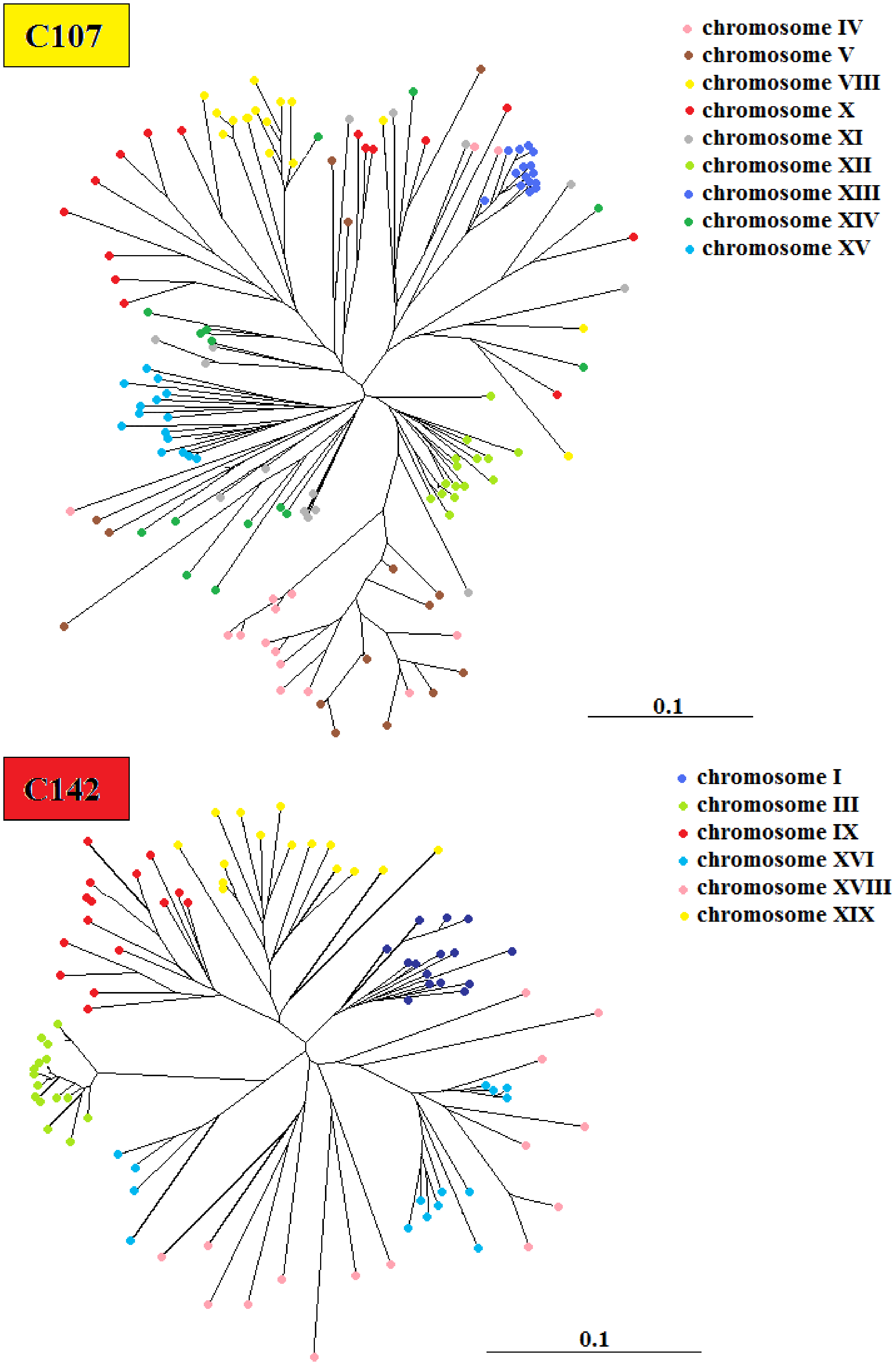


Fig. 2. Neighbor joining analysis of C107 repeats (15 repeats per 9 chromosomes, left) and C142 repeats (15 repeats per 6 chromosomes, right). Bar represents the nucleotides distance.

et al. 2011). Especially *Gypsy* REs are frequent in C142 or C107 repeat rich chromosome sites, confirming the association between centromeric repeats and *Gypsy* LTR-REs. It is, however, to be recalled that the definition of the centromere position requires biochemical and cytological validation, for example by BAC in situ hybridization (Islam-Faridi et al. 2009).

Neighbor joining analysis was performed comparing fifteen C107 and fifteen C142 repeats per *P. trichocarpa* chromosome. In Fig. 2, C142 tree shows a strong diversification among chromosomes, while C107 repeats are more similar, independently of chromosomes (except for chromosomes XII and XV).

Neighbor joining analysis was also performed comparing one hundred C107 repeats within chromosome XII and, subsequently, relating their position on the tree to their position along the chromosome. The same analysis was performed on one hundred C142 repeats within chromosome IX. Sequences were spatially numbered from 1 to 100 and grouped according to their similarity (Fig. 3). It can be observed that, especially for C107, adjacent repeats are more similar compared to repeats laying far from each other.

Two Illumina libraries of cDNA from mRNA isolated from xylem tissues, available at NCBI website under the accession numbers SRX031107 and SRX031105, were mapped to putative centromeric sequences using CLC BIO. No read was found mapping neither to C142 nor to C107, however further analyses are necessary before excluding the transcription of these sequences. In fact it is reported that centromeric satellite DNAs are transcribed (Topp et al. 2004; May et al. 2005)

Interspecific comparison of putative centromeric repeats

The occurrence of C107 and C142 repeats was surveyed also in two other species of poplar, *P. deltoides* and *P. nigra*. DNA was isolated from leaves and Illumina sequencing was performed. Illumina reads were assembled and the occurrence of C142 and C107 repeat units was surveyed in the resulting contigs. This analysis showed that such sequences occur also in these two species. The relative redundancy of C142 and C107 repeats in the three poplar species was estimated by mapping Illumina reads to the putative centromeric consensus repeats of *P. trichocarpa*. To estimate the relative redundancy of C142 and C107 repeats in *P. trichocarpa*, the sequenced genome of this species was splitted into 75mers with a 25 nt overlapping and 75mers were mapped in their turn to putative centromeric repeats. Significant redundancy differences were observed in the three poplar species (Table 1). In *P. deltoides* C142 is by far more redundant than in the other two species; in *P. nigra*, both putative centromeric sequences are less represented than in the other two poplars.

Phylogenetic relationship between putative centromeric repeats of the three poplar species were investigated by a neighbor joining analysis of ninety C107 and ninety C142 repeats isolated from *P. trichocarpa*, *P. nigra*, and *P. deltoides* (Fig. 4). Separation between sequences isolated from the three species is observed. However, the occurrence of clusters including sequences from the three species indicates an incomplete diversification between centromeres of these species.

Discussion

The current *Populus* genome sequence does not annotate the centromeric regions (Klevebring et al. 2009). Moreover, a complete cytogenetic map of the poplar, based on linkage groups as determined by whole genome sequencing, is still to be established (see Islam-Faridi et al. 2009).

We described two putative centromeric sequences, 142 and 107 bp in length, in the *P. trichocarpa* genome. The length of poplar putative centromeric sequences is in the range of that observed in many species: usually, these repeats are in the range of 150–180 bps unit length, but some species have microsatellite repeats (Birchler et al. 2011). The position of poplar putative centromere repeats overlaps that of *Gypsy* retrotransposons, confirming the association between tandem repeats and these retroelements in centromeres. For example, in maize a particular family of retrotransposons called Centromeric Retrotransposons of Maize (Ma et al. 2007, Nagaki et al. 2004, Yan et al. 2005, Wu et al. 2004) was found, composed of four subfamilies (Sharma et al. 2008).

The two sequences, C142 and C107, are organized as tandem repeats in two different groups of chromosomes. No putative centromeric sequences were found in four out of 19 poplar chromosomes, possibly because of underrepresentation of repetitive sequences in the currently available poplar genome sequence (Klevebring et al. 2009). On the other hand, the presence of sequence repeats does not appear an essential feature of centromere, especially in the plant kingdom. For example, a chromosome was found in barley that possessed no canonical centromere repeats but that would function normally (Nasuda et al. 2005). A fragment of a maize chromosome in oat showing acquisition of centromere activity over unique sequences has also been described (Topp et al. 2009). Both of these cases represent the gain of centromere activity without detectable centromeric DNA being

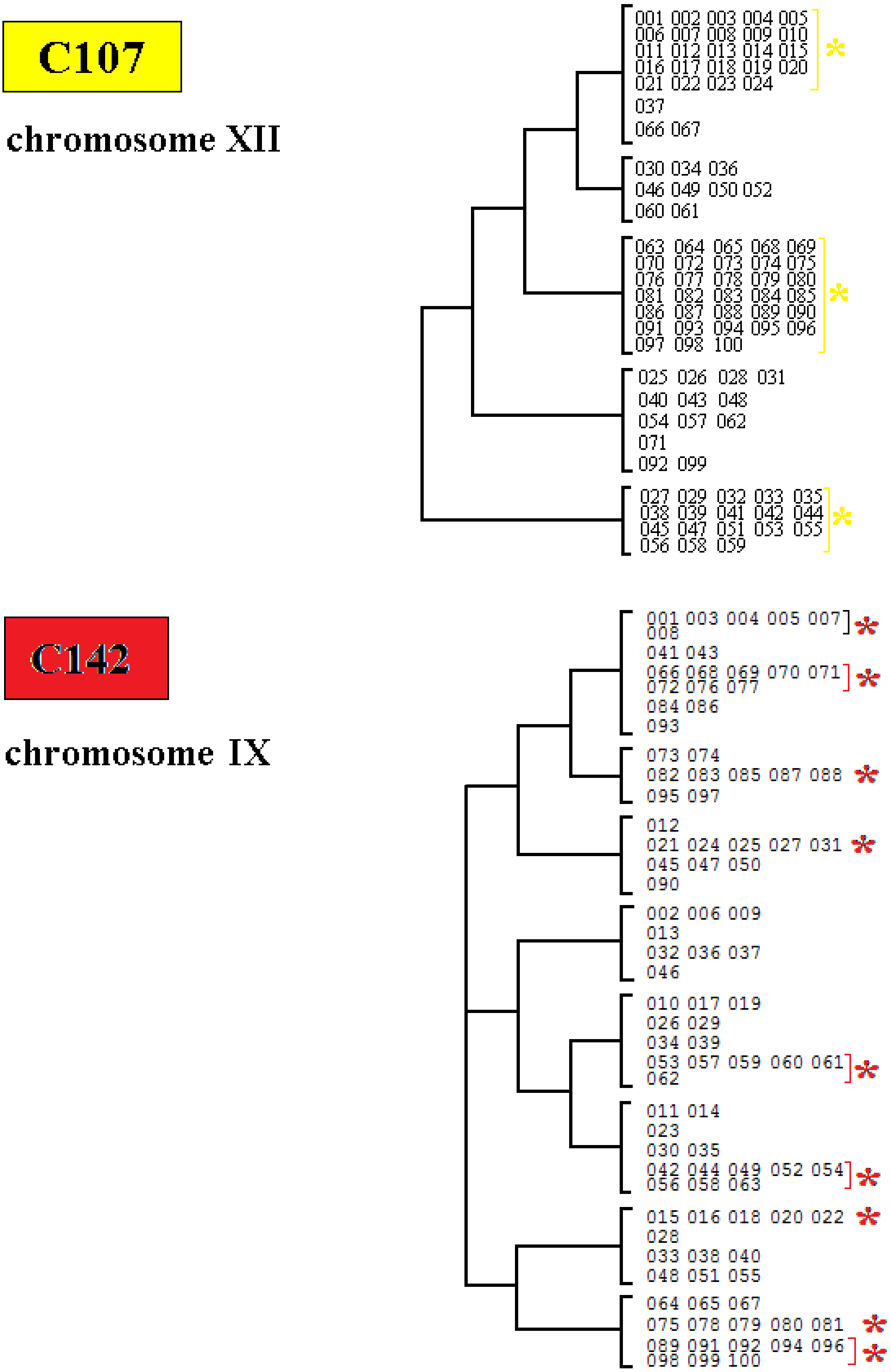


Figure 3. Neighbor joining analysis of 100 C107 repeats within chromosome XII (left) and 100 C142 repeats within chromosome IX (right) in relation to their relative position along the chromosome (sequences are numbered consecutively from 1 to 100). Asterisks indicate clusters of spatially grouped repeats.

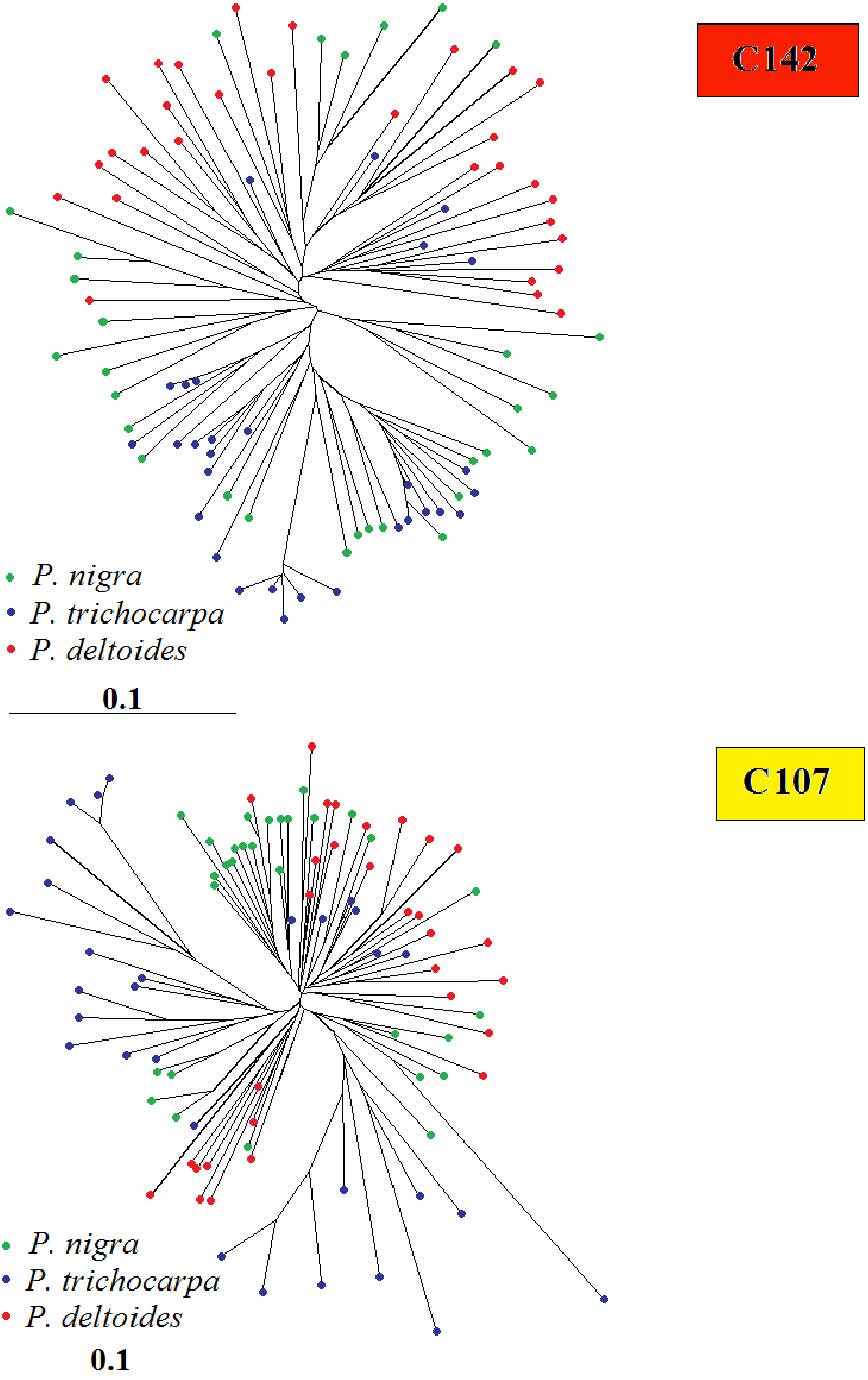


Fig. 4. Neighbor joining analysis of 90 C107 and 90 C142 repeats isolated from *P. trichocarpa*, *P. nigra*, and *P. deltoides*. Bar represents the nucleotides distance.

present.

It appears that epigenetic mechanisms, which are broadly referred to inherited states not conditioned by DNA sequence (Karpen et al. 1997), establish active centromeres on chromosomes, independent of their sequence (Jiang et al. 2003; Morris and Moazed 2007). The specification of centromeres in some organisms is referred to as being totally epigenetic (Birchler et al. 2011)

When centromere DNA repeats are found, they are considered as species-specific: centromeric sequences in different species are highly divergent and show considerable size variation (Ma et al. 2007, Zhong et al. 2002; Han et al. 2010; Wang et al. 2009). The presence of two different centromeric repeats in two groups of chromosomes might be related to an ancient interspecific hybridization occurred during *P. trichocarpa* evolution, or to the formation of neocentromeres. (Amor and Choo 2002).

Neighbour joining analysis of 15 C142 repeats per chromosome revealed a complete differentiation among chromosomes; on the contrary, C107 repeats from nine chromosomes resulted much more similar. These results suggest that: i) the spreading of C142 sequence over chromosomes has occurred before its diversification and amplification; ii) the amplification of C142 repeats should be occurred more recently than that of C107. In fact, the C107 repeats resulted more similar among chromosomes and the few differences among repeats should derive from mutations occurred after amplification. C107 repeats cluster together only in chromosomes XII and XV; probably C107 amplification in these chromosomes have occurred later than in the other chromosomes.

An analysis of spatial distribution of putative centromeric repeats within a chromosome should give an indication on the amplification model. It can be observed that repeats laying adjacent form subgroups of very similar sequences, especially in the chromosome XII, suggesting that the amplification added new repeats that stay close to the ancestor repeat. Moreover, the occurrence of different blocks of repeats with similar sequence suggest multiple sites in which redundancy has started.

Other analyses were carried on other poplar species, *P. deltoides* and *P. nigra*. High-throughput Illumina sequencing of genomic DNAs and subsequent assembly allowed to establish that both C142 and C107 repeats occur at high frequency also in these two species. Centromeric repeats are often reported to be species-specific, however similar centromeric sequences can be found in closely related species within a genus (Zhong et al. 2002; Han et al. 2010; Wang et al. 2009). The relatedness of poplar species is shown by the relative ease with which vigorous interspecific hybrids are obtained (see for example Dillen et al. 2008)

Neighbour-joining analysis of 90 C107 and C142 repeats from the three poplar species shows both repeats were present in chromosomes before poplar speciation. However, the occurrence of clusters composed by repeats from one and the same species, and the different redundancy of C107 and C142 observed in the three species strongly indicates that centromere evolution has proceeded at high rate also after poplar speciation.

Further analyses are in progress to verify by in situ hibridisation the centromeric nature of C107 and C142 repeats. Moreover we are currently comparing poplar centromere evolution with centromeres of other species to determine if and how the perennial habit affects the structure of centromeres.

References

Amor DJ, Choo KH (2002). Neocentromeres: role in human disease, evolution, and centromere study. Am J Hum Genet **71**: 695–714.

Ananiev EV, Phillips RL, Rines HW (1998). Chromosome specific molecular organization of maize (Zea mays L.) centromeric regions. PNAS **95**: 13073–13078.

Benson G (1999). Tandem repeats finder: a program to analyze DNA Sequences. Nucleic Acids Research **2**: 573–580

Birchler JA, Gao Z, Sharma A, Presting GG, Han F (2011). Epigenetic aspects of centromere function in plants. Current Opinion in Plant Biology **14**: 217–222.

Cheng Z, Dong F, Langdon T et al. (2002). Functional rice centromeres are marked by a satellite repeat and a centromere specific retrotransposon. Plant Cell **14**: 1691–1704.

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 Genetics & Genomes **8**: 61-75. DOI: 10.1007/s11295-011-0421-3

Dillen SY, Marron N, Koch B, Ceulemans R (2008). Genetic variation of stomatal traits and carbon isotope discrimination in two hybrid poplar families (*Populus deltoides* 'S9-2' x *P. nigra* 'Ghoy' and *P. deltoides* 'S9-2' x *P. trichocarpa* 'V24'). Ann Bot 102: 399-407.

Doyle JJ, Doyle JL (1989). Isolation of plant DNA from fresh tissue. Focus **12**: 13-15.

Felsenstein J (1989). PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* **5**: 164–166.

Han YH, Wang GX, Liu Z et al. (2010). Divergence in centromere structure distinguishes related genomes in *Coix lacryma-jobi* and its wild relative. *Chromosoma* **119**: 89–98.

Henikoff S, Ahmad K, Malik HS (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293:** 1098-1102.

Henikoff S, Dalal Y (2005). Centromeric chromatin: what makes it unique? *Curr Opin Genet Dev* **15:** 177–184.

Hosouchi T, Kumekawa N, Tsuruoka H, Kotani H (2002). Physical mapbased sizes of the centromeric regions of *Arabidopsis thaliana chromosomes 1, 2, and 3. DNA Res.* **9: 117–121.**

Islam-Faridi MN, Nelson CD, DiFazio SP, Gunter LE, Tuskan GA (2009). Cytogenetic analysis of *Populus trichocarpa* – ribosomal DNA, telomere repeat sequence, and marker-selected BACs. *Cytogenet Genome Res.* **125**: 74-80.

Jin W, Melo JR, Nagaki K et al. (2004). Maize centromeres: organization and functional adaptation in the genetic background of oat. *Plant Cell* **16:** 571-581.

Jiang J, Birchler JA, Parrott WA, Dawe RK (2003). A molecular view of plant centromeres. *Trends Plant Sci.* **8**: 570–575.

Karpen GH, Allshire RC (1997). The case for epigenetic effects on centromere identity and function. *Trends Genet* **13:** 489-496.

Klevebring D, Street N R, Fahlgren N et al. (2009). Genome-wide profiling of *Populus* small RNAs. *BMC Genomics* **10**: 620.

Kulikova O, Geurts R, Lamine M et al. (2004). Satellite repeats in the functional centromere and pericentromeric heterochromatin of *Medicago truncatula*. *Chromosoma* **113:** 276–283.

Kumekawa N, Hosouchi T, Tsuruoka H, Kotani H (2000). The size and sequence organization of the centromeric region of *Arabidopsis thaliana chromosome 5. DNA Res.* **7**: 315–321.

Kumekawa N, Hosouchi T, Tsuruoka H, Kotani H (2001). The size and sequence organization of the centromeric region of *Arabidopsis thaliana chromosome 4. DNA Res.* **8**: 285–290.

Lamb JC, Theuri J, Birchler JA (2004). What's in a centromere? *Genome Bio* **5:** 239.

Lim KB, de Jong H, Yang TJ et al. (2005). Characterization of rDNAs and tandem repeats in heterochromatin of *Brassica rapa. Mol Cells* **19:** 436–444.

Ma J, Wing RA, Bennetzen JL, Jackson SA (2007). Plant centromere organization: a dynamic structure with conserved functions. *Trends Genet* **23**: 134-139.

Malik HS, Henikoff S (2002). Conflict begets complexity: the evolution of centromeres. *Curr Opin Genet Dev* **12:** 711–718.

May BP, Lippman ZB, Fang Y, Spector DL, Martienssen RA (2005). Differential regulation of strand-specific transcripts from *Arabidopsis* centromeric satellite repeats. *PLoS Genet* **1**: e79.

Morris CA and Moazed D (2007). Centromere assembly and propagation. Cell 128: 647-650.

Nagaki K, Song J, Stupar RM et al. (2003). Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres. Genetics **163**: 759–770.

Nagaki K, Cheng Z, Ouyang S et al. (2004). Sequencing of a rice centromere uncovers active genes. Nat Genet **36**: 138-145.

Nasuda S, Hudakova S, Schubert I, Houben A, Endo TR (2005). Stable barley chromosomes without centromeric repeats. PNAS **102**: 9842-9847.

Neumann P, Navrátilová A, Koblízková A et al (2011). Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mobile DNA **2**: 4.

Page RDM (1996). TREEVIEW: an application to display phylogenetic trees on personal computers. Comput Applic Biosci **12**: 357-358.

Rajagopal J, Das S, Khurana DK, Srivastava PS, Lakshmikumaran M (1999). Molecular characterization and distribution of a 145-bp tandem repeat family in the genus *Populus*. Genome **42**: 909-918,

Round EK, Flowers SK, Richards EJ (1997). Arabidopsis thaliana centromere regions: genetic map positions and repetitive DNA structure. Genome Res **7**: 1045–1053.

Sharma A, Presting GG (2008). Centromeric retrotransposon lineages predate the maize/rice divergence and differ in abundance and activity. *Mol Genet Genom* **279**: 133-147.

Thompson H, Schmidt R, Brandes A, Heslop-Harrison JS, Dean C (1996). A novel repetitive sequence associated with the centromeric regions of *Arabidopsis thaliana* chromosomes. Mol Gen Genet. **253**: 247–252.

Topp CN, Zhong CX, Dawe RK (2004). Centromere-encoded RNAs are integral components of the maize kinetochore. *PNAS* **101**: 15986-15991.

Topp CN, Okagaki RJ, Melo JR, Kynast RG, Phillips RL, Dawe RK (2009). Identification of a maize neocentromere in an oat-maize addition line. *Cytogenet Genome Res* **124**: 228-238.

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

Wang GX, Zhang XY, Jin WW (2009). An overview of plant centromeres. *J Genet Genomics* **36**: 529–537.

Wolfgruber TK, Sharma A, Schneider KL et al. (2009). Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. *PLoS Genet* **5**: e1000743. doi:10.1371/journal.pgen.1000743

Wu J, Yamagata H, Hayashi-Tsugane M et al. (2004). Composition and structure of the centromeric region of rice chromosome 8. *Plant Cell* **16**: 967-976.

Yan H, Jin W, Nagaki K (2005). Transcription and histone modifications in the recombination-free region spanning a rice centromere. *Plant Cell* **17**: 3227-3238.

Yin W, Birchler JA, Han F (2011). Maize centromeres: where sequence meets epigenetics. *Front. Biol.* **6**: 102-108.

Zhong CX, Marshall JB, Topp C et al. (2002). Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. *Plant Cell* **14**: 2825-2836.

Paper III

High-throughput analysis of LTR retrotransposon expression in poplar hybrids *Populus deltoides* x *Populus nigra*

Abstract. Though LTR-retrotransposons represent a major component in plant genomes, they are often considered as silent, and their expression is rarely studied. The next generation sequencing methods offer an unprecedented level and unexplored potentialities of analysis, allowing a comprehensive study of the RNA expressed in given tissues and growth conditions. We evaluated the expression of LTR-retrotransposons in a poplar interspecific hybrid, *P. deltoides* x *P. nigra*, using Illumina RNAseq and an LTR-retrotransposon database of *P. trichocarpa*. First, the occurrence and redundancy of *P. trichocarpa* retrotransposons was studied in *P. deltoides* and *P. nigra*. Then, RNA was isolated from leaves of two hybrids obtained from the same parents, cultivated in control conditions or subjected to moderate or severe drought stress.

Nearly all *P. trichocarpa* retrotransposons were found in *P. deltoides* and *P. nigra*, however large differences in retrotransposon redundancy occur between the two species. The majority of retrotransposons were not expressed in the hybrids, however a few of them resulted highly transcribed, with differences during drought stress. The two hybrids, that are genetically different (being parents heterozygous), show different expression of retrotransposons. Such differences between hybrids are larger in drought stressed plants than in controls. *Gypsy* retrotransposons are less transcribed than *Copia*; the most expressed LTR-retrotransposons do not belong to any described superfamily and can be defined as LARDs or TRIMs. Unknown retroelements show similar expression levels in control and stressed plants, contrary to *Gypsy* and *Copia* elements that are more induced by drought. Drought-related motifs are found in higher number in LTRs of active retroelements than in those of inactive ones.

Introduction

The mobile component of the genome is represented by sequences, called transposable elements (TEs), potentially able to change their chromosomal location (transposition) through different mechanisms. TEs are subdivided into two main classes accordingly to their mechanism of transposition: retrotransposons (REs; class I) and DNA transposons (class II). Class II elements transpose by a "cut and paste" mechanism while class I elements are represented by TEs that can transpose through a replicative mechanism which involves an RNA intermediate. Such a "copy and paste" mechanism has been largely successful during evolution of eukaryotes in which class I elements represent the largest portion of genomes. For example, in the case of Oryza australiensis RE amplification doubled the genome size (Piegu et al. 2006).

Retrotransposons are divided into autonomous and non autonomous elements, according to the presence of ORFs that encode for TEs enzymes. Non autonomous elements do not carry enough coding capacity to allow them to transpose autonomously, nevertheless they are able to move using enzymes encoded by other elements (Tanskanen et al. 2006).

The genome of autonomous REs is organized in two domains: the gag domain, which is committed towards the production of virus like particles (VLPs), and the pol domain, whose encoded enzymes are used for processing RE-mRNA and obtaining a double stranded DNA to be integrated into the genome. The occurrence of long terminal repeats (LTRs) flanking the retrotransposon genome distinguish REs in two main classes, namely LTR- and non-LTR-retrotransposons. LTRs carry promoter elements, polyadenilation signals and enhancers regulating retroelements transcription (Bennetzen 2000).

Gypsy and *Copia* LTR-REs are two ubiquitous superfamilies (Voytas et al. 1992, Suoniemi et al. 1998) of plant REs that differ by the order of genes encoded by pol. *Gypsy* and *Copia* elements resemble retroviruses in their structure due to the presence of LTRs and internal ORFs. In the last decade, LTR-REs lacking internal coding domains, such as TRIMs (Terminal-Repeat Retrotransposons In Miniature, Witte et al. 2001) and LARDs (LArge Retrotransposons Derivatives, Kalendar et al. 2004) were described. TRIMs, formerly discovered in Solanum tuberosus and *Arabidopsis*, have been reported in monocots and dicots (Witte et al. 2001); LARDs, which were proved to be transcribed, have been discovered in Triticeae (Witte et al. 2001, Kalendar et al. 2004) and, recently, in sunflower and poplar (Buti et al. 2011, Cossu et al. 2012). TRIMs and LARDs can be identified only when the complete genome sequence or, at least, large DNA sequences are available. Their species-specific sequence and the absence of coding regions can explains their relative rarity in the literature. However, when analysed surveying complete genomes and using structural features as diagnostic (for example the occurrence of LTRs), they have been proved to form a major component in the TE fraction of the genome (Cossu et al. 2012).

In the last decade, the expression of retrotransposons has been reported in a number of plants, especially after exposition to various stresses (Vicient et al. 2001, Rico-Cabanas and Martinez-Izquierdo 2007, Ramallo et al. 2008, Buti et al. 2009, Kawakami et al. 2011). Only in a few cases, however, RE transcription has been shown to determine new insertions in the genome: Tnt1 and Tto1 in Nicotiana and Tos17 in rice showed stress induced (by tissue culture) transcription and transposition (Hirochika 1993, Hirochika et al. 1996, Grandbastien 1998) while these elements are not transcribed in standard culture conditions. A remarkable example of RE dynamics as an evolutionary adaptive mechanism within an ecological system is offered by BARE1 elements in wild barley (Kalendar et al. 2000). Recently, RE activity has been reported for a *Copia* element of sunflower for which RNA expression and subsequent insertion in the genome was shown (Vukich et al. 2009).

Large genome sequencing of grass plants showed that REs are responsible for extensive changes in genome structure and, surprisingly, dramatic differences were reported even among individuals belonging to the same species. It has been proposed that REs restructuring action plays a role in regulating gene expression: for example, allelic variation in non-genic (regulatory) sequence was proposed to be involved in heterosis, i.e. the superior performance of hybrids in respect of their parents (Brunner et al. 2005, Morgante et al. 2007). In this sense, the old epithet of “junk” for repeated sequences, which have affected genome structure and function, is becoming obsolete.

Interaction between REs and host genome has been successful allowing genome expansion and then the evolution of a complex network regulating gene expression (Feschotte 2008). Nevertheless, only few elements have been shown to transpose autonomously and data from EST libraries in grasses indicate that most are poorly transcribed (Meyers et al. 2001, Vicient et al. 2001, Vicient and Schulman 2002). It is conceivable that the activity of REs should be limited by the host genome because of their potential mutagenic action.

The first mechanism of control for mobile elements relies in chromatin structure, since heterochromatin is made of “silent” DNA. Mechanisms underlying chromatin packing in plants act through methylation of histones and cytosines in CG and CNG combinations (Dieguez et al. 1998). More emphasis about the importance of an epigenetic control of TEs is supported by the role of RNA silencing which determines chromatin specific methylation and RNA degradation mediated by small non coding RNAs which may derive from a number of different precursors (Slotkin and Martienssen 2007, Lisch 2009). In the fission yeast Schizosaccharomyces pombe, a basal level of transcripts matching centromeric repeats is substrate for dsRNA synthesis that is involved in preserving heterochromatin structure through histone methylation mediated by RNA silencing (Volpe et al. 2002). A silencing pathway driven by anti-sense small RNAs is responsible of REs and repetitive sequences silencing in the Drosophila germline (Vagin et al. 2006).

Retrotransposon dynamics has been mainly investigated in grasses and other monocotyledons. Dicotyledons have in general been given minor attention, despite their great economic importance.

Recently, we performed a survey of LTR-REs in the genome of *Populus trichocarpa* (Cossu et al. 2012). A computational approach based on detection of conserved structural features, on building multiple alignments, and on similarity searches allowed to identify 1,479 putative full-length LTR-REs. Ty1-*copia* elements were more numerous than Ty3-*gypsy*. However, many LTR-REs were not assigned to any superfamily because lacking of diagnostic features and non-autonomous. LTR-RE remnants were by far more numerous than full-length elements, indicating that during the evolution of poplar, large amplification of these elements was followed by DNA loss. Retrotransposition occurred with increasing frequency following the separation of *Populus* sections, with different waves of retrotransposition activity between Ty3-*gypsy* and Ty1-*copia* elements. Recently inserted elements appear more frequently expressed than older ones.

Next Generation Sequencing (NGS) procedures provide unprecedented levels of sequencing coverage, in short time and at relatively low cost, allowing whole-genome expression analyses. We have applied such techniques to study the transcription of the entire set of poplar full-length LTR-REs in different *P. deltoides* x *P. nigra* interspecific hybrids in control conditions and in plants subjected to drought stress.

Materials and Methods

Plant materials

Rooted cuttings of *P. deltoides* and *P. nigra*, and rooted cuttings from two of their hybrids, produced at INRA, Orleans (France), were cultivated in 20 x 20 cm^2 pots in the open. Leaves of *P. deltoides* and *P. nigra* were used to isolate genomic DNA.

In the late spring 2011, some hybrid plants of 50 cm in height were normally watered and others were subjected to drought by suspending watering. One leaf was collected from each plant. Leaves were subdivided into two portions: one was used for

RNA isolation, the other one was used to measure tissue hydration by determining the relative water content [RWC = 100 (FW-DW)/(TW-DW)], where FW is the fresh weight, DW the dry weight and TW the turgid weight. The experimental design was as follows: 2 clones (biological replicates) x 3 treatments (control, moderate, and severe drought stress) x 2 hybrids (obtained from the same parents).

DNA and RNA isolation and Illumina libraries preparation

Genomic DNA of *P. deltoides* and *P. nigra* was extracted from leaflets of single plants (0.5 g fresh weight) as described by Doyle and Doyle (1989). Genomic libraries were prepared from 5 μg of genomic DNA from *P. deltoides* or *P. nigra* leaves using the Illumina PE DNA Sample Prep kit according to the manufacturer. After spin column extraction and quantification, libraries were loaded on Cluster Station to create CSMA (clonal single molecular array) and sequenced at ultra-high throughput on the Illumina's Genome Analyzer IIx platform to produce 100-bp reads.

Total RNA was isolated from leaves of single plants of *Populus deltoides* x *P. nigra* hybrids with different RWC according to the method described by Logemann et al. (1987) followed by DNAse I (Roche) treatments according to the manufacturer's instructions to completely remove genomic DNA contamination.

RNA-Seq library was generated using the TruSeq RNA-Seq Sample Prep kit according to the manufacturer's protocol (Illumina Inc., San Diego, CA). In short, poly-A RNA was isolated from total RNA and chemically fragmented. First and second strand synthesis were followed by end repair, and adenosines were added to the 3'-ends. Adapters were ligated to the cDNA and 200 ± 25 bp fragments were gel purified and enriched by PCR. The library was quantified using Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA) and run on the Illumina HiSeq2000 (Illumina Inc.) using version 3 reagents. Single-read sequences of length 51 bp were collected.

Alignment and analysis of Illumina reads against the P. trichocarpa *RE database*

Genomic DNA and cDNA sequence alignments were generated with CLC-BIO Genomic Workbench 4.9. The *P. trichocarpa* full-length LTR-RE database (Cossu et al. 2012) is available on the University of Pisa Plant Genetics and Genomics Lab site (http://www.agr.unipi.it/Sequence-Repository.358.0.html). By using CLC-BIO Genomic Workbench, we also mapped 51-nt cDNA reads to 12 *P. trichocarpa* sequences putatively encoding actin as a control.

The evaluation of REs in the genomic DNA was carried out by mapping *P. deltoides* and *P. nigra* DNA reads to poplar RE database, using the following parameters, established by preliminary experiments: mismatch cost = 1, deletion cost = 1, insertion cost = 1, similarity = 0.7, length fraction = 0.7. The use of such low stringency parameters is related to the nature of the sequences analysed in this study: repeated, "non coding" sequences are subject to more rapid evolution than gene sequences and large sequence variability can be expected. These parameters allow to reduce undercounting of each family (data not shown). In fact, using higher costs for mismatch, deletion, and insertion resulted in mapping less than 1/10 nucleotide to the database. The number of mapped reads was reported as average coverage as follows:

$$\text{Average Coverage} = (N)/REFL$$

where N = sum of bases of the aligned part of all the reads, REFL = reference sequence length.

The evaluation of gene expression in *P. deltoides* X *P. nigra* hybrids was performed with the same software, that reports the number of mapped reads per kilobase per million mapped reads, measuring the transcriptional activity for each gene. CLC-BIO Genomic Workbench computes this normalized gene locus expression level (named RPKM) by assigning reads to a sequence in the database and counting them. In the case of reads that match equally well to several sites, the software assigns them to both.

The RPKM value (Mortazavi et al. 2008) estimates the number of reads falling in a given gene locus as follows:

$$RPKM = N/(L \times Ntot \times 10\text{-}6)$$

where N = number of mapping reads at a given gene locus, L = estimated length (Kbp) of the coding portion of the gene, Ntot = number of total mapping reads.

Expression profiles were evaluated considering RPKM values in control, moderately, and severely drought stressed plants using Baggerly's test (Baggerly et al. 2003). Expression values were reported as RPKM ratio between moderately or severely drought stressed plants and control plants when RPKM was higher in stressed than in control plants and with the negative reciprocal ratio when RPKM was higher in control than in stressed plants, thus leading to a '+' value in case of above-average expression levels and a '-' value in case of below-average expression levels.

Baggerly's test compares the proportions of counts in a group of samples against those of another group of samples, and is suited to cases where replicates are available in the groups. The samples are given different weights depending on their sizes (total counts). The weights are obtained by assuming a Beta distribution on the proportions in a group, and estimating these, along with the proportion of a binomial distribution, by the method of moments. The result is a weighted t-type test statistic.

In other analyses, consensus 5'-LTR sequences of *P. deltoides* and *P. nigra* were obtained using CLC-BIO by extracting, using an in-house perl script, 5'-LTRs from *P. trichocarpa* RE database, then by mapping *P. deltoides* and *P. nigra* Illumina DNA reads to *P. trichocarpa* LTRs, with the following parameters: Similarity = 0.7, Length fraction = 0.7, Insertion cost = 1, Deletion cost = 1, Mismatch cost = 1. Consensus LTR sequences were automatically extracted by CLC-BIO and subjected to motif search by CLC-BIO using a list of putative drought responsive motifs obtained modifying the motif list downloaded from PLACE website (http://www.dna.affrc.go.jp/PLACE/index.html; Higo et al. 1999).

Results

LTR-Retrotransposons in **P. deltoides** *and in* **P. nigra**

We have sequenced two genomic libraries of *P. deltoides* and *P. nigra* to obtain a 24.46x and 23.64x genome equivalents, respectively (Table 1). DNA reads were aligned to the *P. trichocarpa* RE reference dataset (Cossu et al. 2012), that includes 1,479 LTR-REs, using CLC BIO Genomic Workbench software.

The percentage of nucleotides mapping to the whole RE database in the two species resulted similar (Table 1) ranging from 21.19 to 22.46% of the genome. We have conducted a similar experiment mapping the RE database with "artificial" Illumina reads of *P. trichocarpa* (obtained splitting the sequenced genome into 75mers, with 25 nt overlapping, with a total of 2.59x coverage) and the percentage of nucleotides mapping to the database was 24.66%, i.e. similar to that of the other species.

Figure 1 shows the distribution of *P. nigra* and *P. deltoides* LTR-REs in relation to their average coverage with Illumina reads. The vast majority of REs in the database show low redundancy in the genome, however a few REs are highly redundant, either in *P. nigra*, in *P. deltoides*, and in both species.

The majority of highly redundant REs cannot be classified as belonging to the main LTR-RE superfamilies, i.e. *Gypsy* and *Copia*, because apparently lacking of coding sequences, hence they should be considered as LARDs or TRIMs (see Cossu et al. 2012).

The number of nucleotides matching to each RE of the database was compared between the two species (Figure 2). It can be found that, between *P. deltoides* and *P. nigra*, 493 REs show large differences, suggesting that many REs have experienced amplification or loss after the separation of these species. Such a comparison was performed also keeping separated *Copia*, *Gypsy*, and Unknown REs of the database (Figure 3). Regression lines are in all cases deviated towards *P. nigra*, indicating that, in general, all superfamilies are more represented in that species. Moreover, the regression coefficients suggest that the largest variation between the two species occur for Unknown REs, followed by *Gypsy* REs.

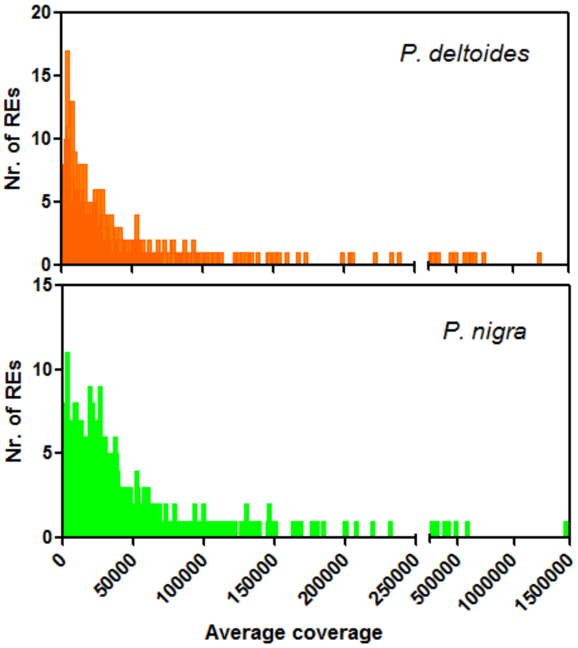


Figure 1. Distribution of REs in *P. deltoides* and in *P. nigra* according to their average coverage with Illumina reads.

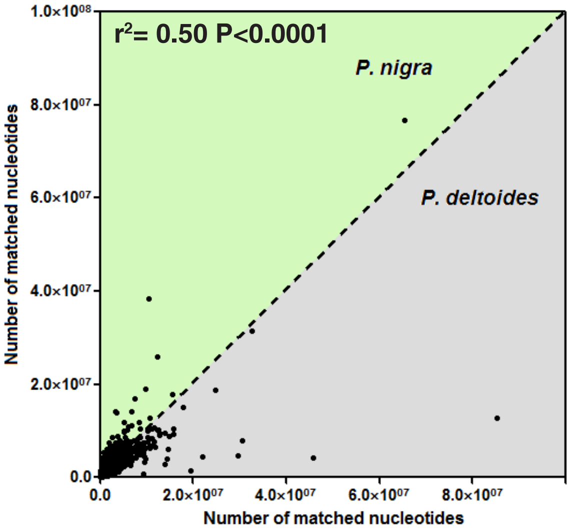


Figure 2. Pairwise comparison of number of nucleotides matching to each RE of the poplar RE database between Illumina read sets of *P. deltoides* and *P. nigra*.

Table 1. Mapping Illumina genomic DNA reads on *P. trichocarpa* RE database

Species	1C DNA		Nr. of reads	Average length (nt)	Coverage	% nucleotides mapped to RE database
	pg	Mbp				
P. deltoides	0.53	523	127,923,016	100	24.46x	21.19
P. nigra	0.54	528	144,774,190	86	23.64x	22.46

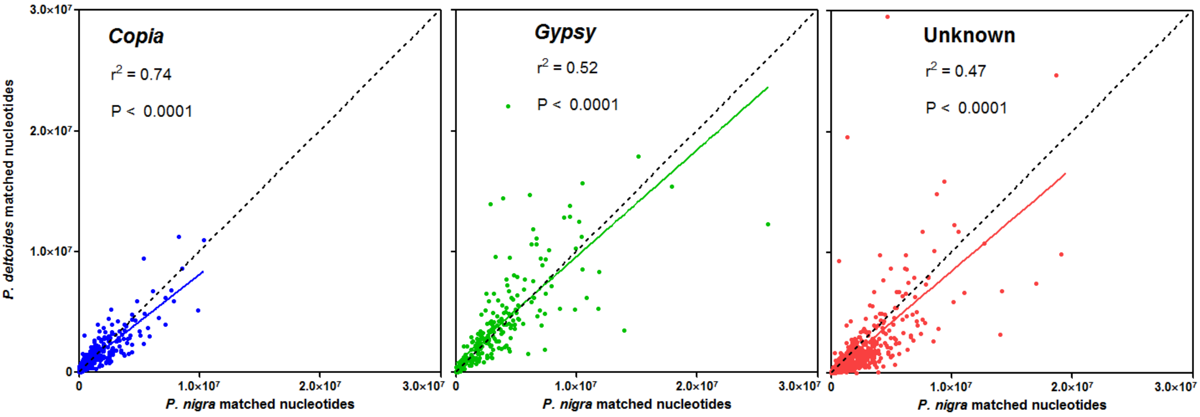


Figure 3. Pairwise comparison of number of nucleotides matching to each *Copia*, *Gypsy*, or Unknown RE of the poplar RE database between Illumina read sets of *P. deltoides* and *P. nigra*.

The dotted line represents the hypothetical relation between variables if REs were equally represented in the genomes of the two species. The plain line is the actual regression line.

On the other hand, it is to be reported that all *P. trichocarpa* REs in the database are mapped by *P. nigra* read set, and all but two by *P. deltoides* read set, suggesting a very small diversification in the RE pool within *Populus* genus as far RE families are concerned.

Expression of LTR-REs in leaves of control (unstressed) **P. deltoides** *x* **P. nigra** *hybrids*

We have prepared and sequenced by Illumina different libraries of cDNA isolated from RNA purified from leaves of control and moderately or severely drought-stressed *P. deltoides* x *P. nigra* hybrids (Table 2). Moderately stressed (S1) leaves showed a RWC of about 85%, severely stressed (S2) leaves RWC was about 57% (Table 2).

We generated 76,635,449 Illumina sequence reads, each 51 nt in length, encompassing 3.9 Gb of sequence data (Table 2). Each stress condition was represented by at least 20 million reads, a tag density sufficient for quantitative analysis of gene expression (Morin et al. 2008).

The sequence reads were aligned to the *P. trichocarpa* retrotransposon reference dataset (Cossu et al. 2012), using CLC BIO Genomic Workbench software set to allow two base mismatches.

We measured the number of Reads Per sequence Kilobase per Million mapped sequence reads (RPKM), a normalized measure of read density that allows transcript levels to be compared both within and between samples.

To evaluate the expression level of retrotransposons, we determined RPKM values also for 12 house-keeping actin encoding genes, that are usually used as reference in transcription analyses: all REs showing RPKM values higher than the most expressed actin gene were considered as highly transcribed.

Since in certain cases the occurrence of RE sequences in cDNA libraries can be related to DNA contamination of the mRNAs to be retrotranscribed, we have analysed the RPKM value of each RE in control leaves, in relation to its average coverage in an “artificial” hybrid composed by an equal

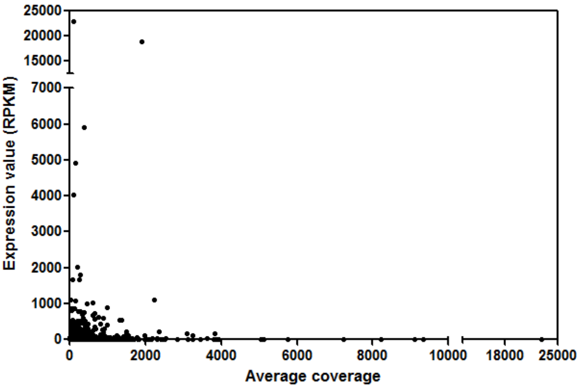


Figure 4. Relationship between RPKM expression values of each REs in the poplar RE database and their average coverage obtained after mapping RE database with Illumina DNA reads of an “artificial” hybrid composed of equal amounts of P. deltoides and *P. nigra* reads.

Table 2. Relative water content in the leave samples of *P. deltoides* x *P. nigra* hybrids subjected or not (C) to drought stress (S1 and S2), RNAseq data of each sample and percentage of nucleotides mapped to RE database for each sample.

Sample	RWC	Nr. of reads	Nr. of nucleotides	% nucleotides mapped to RE database
Hybrid 85/3 (C)	95.51	15,327,554	781,705,254	1.11
Hybrid 85/4 (C)	92.58	4,316,064	220,119,264	0.77
Hybrid 89/6 (C)	95.75	8,927,114	455,282,814	1.03
Hybrid 89/8 (C)	95.40	3,963,712	202,149,312	0.89
Hybrid 85/12 (S1)	86.31	5,487,345	279,854,595	1.25
Hybrid 85/24 (S1)	85.64	5,003,314	255,169,014	1.21
Hybrid 89/10 (S1)	84.89	6,123,484	312,297,684	0.96
Hybrid 89/15 (S1)	86.30	7,355,080	375,109,080	1.48
Hybrid 85/42 (S2)	54.78	3,985,186	203,244,486	1.31
Hybrid 85/45 (S2)	61.83	5,715,359	291,483,309	1.26
Hybrid 89/20 (S2)	52.78	4,575,904	233,371,104	0.69
Hybrid 89/35 (S2)	59.69	5,855,333	298,621,983	1.41
Total		76,635,449	3,908,407,899	1.13

number of *P. deltoides* and *P. nigra* DNA reads (Figure 4). The most redundant REs are not expressed and the most expressed REs are poorly represented in the "artificial" hybrid genome, suggesting that contamination by genomic DNA in the cDNA libraries can be largely ruled out. On the other hand, this result is in agreement with other showing that REs are transcriptionally active when low redundant (Meyers et al. 2001; Yamazaki et al. 2001).

Concerning control (unstressed) plants, the distribution of REs in the poplar database in relation to their expression activity is reported in Figure 5, keeping separated *Gypsy*, *Copia* and Unknown superfamilies. 633 REs were not active in both hybrids. Fifty per cent of *Gypsy* REs, 55.9% of *Copia* REs and 60.0% of unknown REs were transcribed. Unknown REs are by far the most active. Compared to the expression level of the most expressed actin gene, all REs with higher expression than actin gene belong to the unknown superfamily, except one *Copia* and two *Gypsy* REs.

Being the parents heterozygous, poplar interspecific hybrids are expected to be genetically different. Hence, RE expression was compared between hybrids and between clones of one and the same hybrid by pairwise comparison of log RPKM for each retrotransposon (Figure 6). Regression is in both cases highly significant, indicating the same retrotransposons are expressed in different genetic backgrounds. However, the regression significance is higher between clones than between hybrids. This difference should be related to differences in genetic backgrounds between hybrids. The difference between RE expression of clones should be related to experimental casualty and represent the experimental error, though a (micro)environmental effect causing real expression differences cannot be totally ruled out.

Expression of LTR-REs in leaves of poplar hybrids subjected to stress

RNAs isolated from leaves of *P. deltoides* x *P. nigra* hybrids subjected to moderate (i.e., around 85% leaf RWC, indicated as S1) or severe drought stress (i.e., around 57% leaf RWC, S2) were retrotranscribed and cDNA was sequenced using the Illumina procedure (see above, Table 2).

Compared to leaves of control plants, the mean expression of *Gypsy* REs appears stable or even reduced, especially in S2 (Table 3); on the contrary, that of *Copia* and unknown elements is somewhat increased under severe drought stress, and *Copia* REs mean expression becomes higher than that of *Gypsy* REs (Table 3).

The vast majority of REs is expressed in control, moderately stressed and severely stressed leaves (Figure 7). Relevantly, 442 over 585 active Unknown elements (75.6%) are always transcribed, compared to 154 over 257 (59.9%) active *Copia* REs, and 82 over 174 (47.1%) active *Gypsy* REs.

We subdivided the set of 1,479 REs into 9 expression profiles: those remaining constant, those increasing their expression in S1 or in S2 or in both stress levels, those reducing their expression in S1 or S2 or in both stress levels, those increasing their expression in S1 and reducing in S2 and vice versa (Table 4). The expression profiles of the differentially expressed genes were determined by calculating the RPKM ratio between moderately or severely stressed and control leaves, estimating a significant difference when ratio was higher than 2.0 or lower than -2.0. 463 REs were never active in transcription, especially those belonging to the *Gypsy* superfamily. 74 REs did not change their expression level after exposition to drought. The remaining 942 REs changed their expression during drought stress, increasing (507 REs) or reducing (435 REs) their expression level in S1 and/or in S2.

Among *Copia* REs, the most diffuse expression pattern shows an increase of transcription at the highest stress level. On the contrary, for *Gypsy* REs, the most diffuse pattern shows expression level increasing in the first stage of stress (S1) and returning to normal value in S2 (Table 4).

RE expression was compared between hybrids at different stress level by pairwise comparison of log RPKM for each retrotransposon (Figure 8). Though the regression maintains highly significant in both drought stress level, the regression significance progressively reduces from control plants to S1 and to S2 plants. Such difference should suggests that the two hybrids (that are genetically different) respond differently to the stress in terms of retrotransposon activation.

Considering the REs that are more transcribed than the most active actin gene in at least one hybrid and one stress condition, only 4 out of 44 are highly expressed in both hybrids and in all tested conditions. Fourteen are highly expressed in both control hybrids, and 5 are highly expressed in both drought stressed hybrids.

Highly expressed REs and untranscribed elements were then compared as to the occurrence, in their 5'-LTR, of sequence motifs recognizable by transcription factors activated by

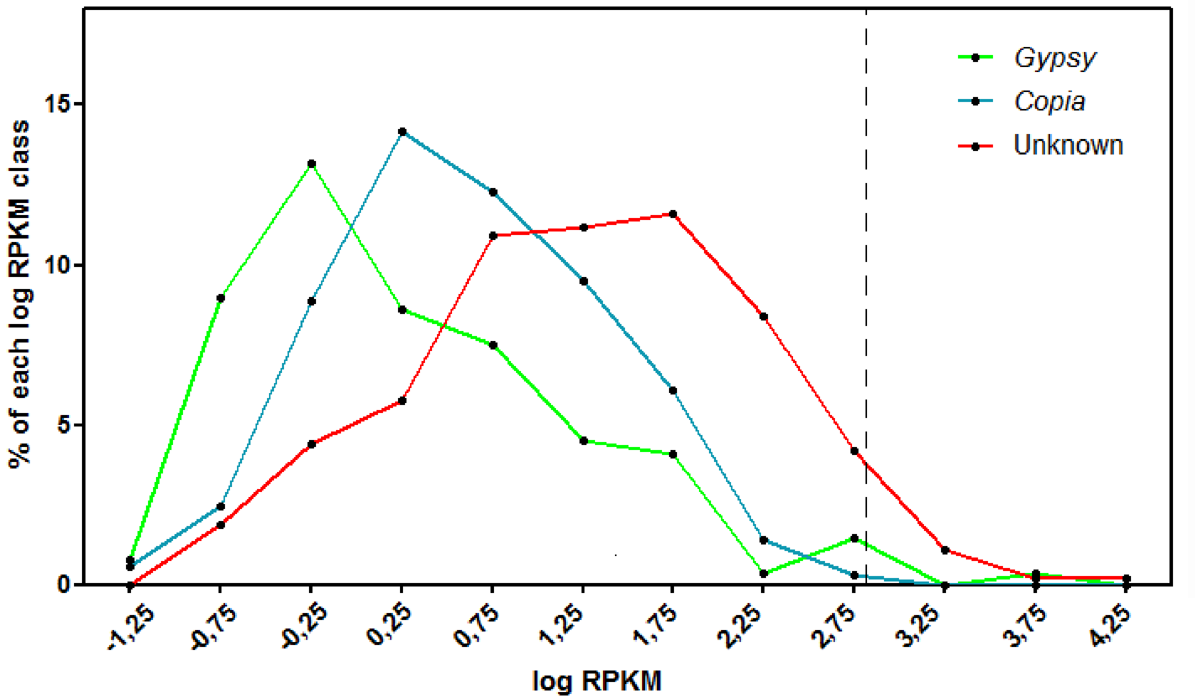


Figure 5. Distribution of REs in the poplar RE database according to their expression value (log RPKM). Vertical dotted line indicates the value corresponding log RPKM of the most expressed actin encoding gene.

Table 3. Number of active and inactive REs in leaves of control (C), moderately (S1) and severely (S2) drought stressed plants. The mean RPKM is reported for active REs only and for all REs of each superfamily.

Super-family	Control				S1				S2			
	Nr. of active REs	Nr. of inactive REs	Active RE Mean RPKM	Total Mean RPKM	Nr. of active REs	Nr. of inactive REs	Active RE Mean RPKM	Total Mean RPKM	Nr. of active REs	Nr. of inactive REs	Active RE Mean RPKM	Total Mean RPKM
Copia	200	158	20.99	11.73	202	156	15.54	8.77	207	151	25.24	14.59
Gypsy	133	133	54.58	27.29	134	132	38.13	19.21	115	151	27.54	11.91
Unknown	513	342	206.35	123.81	514	341	200.11	120.30	514	341	210.04	126.27
Total	846	633	138.67	79.32	850	629	130.71	75.12	836	643	139.18	78.67

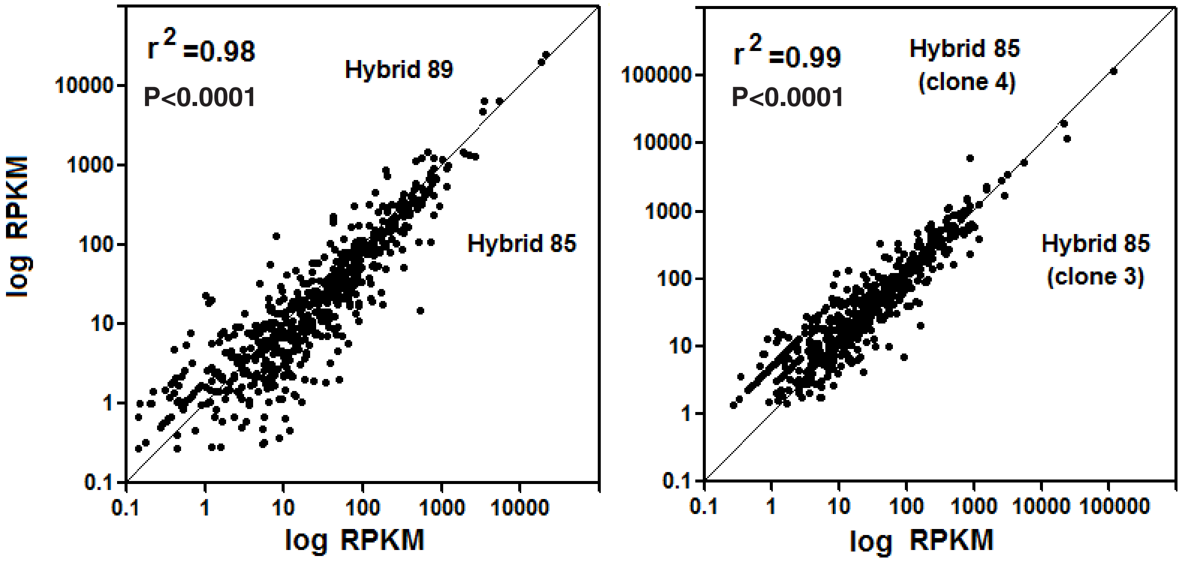


Figure 6. Two-dimensional representations of RE expression estimated log RPKM in leaves from two *P. deltoides* x *P. nigra* hybrids (left) and from two clones of one and the same hybrid (right). The line represents the hypothetical relation between variables if expression values were the same.

drought. A list of such motifs was compiled from the database downloaded from PLACE website (http://www.dna.affrc.go.jp/PLACE/index.html, Higo et al. 1999, Supplementary material #1). *P. deltoides* and *P. nigra* consensus 5'-LTRs were obtained for each highly expressed and for each non transcribed element by mapping Illumina DNA reads of the two species to *P. trichocarpa* 5'-LTRs. Then, the occurrence of drought related cis-regulatory motifs in LTRs was surveyed (Table 5). We observed that many ABREs (abscisic acid responsive elements), DREs (drought responsive elements) are present in the consensus LTRs of both *P. deltoides* and *P. nigra*. It is known that such motifs are crucial for expression of genes induced by dehydration (Yamaguchi-Shinozaki and Shinozaki 2006). The LTRs of highly expressed REs showed higher frequencies of such cis-regulatory motifs than inactive REs (Table 5). However, such motifs occurred with a certain frequency also in untranscribed REs. It is worth noting that we analysed one consensus LTR per RE of which more than one copy are presumably present in the genome with possible differences in activity: since *P. deltoides* and *P. nigra* genomes are being completely sequenced, a more precise analysis will be performed in the future on different REs belonging to the same family.

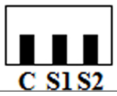
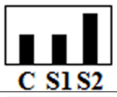
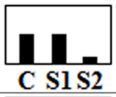
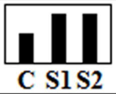
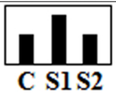
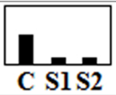
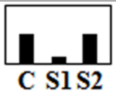
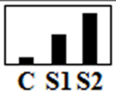
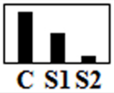
Discussion

For the evaluation of LTR-RE expression in poplar hybrids between *P. deltoides* and *P. nigra*, we have used an LTR-RE database of *P. trichocarpa*, containing 1,479 complete retroelements (Cossu et al. 2012). Obviously, it is presumable that *P. trichocarpa* LTR-REs are partly different in sequence and in redundancy among these species, however, preliminary data to be found in the literature showed that sequence differences are not so common among poplar species, at least for genes (Maestrini et al. 2009, Cossu et al. 2012) and, *P. trichocarpa* genome was sequenced (Tuskan et al. 2006) as a model species for all poplars.

DNAseq analysis by mapping Illumina reads to RE database showed that for the three species nearly the same percentage of nucleotides mapped (ranging from 21 to 24%) indicating similar levels of representation of such REs. Moreover, with the exception of two REs that are not mapped by *P. deltoides* reads, all other REs were proved to be present in the genomes of the three poplar species. We have also de novo assembled DNA reads of *P. deltoides* and *P. nigra* and, after masking the resulting contigs with the RE database, we were not able to recover any known RE in the remaining contigs (data non shown). Hence the *P. trichocarpa* RE database apparently constitutes an exhaustive sample of poplar REs.

Variations among the three species were rather observed in the redundancy of many REs, especially of unknown superfamily, indicating that genome differentiation has occurred at relatively high rates after species separation. On the other hand, previous data in *P. trichocarpa* suggested large activity of LTR-REs during the last MYRs, as indicated by the analysis of RE putative insertion dates (Cossu et al. 2012), a period subsequent to the separation of poplar lines that originated *P. trichocarpa*, *P. deltoides* and *P. nigra*. In fact, recent data based on dating polyploidization events in different *Populus* species indicates that

Table 4. Distribution (%) of RE superfamilies based on corresponding expression patterns observed in leaves of *P. deltoides* x *P. nigra* hybrids in control conditions (C) or during moderate (S1) and severe drought stress (S2).

Expression pattern	Nr. of REs (%)	Nr. of Copia REs (%)	Nr. of Gypsy REs (%)	Nr. of Unknown REs (%)
No expression	463 (31.30)	101 (28.21)	92 (34.59)	270 (31.58)
	74 (5.00)	16 (4.47)	7 (2.63)	51 (5.9)
	163 (11.02)	53 (14.80)	23 (8.65)	87 (10.18)
	108	21 (5.87)	17 (6.39)	70 (8.19)
	55 (3.72)	19 (5.31)	9 (3.38)	27 (3.16)
	189 (12.78)	44 (12.29)	47 (17.67)	98 (11.46)
	115 (7.78)	32 (8.94)	31 (11.65)	52 (6.08)
	147 (9.94)	38 (10.61)	20 (7.52)	89 (10.41)
	100 (6.76)	24 (6.70)	12 (4.51)	64 (7.49)
	65 (4.39)	10 (2.79)	8 (3.01)	47 (5.50)

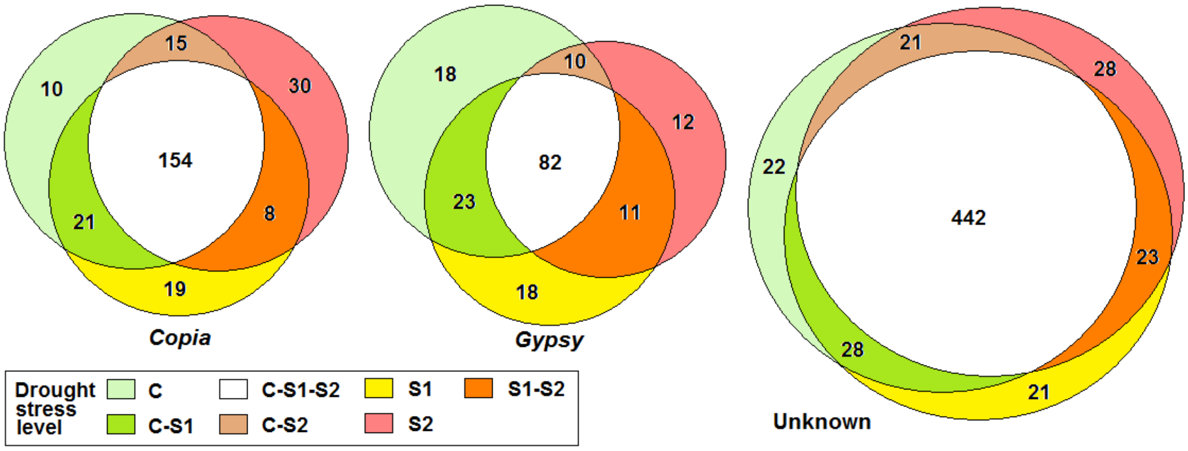


Figure 7. Venn diagrams showing the *Copia*, *Gypsy*, and unknown REs expressed in each of the three conditions studied, i.e. in leaves from P. deltoides x *P. nigra* hybrids normally watered (C), moderately (S1) and severely drought stressed (S2).

genus speciation occurred 8–13 MYRs ago (Sterck et al. 2005, Tuskan et al. 2006).

It is worth noting that a difference can be observed between *P. deltoides* and *P. nigra* LTR-RE accumulation during their evolution: in *P. deltoides* a higher number of REs show the highest redundancy levels than in *P. nigra*; however, in *P. nigra*, a lower number of REs result at the lowest redundancy levels (Figure 1). In other words, it seems that in *P. deltoides* some REs have hugely amplified, while in *P. nigra*, more REs have amplified though at minor levels than in *P. deltoides*.

Genome differentiation between poplar species should be related especially to activity of unknown elements, as suggested by the lowest pairwise correlation coefficient that can be found for this superfamily (Figure 3).

LTR-retrotransposons are apparently expressed in *P. deltoides* x *P. nigra* hybrids. In the various RNAseq samples, the percentages of cDNA nucleotides that map to the database range from 0.69 to 1.48%, i.e. LTR-RE RNAs are only a small fraction of total RNA. Additionally, if one refers only to *Gypsy* and *Copia* REs (i.e., REs that are at least partly recognizable by their sequence) the percentages are further reduced, ranging from 0.06 to 0.18% (data not shown). Small levels of transcription of repeated sequences are often attributed to DNA contamination of RNA samples. In the experiment described here, RE redundancy and transcription resulted totally uncorrelated (Figure 4): redundant REs are not or only slightly expressed, single REs are actively transcribed, suggesting that the presence of RE sequences in the cDNA library is real and not due to DNA contamination, as already indicated for *P. trichocarpa* (Cossu et al. 2012). Such a lack of correlation between RE redundancy and activity was expected because it is known that redundant elements are more easily recognized and subjected to RNA silencing (Lisch 2009).

The different RE superfamilies are differently active in transcription, the least active belonging to the *Gypsy* superfamily and the most active to the group of Unknown elements. More specifically, though as much as 850 over 1479 REs resulted somehow transcribed, only 32 are expressed at high levels, i.e. higher than actin genes. Among these, 29 belong to the Unknown superfamily.

Unknown elements were defined in the database as those REs showing two LTRs, often also the PBS, but always lacking of the RE enzymes coding genes (Cossu et al. 2012). They are non autonomous elements, that use RE enzymes produced by autonomous elements and have been called LARDs (LArge Retrotransposon Derivatives; Kalendar et al. 2004) when longer than 4 kbp, and TRIMs (Terminal-repeat Retrotransposons In Miniature; Witte et al. 2001) when shorter (Wicker et al. 2007).

Such elements are difficult to recognize if large sequences (for example BAC sequences) are not available. They are species-specific and highly variable in sequence, and have probably evolved at high rates except in the promoter sequences that should have been conserved. The reason for LARDs and TRIMs large expression in poplar hybrids may be related to their specificity. We can speculate that Unknown REs, lacking sequences shared within the RE superfamily, might be more prone to escape RNA silencing.

The two analysed hybrids show some differences concerning the level of expression of specific REs. The correlation of RE expression between two hybrids is obviously highly significant, however it is relatively less significant than that between two clones of a same hybrid (Figure 6). The observed differences between clones should be related to experimental casualties and/or to undetermined local environmental differences (soil, light) in the culture conditions. It is presumable that the larger difference observed between hybrids is related to genetic differences between them, that originated from the same heterozygous parents, while the two clones are genetically identical. This result indicates that different interspecific hybrids behave differently concerning expression and - putatively - amplification of transposons and suggests that, during evolution, the same hybridization event can determine different results in terms of genome structure of the resulting species. For example,

Table 5. Mean number of drought related cis-regulatory motifs in the consensus LTRs of *P. deltoides* and *P. nigra* obtained mapping *P. trichocarpa* LTRs with genomic DNA reads of *P. deltoides* and *P. nigra*. Two groups of REs are analysed, those whose expression was higher than the most expressed actin gene, and not expressed REs.

RE group	Species	Nr. of analysed REs	Mean nr. of drought-related cis-regulatory motifs		
			All	of which, ABRE	of which, DRE
Highly expressed	*P. deltoides*	32	19.94	1.94	7.12
	P. nigra	32	19.50	2.09	6.81
Not expressed	*P. deltoides*	463	12.17	1.14	3.68
	P. nigra	463	12.43	1.16	3.93

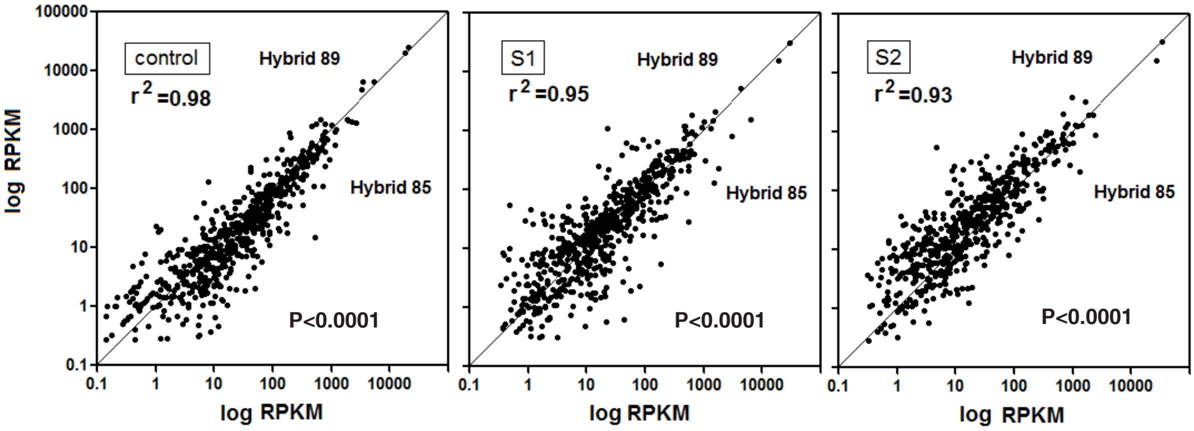


Figure 8. Two-dimensional representations of RE expression estimated log RPKM in leaves from two P. deltoides x *P. nigra* hybrids normally watered (control), moderately (S1) and severely drought stressed (S2). The line represents the hypothetical relation between variables if expression values were the same.

it is known that Helianthus anomalus, Helianthus deserticola, and Helianthus paradoxus derive from interspecific hybridization between H. annuus and H. petiolaris, but the RE portions of their genome are very different (Ungerer et al. 2009).

The large level of expression (and possibly amplification) of REs in the poplar interspecific hybrid analysed in our study might be related to the so called "genomic shock", a process related to the introduction of alien genetic material into a new genetic background (McClintock 1984). No data are available on RE expression in parental species, *P. deltoides* and *P. nigra*. We are currently determining RE expression in parental trees to clarify if (and what proportion of) RE transcription can be ascribed to genomic shock following interspecific hybridization.

On the whole, retrotransposon transcription appears stable during drought stress (Table 3). Many elements are expressed in control and stressed leaves, especially of the Unknown superfamily. Interestingly, many *Gypsy* elements are expressed only in one or two treatments, suggesting a larger specificity of this superfamily in the response to changes in environmental conditions. Alternatively, one can suppose that basal transcription of Unknown elements occurs in more cases because of general reduced silencing efficiency for these elements. On the contrary, more conserved *Gypsy* and, at a minor extent, *Copia* elements might be transcribed only in particular environmental conditions, in which their silencing mechanism might be less efficient.

Finally, searching for drought-related motifs in the promoters evidenced a difference between most active and inactive REs. However, a significant number of drought related cis-regulatory elements is present also in LTRs of inactive REs. Hence the activity of retrotransposons might be more affected by condensation/decondensation of chromatin (determined by siRNAs) than by the presence/absence of motifs in the promoters in their LTRs.

Acknowledgements

This research work supported by PRIN-MIUR, Italy, Project "Verso la delucidazione delle basi molecolari dell'eterosi nelle piante coltivate: variazione cis-regolatoria ed espressione genica in ibridi di pioppo."

References

Baggerly KA, Deng L, Morris JS, Aldaz CM (2003). Differential expression in SAGE: Accounting for normal between-library variation. Bioinformatics 19: 1477-1483.

Bennetzen JL (2000). Transposable elements contributions to plant gene and genome evolution. *Plant Mol Biol* **42**: 251-269.

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* **17**: 343–360.

Buti M, Giordani T, Vukich M et al. (2009). HACRE1, a recently inserted *copia*-like retrotransposon of sunflower (*Helianthus annuus* L.). *Genome* **11**: 904–911.

Buti M, Giordani T, Cattonaro F et al. (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.

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 Genetics & Genomes* **8**: 61-75. DOI: 10.1007/s11295-011-0421-3

Dieguez MJ, Vaucheret H, Paszkowski J, Mittelsten Scheid O (1998). Cytosine methylation at CG and CNG sites is not a prerequisite for the initiation of transcriptional gene silencing in plants, but it is required for its maintenance. *Mol*

Gen Genet. **259**: 207-215.

Doyle JJ, Doyle JL (1989). Isolation of plant DNA from fresh tissue. *Focus* **12:** 13-15.

Feschotte C (2008). Transposable elements and the evolution of regulatory networks. *Nat Rev Genet.* **9**: 397–405.

Grandbastien MA (1998). Activation of plant retrotransposons under stress conditions. *Trends Plant Science* **3:** 181-189.

Higo K, Ugawa Y, Iwamoto M, Korenaga T *(1999). Plant cis-acting regulatory DNA elements (PLACE) database. Nucleic Acids Res.* ***27:*** *297–300.*

Hirochika H (1993). Activation of tobacco retrotransposons during tissue culture. *EMBO J.* **12**: 2521-2528.

Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M (1996). Retrotransposons of rice involved in mutations induced by tissue culture. *PNAS* **93:** 7783-7788.

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH (2000). Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. PNAS 97: 6603-6607.

Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoy A, Schulman AH (2004). LARD retroelements: conserved, non-autonomous components of barley and related genomes. Genetics 166: 1437-1450.

Kawakami T, Dhakal P, Katterhenry AN, Heatherington CA, Ungerer MC (2011). Transposable element proliferation and genome expansion are rare in contemporary sunflower hybrid populations despite widespread transcriptional activity of LTR-Retrotransposons. Genome Biol Evol. 3: 156–167.

Lisch D (2009). Epigenetic regulation of transposable elements in plants. Annu Rev Plant Biol 60: 43–66.

Logemann J, Schell J, Willmitzer L (1987). Improved method for the isolation of RNA from plant tissues. Anal Biochem. 163: 16–20.

Maestrini P, Cavallini A, Rizzo M, Giordani T, Bernardi R, Durante M, Natali L (2009). Isolation and expression analysis of low temperature-induced genes in white poplar (*Populus alba*). Journal of Plant Physiology 166: 1544-1556.

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

Meyers BC, Tingey SV, Morgante M (2011). Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11: 1660-1676.

Morgante M, De Paoli E, Radovic S (2007). Transposable elements and the plant pan-genomes. Curr Opin Plant Biology 10: 149-155.

Morin RD, O'Connor MD, Griffith M et al. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res. 18: 610–621.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621–628.

Piegu B, Guyot R, Picault N et al. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res.16: 1262–1269.

Ramallo E, Kalendar R, Schulman AH, Martínez-Izquierdo JA (2008). Reme1, a *Copia* retrotransposon in melon, is transcriptionally induced by UV light. Plant Mol Biol 66: 137-150.

Rico-Cabanas L, Martinez-Izquierdo JA (2007). CIRE1, a novel transcriptionally active Ty1-*Copia* retrotransposon from Citrus sinensis. Mol Genet Genomics 277: 365–377.

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

Sterck L, Rombauts S, Jansson S, Sterky F, Rouzé P, Van de Peer Y (2005). EST data suggest that poplar is an ancient polyploid. New Phytol 167: 165–170.

Suoniemi A, Tanskanen J, Schulman AH (1998). *Gypsy*-like retrotransposons are widespread in the plant kingdom. Plant J 13: 699-705.

Tanskanen JA, Sabot F, Vicient C, Schulman AH (2006). Life without GAG: The BARE-2 retrotransposon as a parasite's parasite. Gene 390: 166-74.

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

Ungerer MC, Strakosh SC, Stimpson KM (2009). Proliferation of Ty3-*gypsy*-like retrotransposons in hybrid sunflower taxa inferred from phylogenetic data. BMC Biol 7: 40-52. doi:10.1186/1741-7007-7-40

Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. Science 313: 320-324.

Vicient CM, Jaaskelainen MJ, Kalendar R, Schulman AH (2001). Active retrotransposons are a common feature of grass genomes. Plant Physiol. 125: 1283-1292.

Vicient CM, Schulman AH (2002). *Copia*-like retrotransposons in the rice genome: few and assorted. Genome Lett 1: 35-47.

Volpe T, Kidner C, Hall I, Teng G, Grewal S, Martienssen R (2002). Heterochromatic silencing and histone H3 lysine 9 methylation are regulated by RNA interference. Science 297: 1833-1837.

Voytas DF, Cummings MP, Konieczny A, Ausubel FM, Rodermel SR (1992). *Copia*-like retrotransposons are ubiquitous among plants. PNAS 89: 7124-7128.

Vukich M, Schulman AH, Giordani T, Natali L, Kalendar R, Cavallini A (2009). *Copia* and *Gypsy* retrotransposons activity in sunflower (Helianthus annuus L.). BMC Plant Biol 9: 150.

Wicker T, Sabot F, Hua-Van A et al. (2007). A unified classification system for eukaryotic transposable elements. Nature Rev Genet 8: 973–982.

Witte CP, Le QH, Bureau T, Kumar A (2001). Terminal-repeat Retrotransposons In Miniature (TRIM) are involved in restructuring plant genomes. PNAS 98: 13778-13783.

Yamaguchi-Shinozaki K, Shinozaki K (2006). Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. Annu Rev Plant Biol. 57: 781-803.

Yamazaki M, Tsugawa H, Miyao A et al. (2001). The rice retrotransposon Tos17 prefers low-copy-number sequences as integration targets. *Mol Genet Genomics* **265**: 336-344.

Paper IV

High-throughput analysis of transcriptome variation during drought stress in a poplar hybrid

Abstract. Poplar interspecific hybrids are one of the most important forest crops. To contribute new data on molecular response of forest trees to drought, Illumina next generation sequencing technology was used to determine the sequence of most gene transcripts. Such an approach allowed to identify genes and gene networks that contribute to poplar tolerance to water-limiting environments, with a long-term aim of developing strategies to improve plant productivity under drought. We generated 76,635,449 sequence reads, each 51 nt in length, encompassing 3.9 Gb of sequence data from 12 cDNA libraries obtained from leaves of plants of *P. deltoides* x *P. nigra* subjected or not to moderate or severe drought stress. The expression of 45,033 poplar genes included in *P. trichocarpa* Phytozome database was studied by mapping Illumina cDNA reads at various stress stages on poplar unigene models. Expressed genes were characterized by gene ontology and by determining the metabolic pathway to which they belong. Most genes resulted expressed in control and drought stressed plants, however a number of genes was observed significantly induced or repressed by drought. Analysis of expression profiles revealed that only genes involved in the biological process of stress response showed, in the majority, a precocious induction at moderate drought stress (RWC around 85%). On the contrary, induction or repression of most of other genes was more common after severe stress (RWC around 55-60%), even for genes that usually respond promptly to changes in environmental conditions, as those encoding transcription factors. The dataset of expression profiles will be useful for future studies on other stresses and for crop improvement in poplar.

Introduction

A typical plant cell has more than 30,000 genes and an unknown number of proteins, which can have more than 200 known post-translational modifications (PTMs). The molecular responses of cells (and plants) to their environment are extremely complex (Cramer et al. 2011).

Recent advances in biotechnology have dramatically changed our capabilities for gene discovery and functional genomics. High throughput "omics" technologies are facilitating the identification of new genes and gene function. In addition, network reconstructions at the genome-scale are keys to quantify and characterize the genotype to phenotype relationships (Feist and Palsson 2008). Such a "systems biology" approach allows a deeper understanding of physiologically complex processes and cellular functions.

Boyer (1982) indicated that environmental factors may limit crop production by as much as 70%. A 2007 FAO report stated that only 3.5% of the global land area is not affected by some environmental constraint. It is evident that abiotic stress continues to have a significant impact on plants; yields of the "big 5" food crops are expected to decline in many areas in the future due to the continued reduction of arable land, reduction of water resources and increased global warming trends and climate changes (Lobell et al. 2011).

Abiotic stress is defined as environmental conditions that reduce growth and yield below optimum levels. The plant responses to stress are dependent on the tissue or organ affected by the stress. For example, transcriptional responses to stress are tissue or cell specific in roots and are quite different depending on the stress involved (Dinneny et al. 2008). In addition, the level and duration of stress (acute vs chronic) can have a significant effect on the complexity of the response (Tattersall et al. 2007, Pinheiro and Chaves 2011).

Water deficit inhibits plant growth by reducing water uptake into the expanding cells, and alters enzymatically the rheological properties of the cell wall, for example, by the activity of reactive oxygen species (ROS) on cell wall enzymes (Skirycz and Inzé 2010). In addition, water deficit alters the cell wall nonenzymatically, for example, by the interaction of pectate and calcium (Boyer 2009). Furthermore, water conductance to the expanding cells is affected by aquaporin activity and xylem embolism (Boursiac et al. 2008). The initial growth inhibition by water deficit occurs prior to any inhibition of photosynthesis or respiration (Hummel et al. 2010).

Growth is limited by the plant's ability to osmotically adjust or conduct water. The epidermal cells can increase the water potential gradient by osmotic adjustment, which may be largely supplied by solutes from the phloem. Such solutes are supplied by photosynthesis that is also supplying energy for growth and other metabolic functions in the plant. With long-term stress, photosynthesis declines due to stomatal limitations for CO_2 uptake and increased photoinhibition from difficulties in dissipating excess light energy (Pinheiro and Chaves 2011).

One of the earliest metabolic responses to abiotic stresses and the inhibition of growth is the inhibition of protein synthesis (Good and Zaplachinski 1994) and an increase in protein folding and processing (Liu and Howell 2010). Energy metabolism is affected as the stress becomes more severe (e.g. sugars, lipids and photosynthesis) (Cramer et al. 2007, Pinheiro and Chaves 2011). Thus, there are gradual and complex changes in metabolism in response to stress.

The plant molecular responses to abiotic stresses involve interactions and crosstalk with many molecular pathways. One of the earliest signals in many abiotic stresses involve ROS and reactive nitrogen species (RNS), which modify enzyme activity and gene regulation (Wilkinson and Davies 2010, Mittler et al. 2011). Hormones are also important regulators of plant responses to abiotic stress. The two most important are abscisic acid (ABA) and ethylene (Goda et al. 2008).

ABA is a central regulator of many plant responses to environmental stresses, particularly osmotic stresses (Chinnusamy et al. 2008, Hubbard et al. 2010). Its signalling can be very fast without involving transcriptional activity; a good example is the control of stomatal aperture by ABA through the biochemical regulation of ion and water transport processes (Kim et al. 2010). There are slower responses to ABA involving transcriptional responses that regulate growth, germination and protective mechanisms. Recently, the essential components of ABA signalling have been identified, and their mode of action was clarified. The current model of ABA signalling includes three core components, receptors (PYR/PYL/RCAR), protein phosphatases (PP2C) and protein kinases (SnRK2/OST1). The PYR/PYL/RCAR – PP2C – SnRK2 complex plays a key role in ABA perception and signalling (Ma et al. 2009, Park et al. 2009, Umezawa 2011).

Studies of the transcriptional regulation of dehydration stress have revealed both ABA-dependent and ABA-independent pathways (Yamaguchi-Shinozaki and Shinozaki 2006). Cellular dehydration under water limited conditions induces an increase in endogenous ABA levels that trigger downstream target genes encoding signalling factors, transcription factors, metabolic enzymes, and others (Yamaguchi-Shinozaki and Shinozaki 2006). In the vegetative stage, expression of ABA-responsive genes is mainly regulated by bZIP transcription factors (TFs) known as AREB/ABFs, which act in an ABA-responsive-element (ABRE) dependent manner (Yamaguchi-Shinozaki and Shinozaki 2006, Yoshida et al. 2010). Activation of ABA signalling cascades result in enhanced plant tolerance to dehydration stress. In contrast, a dehydration responsive cis-acting element, DRE/CRT sequence and its DNA binding ERF/AP2-type TFs, DREB1/CBF and DREB2A, are related to the ABA independent dehydration and temperature responsive pathways (Yamaguchi-Shinozaki and Shinozaki 2006). DREB1/CBFs function in cold-responsive gene expression (Yamaguchi-Shinozaki and Shinozaki 2005), whereas DREB2s are involved in dehydration-responsive and heat-responsive gene expression (Sakuma et al. 2006, Yamaguchi-Shinozaki and Shinozaki 2006).

Ethylene is also involved in many stress responses (Yoo et al. 2009), including drought. There are known interactions between ethylene and ABA during drought (Wilkinson and Davies 2010).

Forest crops are especially susceptible to drought stress, that can seriously affect biomass production. The genus *Populus* is an important crop and a model system to understand molecular processes of growth, development, and responses to environmental stimuli in trees.

The recent development of next-generation sequencing (NGS) technologies are changing the way transcriptomes and genomes are discovered and defined, including the 454-Roche (http://www.454.com; Margulies et al. 2005), ABI-SOLiD (http://www.appliedbiosystems.com; Pandey et al. 2008), and Solexa/Illumina (http://www.illumina.com; Bentley et al. 2008) technologies. The NGS approach has highlighted the benefits of providing a more thorough qualitative and quantitative description of gene expression than the microarray-based assays (Morozova and Marra 2008, Cantacessi et al. 2010, Wu et al. 2010, Wang et al. 2010, Xiong et al. 2011). In particular, the Illumina system can yield millions of short reads and is therefore more suitable for tag-based transcriptome sequencing and digital gene expression analysis. This combination enables the laborious cloning steps to be avoided, and the higher sequencing depth adds further to its potentially superior accuracy and precision compared to older methods (Wu et al. 2010, Wang et al. 2010, Yang et al. 2011). In light of this information, we have performed a genome-wide analysis of the transcriptome in leaves of an interspecific hybrid between *Populus deltoides* and *Populus nigra*, one of the most cultivated poplar in Northern America and Northern Europe, in response to drought stress using high-throughput techniques of cDNA sequencing as Illumina.

Materials and Methods

Sample preparation and sequencing

Rooted cuttings of *Populus deltoides* x *P. nigra* hybrids, produced at INRA, Orleans (France), were cultivated in 20 x 20 cm^2 pots in the open.

In the late spring 2011, some hybrid plants of 50 cm in height were normally watered and others were subjected to drought by suspending watering. Leaf water loss during the experiment was followed by relative water content measurement [RWC = 100 (FW-DW) / (TW - DW)], where FW is the fresh weight, DW the dry weight and TW the turgid weight. One leaf was collected from each plant and divided into two portions: one was used for RNA isolation, the other was used to measure tissue hydration by determining the RWC. The experimental design was as follows: 2 clones (biological replicates) x 3 treatments (control, moderate, and severe drought stress) x 2 hybrids (obtained from the same parents).

Total RNA was isolated from leaves of single plants with different RWC, according to the method described by Logemann et al. (1987), followed by DNAse I (Roche) treatments according to the manufacturer's instructions to completely remove genomic DNA contamination.

RNA-Seq library was generated using the TruSeq RNA-Seq Sample Prep kit according to the manufacturer's protocol (Illumina Inc., San Diego, CA). In short, poly-A RNA was isolated from total RNA and chemically fragmented. First and second strand synthesis were followed by end repair, and adenosines were added to the 3' ends. Adapters were ligated to the cDNA and 200 ± 25 bp fragments were gel purified and enriched by PCR. The library was quantified using Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA) and run on the Illumina HiSeq2000 (Illumina Inc.) using version 3 reagents. Single-read sequences of length 51 bp were collected.

Alignment and analysis of Illumina reads against the **P. trichocarpa** *unigene model database*

Sequence alignments were generated with CLC-BIO Genomic Workbench 4.9, using the *P. trichocarpa* unigene model database (Tuskan et al. 2006), available at the Phytozome site (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Ptrichocarpa/). The following parameters were used for alignments: maximum number of mismatches= 2, minimum number of reads = 10.

For evaluation of gene expression we calculated the number of mapped reads per kilobase per million mapped reads, measuring the transcriptional activity for each gene. CLC-BIO Genomic Workbench computes this normalized gene locus expression level (named RPKM) by assigning reads to a sequence in the database and counting them.

The RPKM value (Mortazavi et al. 2008) estimates the number of reads falling in a given gene locus as follows:

$$RPKM = N / (L \times N_{tot} \times 10^{-6})$$

where N = number of mapping reads at a given gene locus, L = estimated length (Kbp) of the coding portion of the gene, N_{tot} = number of total mapping reads.

Identification of differentially expressed genes and related metabolic pathways

Functional annotation of genes induced or repressed by drought was made according to their expression profile. Expression profiles were evaluated considering RPKM values in control, moderately, and severely drought stressed plants using Baggerly's test (Baggerly et al. 2003). This compares the proportions of counts in a group of samples against those of another group of samples, and is suited to cases where replicates are available in the groups. The samples are given different weights depending on their sizes (total counts). The weights are obtained by assuming a Beta distribution on the proportions in a group, and estimating these, along with the proportion of a binomial distribution, by the method of moments. The result is a weighted t-type test statistic.

The weighted proportions fold changes between treatments were considered as significant when weight of a sample was at least 3-fold higher or lower than another. Gene expression profiles were subdivided into nine groups: those remaining constant, those increasing their expression in S1 or in S2 or in both stress levels, those reducing their expression in S1 or S2 or in both stress levels, those increasing their expression in S1 and reducing in S2 and vice versa.

The annotation of differentially expressed genes was based on the annotation notes reported in the *P. trichocarpa* Phytozome database, after mapping cDNA sequence reads to this database. Phytozome codes were used for the identification and the mapping of gene ontology (GO) terms to transcripts at the site Popgenie (http://www.popgenie.org/).

The web tool GO term classification counter CateGOrizer (http://www.animalgenome.org/bioinfo/tools/countgo/) was used for grouping and counting GO classes using the GO-Slim method (Hu et al. 2008) for each library, without counting the three root classes (Cellular Component, Biological Process and Molecular Function). Then, for sake of simplicity, GO terms of similar function were assigned to a unique functional class (see Supplementary material).

For the identification of metabolic pathways in which induced or repressed are involved, we have used the MapMan 3.5.1R2 tool (Thimm et al. 2004, Usadel et al. 2009). Expression values were treated as follows: data are reported as weighted proportions fold change between moderately or severely drought stressed plants and control plants as above. When values were higher in stressed than in control plants they were reported as positive, when they were higher in control than in stressed plants as negative, thus leading to a '+' value in case of above-average expression levels and a '-' value in case of below-average expression levels.

Only data related to genes with value higher than +3 or lower than -3 in at least one treatment were exported to MapMan, that converts the data values to colour scale: the transcripts not called are represented as grey, transcripts that change by less than MapMan threshold value of 0.5 are white, transcripts increased are blue and transcripts decreased are red.

RT-PCR Validation

Relative RT-PCR experiments were carried out as follows: first-strand cDNA synthesis was performed with 3 μg of total RNA using M-MLV Reverse Transcriptase RNase H (Solis Biodyne), according to the manufacturer's instructions.

Forward and reverse specific primers were designed on three differentially expressed genes (POPTR_0019s10270.1, POPTR_0008s11610.3, POPTR_0006s14720.1) using an actin encoding gene as standard (POPTR_0019s02630.1). The PCR amplification was carried out in non saturating conditions and involved a 95°C cycle hold for 15 min, followed by 30 cycles at 95°C for 30 s, 58°C for 30 s, and 72°C for 30 s. The PCR products were separated in 2% agarose GelRed^{TM} stained.

Results

Global Analysis of Gene Expression

To obtain a global view of the transcriptome during drought stress in a poplar hybrid, we used a Illumina Genome Analyzer to perform high-throughput tag sequencing analysis on cDNAs from 12 libraries (three stress conditions per two hybrids per two clones of the same hybrid [biological replicates], Table 1). We generated 76,635,449 sequence reads, each 51 nt in length, encompassing 3.9 Gb of sequence data. The total number of tags per library ranged from 3.96 to 15.33 millions, a tag density sufficient for quantitative analysis of gene expression (Morin et al. 2008).

The sequence reads were aligned on the *P. trichocarpa* Phytozome unigene database (Tuskan et al. 2006), using the CLC-BIO software set to allow two base mismatches. The distribution of total and distinct tag counts over different tag abundance categories showed very similar tendencies for all libraries (Table 1). Of the total reads, 62.8% matched either to a unique (47.7%) or to multiple (15.1%) unigene sequences; 37.2% of the tags could not be mapped to the gene sequences.

CLC-BIO measures gene expression in reads per exon kilobase per million mapped sequence reads (RPKM) normalized measure of exonic read density that allows transcript levels to be compared both within and between samples (Mortazavi et al. 2008).

As CLC-BIO distributes multireads at similar loci in proportion to the number of unique reads recorded, we included in the analysis both unique reads and reads that occur up to 10 times to avoid undercount for genes that have closely related paralogs (Mortazavi et al. 2008). We evaluated the expression of 45,033 gene models included in the *P. trichocarpa* Phytozome database, during drought treatments.

RWC of leaves from which RNAs were isolated

Table 1. Number of Illumina reads matching to the *P. trichocarpa* unigene database (45,033 CDS sequences) for each library (C, control; S1, moderate stress; S2, severe stress).

Lybrary	Counted fragments			Uncounted
	Total	Unique	Non specific	
Hybrid 85, clone 3 (C)	10,360,767	7,814,681 (75.4)	2,546,086 (24.6)	4,966,787
Hybrid 85, clone 4 (C)	2,624,418	2,012,238 (76.7)	612,180 (23.3)	1,691,646
Hybrid 89, clone 6 (C)	6,057,500	4,548,228 (75.1)	1,509,272 (24.9)	2,869,614
Hybrid 89, clone 8 (C)	2,475,842	1,866,250 (75.4)	609,592 (24.6)	1,487,870
Hybrid 85, clone 12 (S1)	3,134,221	2,378,566 (75.9)	755,655 (24.1)	2,353,124
Hybrid 85, clone 24 (S1)	3,121,951	2,358,175 (75.5)	763,776 (24.5)	1,881,363
Hybrid 89, clone 10 (S1)	3,963,040	2,973,101 (75.0)	989,939 (25.0)	2,160,444
Hybrid 89, clone 15 (S1)	4,780,311	3,583,306 (75.0)	1,197,005 (25.0)	2,574,769
Hybrid 85, clone 42 (S2)	2,228,413	1,728,419 (77.6)	499,994 (22.4)	1,756,773
Hybrid 85, clone 45 (S2)	3,132,978	2,406,895 (76.8)	726,083 (23.2)	2,582,381
Hybrid 89, clone 20 (S2)	2,327,692	1,832,969 (78.7)	494,723 (21.3)	2,248,212
Hybrid 89, clone 35 (S2)	3,892,818	3,061,307 (78.6)	831,511 (21.4)	1,962,515
Total	48,099,951	36,564,135 (76.0)	11,535,816 (24.0)	28,535,498

is reported in Table 2: moderately stressed leaves (S1) showed a RWC ranging from 84.89 to 86,31, severely stressed leaves (S2) RWC ranged from 52.78 to 61.83. The expression in the three culture treatments is also summarized. To minimize false positives and negatives, we estimated that expression of a gene was significant when RPKM value > 1 in at least one of the two clones of the two hybrids. By this way, we could identify more than 28,000 genes that were significantly expressed at least at one stage. We have also calculated the number of genes that are expressed in all treatments or whose expression is detectable only in one or two treatments by comparing the mean RPKM value of each gene and considering as significant only RPKM values > 1 (Figure 1). It can be observed that, out of 28,714 genes that are expressed in at least one treatment, the vast majority (75.4%) is expressed in all treatments. However, a number of genes is specifically expressed in moderately and/or severely drought stressed leaves (on the whole, 3,483 genes, 12.1%) showing a significant change in the transcription pattern. Intriguingly, a low number of genes are expressed in leaves of control and severely droughted plants (1.1%).

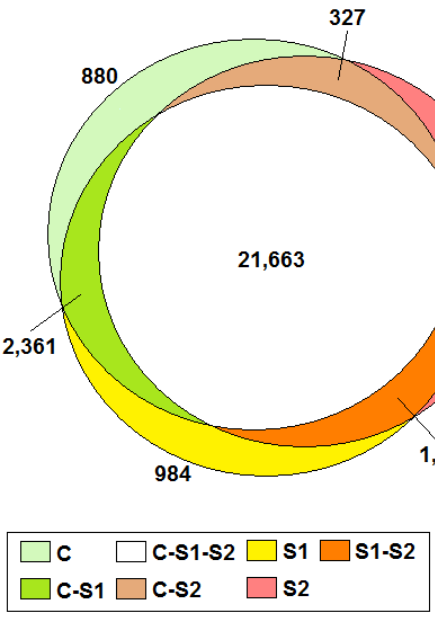


Figure 1. Venn diagrams representing genes expressed (RPKM >1, see text) in leaves of control (C), moderately drought stressed (S1), and severely drought stressed (S2) poplar hybrid plants.

Of the 45,033 poplar gene models, from 21.3 to 24.5% did not have any detectable transcriptional activity across all treatments indicating that either these models are not expressed in any of the developmental stages/tissues examined in the present study or the models do not represent bona fide genes. Additional transcriptome profiling across different developmental stages/tissues (e.g., apical and vegetative meristems, developing inflorescence) and/or different abiotic/biotic environmental variables will be required to further assess the transcriptional activity of these genes.

Stress-induced gene expression

To determine the appropriate read depth criteria for differential gene expression, in preliminary experiments, we established that a 3X ratio cutoff in one of the two samples being compared (i.e., moderate drought stress vs. control) minimizes the rate of false positives while retaining genes of lower expression (data not shown).

The expression profiles of the differentially expressed genes were determined by calculating the weighted proportions fold change (see Materials and Methods) between moderately or severely stressed and control leaves, estimating a significant difference when ratio was higher than 3.0 or lower than -3.0. Genes were subdivided into nine clusters based on their expression modulation (Figure 2). Genes positively or negatively modulated along the whole time course (clusters 1 and 2) are relatively few (only 2.37 and 3.73%, respectively). Genes that are already induced or repressed with moderate stress (clusters 3 and 4) are slightly more frequent (4.83 and 4.69%). Interestingly, clusters 5 and 6, that include genes positively or negatively modulated only by severe stress are the most numerous (10.81 and 16.28%). However, the majority of genes are expressed at the same levels in the three stages (49.43%, cluster 9).

Though the method based on RNAseq has been reported as highly reliable (Zenoni et al. 2010), we have performed reverse transcription PCR on three randomly chosen mRNAs that were differentially expressed in response to moderate or severe stress, validating the differential expression data obtained by RNAseq (data not shown).

To facilitate the global analysis of gene expression, functional categories were assigned

Table 2. Number of *P. trichocarpa* unigene models showing detectable matching to Illumina reads during drought treatments. The extent of expression was estimated calculating RPKM (see text). Each treatment includes two clones of two hybrids. For each treatment the mean leaf RWC is reported.

Treatment	Leaf RWC	Nr. of expressed genes		Nr. of not detectable genes (%)
		RPKM ≥ 1	0 < RPKM < 1	
Control	94.81	28,789	6,659	9,585 (21.3)
Moderate drought	85.78	29,484	5,202	10,347 (23.1)
Severe drought	57.27	28,263	5,722	11,048 (24.5)

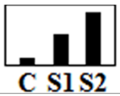
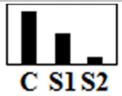
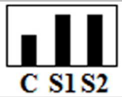
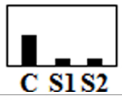
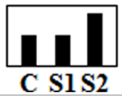
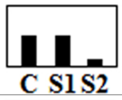
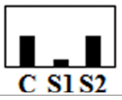
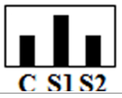
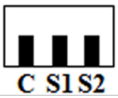
Expression profile	Nr. of genes	%
	229	2.37
	360	3.73
	466	4.83
	453	4.69
	1,04	10.81
	1,57	16.28
	417	4.32
	341	3.53
	4,77	49.43

Figure 2. Number of unigene models per expression profile as schematized on the left (C, control; S1, moderate stress; S2, severe stress). A total of 9,654 unigene models showing at least 3-fold expression variation between control and moderately (S1) or severely (S2) drought stressed *P. deltoides* x *P. nigra* plants were analysed.

to all predicted poplar genes using the GO term assigned to *P. trichocarpa* sequences. Then, the GO term occurrence in the different profiles of expression was calculated using the GO term classification counter CateGOrizer (Hu et al. 2008). The functional category distribution frequency was calculated for each cluster to identify differences in the distribution of genes among the three stress conditions. A summary of the results for each of the three possible expression profiles in the hybrid (i.e., showing increasing expression at least in one stress condition compared to the control, stable, or showing decreasing expression at least in one stress condition) is given in Figure 3, in which, for the sake of simplicity, GO terms of similar function were assigned to a unique functional class (see Supplementary material). Functional classes were sorted by biological process, cellular component, and molecular function. Within each root class differences were seen (Figure 3). Within the molecular function class, "catalytic activity" and "binding" were the most abundant terms, especially in those expression profiles that show changes in transcription rate.

Within the cellular component class, the term "cell part", indicating intracellular processes, was by far the most frequent, in all observed expression profiles, but especially in genes whose expression remained stable. In the biological process class, "metabolism" was the most frequent term; though representing a small portion of gene categories, the term "response to stimula" was observed especially in the profiles showing induction or repression of the transcription.

Biochemical pathways activated by drought

With the objective to display differentially expressed genes onto pathways and to obtain an overview of genes affected in response to drought in *P. deltoides* x *P. nigra*, the MapMan 3.5.1R2 tool was used on 4,391 genes that could be unambiguously treated by MapMan and for which large differential expression values compared to controls (i.e., weighted proportions fold change > 3 or < -3) were observed at least at one stress stage. MapMan allowed the assignment of 4,428 genes, being some of the genes mapped to multiple pathways, into 34 of a total of 35 functional categories. While a large number of genes (1,443) were classified as unknown or not assigned category, the remaining 2,985 genes were identified as belonging to known metabolic pathways or large enzyme families. Among gene categories, 1,874 (62.8%) genes belonged to six categories and had higher proportion of genes comparatively, which include protein metabolism (535 genes), RNA metabolism (493 genes), miscellaneous enzyme families (260 genes), signalling (215 genes), transport (201 genes) and stress (170 genes).

We explored gene categories that are presumably activated during drought response using the Image annotator module of the MapMan application. We selected genes related to transcription regulation, stress responses, energy metabolism, and secondary metabolism, that are well documented to be responsive to wide-array of stresses.

Many genes (396) assigned to transcription factors of different classes were identified and mapped. They are reported in Figure 4. For instance, genes encoding early response to dehydration-related proteins of *Arabidopsis* (code AP2-EREBP) were highly expressed especially after severe drought. Moreover, large expression variations were observed within many transcription families as MYB domain containing family, bHLH family protein, WRKY, Zinc-finger (C2C2, C2H2, C3H) families. Also the NAC domain protein encoding gene family, known as being involved in drought stress response (Ooka et al. 2003, Le et al. 2011) was found to be affected.

Concerning energy metabolism, changes in the magnitudes of enzymes and metabolites of carbon and energy cycles have been documented to play crucial roles in cellular metabolism during the response to stress (Apel and Hirt 2004). The induction of respiratory activities in mitochondria and ATP released during these reactions help to initiate tolerance events under stress conditions, for example during hypoxia (Kreuzwieser et al. 2009). Forty-six genes related to energy metabolism were identified in our analyses (Table 3); of these, 14 were strongly induced with both moderate and severe drought stress, and other 19 were strongly transcribed after severe stress. Genes encoding mitochondrial electron transporters, isocitrate dehydrogenase (3 genes), malate dehydrogenase (4 genes), phosphoenol-pyruvate carboxy kinase, etc. were induced.

Secondary metabolites as flavonoids and

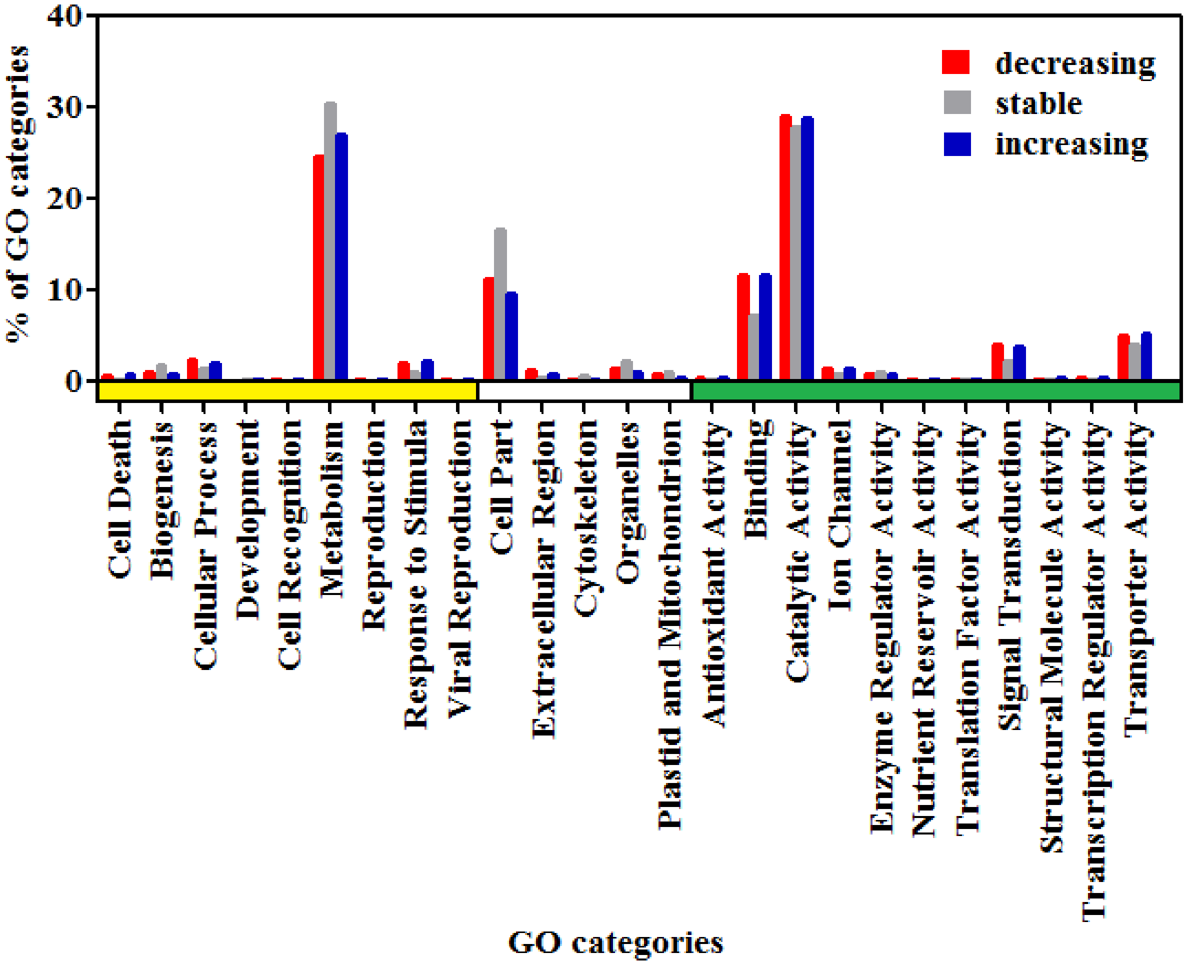


Figure 3. Functional characterization of poplar expressed genes. Genes were categorized hierarchically according to three possible expression behaviours, i) showing a reduction in moderately and/or severe drought compared to the control (profiles 2, 4, 6, 7, see Fig. 2); ii) showing a stable expression (profile 9); showing an increase in moderately and/or severe drought compared to the control (profiles 1, 3, 5, 8); and according to three principal gene ontologies, biological processes, cellular components, and molecular functions (indicated in the x-axis by the horizontal bar, with yellow, white, and green, respectively).

Table 3. Number of *P. trichocarpa* unigene models involved in energy metabolism, secondary metabolism, and stress response according to the Mapman pathway database (see text) that show at least 3-fold expression variation between control and moderately (S1) or severely (S2) drought stressed *P. deltoides* x *P. nigra* plants. Only genes unambiguously matching to AGI *Arabidopsis* unigene model were analysed.

Metabolic pathway	Category	Nr. of identified genes	Highly induced by drought stress	
			S1	S2
Energy metabolism	Respiratory enzymes	19	6	16
	Metabolite transporters	6	6	6
	Mitochondrial electron transporter	21	2	11
Total		46	14	33
Secondary metabolism	Flavonoids	16	3	5
	Phenylpropanoids	18	5	11
	Isoprenoids	17	5	6
	Shikimate pathway	4	1	3
	Wax	2	1	2
	Other pathways	16	1	6
Total		73	16	33
Stress response	Dehydration-related	20	8	8
	HSP and HSP-related	32	10	10
	NBS-LRR class	20	6	5
	Pathogenesis-related	28	6	3
	Other genes	70	15	11
Total		170	45	37

isoflavonoids are known to play a significant role in plant defence responses to pathogens (Dixon and Steele 1999, Uppalapati et al. 2009). The expression of 73 genes related to secondary metabolism were observed in our analysis (Table 3); of these, 16 were induced in response to both moderate and severe stress, other 17 were activated by severe stress. For example, genes related to phenylpropanoids (11 genes, for example the cinnamoyl-CoA reductase family), flavonoids (5 genes, for example two chalcone synthase encoding genes), and isoprenoid metabolism (6 genes, for example encoding mevalonate kinase and decarboxylase) were positively affected, especially by severe drought. Interestingly, also genes involved in wax biosynthesis were strongly induced.

Concerning genes responding to stress factors such as heat shock, anaerobiosis, plant pathogens, oxygen free radicals, heavy metals, water stress and chilling in plants, they have been assessed in various plant species (Matters and Scandalios 1986). In our study, 170 genes with stress-related annotations (either biotic or abiotic), were identified as affected positively or negatively by drought. Differently from the above mentioned classes, induction of stress-related genes is more frequent after moderate than severe stress (Table 3). Genes whose expression was increased by both moderate and severe stress include nine encoding heat shock proteins (HSP) or HSP-binding and other that encode two DNAJ heat shock proteins, a responsive to desiccation 2 (RD2) protein, and an early ERD-related protein. Some dehydrin encoding genes are not affected and others are induced only with moderate stress.

Discussion

Drought-responsive genes were identified using Illumina sequence data generated from leaves of drought stressed plants of two *P. deltoides* x *P. nigra* hybrids. Illumina sequencing has been proved very efficient in the identification of differentially expressed genes (Hoen et al. 2008). Rare and low-abundant transcripts can be detected, resulting in a comparatively greater number of analysed genes than using other technologies. Moreover, differently from other technologies as microarray, cross-hybridization artefacts are avoided and the sequence-based analysis does not require background correction. On the other hand, the appropriate mapping of short reads on annotated regions and assignment of multi-mapping sequences are still critical challenges, though the development of new algorithms for analysis is rapidly overcoming these difficulties (Mortazavi et al. 2008, Shendure 2008).

Poplars are usually sensitive to adverse conditions, in particular to water-limiting environments (Tschaplinski et al. 2006). We used Illumina next generation sequencing technology for determining the sequence of most gene transcripts and to identify genes and gene networks that contribute to poplar tolerance to water-limiting environments with a long-term aim of developing strategies to improve plant productivity under drought.

The expression of 45,033 poplar genes included in *P. trichocarpa* Phytozome database was studied

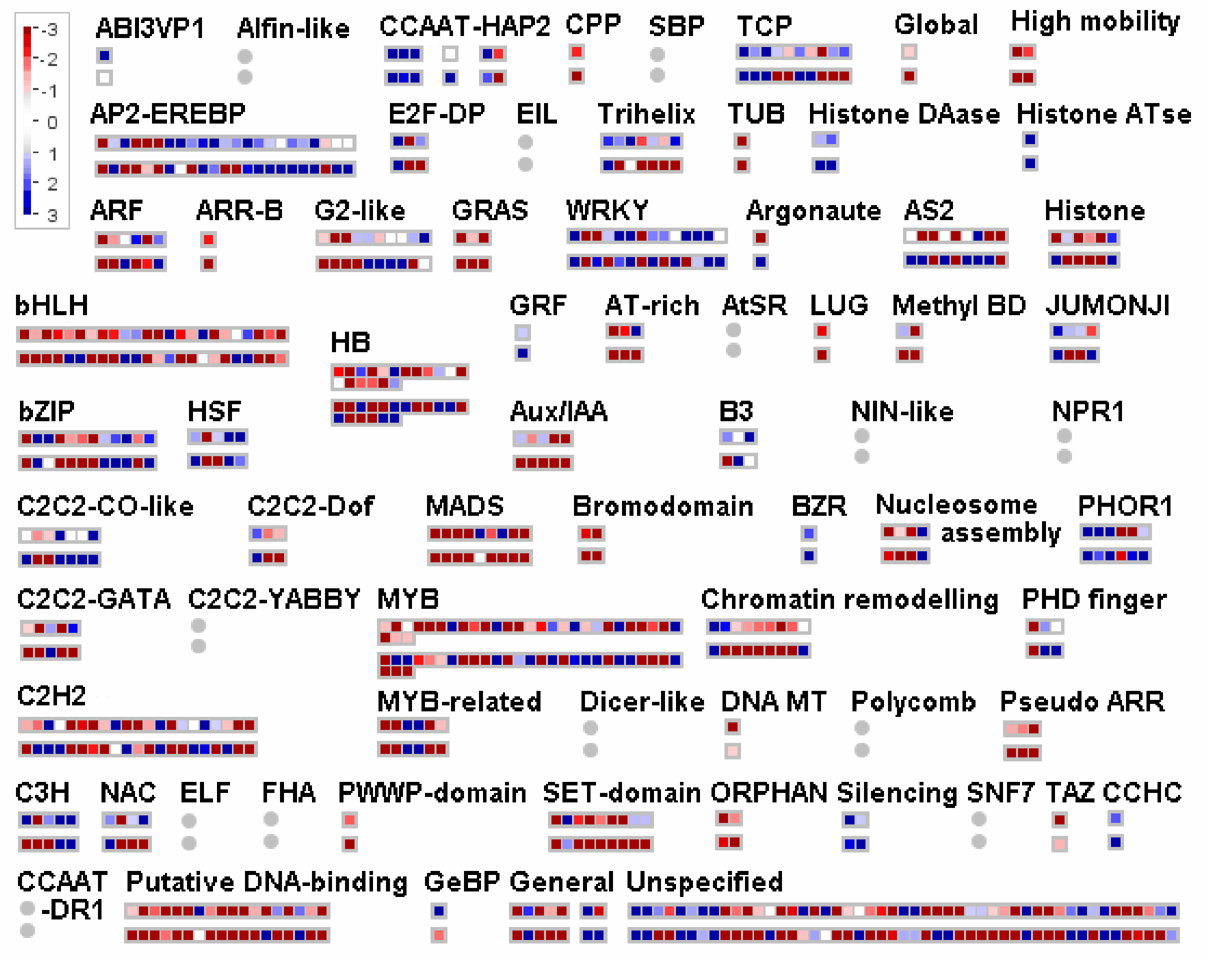


Figure 4. Schematic global representation of expression levels of genes involved in regulation of transcription as obtained using Mapman software on genes induced or repressed (RPKM ratio >3 or <-3) by moderate (upper lines) or severe stress (lower line). Only genes unambiguously matching to the AGI Arabidopsis unigene model are reported. The scale of expression is reported on top left. Each transcription factor family is indicated by its Mapman code.

by mapping Illumina cDNA reads at various stress stages on poplar unigene models.

This global analysis of gene expression provided a comprehensive dataset (Supplementary material 2) in which each gene is represented by its absolute expression level in control, moderately dehydrated and severely dehydrated leaves and by a GO biological process annotation. We observed 9,654 genes with significant expression changes, of which expression profiles during progressive drought stress was established and that were associated with gene ontology annotations. Though limited to 4,391 genes that could be unambiguously treated by MapMan, it was also possible to have a general overview of the metabolic pathways activated or repressed by drought.

Most genes resulted highly expressed in control and droughted plants, suggesting that they were either not affected or only moderately affected by drought, On the other hand, a number of genes was observed significantly induced or repressed by drought and may constitute a useful dataset for further studies.

Analysis of expression profiles of genes that usually respond to a wide array of stresses revealed that only genes involved in the biological process of stress response are, in the majority, precociously induced at moderate drought stress (RWC about 85%). On the contrary, induction or repression of most of other genes was more common after severe stress (RWC 55-60%), even for genes that are usually described as responding promptly to changes in environmental conditions, as those encoding transcription factors. These data indicate that, for many poplar genes, even significant reduction of relative water content (as that observed in S1) is not the signalling event determining change in gene expression.

Systems biology approaches have given us a more holistic view of the molecular responses. The response of the plant to abiotic stress are dynamic and complex and cannot be based only on the analysis of gene expression but the integration of multiple omics studies is necessary (Cramer et al. 2011). Actually, many gene categories as revealed by MapMan analyses, show that, within a family, some genes are repressed and other are induced. Structural genomics studies are necessary to establish if such differences between members of one and the same gene family can be ascribed to differences in cis-regulatory sequences. Moreover, such differences in gene expression can be also related to epigenetic regulation by the environment. Great changes in DNA methylation have been observed among poplar clones, possibly influencing their response to drought (Raj et al. 2011). NGS technologies provide new opportunities to analyze non coding RNAs and can clarify aspects of epigenetic regulation of gene expression (Gregory et al. 2008, Zhang et al. 2006). NGS analyses have delucidated the global transcriptomes of plants exposed to abiotic stresses such as dehydration, cold, heat, high-salinity, osmotic stress, and ABA (Matsui et al. 2008, Zeller et al. 2009), indicating that these stresses increase or decrease transcript abundance from stress-responsive genes, but also from thousands of unannotated nonprotein-coding regions. Matsui et al. (2008) estimated that approximately 80% of previously unannotated upregulated transcripts arise from antisense strands of sense transcripts. In our analyses, 37.2% of Illumina reads could not be mapped to previously annotated genes. It is plausible that many of these do represent antisense transcripts, whose biological function is still to be clarified but can probably be involved in the epigenetic regulation of drought tolerance. A number of drought related microRNAs have been recently identified in *Populus euphratica* (Li et al. 2011). This dataset and specific microRNA datasets of *P. deltoides* x *P. nigra* hybrids will be useful for clarifying the nature and the function of unannotated transcripts.

Acknowledgements

This research work supported by PRIN-MIUR, Italy, Project “Verso la delucidazione delle basi molecolari dell’eterosi nelle piante coltivate: variazione cis-regolatoria ed espressione genica in ibridi di pioppo.”

References

Apel K, Hirt H (2004). Reactive oxygen species: metabolism, oxidative stress, and signal transduction. *Annu. Rev. Plant Biol.* **55**: 373–399.

Baggerly KA, Deng L, Morris JS, Aldaz CM (2003). Differential expression in SAGE: Accounting for normal between-library variation. *Bioinformatics* **19**: 1477-1483.

Bentley DR, Balasubramanian S, Swerdlow HP et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.

Boursiac Y, Boudet J, Postaire O, Luu DT, Tournaire-Roux C, Maurel C (2008). Stimulus-induced downregulation of root water transport involves reactive oxygen species-activated cell signalling and plasma membrane intrinsic protein internalization. *Plant J* **56**: 207-218.

Boyer JS (1982). Plant productivity and environment. *Science* **218**: 443-448.

Boyer JS (2009). Evans Review: Cell wall biosynthesis and the molecular mechanism of plant enlargement. *Funct Plant Biol* **36**: 383-394.

Cantacessi C, Jex AR, Hall RS et al. (2010). A practical, bioinformatic workflow system for large data sets generated by next-generation sequencing. *Nucleic Acids Res* **38**: e171.

Chinnusamy V, Gong Z, Zhu JK (2008) Abscisic acid-mediated epigenetic processes in plant development and stress responses. *J Integr Plant Biol* **50**: 1187-1195.

Cramer GR, Ergul A, Grimplet J et al. (2007). Water and salinity stress in grapevines: early and late changes in transcript and metabolite profiles. *Funct Integr Genomics* **7**: 111-134.

Cramer GR, Urano K, Delrot S, Pezzotti M, Shinozaki K (2011). Effects of abiotic stress on plants: a systems biology perspective. *BMC Plant Biology* **11**: 163.

Dinneny JR, Long TA, Wang JY et al. (2008). Cell identity mediates the response of Arabidopsis roots to abiotic stress. *Science* **320**: 942-945.

Dixon RA, Steele CL (1999). Flavonoids and isoflavonoids – a gold mine for metabolic engineering. *Trends Plant Sci.* **4**: 394–400.

Feist AM, Palsson BO (2008). The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotechnol* **26**: 659-667.

Goda H, Sasaki E, Akiyama K et al. (2008). The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. *Plant J* **55**: 526-542.

Good AG, Zaplachinski ST (1994). The effects of drought stress on free amino acid accumulation and protein synthesis in *Brassica napus*. *Physiol Plant* **90**: 9-14.

Gregory BD, Yazaki J, Ecker JR (2008). Utilizing tiling microarrays for whole genome analysis in plants. *Plant J.* **53**: 636-644.

Hoen PA, Ariyurek Y, Thygesen HH et al. (2008). Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res* **36**: e141.

Hu ZL, Bao J, Reecy JM (2008). CateGOrizer: A web-based program to batch analyze Gene Ontology classification categories. *Online J Bioinform* **9**: 108-112.

Hubbard KE, Nishimura N, Hitomi K, Getzoff ED, Schroeder JI (2010). Early abscisic acid signal transduction mechanisms: newly discovered components and newly emerging questions. *Genes Dev* **24**: 1695-1708.

Hummel I, Pantin F, Sulpice R et al. (2010). *Arabidopsis* plants acclimate to water deficit at low cost through changes of carbon usage: an integrated perspective using growth, metabolite, enzyme, and gene expression analysis. *Plant Physiol* **154**: 357-372.

Kim TH, Bohmer M, Hu H, Nishimura N, Schroeder JI (2010). Guard cell signal transduction network: advances in understanding abscisic acid, CO_2, and Ca^{2+} signaling. *Annu Rev Plant Biol* **61**: 561-591.

Kreuzwieser J, Hauberg J, Howell KA et al. (2009).

Differential response of gray poplar leaves and roots underpins stress adaptation during hypoxia. *Plant Physiol.* **149**: 461–473.

Le DT, Nishiyama R, Watanabe Y, Mochida K, Yamaguchi-Shinozaki K, Shinozaki K (2011). Tran LSP: Genome-wide survey and expression analysis of the plant-specific NAC transcription factor family in soybean during development and dehydration stress. *DNA Res* **18**: 263–276.

Li B, Qin Y, Duan H, Yin W, Xia X (2011). Genome-wide characterization of new and drought stress responsive microRNAs in *Populus euphratica*. *J Exp Bot* **62**: 3765–3779.

Liu JX, Howell SH (2010). Endoplasmic reticulum protein quality control and its relationship to environmental stress responses in plants. *Plant Cell* **22**: 2930-2942.

Lobell DB, Schlenker W, Costa-Roberts J (2011). Climate trends and global crop production since 1980. *Science* **333**: 616-620.

Logemann J, Schell J, Willmitzer L (1987). Improved method for the isolation of RNA from plant tissues. *Anal. Biochem.* **163**: 16–20.

Ma Y, Szostkiewicz I, Korte A et al. (2009). Regulators of PP2C phosphatase activity function as abscisic acid sensors. *Science* **324**: 1064-1068.

Matsui A, Ishida J, Morosawa T et al. (2008). *Arabidopsis* transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a timing array. *Plant Cell Physiol* **49**:1135-1149.

Margulies M, Egholm M, Altman WE et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.

Matters GL, Scandalios JG (1986). Changes in plant gene expression during stress. *Dev. Genet.* **7**: 167–175.

Mittler R, Vanderauwera S, Suzuki N et al. (2011). ROS signaling: the new wave? *Trends Plant Sci.* **16**: 300-309.

Morin RD, O'Connor MD, Griffith M et al. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.* **18**: 610–621

Morozova O, Marra MA (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**: 255–264.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621–628

Ooka H, Satoh K, Doi K et al. (2003). Comprehensive analysis of NAC family genes in *Oryza sativa* and *Arabidopsis thaliana*. *DNA Res* **10**: 239–247.

Pandey V, Nutter RC, Prediger E (2008). Applied biosystems SOLiD™ system: ligation-based sequencing. M. Jantz (Ed.), Next generation genome sequencing: towards personalized medicine, Wiley, Milton, Australia, pp. 29–41.

Park SY, Fung P, Nishimura N et al. (2009) Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of START proteins. *Science* **324**: 1068-1071.

Pinheiro C, Chaves MM (2011). Photosynthesis and drought: can we make metabolic connections from available data? *J Exp Bot* **62**: 869-882.

Raj S, Brautigam K, Hamanishi ET et al. (2011). Clone history shapes *Populus* drought responses. *PNAS* **108**: 12521–12526.

Sakuma Y, Maruyama K, Osakabe Y et al. (2006). Functional analysis of an *Arabidopsis* transcription factor, DREB2A, involved in drought-responsive gene expression. *Plant Cell* **18**: 1292-1309.

Shendure J (2008). The beginning of the end for microarrays? *Nat. Methods* **5**: 585–587.

Skirycz A, Inzé D (2010). More from less: plant growth under limited water. *Curr Opin Biotechnol* **21**: 197-203.

Tattersall EA, Grimplet J, Deluc L et al. (2007). Transcript abundance profiles reveal larger and more complex responses of grapevine to chilling compared to osmotic and salinity stress. *Funct Integr Genomics* **7**: 317-333.

Thimm O, Blasing O, Gibon Y et al. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **37**: 914–939.

Tschaplinski TJ, Tuskan GA, Sewell MM, Gebre GM, Todd DA, Pendley CD (2006). Phenotypic variation and quantitative trait locus identification for osmotic potential in an interspecific hybrid inbred F2 poplar pedigree grown in contrasting environments. *Tree Physiology* **25**: 595-604.

Tuskan GA, Difazio S, Jansson S et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596-604.

Umezawa T (2011). Systems biology approaches to abscisic acid signaling. *J Plant Res* **124**: 539-548.

Uppalapati SR, Marek SM, Lee HK et al. (2009). Global gene expression profiling during *Medicago truncatula*-Phymatotrichopsis omnivore interaction reveals a role for jasmonic acid, ethylene, and the flavanoid pathway in disease development. *Plant Physiol.* **22**: 7–17.

Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A, Stitt M. (2009). A guide to using MAPMAN to visualize and compare omics data in plants: a case study in the crop species, Maize. *Plant Cell Environ.* **9**:1211–1229.

Wang QQ, Liu F, Chen XS et al. (2010). Transcriptome profiling of early developing cotton fiber by deep-sequencing reveals significantly differential expression of genes in a fuzzless/lintless mutant. *Genomics* **96**: 369–376.

Wilkinson S, Davies WJ (2010). Drought, ozone, ABA and ethylene: new insights from cell to plant to community. *Plant, Cell and Environ.* **33**: 510-525.

Wu T, Qin ZW, Zhou XY, Feng Z, Du YL (2010). Transcriptome profile analysis of floral sex determination in cucumber. *J Plant Physiol* **167**: 905–913.

Xiong Y, Li Q, Kang B, Chourey PS (2011). Discovery of genes expressed in basal endosperm transfer cells in maize using 454 transcriptome sequencing. *Plant Mol Biol Report* **4**: 835-847

Yamaguchi-Shinozaki K, Shinozaki K (2005). Organization of *cis*-acting regulatory elements in osmotic- and cold-stress-responsive promoters. *Trends in Plant Scienze* **10**: 88-94.

Yamaguchi-Shinozaki K, Shinozaki K (2006). Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annu Rev Plant Biol.* **57**: 781-803.

Yang SS, Tu ZJ, Cheung F et al. (2011). Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems. *BMC Genomics* **12:** 199.

Yoo SD, Cho Y, Sheen J (2009). Emerging connections in the ethylene signaling network. *Trends Plant Sci* **14**: 270-279.

Yoshida T, Fujita Y, Sayama H et al. (2010). AREB1, AREB2, and ABF3 are master transcription factors that cooperatively regulate ABRE-dependent ABA signaling involved in drought stress tolerance and require ABA for full activation. *Plant J* **61**: 672-685.

Zeller G, Henz SR, Widmer CK et al. (2009). Stress-induced changes in the *Arabidopsis thaliana* transcriptome analyzed using whole-genome tiling arrays. *Plant J* **58**: 1068-1082.

Zenoni S, Ferrarini A, Giacomelli E et al. (2010). Characterization of Transcriptional Complexity during Berry Development in *Vitis vinifera* Using RNA-Seq. *Plant Physiology 152*: 1787–1795.

Zhang X, Yazaki J, Sundaresan A et al. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell* **126**: 1189-1201.

SUPPLEMENTARY MATERIAL # 1

GO functional classes used in our analyses and GO terms included in each class.

GO Class	GO ID
Cell Death	GO:0016265 GO:0008219
Biogenesis	GO:0016043 GO:0007005
Cellular Process	GO:0007049 GO:0008283 GO:0019725 GO:0007154 GO:0006091 GO:0003774 GO:0004872
Development	GO:0007275 GO:0009790 GO:0030154
Cell Recognition	GO:0008037
Metabolism	GO:0008152 GO:0009056 GO:0009058 GO:0019538 GO:0006259 GO:0006629 GO:0005975 GO:0019748 GO:0006412 GO:0006464 GO:0006139
Reproduction	GO:0000003
Response to Stimula	GO:0009628 GO:0009607 GO:0009719 GO:0009605 GO:0006950
Viral Reproduction	GO:0016032
Cell Part	GO:0005623 GO:0005622 GO:0005737 GO:0005618 GO:0005634 GO:0005694 GO:0005886 GO:0005635 GO:0005829 GO:0005654 GO:0016023 GO:0005815 GO:0000228 GO:0005730
Extracellular Region	GO:0030313 GO:0030312 GO:0005578 GO:0005576
Cytoskeleton	GO:0005856 GO:0007010
Organelles	GO:0006996 GO:0005794 GO:0005783 GO:0005773 GO:0005840 GO:0005777 GO:0005811
Plastid and Mitochondrion	GO:0009536 GO:0005739 GO:0009579
Antioxidant Activity	GO:0016209
Binding	GO:0005488 GO:0003676 GO:0003723 GO:0003682 GO:0003677 GO:0000166 GO:0005515 GO:0008289 GO:0030246 GO:0008092 GO:0003779
Catalytic Activity	GO:0003824 GO:0008233 GO:0016787 GO:0016740 GO:0004518
Ion Channel	GO:0005216 GO:0006811
Enzyme Regulator Activity	GO:0030234
Nutrient Reservoir Activity	GO:0045735
Translation Factor Activity	GO:0008135
Signal Transduction	GO:0007165 GO:0004871 GO:0016301 GO:0004672 GO:0004721 GO:0005102 GO:0005509 GO:0019825
Structural Molecule Activity	GO:0005198
Transcription Regulator Activity	GO:0003700 GO:0030528 GO:0040029
Transporter Activity	GO:0006810 GO:0005215 GO:0015031

SUPPLEMENTARY MATERIAL # 2

Dataset of genes surveyed in this study. The Excel file provides for each gene (identified by the *P. trichocarpa* Phytozome unigene model dataset) the expression level (RPKM) in control, moderately dehydrated and severely dehydrated leaves and the GO biological process annotation. (This file will be published on the University of Pisa Plant Genetics and Genomics Lab site (http://www.agr.unipi.it/Sequence-Repository.358.0.html).

Paper V

Whole genome analysis of differential gene and allelic expression in heterotic and non heterotic hybrids

Abstract. High-throughput Illumina RNAseq was used to compare gene and allelic expression in two *Populus deltoides* x *P. nigra* hybrids, showing or not heterosis concerning stem circumference and height. Analyses were performed on RNAs isolated from leaves of plants grown in normal conditions or exposed to moderate (85% leaf RWC) or severe (57% leaf RWC) drought. On the whole, 21.7, 32.3, and 33.4% of genes resulted differentially expressed in the two hybrids in the three treatments. Moreover, the number of genes differentially activated or repressed in the two hybrids increase in response to drought, suggesting that genetic differences can have an important role in stress tolerance. Such an increase is even higher when limiting analysis to genes involved in stress response, in signalling and in transcription regulation. The occurrence of differential allelic expression in the same samples was also analysed in 200 randomly chosen genes. Fifty to sixty percent of these genes, depending on the hybrid and on the treatment, showed equal allelic expression but in the other genes the proportion between two alleles ranged from 60:40 to 90:10, i.e. they showed significant differential allelic expression. The results are discussed in relation to similar studies in the literature and to the importance of such phenomena in generating heterosis.

Introduction

Heterosis refers to the superiority, in biomass and fertility of an hybrid compared to its inbred parents (Schull 1908). Heterosis occurs in many (but not all) inter-varieties or interspecific hybrids, providing a yield advantage to the hybrid ranging between 15 to 50%, depending on the crop (Duvick 2001). Despite a century of investigations, the genetic and the molecular basis of heterosis is still unclear. This is probably due to the polygenic nature of traits such as growth vigor and yield as well as to the complexity of molecular events that take place when merging two divergent genomes in an hybrid. According to one view, dominance is the main cause for heterosis. This model suggests that deleterious alleles in one parent are “backed-up” by a beneficial allele in the other parent. As it is likely that several loci are involved in the heterotic effect, complementation of the deleterious loci of each parent would contribute to the superiority of the hybrid [e.g. F1 (Aa Bb) > parents (aa BB) and (AA bb)]. However, a series of evidences suggest that it does not provide a complete explanation to heterosis (Birchler et al. 2003). In fact, heterotic QTLs were found to be controlled by the interaction between different alleles in ways that cause overdominance [F1 (AA’) > Parents (AA or A’A’)]. Loci that determine heterosis, acting in *cis* or in *trans*, might also mask each other resulting in epistasis. Studies in several crops point to the involvement of dominance, overdominance as well as epistasis, although the relative importance of each mode of control varies in the different studies. The large rearrangements and genic and inter-genic non-colinearity detected among maize inbred lines (Brunner et al. 2005, Fu and Dooner 2002, Morgante et al. 2005) should be taken into account for a better estimation of this parameter and its relation to heterosis.

Studies on gene expression profiles in hybrids have both shown that many genes behave as expected from an additive model (i.e. the simple combination of parental expression patterns). Additivity of gene expression could contribute to the heterotic phenotype, either via the dominance/complementation model, or via phenotypic overdominance. However, many other genes showed novel patterns of expression, which highlights the potential for novelty and plasticity in hybrid genomes.

Heterosis could act on a per locus basis. Two slightly diverse alleles could together enhance the yield of the hybrid in ways that do not exist in each parent separately. This could be the case for similar enzymes with different and complementing properties, or for proteins that work in complexes and that may form new heterodimers as well as homodimers in the hybrid.

Interspecific and even intraspecific comparisons between genic and intergenic regions of different plant species have revealed that genomes of individuals belonging to one and the same species are not always completely colinear as far as their sequence (Brunner et al. 2005, Scherrer et al. 2005). Colinearity is mainly restricted to the genic regions, in intergenic regions large rearrangements can occur, including gene duplications and insertion of retroelements. In maize, comparison of sequences from different inbreds at the same locus showed that most of the nonshared sequences consist of LTR retrotransposons and other mobile elements

(Brunner et al. 2005). Variability was not only found in the composition and length of intergenic regions (mainly composed by retroelement blocks), but also in the gene space, where even several coding sequences were missing (Fu and Dooner 2002, Morgante et al. 2005, He et al. 2009). Complementation of non-shared genes might be one of the factors contributing to heterosis. For example, though non-shared genic sequences appear to be generally non-functional, they could contribute gene silencing of homologous sequences through the production of siRNA (Hamilton and Baulcombe 1999), affecting the phenotype.

Also differences in the repetitive fraction can have a role in heterosis. In several instances, conserved and active alleles in the two inbreds used to produce a hybrid are flanked by different DNA, for example, by non-conserved retrotransposons inserted nearby (Brunner et al. 2005). Such retroelements are known to be potentially induced by various stresses (Kuff and Lueders 1988, Hirochika et al. 1996) and they may affect the transcription of neighbouring genes by producing single, chimeric, or antisense transcripts or by acting as enhancers (see Kashkush et al. 2003). In conclusion, different repetitive sequence environments should affect tissue specificity or temporal regulation of expression of genes. Such differences have been proposed to be one of the causes of heterotic complementation (Birchler et al. 2003, Song and Messing 2003, Springer and Stupar 2007b) and are comparable to allelic interactions proposed by the overdominance theory for explaining hybrid vigour (Crow 1948).

The genus *Populus* is an important crop and a model system to understand molecular processes of growth, development, and responses to environmental stimuli in trees. The interspecific hybrid between *Populus deltoides* and *Popolus nigra*, is one of the most cultivated poplar in Northern America and Northern Europe. Cultivars of this interspecific hybrid show high heterosis. In order to contribute clarifying the molecular bases of heterosis and in view of the sequencing of *Populus deltoides* and *Populus nigra* genomes, that will make us able to evaluate the occurrence of cis-regulatory differences between these two species, we have surveyed the occurrence of different allelic expression in genes of differently heterotic *Populus deltoides* x *P. nigra* interspecific hybrids in control conditions and in plants subjected to drought stress, i.e. a condition that can seriously affect plant productivity through a reduction of biomass production. The occurrence of differences in allelic expression of one and the same gene has to be related to differences in cis-regulatory regions that are presumably frequent between the genomes of two different species. A comparison between highly heterotic and non heterotic hybrids should allow to hypothesize the involvement of differential allelic expression in heterosis.

Materials and methods

Plant materials, sample preparation and sequencing

Analyses were performed on two interspecific hybrids of the DxN 812b family (genotype numbers 661200585 and 661200589, hereafter called 85 and 89) between *Populus deltoides* (genotype L155-079) as female and *P. nigra* (genotype 71077-308) as male, and on their parents, produced at INRA, Orleans (France). Hybrid performance was measured at Orleans referring to total height at 2 years and circumference at 1 m after two years of culture.

Rooted cuttings of *Populus deltoides* and *P. nigra*, and rooted cuttings from two hybrids between them, were cultivated in 20 x 20 cm2 pots in the open. Leaves of *P. deltoides* and *P. nigra* were used to isolate genomic DNA according to the method described by Doyle and Doyle (1989).

In the late spring 2011, some hybrid plants of 50 cm in height were normally watered and others were subjected to drought by suspending watering. Leaf water loss during the experiment was followed by relative water content measurement [RWC = 100 (FW-DW) / (TW - DW)], where FW is the fresh weight, DW the dry weight and TW the turgid weight. One leaf was collected from each plant and divided into two portions: one was used for RNA isolation, the other was used to measure tissue hydration by determining the RWC The experimental design was as follows: 2 clones (biological replicates) x 3 treatments (control, moderate, and severe drought stress) x 2 hybrids (obtained from the same parents).

Total RNA was isolated from leaves of single plants with different RWC according to the method described by Logemann et al. (1987), followed by DNAse I (Roche) treatments according to the manufacturer's instructions to completely remove genomic DNA contamination.

RNA-Seq library was generated using the TruSeq RNA-Seq Sample Prep kit according to the manufacturer's protocol (Illumina Inc., San Diego, CA). In short, poly-A RNA was isolated from total RNA and chemically fragmented. First and second strand synthesis were followed by end repair, and adenosines were added to the 3' ends. Adapters were ligated to the cDNA and 200 ± 25 bp fragments were gel purified and enriched by PCR. The library was quantified using Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA) and run on the Illumina HiSeq2000 (Illumina Inc.) using version 3 reagents. Single-read sequences of length 51 bp were collected.

Alignment and analysis of Illumina reads against the P. trichocarpa *unigene model database*

Sequence alignments were generated with CLC-BIO Genomic Workbench 4.9 (CLC bio, Aarhus, Denmark), using the *P. trichocarpa* unigene model database (Tuskan et al. 2006) available at the Phytozome site (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Ptrichocarpa/). The following parameters were used for alignments: maximum number of mismatches= 2, minimum number of reads = 10.

For evaluation of gene expression we calculated the number of mapped reads per kilobase per million mapped reads, measuring the transcriptional activity for each gene. CLC-BIO Genomic Workbench computes this normalized gene locus expression level (named RPKM) by assigning reads to a sequence in the database and counting them. In the case of reads that match equally well to several sites, the software assigns them proportionally.

The RPKM value (Mortazavi et al. 2008) estimates the number of reads falling in a given gene locus as follows:

$$\mathrm{RPKM} = N / (L \times N_{tot} \times 10^{-6})$$

where N = number of mapping reads at a given gene locus, L = estimated length (Kbp) of the gene locus, N_{tot} = number of total mapping reads.

Expression changes between hybrids were evaluated considering RPKM values in control, moderately, and severely drought stressed plants using Baggerly's test (Baggerly et al. 2003). This compares the proportions of counts in a group of samples against those of another group of samples, and is suited to cases where replicates are available in the groups. The samples are given different weights depending on their sizes (total counts). The weights are obtained by assuming a Beta distribution on the proportions in a group, and estimating these, along with the proportion of a binomial distribution, by the method of moments. The result is a weighted t-type test statistic.

Expression values were treated as follows: data are reported as weighted proportions fold change between hybrids. When values were higher in the first hybrid than in the second they were reported as positive, when they were higher in the second than in the first as negative. The weighted proportions fold changes between hybrids were considered as significant when fold change for a gene was at least 2 or 4.

For the analysis of gene categories, we have selected genes involved in transcription regulation, signalling, and stress response using the annotation tool of MapMan 3.5.1R2 (Thimm et al. 2004, Usadel et al. 2009).

SNP analysis

We used CLC-BIO Genomic Workbench version 4.9 to align sequence reads on the Phytozome gene model database (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Ptrichocarpa/) using alignment parameters of maximum number of mismatches= 2, minimum number of reads = 10, and then call the putative SNPs. SNP calling was performed using SNP detection software included in CLC Genomics Workbench. For SNP detection, the central base quality score of ≥20 and average surrounding base quality score of ≥15 were set to assess the quality of SNPs. Under the criteria of minimum coverage (read depth) of 10 and the minimum variant frequency of 15%, the variations compared to the reference sequence were counted as SNPs.

After calling, SNPs were processed using the following parameters: (1) called SNPs must be covered by at least 100 or 50 reads, depending on the experiments; (2) multiple SNPs must be discarded.

To determine the expression percentage of one of the two alleles of a gene between control and droughted plants, we calculated the mean of the percentages of the most expressed allelic SNPs in control and the mean of the percentages of the same SNPs in the droughted plants.

Results

Analysis of hybrid performance

The performance of the two hybrids between *P. deltoides* and *P. nigra* was measured referring to their total height and circumference at 1 m after 2 years of culture. The height was 514 and 286 cm, and the circumference 13.7 and 5.1 cm, for hybrid 85 and hybrid 89, respectively. Such differences between hybrids are related to large heterozygosity of the parental genotypes, resulting in genetic differentiation between individuals sharing parents. Compared to their parents, the hybrid performance of interspecific hybrids evidence differential heterosis level, i.e., hybrid 85 can be considered as highly heterotic while hybrid 89 display a low level of heterosis.

High-throughput gene expression variation between hybrids

To obtain a global view of the transcriptome differences between the two poplar hybrids in control and drought condition, we used a Illumina Genome Analyzer to perform high-throughput tag sequencing analysis on cDNAs from 12 libraries (two hybrids per three culture conditions [control, moderate and severe drought] per two clones of the same hybrid [biological replicates], Table 1).

Table 1. Number of Illumina fragments matching to the *P. trichocarpa* unigene database (45,033 CDS sequences) for each library (C, control; S1, moderate stress; S2, severe stress). For each library, the RWC of leaves from which RNA was isolated is also reported.

Library	Leaf RWC	Counted fragments			Uncounted
		Total	Unique (%)	Non specific (%)	
Hybrid 85, clone 3 (C)	95.51	10,360,767	7,814,681 (75.4)	2,546,086 (24.6)	4,966,787
Hybrid 85, clone 4 (C)	92.58	2,624,418	2,012,238 (76.7)	612,180 (23.3)	1,691,646
Hybrid 89, clone 6 (C)	95.75	6,057,500	4,548,228 (75.1)	1,509,272 (24.9)	2,869,614
Hybrid 89, clone 8 (C)	95.40	2,475,842	1,866,250 (75.4)	609,592 (24.6)	1,487,870
Hybrid 85, clone 12 (S1)	86.31	3,134,221	2,378,566 (75.9)	755,655 (24.1)	2,353,124
Hybrid 85, clone 24 (S1)	85.64	3,121,951	2,358,175 (75.5)	763,776 (24.5)	1,881,363
Hybrid 89, clone 10 (S1)	84.89	3,963,040	2,973,101 (75.0)	989,939 (25.0)	2,160,444
Hybrid 89, clone 15 (S1)	86.30	4,780,311	3,583,306 (75.0)	1,197,005 (25.0)	2,574,769
Hybrid 85, clone 42 (S2)	54.78	2,228,413	1,728,419 (77.6)	499,994 (22.4)	1,756,773
Hybrid 85, clone 45 (S2)	61.83	3,132,978	2,406,895 (76.8)	726,083 (23.2)	2,582,381
Hybrid 89, clone 20 (S2)	52.78	2,327,692	1,832,969 (78.7)	494,723 (21.3)	2,248,212
Hybrid 89, clone 35 (S2)	59.69	3,892,818	3,061,307 (78.6)	831,511 (21.4)	1,962,515
Total		48,099,951	36,564,135 (76.0)	11,535,816 (24.0)	28,535,498

The total number of tags per library ranged from 3.96 to 15.33 millions, a tag density sufficient for quantitative analysis of gene expression (Morin et al. 2008). In Table 1, the RWC of leaves from which RNAs were isolated is also reported: moderately stressed leaves (S1) showed a RWC ranging from 84.89 to 86,31, severely stressed leaves (S2) RWC ranged from 52.78 to 61.83.

In a first experiment, the sequence reads were aligned on the *P. trichocarpa* Phytozome unigene database (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Ptrichocarpa/), using the CLC-BIO software set to allow two base mismatches. Of the total reads, 62.8% matched either to a unique (47.7%) or to multiple (15.1%) unigene sequences; 37.2% of the tags could not be mapped to the gene sequences.

CLC-BIO measures gene expression in reads per exon kilobase per million mapped sequence reads

(RPKM), a normalized measure of exonic read density that allows transcript levels to be compared both within and between samples (Mortazavi et al. 2008).

As CLC-BIO distributes multireads at similar loci in proportion to the number of unique reads recorded, we included in the analysis both unique reads and reads that occur up to 10 times to avoid undercount for genes that have closely related paralogs (Mortazavi et al. 2008). We evaluated the expression of 45,033 genes included in the *P. trichocarpa* Phytozome database (Tuskan et al. 2006) during drought treatments. To obtain statistical confirmation of the differences in gene expression between hybrids across treatments, we compared the RPKM-derived read count using Baggerly's test (Baggerly et al. 2003). Differences between hybrids were considered significant for $P \leq 0.05$. We observed that the majority of genes are expressed at same extent in both hybrids in the three treatments. We also surveyed genes whose expression was highly induced in one of the two hybrids, considering all those genes whose weighted proportions fold change was higher than |2| or |4| (see Materials and methods).

The differential expression between hybrids in the three treatments is summarized in Table 2. It can be observed that, of the 26-27,000 genes expressed in both hybrids and in at least one treatment, the vast majority show the same expression level. Only 21.7, 32.3, and 33.4% are more expressed or much more expressed in one of the two hybrids, in control, S1 and S2 plants, respectively. It is however apparent that coping with drought determines a reduction of the number of genes that are equally expressed in the two hybrids (Table 2), i.e. the two hybrids show large differences as to which genes are activated by stress.

With the objective to analyse genes belonging to precise metabolic pathways, the MapMan 3.5.1R2 tool was used on 4,391 genes that could be unambiguously treated by MapMan. MapMan allowed the assignment of 4,428 genes, being some of the genes mapped to multiple pathways, into a total of 34 of 35 functional classes. Among these genes, we selected those related to signalling (113 gene models), transcription regulation (331), and stress responses (73), that are well documented to be responsive to wide-array of stresses. As observed for all analysed genes (Table 2), also the vast majority of these genes resulted similarly expressed in the two hybrids (Table 3). Interestingly, the percentage of similarly expressed genes belonging to these metabolic pathways in stressed plants is larger (and even much larger) than for the whole gene dataset, suggesting that different hybrids behave differently in response to stress, especially concerning stress response genes and, hence, probably display different drought tolerance.

Differential allelic expression in hybrids subjected to drought

Phenotypic differences between hybrid can be related not only to differences in gene expression but also to differences in allelic expression. We have analysed allelic expression in genes randomly chosen among 3310 genes that are significantly expressed (RPKM > 1) in both hybrids in all treatments. In particular, we analysed the first 200 heterozygous genes (in alphabetical order according to their Phytozome code) expressed in both control and moderately stressed plants or in both control and severely stressed plants.

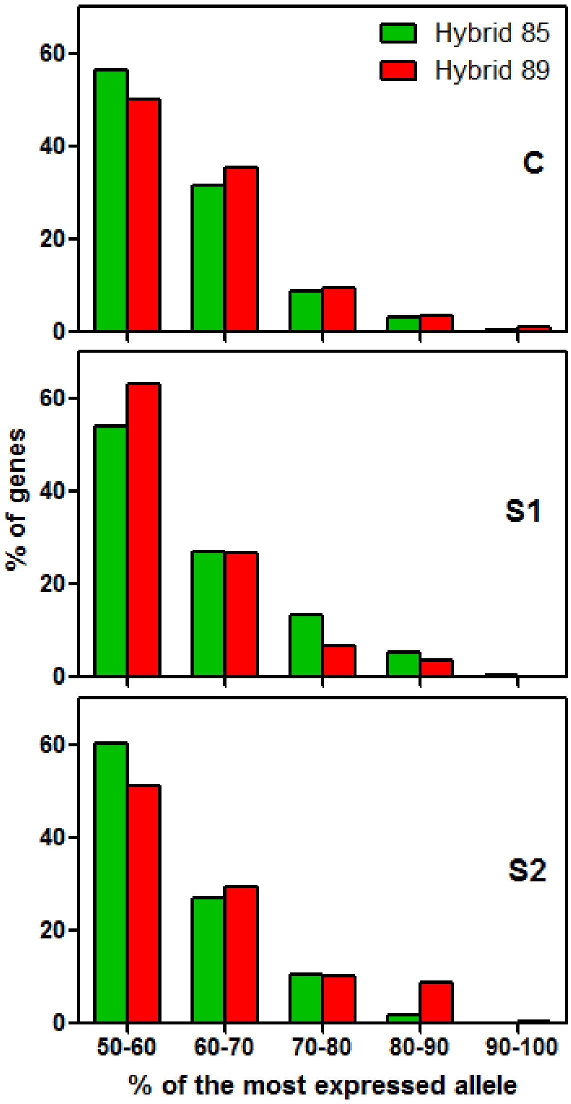


Figure 1. Distributions of the most expressed allele in a sample of 250 genes of hybrids 85 and 89, cultivated under normal hydration (C) or subjected to moderate (S1) or severe (S1) drought.

Firstly, we have verified the occurrence of differential allelic expression in each stage (Figure 1). Most genes show similar expression level of the two alleles. Hybrid 85 shows more genes with different allelic expression in moderately stressed leaves; on the contrary, in the hybrid 89 differential allelic expression is more pronounced in severely stressed leaves (Figure 1), indicating that stress response through changes in allelic expression occurs differently in the two hybrids.

Then, we analysed genes showing differences in allelic expression between control and S1 or S2 in the same sample of 200 heterozygous genes to which at least 100 RNAseq tags could be aligned. Difference in allelic expression was measured as the percentage difference between the percentages of one and the same of the two alleles in RNAseq of control and droughted leaves (S1 or S2) (Table 4). The percentage of genes that show different allelic expression in plants exposed to stress are similar in the two hybrids, except for a strong increase of genes showing large difference in allelic expression in the passage from control to severe stress in the hybrid 89 (Table 4). However, often one of the two alleles of a gene appears to be more expressed in both stress conditions than in the controls. Among genes with large differential allelic expression between control and moderate or severe stress, we can find stress responsive genes, for example encoding a NAC-domain protein, an osmotin, a CCR-like protein, a glutaredoxin, a glutathione lyase.

Interestingly, a number of genes showed a switch in their allelic expression (i.e., the most expressed allele in control leaves is the least expressed in stressed leaves), especially in hybrid 89, in the passage from control to severe stress (Table 5). Analyses are now in progress to verify the occurrence of differential allelic expression in all gene models.

We have also studied differential allelic expression in genes related to signalling (113 gene models), to transcription regulation (331), and to stress responses (73, see above). Of these, considering genes mapped at least by 50 reads, the vast majority is apparently heterozygous (Table 6). As observed for all analysed genes (Table 4), in the hybrid 89 different allelic expression is observed, especially in the passage from control to severe stress condition, also for stress responsive genes (Table 6). On the contrary, genes involved in signal transduction respond by changing their allelic expression already in moderate stress conditions. Differential allelic expression of transcription factors encoding genes strongly increase from control to moderate stress and then to severe stress, in both hybrids.

Table 2. Global analysis of differences in gene expression in the two analysed poplar hybrids (85 and 89). As a reference, the 45,033 gene models in the Phytozome database were used (see text). Differences between hybrids were evaluated by the weighted proportions fold change (same expression level indicates a ratio ranging from -2 to 2; > indicates a ratio between 2 and 4 or -2 and -4; >> indicates a ratio > 4 or < -4).

Treatment	Number of genes expressed in the two hybrids	Number of genes showing (in the two hybrids):				
		Same expression level (%)	85 > 89 (%)	85 >> 89 (%)	89 > 85 (%)	89 >> 85 (%)
Control	26,799	20,976 (78.3)	2,947 (11.0)	1,380 (5.1)	969 (3.6)	527 (2.0)
S1	27,739	18,778 (67.7)	1,847 (6.7)	1,363 (4.9)	3,405 (12.3)	2,346 (8.5)
S2	26,215	17,448 (66.6)	2,255 (8.6)	1,224 (4.7)	3,147 (12.0)	2,141 (8.2)

Table 3. Analysis of differences in gene expression for genes involved in signalling, transcription regulation, and stress response in the two analysed poplar hybrids (85 and 89). Genes were extracted from Phytozome database according to the Arabidopsis corresponding code used in Mapman (see Materials and Methods). Differences between hybrids were evaluated by the weighted proportions fold change (same expression level indicates a ratio ranging from -2 to 2; > indicates a ratio > 2 or < -2; >> indicates a ratio > 4 or < -4).

Metabolic pathway	Treatment	Number of analysed genes	Number of genes showing (in the two hybrids):				
			Same expression level (%)	85>89 (%)	85>>89 (%)	89>85 (%)	89>>85 (%)
Signalling	Control	113	64 (81.0)	11 (13.9)	0 (0.0)	3 (3.8)	1 (1.3)
	S1		49 (60.5)	12 (14.8)	7 (8.6)	10 (12.3)	3 (3.7)
	S2		55 (59.8)	10 (10.9)	6 (6.5)	12 (13.0)	9 (9.8)
Transcription regulation	Control	331	215 (86.0)	21 (8.4)	3 (1.2)	6 (2.4)	5 (2.0)
	S1		205 (75.6)	29 (10.7)	8 (3.0)	21 (7.7)	8 (3.0)
	S2		215 (73.6)	19 (6.5)	11 (3.8)	25 (8.6)	22 (7.5)
Stress response	Control	73	35 (77.8)	5 (11.1)	2 (4.4)	2 (4.4)	1 (2.2)
	S1		27 (49.9)	3 (5.5)	7 (12.7)	9 (16.4)	9 (16.4)
	S2		31 (55.4)	2 (3.6)	0 (0.0)	17 (30.4)	6 (10.7)

Discussion

This study examined gene and allele-specific expression in two *P. deltoides* x *P. nigra* hybrids using Illumina sequencing technology on cDNAs from leaves of plants grown in normal conditions or under drought. This technology allows to obtain a complete survey of gene and allele expression in a given tissue.

The two hybrids, obtained by the same cross, were chosen according to their heterosis level. Hybrid 85 showed large heterotic effects concerning stem circumference and height; by contrast, hybrid 89 did not show heterotic effects for these characters. The two hybrids are genetically different because of large heterozygosity of the parents. Hence, a significant component of heterotic effects can be ascribed to dominance, complementation and epistasis. However, as for all cultivated hybrids, a component of heterosis can be related to overdominance, i.e. the intrinsic superiority of heterozygous condition in respect of both homozygous genotypes.

The genetic differences between analysed hybrids are evidenced by differences in gene expression. On the whole, 21.7 to 33.4% of genes expressed in both hybrids and in the three treatments, are differentially expressed in the two hybrids (see Table 2). This result, considering the large number of genes tested, indicates the large variability that can be obtained by crossing two heterozygous and genetically distant parents. Moreover, the number of genes differently activated or repressed in the two hybrids increase in response to drought. Such an increase is even higher when limiting analysis to genes involved in stress response, in signalling and in transcription regulation, i.e. three classes of genes that are known to be activated by stress (Table 3). These results suggest that genetic differences in this poplar hybrid population can have a potential value to be exploited for stress tolerance.

We have also analysed the occurrence of differential allelic expression in the same samples. Though Illumina technology allows a whole genome analysis of such phenomenon, we have preliminary limited our analysis to 200 randomly chosen genes. Fifty to sixty percent of these genes,

Table 4. Number of genes showing differences in allelic expression between control and S1 or S2 in a sample of 200 heterozygous genes (to which at least 100 RNAseq tags could be aligned). Difference in allelic expression was measured as the percentage difference between the percentages of one and the same of the two alleles in RNAseq of control and droughted leaves (S1 or S2).

Hybrid	Nr. of analysed heterozygous genes	Nr. of genes showing changes in allelic expression			
		between C and S1 (%)		between C and S2 (%)	
		$3 < \Delta\% < 10$	$\Delta\% > 10$	$3 < \Delta\% < 10$	$\Delta\% > 10$
85	200	83 (41.5)	26 (13.0)	88 (44.0)	33 (16.5)
89	200	91 (45.5)	24 (12.0)	87 (43.5)	68 (34.0)

Table 5. Number of genes showing a switch of the most expressed allele between control and drought stressed leaves (S1 or S2) in a sample of 200 heterozygous genes (to which at least 100 RNAseq tags could be aligned). Only switches outside the 40-60% frequency interval were reported.

Hybrid	Nr. of analysed heterozygous genes	Nr. of genes showing significant switch in allelic expression	
		between C and S1	between C and S2
85	200	10	9
89	200	6	33

depending on the hybrid and on the treatment, showed equal allelic expression (Figure 1) but for the other genes the proportion between the two alleles changed from 60:40 to 90:10, i.e. they showed significant differential allelic expression.

The common occurrence of differential allelic expression in poplar hybrids detected in this study confirms that this phenomenon is widespread in plant species. In a previous study in *P. trichocarpa* x *P. deltoides* hybrids, 30 genes were surveyed for their allelic expression, and, using a threshold cutoff of 1.5-fold, 17 of the 30 (57%) genes (Zhuang and Adams 2007) showed variation in allelic expression levels. Besides studies on poplars, allele specific expression has been reported in barley hybrids for 63% of analysed genes (von Korff et al. 2009). In maize hybrids, Springer and Stupar (2007a) showed that half of the analysed genes showed unequal allelic expression. Then, Guo et al. (2008), using massively parallel signature sequencing, reported that 60% of maize genes were subjected to differential allelic expression in the meristems. In Arabidopsis hybrids, the frequencies of allelic imbalance detected were lower than in maize, concerning only 7% of the genes carrying allelic polymorphisms (Kiekens et al. 2006), though such low values can be determined by the high threshold established by the authors to assess the occurrence of allelic imbalance. Though different methodological approaches can explain at least in part the differences in the extent of allele specific expression among species, it is to be considered that frequency, level and functional relevance of such phenomenon is obviously strongly affected by the reproductive strategy of the species, by its domestication history, by differences in genome plasticity and in the levels of sequence variation (von Korff et al. 2009). The apparent higher degree of allelic expression variation in the *Populus* interspecific hybrids and in the maize intraspecific hybrids than in the Arabidopsis hybrids could be related to the highly polymorphic maize genome (Guo et al. 2004) and, similarly, to the genetic divergence between *Populus* species.

The differences in allele expression were also influenced by the drought treatment. Changes in allele specific expression were observed for many genes (54.5 to 77.5%, depending on the hybrid and the treatment, see Table 4) comparing control and moderate or severe drought stress. In 20 and 39 over 200 analysed genes of hybrids 85 and 89, respectively, changes in allelic expression resulted in a switch, i.e. the same allele was more expressed in control leaves and less expressed in stressed leaves. This suggests that also changes in allelic expression can be involved in stress tolerance.

Changes in allele specific expression during drought stress have been reported also in other species, for example maize (Guo et al. 2004) and barley (von Korff et al. 2009), indicating that differential allelic expression of stress-related genes may affect drought adaptation of a genotype, though a functional relationship between such differential expression in these genes and phenotypic performance of the hybrid has not been still established.

The high frequencies of *cis*-acting regulatory variations in maize, barley and poplar have been attributed to high levels of genetic diversity, and proposed as a potential molecular basis for heterosis (Birchler et al. 2006; Springer and Stupar 2007b; Zhuang and Adams 2007). In maize inbred lines and in barley genotypes large differences have been reported in the composition of intergenic regions, related to the presence of different types of retroelements and repetitive sequences (Brunner et al. 2005, Scherrer et al. 2005). Such variations

Table 6. Number of homozygous and heterozygous genes involved in signalling pathway, in regulation of transcription, and in stress response (according to MapMan code, see Materials and Methods) that were represented in our experiments by at least 50 mapping reads. Of the heterozygous genes, the number of those showing differences in allelic expression between different treatments (control, S1, S2) is also reported. Difference in allelic expression was measured as the percentage difference between the percentages of one and the same of the two alleles in RNAseq of control and droughted leaves.

Metabolic pathway	Hybrid	Nr. of analysed genes	Nr. of genes with at least 50 mapping reads			Nr. of genes showing changes in allelic expression	
			Total	Homozygous	Heterozygous	between C and S1	between C and S2
Signalling	85	113	40	1	39	3	3
	89	113	41	2	39	2	3
Transcription factors	85	331	25	1	24	10	17
	89	331	19	0	19	6	10
Stress response	85	73	5	1	4	2	2
	89	73	9	1	8	2	4

should depend on recent bursts of transposition activity that have been ascertained in many plant genomes (Morgante et al. 2007) and also in poplar (Cossu et al. 2012) and may result in *cis*-regulatory variations.

Differential allelic expression is usually related to the occurrence of cis-acting regulatory variation between the two parental genotypes. Previous studies have revealed the predominance of *cis* regulation over *trans* regulation in hybrids. Although variable proportions of complete *cis*-regulation are found in various studies of different organisms, *cis*-effects were consistently involved in most if not all of the assayed genes, and pure *trans*-regulation is rare in *Populus* hybrids (Zhuang and Adams 2007) and in maize hybrids (Stupar and Springer 2006, Guo et al. 2008), and absent in *Drosophila* hybrids (Wittkopp et al. 2004). As *cis*-elements function in an allele-specific manner, allelic expression following *cis*-regulation reflects an inheritance of the regulatory pattern from the two parents to the hybrid.

After hybridization both alleles are exposed to common *trans*-regulators in the same cellular environment, and so *trans*-regulation and combined *cis*- and *trans*-regulation could be induced by hybridization (Landry et al. 2005). There is a hypothesis that *cis*- and *trans*-compensatory evolution is important in leading to novel gene expression and performance in the hybrids (Landry et al. 2005). It has been proposed that reuniting diverged regulatory factors and hierarchies in hybrids can lead to altered gene expression patterns (Riddle and Birchler 2003). However other factors, such as epigenetic variation, might account for the expression changes in those genes.

The *Populus* hybrids used in this study show different levels of heterosis. The altered gene regulation in hybrids observed in this study might be involved in generating the heterotic phenotype observed in one of the two hybrids. We are currently extending our analysis of differential allelic expression to the complete *Populus* transcriptome (45,033 genes). Moreover, high-throughput sequencing techniques are being used to sequence the genomes of *Populus deltoides* and *P. nigra*, allowing to establish an ultimate correspondence between differential allelic expression and *cis*-regulatory sequence variation, and to explore the importance of these phenomena in producing heterosis.

Acknowledgements

This research work supported by PRIN-MIUR, Italy, Project “Verso la delucidazione delle basi molecolari dell’eterosi nelle piante coltivate: variazione cis-regolatoria ed espressione genica in ibridi di pioppo.”

References

Baggerly KA, Deng L, Morris JS, Aldaz CM (2003). Differential expression in SAGE: Accounting for normal between-library variation. *Bioinformatics* **19**: 1477-1483.

Birchler JA, Auger DL, Riddle NC (2003). In search of the molecular basis of heterosis. *Plant Cell.* **15**: 2236–2239.

Birchler JA, Yao H, Chudalayandi S (2006). Unraveling the genetic basis of hybrid vigor. *PNAS* **103**: 12957–12958.

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005). Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell***17**: 343–360.

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 Genetics & Genomes* **8**: 61-75. DOI: 10.1007/s11295-011-0421-3

Crow JF (1948). Alternative hypotheses of hybrid vigor. *Genetics* **33**: 477–487.

Doyle JJ, Doyle JL (1989). Isolation of plant DNA from fresh tissue. *Focus* **12**: 13-15.

Duvick DN (2001). Biotechnology in the 1930s: The development of hybrid maize. *Nat. Genet. Rev.* **2**: 69–74.

Fu H, Dooner HK (2002). Intraspecific violation of genetic colinearity and its implications in maize. *PNAS* **99**: 9573–9578.

Guo M, Rupe MA, Zinselmeier C, Habben J, Bowen BA, Smith OS (2004). Allelic variation of gene expression in maize hybrids. *Plant Cell* **16**: 1707–1716.

Guo M, Yang S, Rupe M et al. (2008). Genome-wide allele-specific expression analysis using Massively Parallel Signature Sequencing (MPSS) reveals cis- and trans-effects on gene expression in maize hybrid meristem tissue. *Plant Molecular Biology* **66**: 551–563.

Hamilton AJ, Baulcombe DC (1999). A novel species of small antisense RNA in post-transcriptional gene silencing. *Science* **286**: 950–952.

He L, Dooner HK (2009). Haplotype structure strongly affects recombination in a maize genetic interval polymorphic for Helitron and retrotransposon insertions. *PNAS **106**: 8410–8416.*

Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M (1996). Retrotransposons of rice involved in mutations induced by tissue culture. *PNAS* **93**: 7783–7788.

Kashkush K, Feldman M, Levy AA (2003). Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat. Genet.* 33(1): 102-106

Kiekens R, Vercauteren A, Moerkerke B et al. (2006). Genome-wide screening for cis-regulatory variation using a classical diallele crossing scheme. *Nucleic Acids Research* **34**: 3677–3686.

Kuff EL, Lueders KK (1988). The intracisternal A-particle gene family: structure and functional aspects. *Adv Cancer Res **51**: 183–276.*

Landry CR, Wittkopp PJ, Taubes CH, Ranz JM, Clark AG, Hartl DL (2005). Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophila*. *Genetics* **171**: 1813-1822.

Logemann J, Schell J, Willmitzer L (1987). Improved method for the isolation of RNA from plant tissues. *Anal. Biochem* **163**: 16–20.

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.

Morgante M, De Paoli E, Radovic S (2007). Transposable elements and the plant pan-genomes. *Curr Opin Plant Biology* **10:** 149-155.

Morin RD, O'Connor MD, Griffith M et al. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.* **18**: 610–621.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621–628

Riddle NC, Birchler JA (2003). Effects of reunited diverged regulatory hierarchies in allopolyploids and species hybrids. *Trends Genet* **19**: 597-600.

Scherrer B, Isidore E, Klein P et al. (2005). Large intraspecific haplotype variability at the Rph7 locus results from rapid and recent divergence in the barley genome. *Plant Cell* **17**: 361–374.

Shull GH (1908). The composition of a field of maize. *Am. Breeders Assoc. Rep.* **4**: 296-301.

Song R, Messing J (2003). Gene expression of a gene family in maize based on noncollinear haplotypes. *PNAS* **100**: 9055–9060

Springer NM, Stupar RM (2007a). Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. *Plant Cell* **19**: 2391–2402.

Springer NM, Stupar RM (2007b). Allelic variation and heterosis in maize: how do two halves make more than a whole? *Genome Res.* **17**: 264–275.

Stupar RM, Springer NM (2006). Cis-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. *Genetics* **173:** 2199–2210.

Thimm O, Blasing O, Gibon Y et al. (2004). MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **37**: 914–939.

Tuskan GA, Difazio S, Jansson S et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596-604.

Usadel B, Obayashi T, Mutwil M et al. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ* **32**: 1633–1651.

von Korff M, Radovic S, Choumane W et al. (2009). Asymmetric allele-specific expression in relation to developmental variation and drought stress in barley hybrids. *The Plant Journal* **59**: 14–26.

Wittkopp PJ, Haerum BK, Clark AG (2004). Evolutionary changes in cis and trans gene regulation. *Nature* **430**: 85–88.

Zhuang Y, Adams KL (2007). Extensive allelic variation in gene expression in *Populus* F1 hybrids. *Genetics* **177**: 1987–1996.

Appendix

Temporal dynamics in the evolution of the sunflower genome as revealed by sequencing and annotation of three large genomic regions.

M. Buti · T. Giordani · F. Cattonaro · R. M. Cossu · L. Pistelli · M. Vukich · M. Morgante · A. Cavallini · L. Natali

Received: 12 February 2011 / Accepted: 9 May 2011 / Published online: 7 June 2011

Abstract Improved knowledge of genome composition, especially of its repetitive component, generates important informations in both theoretical and applied research. In this study we provide the first insight into the local organization of the sunflower genome by sequencing and annotating 349,380 bp from 3 BAC clones, each including one single-copy gene. These analyses resulted in the identification of 11 putative gene sequences, 18 full-length LTR retrotransposons, 6 incomplete LTR retrotransposons, 2 non autonomous LTR-retroelements (LINEs), 2 putative DNA transposons fragments, and one putative helitron. Among LTR-retrotransposons, non autonomous elements (the so-called LARDs), that do not carry any protein encoding sequence, were discovered for the first time in the sunflower. The insertion time of intact retroelements was measured, based on sister LTRs divergence. All isolated elements inserted relatively recently, especially those belonging to *Gypsy* superfamily. Retrotransposon families related to those identified in the BAC clones are present also in other species of *Helianthus*, both annual and perennial, and even in other Asteraceae. In one of the three BAC clones we found 5 copies of a Lipid Transfer Protein (LTP) encoding gene within less than 100,000 bp, four of which are potentially functional. Two of these are interrupted by LTR retrotransposons, in the intron and in the coding sequence, respectively. The divergence between sister LTRs of the retrotransposons inserted within the genes, indicate that *LTP* gene duplication started earlier than 1.749 MYRS ago. On the whole, the results reported in this study confirm that the sunflower is an excellent system to study transposons dynamics and evolution.

Introduction

Improved knowledge of genome composition, especially of its repetitive component, generates important information in both theoretical and applied research, for example to improve strategies for genetic and physical mapping of genomes and for the discovery and development of molecular markers. Moreover, knowledge of genome composition is a prerequisite for the annotation steps in sequencing projects both of ESTs (Expressed Sequence Tags) and of genomic regions.

To date, substantial progress has been made in unveiling the structure and organization of plant genomes. In the emerging view of plant evolution it is well-established that angiosperm species radiation has been accompanied, if not promoted, by polyploidization events and differential amplification of a repetitive component of their genomes represented by the long-terminal-repeat (LTR) retrotransposons (REs) (Grover et al. 2008; Soltis and Soltis 1999). LTR-retrotransposons (LTR-REs) are capable of replicating through a copy and paste mechanism and have the potential to increase the genome size of their host in a very short time span (Hawkins et al. 2006; Neumann et al. 2006; Piegu et al. 2006).

Sequencing of several plant genomes have

Communicated by A. Bervillé.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1626-4) contains supplementary material, which is available to authorized users.

M. Buti · T. Giordani · R. M. Cossu · L. Pistelli ·
M. Vukich · A. Cavallini · L. Natali (&)
Department of Crop Plant Biology, University of Pisa, Pisa, Italy
e-mail: lnatali@agr.unipi.it
F. Cattonaro · M. Morgante
Istituto di Genomica Applicata,
Parco ScientiWco e Tecnologico Luigi Danieli, Udine, Italy
M. Morgante
Department of Crop and Environmental Sciences,
University of Udine, Udine, Italy

revealed that the large degree of genomic variation and the occurrence of non-shared genomic sequences in closely allied grass species can be ascribed to the very young age of their extant LTR-REs complement.

The replicative mechanism of LTR-REs, coupled with the error-prone nature of transcription and reverse transcription, determines the generation of different RE families, characterized by sequence variability in both the coding, transcribed portion, and in the LTRs (Beguiristain et al. 2001). RE families have been reported that amplified differentially in different lineages within single plant groups or even within a single species (e.g. in maize) over a time span of less than one million years (Brunner et al. 2005; Wang and Dooner 2006). Similar events have taken place in several cereal species (Scherrer et al. 2005; Piegu et al. 2006; Vitte and Bennetzen 2006; Paterson et al. 2009) and in some dicots as well, even though to a less dramatic extent (Hawkins et al. 2006; Holligan et al. 2006; Neumann et al. 2006; Ungerer et al. 2006). In the recently sequenced sorghum genome, for example, the concomitant action of transposable element insertion and removal by illegitimate recombination or by DNA loss resulted in an average insertion age of 0.8 million years and in 50% of the detected elements having inserted within the last 500,000 years (Paterson et al. 2009).

Among species with large genomes, grasses such as maize, barley and wheat are by far the group of plants for which most information on retrotransposon-related genome structure has been collected. Apart from *Gossypium* species, relatively little attention has been given to large genome sized dicotyledons, despite their great economic importance. For example, studies on the genome composition and organization in the *Asteraceae* family, which is very large and includes very important crop species such as sunflower, are at their very beginning (Cavallini et al. 2010).

Sunflower (*Helianthus annuus* L.) is the most important species belonging to the genus *Helianthus*, whose relatively recent origin ranges between 4.75 and 22.7 million years. Based on the geographic distributions of its closest relatives, the genus *Helianthus* likely originated in Mexico, with subsequent migration through North America (Schilling et al. 1998). Sunflower haploid genome size is around 3,000 Mb. New *Helianthus* species have arisen by interspecific hybridization, some of which have been extensively studied (Rieseberg et al. 2003; Gross et al. 2007).

Sample sequencing of a small-insert genomic library from sunflower provided a set of sequences that were used to analyze the composition of the sunflower genome in terms of types and abundance of repetitive elements (Cavallini et al. 2010). The fraction of repetitive sequences amounted to 62% of the sequences, while the putative functional genes accounted for 4%; the largest component of the repetitive fraction of the sunflower genome was represented by LTR-REs, especially of the *Gypsy* superfamily. Class II elements were barely represented in the library.

The identification of transposable elements was however difficult in sunflower because of the paucity of sequences of previously described and annotated elements. While a fraction of the coding portions of the elements were recognized through the BlastX homology searches, any of the non-coding portions (e.g., the long terminal repeat regions of LTR-REs) were much more difficult to detect due to the high rate of sequence evolution of transposable elements between species (Ma and Bennetzen 2004). Sequencing of large genome regions appears to be more effective for identifying and characterizing repetitive sequences than BLAST homology searches of relatively short sequences. For example, a more accurate dating of amplification events of the LTR-RE component requires a comparison of the two LTR sequences from single elements, that can be obtained from the sequencing of large genomic regions (SanMiguel et al. 1996).

For these reasons, we sequenced and annotated three clones from a sunflower BAC-library, for a total of 349,380 bp. By this analysis, we provide the first insight into the local organization of the sunflower genome showing nests of REs inserted one into each other and allowing the estimation of retroelement insertion ages. Different waves of retroelement mobilization during the evolution of this species and the occurrence of very recent retrotransposition events are suggested.

Materials and methods

BAC-library screening

A bacterial artificial chromosome (BAC) library from sunflower inbred Ha383 was available from the CUGI (USA). We chose 3 genes that bibliographic information and experimental evidences suggested to be in single copy: a Lipid Transfer Protein encoding gene (*LTP*), a Dehydrin encoding gene (*DHN*) and a Z-Carotene Desaturase encoding gene (*DES*).

The three selected genes were used to develop three probes to screen the BAC-library. For each gene we performed PCR, using specific primers: 5’-TGGCAAAGATGGCAATGATG-3’ and 5’-ATCAAAGACACATACACATCCATA-3‘ for LTP; 5’-CAGCATATGGCAAACTACCGAGGAGATAA-3’ and 5’-CGAATTCGTGAAACCACATACAAAACAAAA-3’ for DHN; 5’-GGCAA GCTGCAGGGTTGG-3’ and 5’-AGACTCAGCTCATCAACT-3’ for DES. Sequences were amplified using 100 ng of genomic

DNA as a template; thermocycling was performed at 94 °C for 30 s, 60 °C for 30 s and 72 °C for 60 s, for 30 cycles, using Taq-DNA polymerase (Promega). PCR products were then used as templates for probes construction.

Radioactive ^{32}P probes were prepared with [α-^{32}P] dCTP by a random-primed synthesis with Klenow fragments (Roche) using 25 ng of each PCR product. Probes were purified using ProbeQuant G-50 Micro Columns (GE Healthcare). BAC-library hybridizations with the three probes were carried out in 5x SSC, 5x Denhardt solution, 0.5% SDS, 100 μg/ml salmon sperm DNA for 16 h at 65°C and the nylon filters were washed with 0.3 x SSC,0.1% SDS at 65°C. Filters are exposed for two days to a multipurpose phosphor storage screen (Cyclone Storage Phosphor System, Packard, CT, USA) in order to obtain a digital image of the radioactivity distribution. The obtained digital images were then analysed using a phosphoimager (Cyclone Storage Phosphor system, Packard).

To avoid false positive results, hybridization-positive clones were submitted to a PCR amplification using the specific primers reported above: by this way we could verify if the selected gene is actually included in the clone.

Among the hybridization-positive, PCR-positive BAC-clones, we selected one clone per gene to be sequenced and analyzed (*DES*: clone 0516 M24; *DHN*: clone 0340 D07; *LTP*: clone 0148 M20).

BAC-clones sequencing

The three selected BAC-clones were sequenced using a shotgun strategy (Tarchini et al. 2000) using a standard protocol at 11-12x redundancy (considering only bases of Phred quality ≥ 20). 10 μg of DNA were extracted by two subsequent maxipreps from each of three *Helianthus annuus* genomic BAC clones. BAC DNAs were treated with Plasmid-Safe™ ATP-Dependent DNase (Epicentre) in order to remove contaminating bacterial chromosomal DNA.

DNA was sheared by Hydroshear (Genomics Solution) at the following setting parameters: DNA volume: 200μl, # of cycles=15, Speed Code=13.

DNA was purified and concentrated by using filter columns (QIAquick PCR Purification Kit, QIAGEN™) and resuspended in 40 μL of double-distilled water. Uncompleted ends were repaired in a 50 μL reaction mix using the End-It^{TM} DNA End-Repair Kit (Epicentre^{TM}), following the indications of the manufacturer. End-repaired DNA was run on a 1% agarose gel. Fragments in the size range of 2.5-4.0 kb were selected and DNA was purified from the gel using the QIAquick Gel Extraction Kit (QIAGEN™) and ligated into pSmart-LC plasmid using the CloneSmart LCAmp Blunt Cloning Kit (Lucigen™) according to the manufacturer's protocol. 1 μL of this ligation mix was then used to transform *E. coli* strain DH10β using the OF10G Supreme^{TM} Electrocompetent Cells (Lucigen™) and a Bio-Rad Gene Pulser II electroporator. Recombinants were selected on Luria-Bertani plates with ampicillin.

Mate-paired reads were produced by sequencing with BigDye Terminator Cycle Sequencing Kit (Applied Biosystems™) and the SL1 and SR2 primers. The samples were purified by ethanol precipitation and were subsequently run on an ABI 3730xl capillary sequencer, starting from minipreps prepared with the MultiScreen Plasmid$_{384}$ system (Millipore). The total number of sequences (1536 mate-paired per clone, 700 bp read length on average) was then trimmed using PHRED and assembled using PHRAP (http://www.phrap.org) and PCAP. PCR primers were designed to walk across the sequence gaps by extracting the non repetitive ends of the relevant contig sequences and importing them together into the Primer 3.0 program (Rozen and Skaletsky 2000). Subcontigs robustly connected by clone mates were merged manually where the sequencing failed. Merged sequences were further confirmed by PCR on genomic DNA.

Sequences are deposited at EMBL database, under the accession numbers JN021934-36 and at the Dept. of Crop Plant Biology of Pisa University repository website (http://www.agr.unipi.it/Sequence-Repository.358.0.html).

Sequence analysis

The method used for BAC sequence annotation and transposable elements identification was partially based on an automatic pipeline for BLAST searches. Customized PERL scripts were utilized to fragment the complete sequences of both BAC clones into several partially overlapping 2500-bp-long regions which were subsequently analyzed by automatic BLASTX and BLASTN searches with MPI BLAST software (http://mpiblast.lanl.gov) against public non redundant databases at GenBank. BLAST results for each fragment were later recombined into a single file after automatic correction of nucleotide coordinates. Since the number of BLAST hits that can be provided in a single search is limited and highly conserved motifs are redundant, this procedure increased the number of matches along the whole BAC sequences by allowing for detection of additional weaker but still significant homologies. To limit false positive detection, we used a fixed E-value threshold of $E < 10^{-5}$ for BLASTN and $E < 10^{-10}$ for BLASTX.

Repetitive DNA content of each BAC clone was estimated by masking sequences using BLAST software against the RepBase (Jurka 2000) and the sunflower small insert genomic library (Cavallini et al. 2010).

In order to identify homologies to conserved features of already known retroelements, the complete sequences from each of the three BAC clones were used to conduct BLASTX and BLASTN searches against non redundant databases at GenBank and screened for similarity matches to either REs *gag-pol* polyprotein or transposase or other characterized gene products typically encoded by transposable elements. LTR retroelements were also identified using LTR FINDER (Xu and Wang 2007) and DOTTER softwares (Sonnhammer and Durbin 1995). LTR-FINDER uses a suffix-array based algorithm to construct all exact match pairs that are extended to long highly similar pairs. Alignment boundaries are obtained adjusting the ends of LTR pair candidates using the Smith-Waterman algorithm. These boundaries are re-adjusted, based on the occurrence of typical LTR-RE features such as being flanked by the dinucleotides TG and CA, at 5' and 3' ends, respectively, the presence of a target-site duplication (TSD) of 4-6 bp, of a putative 20-25 bp long primer binding site (PBS), complementary to a tRNA at the end of putative 5'-LTR, and of a 20-25 bp long polypurine tract (PPT) just upstream of the 5' end of the 3' LTR.

For *LTP* gene copies analysis, sequences were aligned using ClustalW (Thompson et al. 1994), then genetic similarity between each sequence was measured using the DNAdist program of the PHYLYP package (Felsenstein 1989). The triangular matrix was imported into NTSYS-pc version 2.01 h package (Rohlf 1998) to construct dendrograms using the UPGMA in the SAHN routine for cluster analysis. The number of synonymous substitutions per site between *LTP* genes was calculated using DnaSP (Rozas and Rozas 1999).

Insertion age calculation of full length retroelements

Retrotransposon insertion age was estimated comparing the 5'- and 3'-LTRs of each putative RE. The two LTRs of a single RE are identical at the time of insertion because they are mostly copied from the same template. The two LTRs were aligned with ClustalW software, indels were eliminated, and the number of nucleotide substitutions per site were calculated using DnaSP (Rozas and Rozas 1999).

Insertion time estimates are based on occurrence of nucleotide substitutions between LTRs using a nucleotide substitution rate of 2.0 x 10^{-8} synonymous substitutions per site per year proposed for sunflower REs by Ungerer et al. (2009). According to this rate, insertion time for each intact RE was estimated.

DNA isolation and hybridization

Seeds of the sunflower HCM line were washed in tap water and germinated on moist paper in Petri dishes and plants were grown in the open air. Young leaves were collected and DNA purification was carried on according to Cavallini et al. (2010). A sunflower small insert library (Cavallini et al. 2010) was used for relative quantification of the transposons identified in the BAC clones. Forty microliters of plasmid DNA from each of the clones of the sunflower small insert library was first linearised by overnight digestion with *Eco*RI (4 units) in a total volume of 50 ml. DNA was then denatured for 10 min at 91 °C and gridded at moderate density (4 x 4) in duplicate using a Beckman Biomek 2000 replicator tool onto Nylon membranes that had been presoaked in denaturation buffer. Filters were then denatured for 3 min in 1.5 M NaCl, 0.5 M NaOH, neutralized for 15 min in 1.5 M NaCl, 0.5 TrisHCl pH8, and rinsed in 5 x SSC. Filters were then exposed to UV light for 2.5 min. The clones arrayed on the membranes were probed using total labeled genomic DNA from *Helianthus annuus*, *H. petiolaris*, *H. argophyllus*, *H. debilis*, *H. ciliaris*, *H. pumilus*, *H. atrorubens*, *H. giganteus*, *H. simulans*, *H. tuberosus*, *Viguiera multiflora*, *Tithonia rotundifolia*, and other Asteraceae (*Xanthium strumarium*, *Calendula officinalis*, *Senecio vulgaris*, *Tagetes erecta*, *Achillea* spp., *Bellis perennis*, *Gerbera* spp., *Leontopodium* spp., *Taraxacum officinalis* and *Cynara scolymus*). Total genomic DNA from each species was isolated

Table 1 Genomic parameters derived from BAC sequences. The number of full length mobile elements is in parentheses

BAC clone	Total BAC length (bp)	GC content	Number of genes	Number of mobile elements	Density of mobile elements (number/kb)
DES	110,201	39.22	3	8 (5)	1/13.8
DHN	103,566	37.40	2	8 (6)	1/12.9
LTP	135,613	37.68	6	13 (8)	1/10.4
Total	349,380	38.08	11	29 (19)	1/12.0

Table 2 Putative genes identified in the three BAC clones sequenced in these experiments

BAC clone	Gene	Exon length (bp)	Intron length (bp)	Exons/Gene
DES	*Acyl Carrier protein*	3,835	335	4
	Z-Carotene Desaturase	1,744	3,329	13
	VAMP-associated protein	1,107	0	1
DHN	*Dehydrin*	770	148	2
	PSII Chlorophill A	2,197	1,402	4
LTP	*Lipid Transfer Protein 1*	357	627	2
	Lipid Transfer Protein 2	351	133	2
	Lipid Transfer Protein 3	351	123	2
	Lipid Transfer Protein 4	351	6,627	2
	Lipid Transfer Protein 5	351	121	2
	UDP-Glu glucosyltransferase	1,406	0	1
	Mean	1,165	1,168	2.5

from young leaves and digoxigenin-labeled by the random primed DNA labeling technique using a DIG DNA Labeling Kit (Roche) according to the manufacturer's recommendations. Hybridization and detection were performed as described by Cavallini et al. (2010). Labeled lambda DNA was also used as control probe. The relative hybridization intensity for each spot in macroarrays was analyzed by eye and quantified in arbitrary units in the range 0–3, where 0 is for not labeled, 1 for slightly labeled, 2 for labeled, and 3 for heavily labeled. For each transposons identified in BAC clones the hybridization intensity was calculated as the mean of intensity of each corresponding clone.

Whole genome shotgun sequencing by Illumina's Sequencing-By-Synthesis (SBS) technology

A genomic library was prepared from 5μg of genomic DNA from the same line of *H. annuus* using the Illumina PE DNA Sample Prep kit according to the manufacturer. After spin column extraction and quantification, the library was loaded on Cluster Station to create CSMA (Clonal Single Molecular Array) and sequenced at ultra-high throughput on the Illumina's Genome Analyser IIx platform to produce 75 bp paired-end reads. Then, alignments to BAC sequences were performed at 1,000 bp intervals using the program Genomics Workbench 3.0 (CLC Bio) and the number of Illumina hits was calculated along the BAC sequences.

Results

BAC sequencing and annotation

We chose 3 genes that bibliographic information and experimental evidences suggested to be in single copy: a Lipid Transfer Protein encoding gene (*LTP*), a Dehydrin encoding gene (*DHN*) and a Z-Carotene Desaturase encoding gene (*DES*). The three selected genes were used as probes to screen a BAC-library. Three selected BAC-clones were sequenced, yielding the nucleotide sequence of three large genomic regions of 135,613 bp (LTP-clone), 110,201 bp (DES-clone), and 103,566 bp (DHN-clone). Sequencing of 3 BAC clones provides significant new insights into sunflower genomic organization (Table 1). BLASTX and BLASTN searches against non redundant databases at GenBank identified, beside *LTP*, *DES*, and *DHN* genes, other eight protein-encoding genes (Table 2). The BAC clone carrying the *LTP* gene revealed that this gene is present in five copies of different length and sequence (see below).

The pairwise comparison between the three BAC clones resulted into a low percentage of significant homology, ranging from 2.9 to 12.2% of each clone sequence, indicating no excessive redundancy between the three regions.

Eleven gene sequences (accounting for 21,525 bp, Table 2) were found in the three BAC sequences (accounting for 349,380 bp), i.e. gene sequences account for 6.16% of the BAC

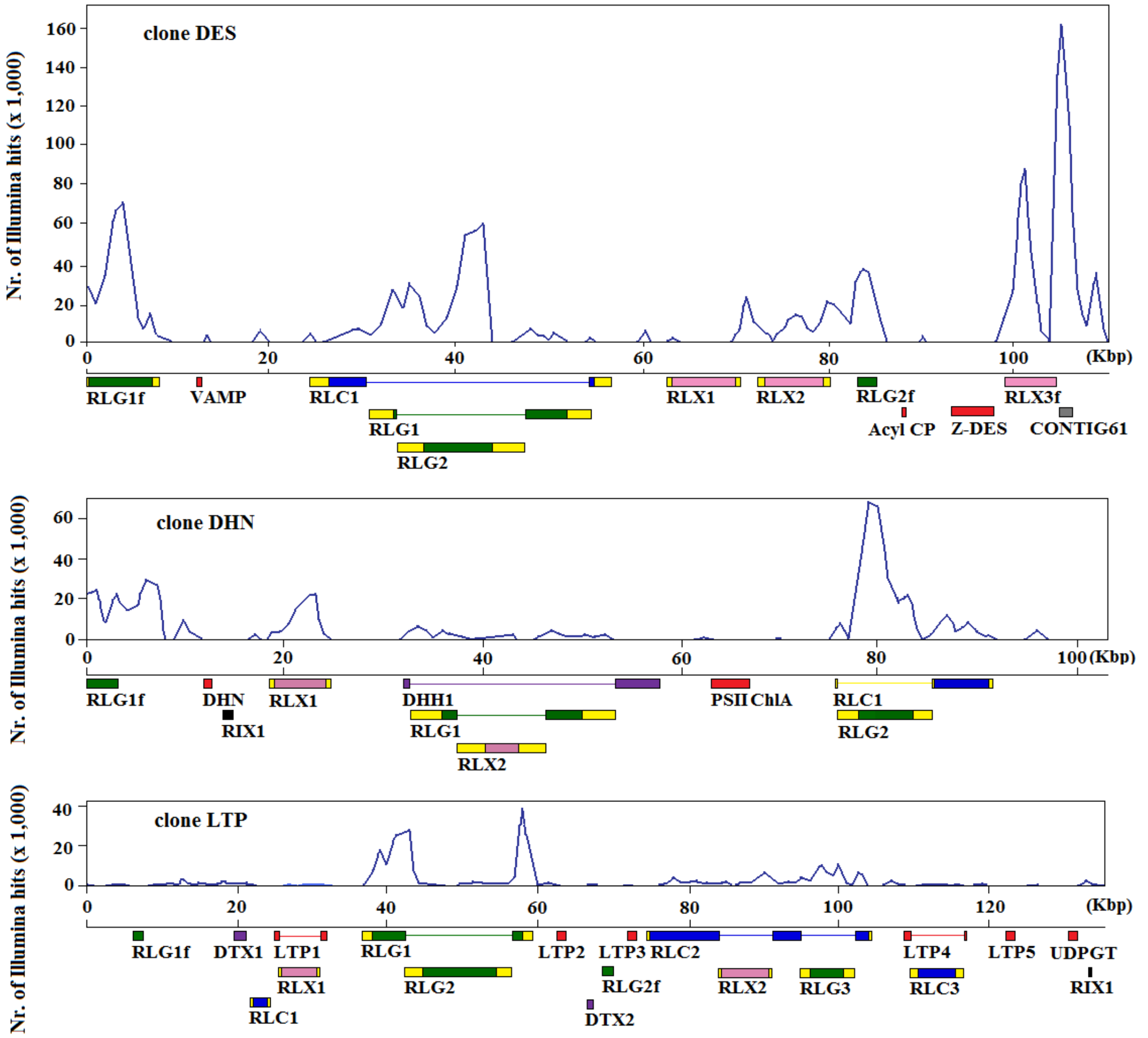


Fig. 1 Annotation of DES, DHN and LTP BAC clones and number of Illumina hits matching to BAC sequences. Transposon sequences are indicated according to Wicker et al. (2007). Incomplete LTR-REs are indicated with the letter f in their code.

sequences. For comparison, it may be observed that in the sunflower small insert library (Cavallini et al. 2010), identified gene sequences (700 bp long, on average) were 64 over 1638 of the whole library, i.e. 3.91%. Consequently, gene sequence content appears overestimated in the BAC clones selected for sequencing, as expected because clones that contain genes (therefore probably corresponding to genic regions) were specifically chosen.

Performing BLASTX, JDOTTER and LTR-FINDER analyses resulted in the identification of 18 full-length LTR-REs, namely with intact ends, irrespective of whether these elements were potentially functional or contained inactivating mutations in their internal sequence (Tables 1 and 3). Seven of them belong to the *Gypsy* superfamily, five to the *Copia* superfamily and six are putative LARDs, i.e., non-autonomous retroelements. We also found 8 incomplete REs (5 *Gypsy*, 1 LARD, and 2 LINEs) that exhibited ill-defined or truncated boundaries. Moreover, two putative DNA transposons fragments, and a putative helitron, interrupted by two LTR-REs, were present.

The arrangement of REs denoted extensive transposition activity in the regions and, similar to that observed in maize (SanMiguel et al. 1996), in many cases elements inserted into others; in one case, two different retroelements are inserted in a single element. On the whole, 15 out of 29 transposons found in the BAC sequences were single, namely adjacent to sequences of the host genome.

All the putatively intact LTR-REs are annotated in Table 4. Twenty-one out of the 29 transposons identified in the BAC clones were also detected in the small-insert library by homology searches (BLAST E-value smaller or equal to $1x10^{-10}$). The annotated map of DES, DHN, and LTP BAC clones are reported in Fig. 1.

To improve BAC annotation, 55 millions of 75-mers obtained by Illumina SBS were aligned to

BAC sequences (Fig. 1). Peaks of Illumina 75-mers occurred in regions corresponding to LTR-REs, especially *Gypsy* elements and LARDs, while *Copia* elements resulted less represented. However, extensive variation in redundancy, as determined by Illumina library alignment, can be observed within superfamilies. For example, DESRLG1f, DESRLG2, DESRLG2f, DESRLX3f, DHNRLG2, and LTPRLG1 show the largest redundancy, with 40,000 Illumina hits or more.

Only a few regions (at 5'-end of the DHN clone and at 3'-end of the DES clone) show high Illumina redundancy and could not be annotated by BLAST analysis, confirming that most of the repetitive component of the sunflower genome is represented by retrotransposons. Interestingly, at the 3'-end of DES clone the highest peak of Illumina hits is found, with more than 160,000 hits, in a region corresponding to the sunflower most repetitive family (named Contig 61), whose nature was unknown, found in the previous study based on the small insert library (Cavallini et al. 2010). Unfortunately, not even the present analyses allow establishing the nature of this repeat, which therefore remains unknown.

It is also to be noted that, in nested elements, inserted elements are often differently redundant than host elements. For example, in the LTP clone the *Gypsy* element LTPRLG1, interrupted by another *Gypsy* element (LTPRLG2), is highly redundant, contrary to the nested element. The opposite trend is observed for DHNRLG2 inserted into DHNRLC1 (Fig. 1).

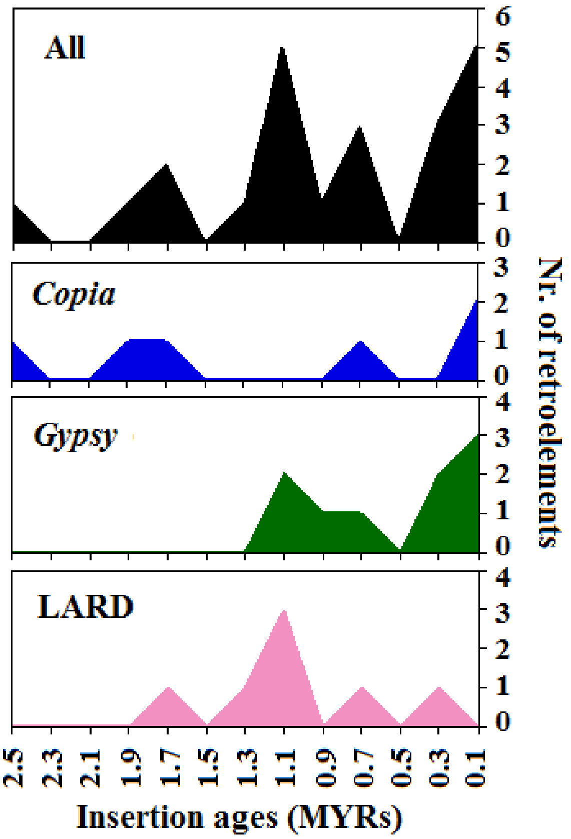


Fig. 2 Distributions of *Copia*, *Gypsy*, and LARD full-length elements identified in the three sequenced BAC clones according to their estimated insertion ages (MYRS).

Transposon dynamics

Twenty-six out of 29 transposons identified in the three BAC clones are retroelements. LTR-REs (*Gypsy*, *Copia*, and LARDs) are 24. In many cases (18 REs) they are complete elements (Table 3).

A complete element can be defined as one that shows two relatively intact LTRs and identified PPT and PBS sites and is also flanked by TSDs. They were first classified as belonging to *Gypsy* (RLG, Wicker et al. 2007), *Copia* (RLC) or LARD (RLX) superfamilies according to BLAST similarity of their internal (i.e., between LTRs) portion to NCBI and REPBASE (Jurka 2000) databases. The coordinates and the characteristics of the complete LTR-REs are reported in Table 4.

The time of insertion of intact retroelements was estimated, based on sister LTR divergence. Indeed, at the time an element inserts into the genome, the LTRs are usually 100% identical since the retroelement transcription starts from the R region in 5' LTR and terminates at the end of the R region in 3' LTR, thus including only one copy of each U5 and U3 regions. Combination of single copy U5 and U3 regions with a hybrid R region during reverse transcription into cDNA yields two identical LTRs at both termini of retroelements prior to integration (Kumar and Bennetzen, 1999). As time passes, mutations occur within the LTRs at a rate that has been proposed to be higher than that of single copy regions, at least in rice (Ma and Bennetzen 2004). Hence LTR retroelements have a built-in clock that can be used to estimate the insertion age (SanMiguel and Bennetzen 1998).

It is to be recalled that the estimation of insertion time by the number of mutations in sister LTRs is subject to error because it assumes the same mutation rates in all retroelements and chromosome positions (Cossu et al., in preparation). Anyway, this method appears as the most suitable to study RE dynamics.

Eighteen LTR pairs, logically identified in full-length elements by JDOTTER and homology analyses, were aligned and nucleotide distance was assessed. The same analysis was performed to four complete LTR-REs (one *Copia*, two *Gypsy*, and one LARD) found in the sequence of two other BAC clones available in GenBank (FJ269356 and GU074383). Insertion age was calculated using the substitution rate of 2.0 x 10^{-8} reported for sunflower REs by Ungerer et al. (2009) according to a personal communication by M. Barker and L. Rieseberg, University of British Columbia. Insertion time estimates based on LTR divergence were consistent with the relative layering of nested REs.

Table 3 Mobile elements found in the three BAC clones. The number of putatively complete elements is in parentheses

BAC clone	Retrotransposons				DNA transposons
	Gypsy	*Copia*	LARD	LINE	
DES	4 (2)	1 (1)	3 (2)	0 (-)	0 (-)
DHN	3 (2)	1 (1)	2 (2)	1 (-)	1 (1)
LTP	5 (3)	3 (3)	2 (2)	1 (-)	2 (-)
Total	12 (7)	5 (5)	7 (6)	2 (-)	3 (1)

We observed a peak of elements with LTR divergence between 1.0 and 1.2 MYRS (Fig. 2); another peak is observed within the last 200,000 yrs, and a *Copia* RE does not show variations in its LTRs, suggesting that its insertion should be occurred between 0 and 54.000 years, i.e. the retrotransposition process could be still active.

The three superfamilies show different time span activity, that overlapped only partially. *Gypsy* elements are by far the most recently inserted, followed by LARDs; *Copia* elements transposition is scattered, from relatively ancient to very recent (Fig. 2).

Genome expansion related to the amplification of *Gypsy* and *Copia* retroelements has been shown to occur in the evolution of three *Helianthus* hybrid species adapted to extreme environments (Ungerer et al. 2009). In agreement to the results reported by Ungerer et al. (2009), our data show that mobilization waves of REs in sunflower are very recent, compared to other species (see for example Baucom et al. 2009; Bennetzen 2007; Ma et al. 2004).

To analyse the conservation of transposons (complete and fragmented) contained in the BACs within the genus *Helianthus* and other Asteraceae, we hybridized genomic DNA from 4 annual and 6 perennial *Helianthus* species, from *Viguiera multiflora* and *Tithonia rotundifolia* (two *Helianthus* related species) and from other 10 Asteraceae species (see Materials and Methods) to a panel of 1,344 clones from a small insert library of sunflower spotted on nylon membranes (Cavallini et al. 2010) and analyzed clones sharing their sequence with REs identified in the BAC clones.

The signals detected in many spots indicated that the repetitive sequences occurring in the BAC clones are present in high copy number in *H. annuus* and conserved enough in sequence to be detected by hybridization in the other species (Fig. 3). The conservation of transposon families is clearly evident not only within *Helianthus*, but also in other Asteraceae, despite their estimated evolutionary distance.

The three superfamilies show different pattern of hybridization in different groups of species of *Helianthus* (Fig. 3): *Copia* elements are equally redundant in annuals and perennials, while *Gypsy* REs are generally much more frequent in annual species than in perennial. Interestingly, LARDs are generally much more redundant in perennial species than in annual, despite being identified in *H. annuus* (i.e. an annual species).

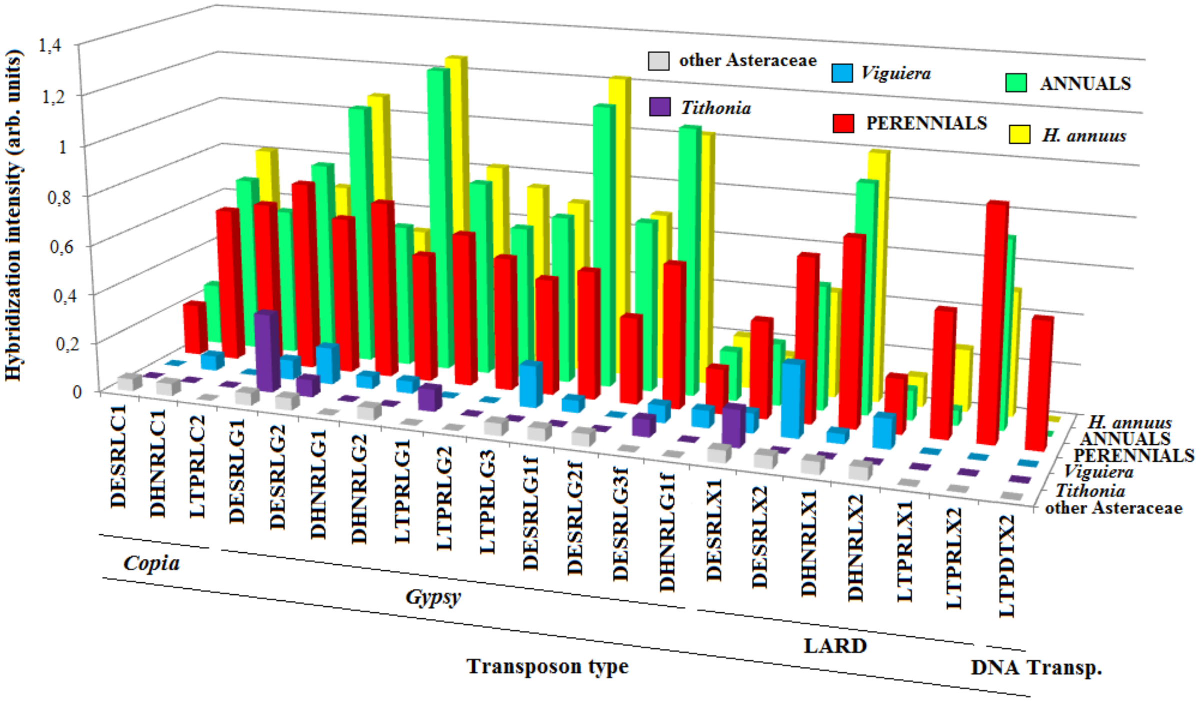


Fig. 3 Mean hybridization intensity of clones from a small insert library and with sequence similarity to 21 transposons identified in three BACs, spotted on nylon membrane and hybridized with labeled genomic DNAs of *H. annuus*, four annual and six perennial *Helianthus* species, *Viguiera multiflora*, *Tithonia rotundifolia* and other ten Asteraceae species. Hybridization signal intensity of each clone was evaluated in arbitrary units: 0, lack of signal; 1, low-intensity signal; 2, medium intensity signal; and 3, strong-intensity signal. For each transposons is reported the mean of labelling intensities of small insert clones corresponding to that transposon.

These different redundancy patterns suggest that the REs identified in the three BAC clones occurred in the progenitor of the genus before splitting of annuals and perennials, however, LARDs have increased their number especially in perennials and *Gypsy* elements especially in annuals. This is consistent with the recent burst of transposition observed for *Gypsy* elements in the sequenced BACs.

Concerning DNA transposons, those containing a transposase gene are fragmented, indicating that they were subjected to large mutations and/or deletions. The third is probably a helitron, because of the occurrence of putative diagnostic features (Du et al. 2008). Such features include i) a putative helicase encoding sequence; ii) many ATC trinucleotides in the 5' helicase flanking region; iii) two CTRRT sequences, preceded (at -11 nucleotides) by putative hairpin sequences in the 3' helicase flanking region. The helicase gene resulted interrupted by the insertion of a *Gypsy* element, on its turn interrupted by a LARD. This putative helitron sequence is the first to be described in sunflower. The insertion of the *Gypsy* element into the helitron can be dated to 1.14 MYRS ago; accordingly, the putative helitron inserted before that date.

The *LTP* locus

Sequencing of the BAC clone highlighted that the *LTP* locus comprises 5 copies of the *LTP* gene, named *LTP1* to *LTP5*; three of these *LTP* gene copies are forward oriented (*LTP2*, *4*, and *5*), two are reverse oriented (*LTP1* and *3*, Fig. 4). All copies show two exons and one intron. *LTP1* is interrupted by a non autonomous RE in its coding region and it is presumably inactivated. Also *LTP4* is interrupted by a LTR-RE (of the *Copia* superfamily); in this case, however, the retroelement is inserted into the intron, therefore functionality of *LTP4* cannot be ruled out. In fact, the coding regions of *LTP4*, as also those of *LTP2*, 3, and 5, do not show stop codons, indicating the possibility that all these gene copies encode functional protein sequences.

Considering *LTP* gene copies without inserted REs, the coding portion is always 351 bp; intron

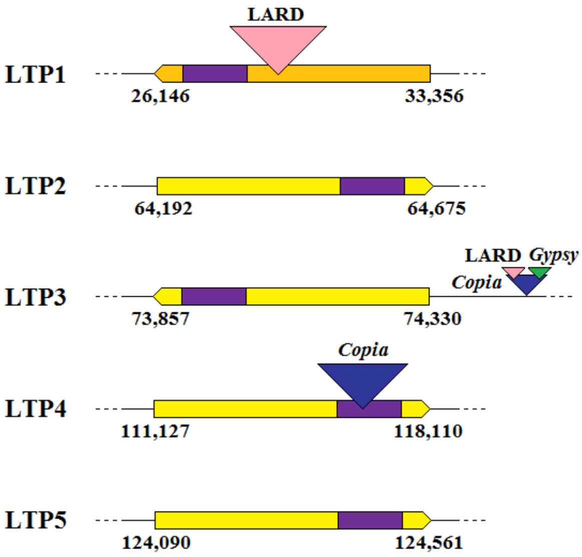


Fig. 4 Schematic representation of five copies of the LTP-encoding gene in the LTP BAC clone. Exons are indicated in yellow and introns in violet. REs interrupting or strictly adjacent to LTP genes are represented as triangles. Numbers indicate the coordinates of each gene in the LTP-BAC sequence.

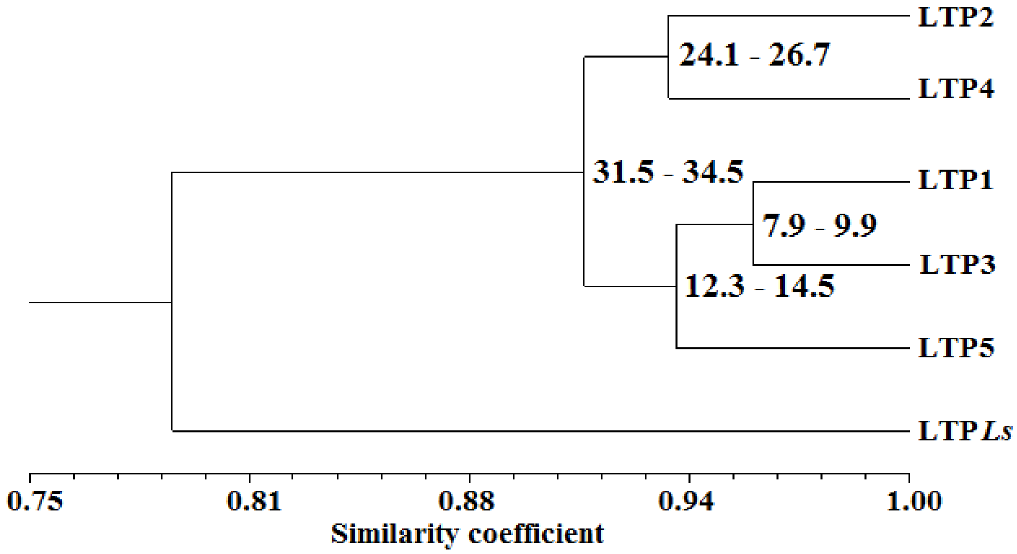


Fig. 5 Dendrogram obtained from UPGMA cluster analysis of 5 LTP gene copies in the LTP BAC clone. LTP Ls indicates a LTP coding sequence of *Lactuca sativa* used as the outgroup. For sunflower sequences, the putative time interval of duplication (in MYRS) is indicated at each node, based on a synonymous substitution rate per year of 1 x 10^{-8}.

Table 4 Characteristics of 18 putatively complete retroelements identified in the three BAC clones

BAC clone	Super-family	Code	RT length (bp)	Verso	Start	5' LTR length (bp)	3' LTR length (bp)	TSR	Illumina Reads	Putative PPT	Putative insertion period (MYRs)
DES	*Copia*	DESRLC1	8196	-	24506	1898	1898	CCCAT	23998	GAGTAAGTGTGGGGA	0.05-0.11
	Gypsy	DESRLG1	9560	+	30395	2447	2412	ATGGT	47971	TAAAGGAGGGGATAC	0.00-0.04
	Gypsy	DESRLG2	14210	+	33233	3537	3537	ACGAG	228302	AAGGGGGTGAGGA	0.00-0.03
	LARD	DESRLX1	7902	+	63006	586	558	-	6234	ACCCCGTGCGTAGG	1.08-1.26
	LARD	DESRLX2	7720	+	72735	773	773	-	82218	AGGGGGAGATTA	1.10-1.23
DHN	LARD	DHNRLX1	5917	+	19117	466	466	TTTAG	118020	AAGGGGGAG	1.18-1.39
	Gypsy	DHNRLG1	11720	+	33080	3391	3440	ATTTG	67303	TCAAGGGGGAGT	1.12-1.15
	LARD	DHNRLX2	8946	-	37930	2858	2866	CTTAT	20829	ATGAAGGAAAAGGGT	0.65-0.68
	Gypsy	DHNRLG2	9788	-	80058	2414	2417	TTGAT	21706	AAAACTTGGGGATAA	0.99-1.04
	Copia	DHNRLC1	7305	-	79978	404	404	TTTTA	96451	ATCCAAGGGGGAG	1.73-1.98
LTP	*Copia*	LTPRLC1	1685	+	23786	183	184	-	195	TTAGGAGGGGGG	2.19-2.73
	LARD	LTPRLX1	6222	+	26828	486	486	GGATG	10507	GATAAGGGGGAG	1.65-1.85
	Gypsy	LTPRLG1	8688	+	38188	1444	1442	-	202048	GAAATGAAAAAGAAA	0.66-0.73
	Gypsy	LTPRLG2	13951	-	43824	1765	1765	ATGAG	21706	AGGACGAAAAAAAGA	0.25-0.31
	Copia	LTPRLC2	16150	+	75543	182	182	CAATA	30611	AGCTTGAGGGGGAG	1.37-1.92
	LARD	LTPRLX2	7053	+	85414	454	453	CCTGT	30201	AAGTTATGAAGACAA	0.22-0.44
	Gypsy	LTPRLG3	7013	-	96713	1478	1453	TGACA	84467	GAAATAAGGTGAAAA	0.93-1.00
	Copia	LTPRLC3	6511	+	111496	931	919	TCATG	5632	AAACACAAAATAAAA	0.00-0.05

length is more variable, ranging from 121 to 627 bp. *LTP1* and *LTP4* have a RE inserted in their coding portion and intron, respectively; excluding inserted REs from their sequence, *LTP1* coding portion is 357 bp long and *LTP4* intron is 6627 bp long.

Dot plot analysis shows that only coding portion are repeated, while regions adjacent to each gene copy seem to be specific to each gene. In fact, extensive variability is found in the putative proximal promoter regions; a 2,000 pb region, upstream of each gene, was scanned for regulatory *cis*-elements against the PLACE database (Higo et al. 1999): a number of putative regulatory elements were found, of different types and in different number for the different gene copies. The number of some *cis*-elements, selected especially among those responsive to environmental changes, show large variability (see Supplementary Materials), suggesting that each gene follows a specific expression pattern. On the contrary, at protein sequence levels, only minor variations are observed, that probably do not affect LTP function. Actually, K_a (the number of non synonymous substitutions per site) ranges from 0.01 to 0.04. Such values are very low compared to K_s (the number of synonymous substitutions per site), ranging from 0.1 to 0.3, i.e. ten-fold the K_a. This suggests conservative selection for LTP gene sequences.

A phylogenetic analysis, by the neighbour-joining method, of the 5 *LTP* gene copies was performed using a LTP encoding sequence of *Lactuca sativa* (GenBank accession number EF101532) as outgroup (Fig. 5). The dendrogram allows deducing a first duplication originating two ancestor sequences that on their turn duplicated once and twice, respectively. The occurrence of intact REs within *LTP1* and *LTP4* allows at least partially to elucidate the time course of *LTP* gene duplications. According to divergence between sister LTRs, the *Copia* element interrupting *LTP4* inserted recently, because no nucleotide substitutions were observed between LTRs. On the contrary, insertion date of the LARD nested into *LTP1* amounts to 1.749 MYRS. Therefore, it can be concluded that gene duplication started before 1.749 MYRS ago.

Actually, hypothesizing that duplicated *LTP* genes originated from a unique ancestor, the number of synonymous substitutions per site between *LTP* gene copies should allow to date each duplication event. Based on the synonymous substitution rate of 1.0 x 10^{-8} proposed for sunflower genes by Barker and Rieseberg (see above), we have calculated the putative dates of duplication events (Fig. 5). It can be supposed that duplications started between 31.5 and 34.5 MYRS ago and that the last duplication (involving *LTP1* and *LTP3* genes) occurred between 7.8 and 10.0 MYRS ago, i.e., before the insertion of REs within two of *LTP* genes, as expected.

Discussion

Sequencing large genomic regions allowed improving the characterization of the sunflower genome, beyond available biochemical, cytological and molecular data.

The repetitive component of the *H. annuus* genome amounts to more than 60% (Cavallini et al. 2010). LTR-RE redundancy is very large and has been described in a number of studies (Santini et al. 2002; Natali et al. 2006; Ungerer et al. 2009; Cavallini et al. 2010). As in the genome of other plant species, LTR-REs are the vast majority, with large prevalence of *Gypsy* over *Copia* elements. In each of the three selected BAC clones we could find nested REs, suggesting that transposition is pervasive of the whole genome.

In the three BAC clones, we have isolated and characterized a number of complete retroelements, adding numerous sequences to the only complete retroelement till now described in the sunflower, HACRE1 (Buti et al. 2009). Both the number of retroelements in the sequenced BAC clones and Illumina data confirm that *Gypsy* elements are prevalent over *Copia* ones in the sunflower genome (see Cavallini et al. 2010), similar to other plant species. For example, in angiosperms, *Gypsy* superfamily is more represented than *Copia* superfamily in the genomes of papaya, (with respective ratio of 5:1, Ming et al. 2008), in *Sorghum* (4:1, Paterson et al. 2009), in rice (3:1, The International Rice Genome Sequencing Project 2005), and in poplar (Tuskan et al. 2006). On the contrary, *Copia* elements are prevalent over *Gypsy* ones in grapevine (2:1, The French-Italian Public Consortium for Grape Genome Characterization 2007). Maize genome shows a similar abundance of the two classes (Meyers et al. 2001), with *Gypsy* elements especially concentrated in gene-poor regions and *Copia* REs overrepresented in gene-rich ones (Baucom et al. 2009; Schnable et al. 2009). Similar data are reported for other cereal species with large genomes such as wheat and barley (Vicient et al. 2005; Paux et al. 2006). Species of the *Gossypium* genus show a variable proportion of *Gypsy* versus *Copia* elements with *Gypsy* elements prevailing in species with larger genome sizes (Hawkins et al. 2006). Such a comparison, though referred to superfamilies, confirms that the dynamics of retrotransposons are different in different species. Further data would be necessary to evaluate if different RE families have undergone different transposition waves, as for example observed in poplar (Cossu et al., submitted).

It is worth noting that, for the first time, putatively complete non-autonomous elements (the so-called

LARDs, Kalendar et al. 2004) have been identified in the sunflower; in fact, this class of REs can be identified only when their complete sequence is available, allowing to recognize the occurrence of LTRs. The number of intact LARDs is the same of intact *Copia* elements, suggesting that the redundancy of LARD superfamily is similar to that of *Copia* one.

Most of the identified REs appear to be specific to *Helianthus*, as already suggested by previous studies (Natali et al. 2006). The redundancy of each element was estimated using an Illumina library of the same sunflower line. Illumina 75mers were aligned to the three BAC sequences and showed a strict correspondence to the annotation: peaks of redundancy are observed in the regions containing REs; moreover, differences can be found among different elements confirming the possibility of using SBS technologies for relative quantification purposes, as reported by Swaminathan et al. (2007).

Concerning retrotransposon dynamics, the identification of sister LTRs allowed for the first time to date the insertion of retroelement in the sunflower genome using this method, established by Ma et al. (2004) in maize or barley. An analysis of insertion age based on comparison of RT-coding sequences of sunflower was carried out by Ungerer et al. (2009), that reported large and recent activity of elements in *Helianthus* species derived from interspecific hybridization between *H. annuus* and *H. petiolaris*. All the REs identified in the three BAC clones show a relatively recent insertion time, in a time span of 0 to 2.6 MYRS. These data indicate that in the sunflower, as in maize (Brunner et al. 2005; Wang and Dooner 2006), retrotransposon burst is very recent and probably still occurring, as already suggested by Cavallini et al. (2010), Ungerer et al. (2009), and Vukich et al. (2009a). On the other hand, it has been recently demonstrated that many sunflower elements are transcribed even in the absence of environmental stimuli (Buti et al. 2009; Vukich et al. 2009b). Vukich et al. (2009b) also showed that, even at a very low rate, transcription of retroelement is followed by insertion in another chromosomal site, i.e. it results in an increase of retrotransposon number.

As far as LTR-RE superfamilies, some differences can be observed in the insertion time between *Copia* and *Gypsy* elements. Also in other species, LTR-RE superfamilies are subjected to different amplification histories during the evolution of the host; for instance, in wheat, *Copia* and *Gypsy* superfamilies are differently represented in the A and B genome (Charles et al. 2008). An example of different amplification histories among RE families was reported for *Copia* elements of *Vitis vinifera* (Moisy et al. 2008) and *Populus trichocarpa* (Cossu et al., submitted).

It has been suggested that the capacity to transpose of a LTR-RE is related to its redundancy, i.e. low redundant REs are more active than high redundant ones because these are more commonly subject to inactivation by small RNAs. In this sense, the few elements in plants for which new insertion events were shown, are three *Copia*-like elements, *Tnt1*, *Tto1*, and *Tos17*, present in a relatively low copy number (< 1,000) per haploid genome (see Yamazaki et al. 2001) and a low redundant *Gypsy* element of sunflower (Vukich et al. 2009b). Interestingly, in the BAC clones sequenced here, Illumina analysis shows two cases in which the inserted elements are much less redundant than the interrupted ones. However, in other cases, especially when *Copia* REs are interrupted by *Gypsy* ones, these are more redundant, suggesting that the negative correlation between RE transposition and redundancy is not a general rule.

Beside recent retrotransposon activity, occurrence of past activity is indicated by the hybridization of genomic DNA of annual and perennial species of *Helianthus* to clones of the sunflower small insert library described by Cavallini et al. (2010). Clones homologous to sequences of the REs identified in the BACs show hybridization signals in both *Helianthus* sections, indicating that such retroelements were already present in the *Helianthus* ancestor, before splitting between annuals and perennials. Then, variations (either increases or decreases) have occurred in the extant species. It is known that the rates of both genome expansion and genome contraction processes appear to vary between species (Bennetzen et al. 2005; Vitte and Bennetzen 2006), allowing some genomes to shrink while others expand. Rearrangements, illegitimate and unequal homologous recombination are the processes driving DNA removal in plants by multiple mechanisms, including repair of double-strand breaks (nonhomologous end-joining) and slipstrand mispairing (Ma and Bennetzen 2004). Therefore, as in other genera, retrotransposon activity seems to be a major force acting in the diversification of species (Ungerer et al. 2006; 2009).

As far as the structure of sequenced loci, the *LTP* locus appears the most interesting, with 5 copies of the *LTP* gene within less than 100,000 bp, four of which are potentially functional, being *LTP1* probably inactivated by a retroelement insertion. Sequence analysis of the proximal putative promoter sequence suggest the mode by which the plant uses gene redundancy: the promoter sequences of sunflower *LTP* genes are very different and should ensure large differences in the regulation pattern of each copy. Such differences have been observed in other species such as grapevine (Falginella et al. 2010). On the other hand, only minor differences may be observed as to the proteins encoded by the four LTP putatively functional genes. It can be concluded that the major specificities of the 5 LTP genes (or at least of the four putatively functional ones) stand in their regulation pattern rather than in their biochemical function.

Finally, it can be observed that *LTP1* inactivation by the *Copia* retroelement has occurred very recently (as indicated by complete similarity between sister LTRs), further suggesting that sunflower is still evolving at high rate.

Actually, a relative incompleteness of species differentiation within *Helianthus* is indicated by cross compatibility between *H. annuus* and annual *Helianthus* species and sometimes also between *H. annuus* and perennial species (Whelan 1978). On the whole, the results reported in this study confirm that the sunflower is an excellent system to study plant genome evolution.

Acknowledgements

Research work supported by PRIN-MIUR, Italy, Project "Variabilità di sequenza ed eterosi in piante coltivate".

References

Baucom RS, Estill JC, Leebens-Mack J, Bennetzen JL (2009) Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Res 19: 243-254

Beguiristain T, Grandbastien MA, Puigdomenech P, Casacuberta JM (2001) Three Tnt1 subfamilies show different stress-associated patterns of expression in tobacco. Consequences for retrotransposon control and evolution in plants. Plant Physiol 127: 212–221

Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot 95: 127–132

Bennetzen JL (2007) Patterns in grass genome evolution. Curr Opin Plant Biol 10: 176-181

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell 17: 343–360

Buti M, Giordani T, Vukich M, Gentzbittel L, Pistelli L, Cattonaro F, Morgante M, Cavallini A, Natali L (2009) HACRE1, a recently inserted *copia*-like retrotransposon of sunflower (*Helianthus annuus* L.). Genome 11: 904-911

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

Charles M, Belcram H, Just J, Huneau C, Viollet A, Couloux A, Segurens B, Carter M, Huteau V, Coriton O, Appels R, Samain S and Chalhoub B (2008) Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics 180: 1071–1086

Du C, Caronna J, He L, Dooner HK (2008) Computational prediction and molecular confirmation of Helitron transposons in the maize genome. BMC Genomics 9: 51

Falginella L, Castellarin SD, Testolin R, Gambetta GA, Morgante M, Di Gaspero G (2010) Expansion and subfunctionalisation of flavonoid 3',5'-hydroxylases in the grapevine lineage. BMC Genomics 11: 562.

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166

Gross BL, Turner KG, Rieseberg LH (2007) Selective sweeps in the homoploid hybrid species *Helianthus deserticola*: evolution in concert across populations and across origins. Mol Ecol 16: 5246–5258

Grover C, Hawkins J, Wendel J (2008) Phylogenetic insights into the pace and pattern of plant genome size evolution. In: *Volff J-N (ed) Plant Genomes. Genome Dynamics. Vol. 4, pp 57-68, Karger, Basel (Switzerland)*

Hawkins JS, Kim HR, 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

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database. Nucl Acids Res 27: 297-300

Holligan D, Zhang X, Jiang N, Pritham EJ, Wessler SR (2006) The transposable element landscape of the model legume *Lotus japonicus*. Genetics 174: 2215–2228

Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 16: 418-420

Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoy A, Schulman AH (2004) Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics 166: 1437–1450

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

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

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

Meyers BC, Tingey SV, Morgante M (2001)

Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11: 1660-1676

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Albert H, Suzuki JY, Tripathi S, Moore PH, Gonsalves D (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452: 991-997

Moisy C, Garrison KE, Meredith CP, Pelsy F (2008) Characterization of ten novel Ty1/*Copia*-like retrotransposon families of the grapevine genome. BMC Genomics 9: 469

Natali L, Santini S, Giordani T, Minelli S, Maestrini P, Cionini PG, Cavallini A (2006) Distribution of Ty3-*Gypsy*- and Ty1-*Copia*-like DNA sequences in the genus *Helianthus* and other Asteraceae. Genome 49: 64–72

Neumann P, Koblizkova A, Navratilova A, Macas J (2006) Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. Genetics 173: 1047–1056

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev I, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Rahman M, Ware D, Westhoff P, Mayer KFX, Messing M and Rokhsar DS (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457: 551-556

Paux E, Roger D, Badaeva E, Gay G, Bernard M, Sourdille P, Feuillet C (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J 48: 463–474

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

Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, Durphy JL, Schwarzbach AE, Donovan LA, Lexer C (2003) Major ecological transitions in wild sunflowers facilitated by hybridization. Science 301: 1211-1216

Rohlf FJ (2008) NTSYSpc: Numerical Taxonomy System, ver. 2.00. Exeter Publishing Ltd, Setauket, NY

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15: 174-175

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds), Bioinformatics methods and protocols: methods in molecular biology, Humana Press, Totowa, NJ, pp. 365–386

SanMiguel P, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot 82: 37-44

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

Santini S, Cavallini A, Natali L, Minelli S, Maggini F, Cionini PG (2002) Ty1/*Copia*- and Ty3/*Gypsy*-like DNA sequences in *Helianthus* species. Chromosoma 111: 192–200

Scherrer B, Isidore E, Klein P, Kim JS, Bellec A, Chalhoub B, Keller B, Feuillet C (2005) Large intraspecific haplotype variability at the Rph7 locus results from rapid and recent divergence in the barley genome. Plant Cell 17: 361-374

Schilling EE, Linder CR, Noyes RD, Rieseberg LH (1998) Phylogenetic relationships in *Helianthus* (Asteraceae) based on nuclear ribosomal DNA internal transcribed spacer region sequence data. Syst Bot 23: 177–187

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112-1115

Soltis DE, Soltis PS (1999) Polyploidy: recurrent formation and genome evolution. Trends Ecol Evol 9: 348-52

Sonnhammer EL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167: GC1–GC10

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

Tarchini R, Biddle P, Wineland R, Tingey S, Rafalski A (2000) The complete sequence of 340 kb of DNA around the rice Adh1-adh2 region reveals interrupted colinearity with maize chromosome 4. Plant Cell 12: 381–391

The French-Italian Public Consortium for Grape Genome Characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449: 463-467

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

Thompson JD, Desmond G, Gibson H, Gibson TJ (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 22: 4673–4680

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

Ungerer MC, Strakosh SC, Stimpson KM (2009) Proliferation of Ty3/*gypsy*-like retrotransposons in hybrid sunflower taxa inferred from phylogenetic data. BMC Biology 7: 40

Ungerer MC, Strakosh SC, Zhen Y (2006) Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. Curr Biol 16: R872-873

Vicient CM, Kalendar R, Schulman AH (2005) Variability, recombination, and mosaic evolution of the barley BARE-1 retrotransposon. J Mol Evol 61: 275–291

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA 103: 17638-17643

Vukich M, Schulman AH, Giordani T, Natali L, Kalendar R, Cavallini A (2009a) Genetic variability in sunflower (*Helianthus annuus* L.) and in the *Helianthus* genus as assessed by retrotransposon-based molecular markers. Theor Appl Genet 119: 1027-1038

Vukich M, Schulman AH, Giordani T, Natali L, Kalendar R, Cavallini A (2009b) *Copia* and *Gypsy* retrotransposons activity in sunflower (*Helianthus annuus* L.). BMC Plant Biology 9:150

Wang Q, Dooner HK (2006) Remarkable variation in maize genome structure inferred from haplotype diversity at the *bz* locus. Proc Natl Acad Sci USA 103: 17644–17649

Whelan EDP (1978) Cytology and interspecific hybridization. In: Carter JF (ed), Sunflower Science and Technology. Am. Soc. Agronomy, pp. 339-370

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 system for eukaryotic transposable elements. Nature Rev Genet 8: 973-982

Wilson RK, Mardis ER (1997) Shotgun sequencing. In: Birren B, Green ED, Klapholtz S, Myers RM, Roskams J (eds) Genome Analysis: A laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucl Acids Res 35: W265-W268

Yamazaki M, Tsugawa H, Miyao A., Yano M, Wu J, Yamamoto S, Matsumoto T, Sasaki T and Hirochika H (2001) The rice retrotransposon Tos17 prefers low-copy-number sequences as integration targets. Mol Genet Genom 265: 336-344

Conclusions

The results obtained during the three-year doctorate course and reported in this work can be subjected to different considerations.

Methodological aspects

A classical area of genetic investigation has been explored using novel methods that imply next generation sequencing and bioinformatic analyses. Next generation sequencing apparently represents a step change and a starting point for genetical and biological research in the XXIst century, providing the opportunity to produce genomic sequences at ever increasing speed and decreasing costs. Using next-generation sequencing technologies it is possible to resequence entire plant genomes or sample entire transcriptomes more efficiently and economically and in greater depth than ever before. Obviously, this step change in technology entails the need for efficient tools and strategies to rapidly narrow down and to accurately describe DNA regions or cDNA transcripts of interest in an ocean of unorganised sequence data.

After a whole genome computational analysis of retrotransposons of *Populus trichocarpa*, performed by different softwares (paper I), the stage at Arizona Genomics Institute allowed to practice protocols for assembling, mapping and analysing NGS data. In particular, three applications of NGS have been exploited: i) the analysis of retrotransposons component in the poplar genome and the comparison among different *Populus* species (paper I and III); ii) the isolation of a number of repeated sequences belonging to two specific families and their relation to the structure of poplar genome (paper II); iii) the analysis, by RNAseq, of the expression of repeated sequences, in particular retrotransposons (paper III); iv) the study of gene expression by RNAseq (paper IV); v) the analysis of differential allelic expression in poplar interspecific hybrids (paper V).

Concerning the analysis of the repetitive component of poplar genome, NGS allowed a complete survey. Obviously, such an analysis can conveniently be improved if the complete genome of the analysed species is available. In the case of poplar species *P. nigra* and *P. deltoides*, genome sequencing has been started and should be completed during 2012, allowing a deeper analysis, with regard to the origin and localization of retroelements in the chromosomes.

The analysis of transcripts by NGS has been proved to attain an unprecedented depth, allowing the isolation of even rarely expressed genes.

The analysis of allelic expression (paper V) is of special interest. Allele Specific Expression (ASE) assays allow to test the effect of structural variations in intergenic regions onto the expression of flanking genes, providing an indirect measure for quantifying cis-regulatory effects by determining the relative proportions of alleles present in the transcripts pool of heterozygous individuals. As both alleles in the heterozygote are expressed in the same cell and are exposed to common regulatory factors, genes exhibiting asymmetric allele expression are inferred to be controlled by cis-acting regulatory variation. Detection of ASE in heterozygous cells offers the advantage that the two alleles are compared under identical circumstances within a single individual genotype, providing an internal control for confounding factors such as differences in mRNA preparation and quality, and environmental and trans-acting factors. ASE is usually performed with allele-specific SNP assays performed on cDNA from the relevant tissues. Here we have used a novel application of the RNA-Seq technology to measure gene expression and the SNPs present within genes to detect a subset of genes where we can attribute RNA-Seq reads to the respective alleles to estimate relative allele abundance in the transcripts pool.

On the whole, we have applied only a few of the potentialities of NGS technology. Rather than sequencing individual genomes or transcriptomes, we envision the sequencing of hundreds or even thousands of related genomes to sample genetic diversity within and between germplasm pools. Identification and tracking of genetic variation are now so efficient and precise that thousands of variants can be tracked within large populations. Such variants can be exploited for a number of applications, as linkage mapping, association mapping, wide crosses and alien introgression, epigenetic modifications, transcript profiling, and population genetics. Such studies are expected to greatly advance crop genetics and breeding, leading to crop improvement.

Analysis of repetitive sequences

Though poplar genome has been sequenced since 2007, the repetitive component of the genome has received little attention. Poplar, as grapevine and Arabidopsis, has a small genome, but, differently from these two species, data on poplar retrotransposons were rare.

The origin, the evolution, and the putative function of repetitive sequences can be different between species, depending on the size of the genome. A huge amount of data on the repetitive component is available especially in plant species with large genome size, in which the repetitive fraction is much larger than in small sized genomes. Studies on retrotransposon dynamics during the evolution of a species, have allowed to evaluate the equilibrium between amplification and loss of these sequences. For example, in maize, rice, sunflower, relatively recent burst of retrotransposons activity has been evidenced, determining large increase of genome size, at least in some species within these genera.

In poplar, no analysis had been performed to characterize retrotransposon sequences regarding to their function and age. We have observed a relatively recent burst of retrotransposons activity, though counterbalanced by high levels of DNA loss (paper I).

Another important result is represented by the huge fraction of retrotransposons of unknown superfamily, the so called LARDs and TRIMs, that account for near half of the retrotransposon fraction of the genome (paper I). They are non-autonomous retrotransposons because lacking coding capacity and cannot be evidenced across species by analyses of sequence similarity. It has been observed that these elements are especially expressed in poplars, possibly because their sequence specificity make them more difficult to be silenced (paper III). The occurrence of “unknown” retrotransposons in plant genomes is probably underestimated and should be surveyed in many other species. Though the datum is referred only to around 300,000 bp (i.e. a very small fraction of the genome), we have observed that one third of retrotransposons individuated in sunflower are LARDs (appendix I), confirming the large frequency of such unknown retrotransposons. LARDs are often seen as “parasites” of other LTR-retroelements because using the enzyme machinery of autonomous elements for their reproduction. However, the large sequence variability of non autonomous elements could be conveniently used by the host species. For example, it is supposed (and, in some cases, demonstrated) that insertion or loss of retrotransposons can alter expression patterns of surrounding genes so contributing to the fine tuning of gene activity. If so, the insertion or loss of hugely variable LARDs can allow a much bigger extent of variation in this respect.

A quite new use of bioinformatic and NGS approach for the analysis of repetitive sequences consisted in the individuation of poplar centromeres (paper II). Though other analyses are to be performed, it appears interesting the occurrence of two distinct centromeric repeats. Mechanisms are supposed to exist in eukaryotes leading to homogenization of centromeric repeats, whose sequence is usually species-specific but conserved in all chromosomes of a species (in which only small variants of the repeats are observed). The occurrence of two different centromeric repeats can be related to an ancient interspecific hybridization from which poplars species have originated. The lack of homogeneization among chromosomes might indicate that such hybridization has occurred in recent times or, more probably, it could be related to the perennial habit of poplar, and hence its very long generation time, so that even an ancient allopolyploidy is still recognizable.

The results on retrotransposon expression (paper III) are also interesting. Retrotransposons are often silenced at DNA level, by siRNA-directed chromatin inactivation. The large number of LTR-retrotransposons transcripts indicates that many poplar retroelements are not silenced at chromatin level. However the action of RE-specific siRNAs in degrading transcripts cannot be excluded, leading to post-transcriptional RE inactivation of these retroelements. It might be supposed that the observed expression of retroelements is the result of a misfunction of silencing apparatus, possibly related to the interspecific origin of the analysed plants. It is known, in fact, that interspecific hybridization is one of the primary causes of genomic shock, a process leading to the production of new genetic variability, through unveiling epigenetic modifications and/or activation of transposons. We are currently analysing retrotransposons expression in the parents of the interspecific hybrids used in our experiments, to verify if any (or all) retrotransposon activated in hybrids is inactive in the parents. The other important aspect that is to be exploited is what consequences derive from the observed RE expression, i.e., to determine if REs expression is or not followed by retrotranscription and subsequent insertion of RE-cDNA into the genome.

On the whole, poplar species appear to be prone to large variability with regard to retrotransposon sequence and redundancy, with possible consequences on the regulation of the activity of protein coding genes.

Gene expression and heterosis

The analysis of gene expression, performed mapping RNAseq data to the complete poplar transcriptome, allowed to establish a reference expression dataset to be used for studying drought related gene expression (paper IV).

The two interspecific hybrids used in our experiments exhibit different levels of heterosis, i.e., their productivity (in terms of biomass production) is for one genotype much larger than that of parents and, for the other genotype, is similar to that of parents. We have analysed expression data in control and droughted plants also in relation to genetic differences between the two hybrids and, possibly, to different heterosis level (paper V).

In several instances, conserved and active alleles, in the two parents used to produce a hybrid, are flanked by different DNA sequences, for example, by non-conserved retrotransposons inserted nearby. Such retroelements are known to be potentially induced by various stresses (as observed in our experiments also) and they may affect the transcription of neighbouring genes by producing single, chimeric, or antisense transcripts or by acting as enhancers. In conclusion, different repetitive sequence environments should affect tissue specificity or temporal regulation of expression of genes. Such differences have been proposed to be one of the causes of heterotic complementation and are comparable to allelic interactions proposed by the overdominance theory for explaining hybrid vigour.

In our experiments, allele variation of expression level has been observed between poplar hybrids for 200 randomly chosen genes. Although this number is much higher than the number of genes usually analysed in this kind of studies, we are currently extending our analysis to the complete poplar transcriptome. In hybrids, the alleles are exposed to a common genetic and environmental context, so allelic expression variation, in different tissues and in responding to environmental stresses, should necessarily derive from cis-regulatory variation.

The allele-specific expression variation in different tissue types, environments, and stress conditions suggest a differential role for the two alleles during hybrid growth and in its interaction with the environment. It is possible that the functional diversity of the two parental alleles in the hybrid may have an impact on hybrid performance through allelic complementation. The availability of the complete genome sequence of parental *P. deltoides* and *P. nigra* genotypes, that should be completed within this year, should allow to discover cis-regulatory variations associated to different allelic expression confirming the importance of such variation, usually due to retrotransposon variability, in generating heterosis.